Latent variable models in biology and ecology

Chapter 5: Bayesian inference for Hidden Markov Models

Sophie Donnet. INRAE

Master 2 MathSV. March 1, 2022

université PARIS-SACLAY

AgroParisTech
INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT
PARIS INSTITUTE OF TECHNOLOGY FOR LIFE, FOOD AND ENVIRONMENTAL SCIENCES

Introducing example

Basics in probability

- Data of Alzheimer symptoms [Moran et al., 2004]
- Presence or absence of 6 symptoms of Alzheimer's disease (AD) in 240 patients diagnosed with early onset AD conducted in the Mercer Institute in St. James's Hospital, Dublin.
- Studied symptoms: Hallucination, Activity, Aggression, Agitation, Diurnal and Affective
- Final goal:We want to know if we can make groups of patients suffering from the same subset of symptoms
- HERE: we only study the presence of hallucinations.
- Data :
  - Vector of size $n = 240$ rows: $(y_i)_{i=1\dots n}$.
  - $y_i = 1$ denotes the presence of hallucinations for patient i, $y_i = 0$ is the absence.

y_i is the realisation of a random variable Y_i

Assumptions

The Y_i's are independent and identically distributed

Statistical model: $\forall i = 1 \ldots n$,

$$\begin{cases} \mathbb{P}(Y_i = 1) & = & \theta \\ \mathbb{P}(Y_i = 0) & = & 1 - \theta \end{cases}$$

$$\Updownarrow$$

$$P(Y_i = y_i|\theta) = \theta^{y_i}(1-\theta)^{1-y_i}, y_i \in \{0, 1\}$$

$$\Updownarrow$$

$$Y_i \sim_{i.i.d} \mathcal{B}ern(\theta)$$

Unknown

$$\theta$$

First estimator of θ: empirical estimator

From the observations $y_1, \ldots, y_n$:

$$\widehat{\theta} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{n_1}{n}$$

- where n_1 is the number of individuals suffering from hallucinations
- Here it's easy to propose one.
- But what if one considers a more complex model (see later)?

Second estimator: maximum likelihood i

Likelihood function

The likelihood of a (set of) parameter value(s), θ, given observations $\mathbf{y}$ is equal to the probability of observing these data $\mathbf{y}$ assuming that θ was the generating parameter.

Second estimator: maximum likelihood ii

- Here:

$$
\begin{aligned}
\ell(\mathbf{y};\theta) &= P(Y_1 = y_1, \dots Y_n = y_n|\theta) \\
&= \prod_{i=1}^{n} P(Y_i = y_i|\theta) \\
&= \prod_{i=1}^{n} \theta^{Y_i}(1-\theta)^{1-Y_i} \\
&= \theta^{\sum_{k=1}^{n} Y_i}(1-\theta)^{\sum_{i=1}^{n} 1-Y_k} \\
&= \theta^{n_1}(1-\theta)^{n-n_1}
\end{aligned}
$$

Second estimator: maximum likelihood iii

- Maximum likelihood estimator : Calculate the "better" parameter θ, i.e. the one maximizing the likelihood function (derivation with respect to θ)

$$\widehat{\theta}^{MLE} = \arg\max_{\theta} \ell(\mathbf{y};\theta)$$

- Here maximum likelihood estimator (estimation)

$$\begin{aligned}
\arg\max_{\theta} \ell(\mathbf{y};\theta) &= \arg\max \log \ell(\mathbf{y};\theta) \\
&= \arg\max_{\theta} \log \theta^{n_1}(1-\theta)^{n-n_1} \\
&= \arg\max_{\theta} [n_1 \log\theta + (n-n_1)\log(1-\theta)]
\end{aligned}$$

Second estimator: maximum likelihood iv

$$\frac{\partial \log \ell(\mathbf{y};\theta)}{\partial\theta} = 0 \Leftrightarrow \frac{n_1}{\theta} - \frac{n-n_1}{1-\theta} = 0 \Leftrightarrow$$
$$(1-\theta)n_1 = (n-n_1)(1-\theta) \Leftrightarrow \theta = \frac{n_1}{n}$$

$$\text{Estimator} : \frac{\sum_{i=1}^n Y_i}{n}, \qquad \text{Estimation} : \frac{\sum_{i=1}^n y_i}{n}$$

- Comments
  - Automatic estimation method
  - Theoretical properties well known when the number of observations n is big
  - The maximization can be difficult

Classical (frequentist) statistics: confidence interval

- Confidence interval: finding two bounds depending on the observations such that this interval $[u(\boldsymbol{Y}), v(\boldsymbol{Y})]$ contains the true parameter θ with high probability.

$$\mathbb{P}_{\boldsymbol{Y}}\left(\theta \in [u(\boldsymbol{Y}), v(\boldsymbol{Y})]\right) = 1 - \alpha$$

- Here :

$$\mathbb{P}_{\boldsymbol{Y}}\left(p \in \left[\widehat{\theta} - \frac{q_{0.05/2}}{\sqrt{n}}\sqrt{\widehat{p}(1-\widehat{\theta})}, \widehat{\theta} + \frac{q_{0.05/2}}{\sqrt{n}}\sqrt{\widehat{\theta}(1-\widehat{\theta})}\right]\right) = 0.95$$

- Interpretation (wikipedia) *"There is a $(1-\alpha)$% probability that the calculated confidence interval from some future experiment encompasses the true value of the population parameter θ."*
- It is a probability over $\boldsymbol{Y}$: $\boldsymbol{Y}$ is random.

Bayesian inference

Main idea

1. Model: $\mathbf{y}$ is the realisation of $\mathbf{y} \sim P(\boldsymbol{Y}|\theta)$
2. The unknown parameter θ is a random object and so we give him a prior probability distribution :

$$\theta \sim \pi(\theta)$$

3. Remember the Bayes Formula:

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)}$$

$$\theta \leftrightarrow B \qquad \mathbf{y} \leftrightarrow A$$

$$p(\theta|\mathbf{y}) = \frac{P(\mathbf{y}|\theta)\pi(\theta)}{P(\mathbf{y})} = \frac{\ell(\mathbf{y}|\theta)\pi(\theta)}{P(\mathbf{y})}$$

4. $p(\theta|\mathbf{y})$ is the posterior probability distribution

Remarks about the Bayesian evidence $P(\mathbf{y})$

$$p(\theta|\mathbf{y}) = \frac{\ell(\mathbf{y}|\theta)\pi(\theta)}{P(\mathbf{y})}$$

- $p(\theta|\mathbf{y})$ is a probability density so its "sum" over all the possible values of θ is equal to 1 i.e. :

$$\int_\theta p(\theta|\mathbf{y})d\theta = 1$$

- Leading to:

$$\frac{\int_\theta \ell(\mathbf{y}|\theta)\pi(\theta)d\theta}{P(\mathbf{y})} = 1 \Leftrightarrow \int_\theta \ell(\mathbf{y}|\theta)\pi(\theta)d\theta = P(\mathbf{y})$$

- $P(\mathbf{y})$ is only a normalization constant also called the marginal likelihood (because it is the likelihood integrated over the prior distribution). The form on θ is given by $\ell(\mathbf{y}|\theta)\pi(\theta)$.

Consequence

As a consequence

$$p(\theta|\mathbf{y}) \propto \ell(\mathbf{y}|\theta)\pi(\theta)$$

where $\propto$ should not hide factors that depend on θ

Alternative notation

$$p(\theta|\mathbf{y}) = [\theta|\mathbf{y}] = \frac{[\mathbf{y}|\theta][\theta]}{[\mathbf{y}]} = \frac{\ell(\mathbf{y}|\theta)\pi(\theta)}{P(\mathbf{y})}$$

First example

- $\theta \in [0, 1]$
- Prior distribution

$$\pi(\theta) = \mathbf{1}_{[0,1]}(\theta)$$

- Posterior distribution

$$\begin{aligned} [\theta|\mathbf{y}] &= \frac{[\mathbf{y}|\theta][\theta]}{[\mathbf{y}]} \propto [\mathbf{y}|\theta][\theta] \\ &\propto \theta^{n_1}(1-\theta)^{n-n_1}\mathbf{1}_{[0,1]}(\theta)^1 \\ &\propto \theta^{n_1+1-1}(1-\theta)^{n-n_1+1-1}\mathbf{1}_{[0,1]}(\theta) \end{aligned}$$

We “recognize” a Beta distribution (See Wikipedia)

R Code

```
n  <- length(Y)
n_1<- sum(Y[,1])
a <- 1
b <- 1
curve(dbeta(x,a+n_1,b+n-n_1),0,0.4,ylab="",xlab="p",
lwd=2,col=2,ylim=c(0,20))
```

Posterior distributions for various n

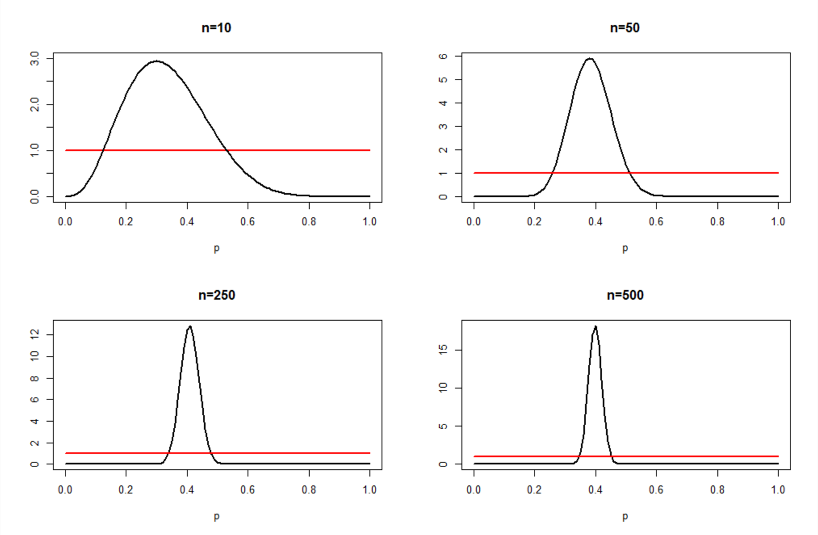


Figure 1: Posterior densities for θ, for various sizes of the sample n (prior distribution in red)

Questions

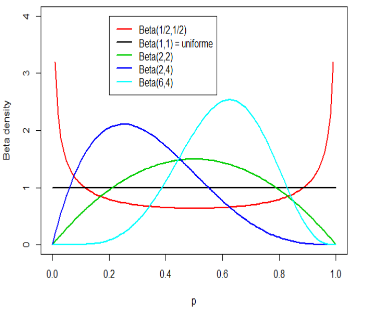
- How to choose the prior ditribution?
- How to summarize the posterior distribution? How to do take decisions with the posterior distribution?
- Is it always easy to determine the posterior distribution?

Choice of the prior distribution

- $\theta \in [0,1]$: $\theta \sim \mathcal{B}eta(a,b)$
- (a,b) are hyperparameters
- $[\theta] \propto \theta^{a-1}(1-\theta)^{b-1}\mathbf{1}_{[0,1]}(\theta)$
- How to chose (a,b) ?
  - If I don't know anything, $a = b = 1$: uniform distribution on $[0,1]$

$$[\theta] \propto \mathbf{1}_{[0,1]}(\theta)$$

  - By tuning a and b, " a priori" give advantage to some values : include knowledge coming from previous studies or experts.



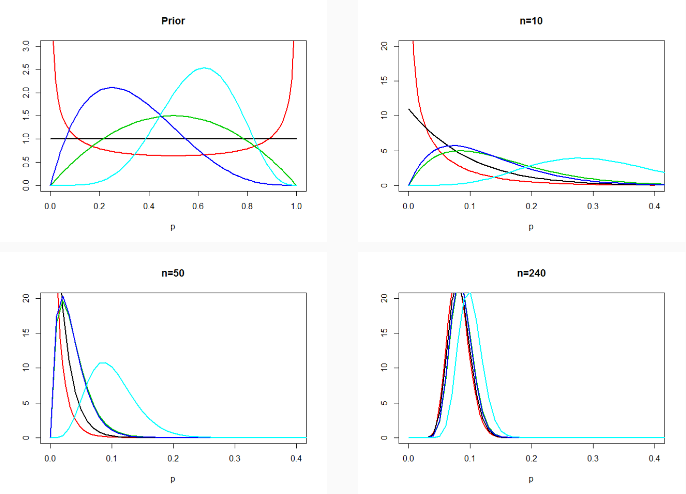
Posterior distribution

$$
\begin{aligned}
[\theta|\mathbf{Y}] &= \frac{[\mathbf{Y}|\theta][\theta]}{[\mathbf{Y}]} \propto [\mathbf{Y}|\theta][\theta] \\
&\propto \theta^{n_1}(1-\theta)^{n-n_1}\theta^{a-1}(1-\theta)^{b-1}\mathbf{1}_{[0,1]}(\theta) \\
&\propto \theta^{a+n_1-1}(1-\theta)^{b+n-n_1-1}\mathbf{1}_{[0,1]}(\theta)
\end{aligned}
$$

We recognize

$$\theta|\mathbf{Y} \sim \mathcal{B}eta(a+n_1, b+n-n_1)$$

Posterior distributions for various prior and n



Comments (1)

- The prior distribution on θ is updated into a posterior distribution using the data
- The posterior/prior distributions quantifies my incertitude on θ
- Posterior: compromise between the prior distribution and the data

$$
\begin{aligned}
p(\theta|\mathbf{Y}) &\propto \pi(\theta)\ell(\mathbf{y}|\theta) \\
\log p(\theta|\mathbf{y}) &= \log \pi(\theta) + \log \ell(\mathbf{y}|\theta) + C \\
\log p(\theta|\mathbf{y}) &= \log \pi(\theta) + \sum_{i=1}^{n} \log \ell(y_i|\theta) + C
\end{aligned}
$$

- The prior distribution has an influence on the posterior distribution if the number of observations n is small
- This influence vanishes if the number of observations increases

Comments (2)

The prior distribution quantifies the prior (un)knowledge on θ.

- In case of complete prior incertitude : non-informative prior (Jeffreys: automatic construction. Improper prior)
- In case of external knowledge (previous experiments, experts) : informative prior

Non informative prior

If we do not know anything about θ

- Use an uniform prior as we did $\theta \sim \mathcal{U}_{[0,1]}$
- The prior distribution can be improper i.e $\int \pi(\theta)d\theta = \infty$ provided the posterior distribution is a probability density
- Method to create an informative prior automatically: **Jeffreys's prior**

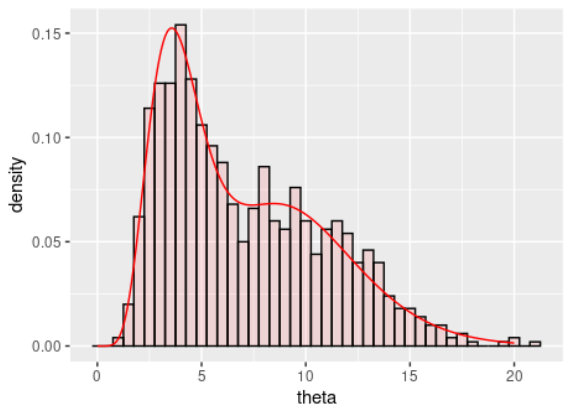
  $$\pi(\theta) \propto \sqrt{\det(I(\theta))}$$

  where $I(\theta)$ is the Fisher information (i.e. is big when the data contain a lot of information on the parameters)

  - The prior gives more importance to values such that the data give a lot of informations about it: minimizes the influence of the prior

Statistics for decisions

From my posterior distribution



- Parameter estimation
- Credible interval
- Hypothesis testing [2]
- Model selection[2]

Bayesian estimator

Aim

Give an estimated value to θ

Once we have the posterior distribution:

- Posterior expectation:

$$E[\theta|\mathbf{Y}] = \int_\theta \theta[\theta|\mathbf{Y}]d\theta$$

- Posterior median:

$$\mathbb{P}(\theta \leq q_{0.5}|\mathbf{Y}) = 0.5$$

- Maximum a posteriori MAP: $\arg\max_\theta\ [\theta|\mathbf{Y}]$

$$\begin{aligned}
\arg\max_\theta\ [\theta|\mathbf{Y}] &= \arg\max_\theta\ \log \ell(\mathbf{Y}|\theta) + \log \pi(\theta) - \cancel{\log P(\mathbf{Y})} \\
&= \arg\max_\theta\ \log \prod_{i=1}^{n} \mathbb{P}(Y_i|\theta) + \log \pi(\theta)
\end{aligned}$$

Bayesian estimator in our example

$$\theta \sim \mathcal{B}eta(a, b), \qquad \theta|\mathbf{Y} \sim \mathcal{B}eta(a + n_1, b + n - n_1)$$

- Posterior expectation

$$E[\theta|\mathbf{Y}] = \frac{a + n_1}{a + n_1 + b + n - n_1} = \frac{a + n_1}{a + b + n}$$

- MAP

$$\arg\max_{\theta}[\theta|\mathbf{Y}] = \frac{a + n_1 - 1}{a + n_1 + b + n - n_1 - 2} = \frac{a + n_1 - 1}{a + b + n - 2}$$

- Posterior median : no explicit expression

$$\approx \frac{a + n_1 - \frac{1}{3}}{a + n_1 + b + n - n_1 - \frac{2}{3}} = \frac{a + n_1 - \frac{1}{3}}{a + b + n - \frac{2}{3}}$$

Aim

Finding the shortest (if possible) interval such that

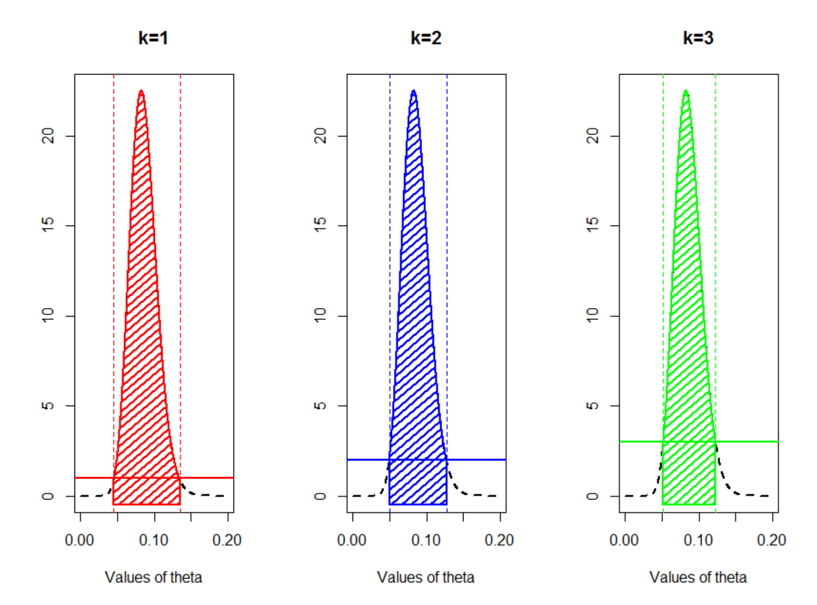
$$\mathbb{P}(\theta \in [a, b]|\mathbf{Y}) = 1 - \alpha$$

Several ways to define it :

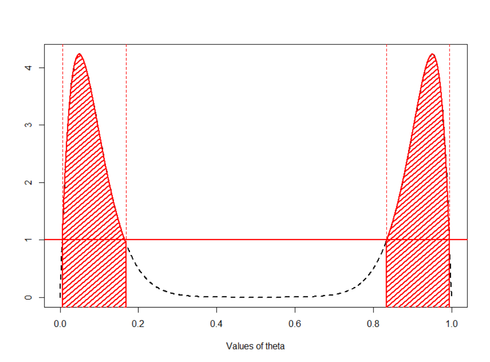
Highest posterior density interval (HPD)

It is the narrowest interval, which for a unimodal distribution will involve choosing those values of highest probability density including the mode. $\mathcal{C} = \{\theta; \pi(\theta|\mathbf{Y}) \geq k\}$ where k is the largest number such that $\int_{\theta;\pi(\theta|\mathbf{Y})\geq k} \pi(\theta|\mathbf{Y}) d\theta = 1 - \alpha$

Credible interval ii



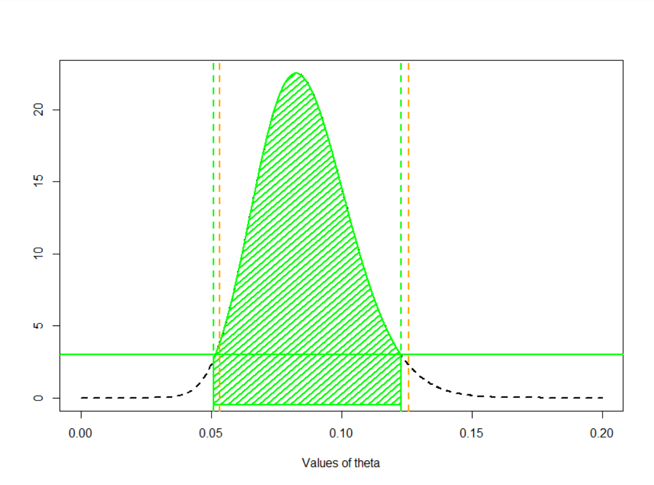
Highest posterior density region



- Be careful : if the posterior density is multi-modal, one can get the union of 2 intervals.
- Difficult to get in practice because we have to invert the density function

Equal-tailed interval

Choosing the interval where the probability of being below the interval is as likely as being above it. This interval will include the median.



Take home messages

- Bayesian statistics are only related to statistical inference (estimation, hypothesis testing...)
- A statistical model is not Bayesian per se (except in neurosciences where some of them consider that the brain is ITSELF Bayesian)
- Bayesian inference is based on a prior distribution on the unknown quantities (parameters, models...)
- The prior distribution quantifies the knowledge on the unknown quantities BEFORE the experiment. We can know nothing (non-informative prior) or something from previous studies, from experts (informative prior).
- The sensibility to the prior has to be analysed to be aware of this influence

Focus on this class

- Bayesian decision is a large topic.
- Focus of this course on the methods to obtain the posterior distribution.

Conjugate prior : easy case

In our example : beta prior $\rightarrow$ beta posterior

- We talk about conjugate prior when the prior and the posterior distributions are in the same family
- Examples

$[y\|\theta]$	$[\theta]$	$[\theta\|y]$	$\mathbb{E}[\theta\|y]$
$\mathcal{N}(\theta, \sigma^2)$	$\mathcal{N}(\mu, \tau^2)$	$\omega^2 = [\frac{1}{\sigma^2} + \frac{1}{\tau^2}]^{-1}$ $\mathcal{N}(\omega^2(\frac{y}{\sigma^2} + \frac{\mu}{\tau^2}), \omega^2)$	$\omega^2(\frac{y}{\sigma^2} + \frac{\mu}{\tau^2})$
$\Gamma(n, \theta)$	$\Gamma(\alpha, \beta)$	$\Gamma(\alpha + y, \beta + n)$	$\frac{\alpha+x}{\beta+n}$
$\mathcal{B}in(n, \theta)$	$\mathcal{B}(\alpha, \beta)$	$\mathcal{B}(\alpha + y, \beta + n - y)$	$\frac{\alpha+y}{\alpha+n+\beta}$
$\mathcal{P}(\theta)$	$\Gamma(\alpha, \beta)$	$\Gamma(\alpha + y, \beta + 1)$	$\frac{\alpha+x}{\beta+1}$

See Wikipedia for instance

To go further

- For the exponential family of distributions, we have a conjuguate prior $\rightarrow$ very rare in practice
- Note that the Gaussian regression model is congugate:

$$\mathbf{Y} \sim \mathcal{N}(X\beta, \sigma^2\mathbb{I})$$
$$\beta|\sigma \quad \sim \quad \mathcal{N}(\beta_0, \sigma^2\Omega)$$

  Then,

- For any more complex model, (such as Latent Variable models) the posterior distribution is not explicit

Illustration on the mixture model

In a few words: My data y_i are issued from to populations, each population having its own mean. I do not know to which population each observation belongs.

- Model $Z_i \in \{1, 2\}$

$$
\begin{array}{rcl}
P(Z_i = 1) & = & \pi_1 \\
Y_i | Z_i = k & \sim & \mathcal{N}(\mu_k, 1)
\end{array}
$$

- Parameters: $\theta = (\pi_1, \mu_1, \mu_2)$
- Likelihood:

$$
[\mathbf{Y}|\theta] = \prod_{i=1}^{n} \left[\pi_1 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_i - \mu_1)^2} + (1 - \pi_1) \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_i - \mu_2)^2} \right]
$$

- Prior distribution:

$$
\pi_1 \sim \mathcal{U}_{[0,1]}, \quad \mu_k \sim \mathcal{N}(0, \omega^2), \quad k = 1, 2
$$

Mixture distribution : posterior

$$
\begin{aligned}
[\theta|\mathbf{Y}] &\propto [\mathbf{Y}|\theta][\theta] \\
&\propto \prod_{i=1}^{n}\left[\pi_1 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_i-\mu_1)^2} + (1-\pi_1)\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_i-\mu_2)^2}\right]\mathbf{1}_{[0,1]}(\pi_1) \\
&\quad \frac{1}{\omega\sqrt{2\pi}} e^{-\frac{1}{2\omega^2}\mu_1^2} \frac{1}{\omega\sqrt{2\pi}} e^{-\frac{1}{2\omega^2}\mu_2^2}
\end{aligned}
$$

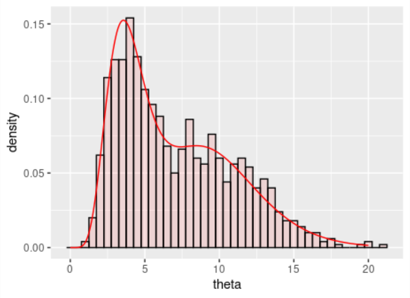
- Non conjuguate model, posterior distribution not explicit.
- How to evalute, for instance the posteriori mean: $\int \theta[\theta|\mathbf{Y}]d\theta$?

How to determine a complex posterior distribution?

- Resort to algorithms to approximate the posterior distribution.
- 2 approaches
  - Sampling methods: supply realizations of the posterior distribution $\theta^{(1)}, \ldots, \theta^{(m)}, \ldots, \theta^{(M)}$.
  - Deterministic methods: approximate the density $p(\theta|\mathbf{Y})$ in a given family of distribution.

Sampling methods

If we can simulate $\theta^{(m)} \sim_{i.i.d.} P(\theta|\mathbf{y})$ for $m = 1, \ldots, M$, then $\frac{1}{M}\sum_{m=1}^{M} \delta_{\theta^{(m)}}(\cdot) \approx p(\cdot|\mathbf{y})$ (Glivenko-Cantelli theorem)



- *Law of large numbers* : $\frac{1}{M}\sum_{m=1}^{M} \theta^{(m)}$ approximates* $E[\theta|\mathbf{y}]$

Monte Carlo Markov Chains methods

Gibbs Sampler, Metropolis-Hastings algorithm...

- *Main idea*: design a Markov Chain such that its stationary distribution is the posterior distribution
- Generic methods
- Supplies asymptotically realizations of the posterior distribution $\theta^{(1)}$, $\ldots$, $\theta^{(m)}$, $\ldots$, $\theta^{(M)}$
- Made the success of the Bayesian inference

Importance samplers

- Simulate “particles” $\theta^{(1)}, \ldots, \theta^{(m)}, \ldots, \theta^{(M)}$ with a “simple” distribution
- Give weights to the particles to correct the discrepancy between the distribution used to simulate and the posterior distribution

Deterministic approximation

Variational Bayes for instance

- Approximate the density $p(\theta|\mathbf{y})$ in a given family of distribution
- Minimizes a divergence with the true posterior density.
- Optimization

Example 1 : non linear model

Assume that we want to explain the presence of hallucination by the patient age and the moment the disease began

- For any individual i, $Y_i = 1$ if we observe hallucinations
- Co-variables: $X_i = (A_i, D_i)$ are the age, and the moment the disease appeared in patient i
- Generalized linear model : Probit regression

$$\begin{aligned} Y_i &\sim \mathcal{B}ern(p_i) \\ p_i &= \Phi(\theta_0 + \theta_1 A_i + \theta_2 D_i) = \Phi({}^t X_i \boldsymbol{\theta}) \end{aligned}$$

where $\boldsymbol{\theta} =^t (\theta_1, \theta_2, \theta_3)$ et $\Phi : \mathbb{R} \mapsto [0,1]$ is the cumulative probability function of a $\mathcal{N}(0,1)$

Likelihood, prior, posterior

- $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2)$
- Likelihood

$$[\mathbf{Y}|\boldsymbol{\theta}] = \prod_{i=1}^{n} \Phi(\theta_0 + \theta_1 A_i + \theta_2 D_i)^{Y_i}(1 - \Phi(\theta_0 + \theta_1 A_i + \theta_2 D_i))^{1-Y_i}$$

- Prior distribution on $\boldsymbol{\theta} \in \mathbb{R}^3$

$$\pi(\boldsymbol{\theta}) \sim \mathcal{N}(0_{\mathbb{R}^3}, \omega \mathbb{I}_3), \quad \text{or} \quad \pi(\boldsymbol{\theta}) \propto 1$$

- Posterior distribution on $\boldsymbol{\theta}$

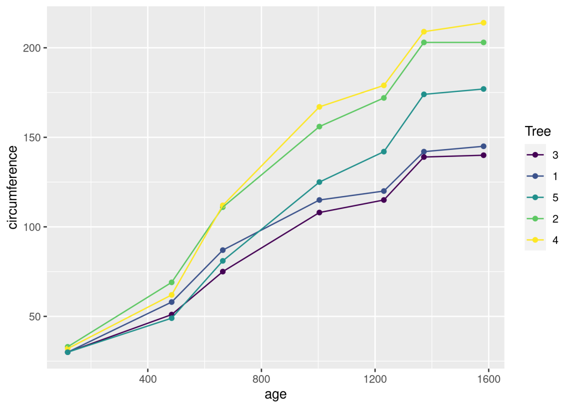
$$\begin{aligned} [\boldsymbol{\theta}|\mathbf{Y}] &\propto [\mathbf{Y}|\theta][\boldsymbol{\theta}] \\ &\propto \prod_{i=1}^{n} \Phi(\theta_0 + \theta_1 A_i + \theta_2 D_i)^{Y_i}(1 - \Phi(\theta_0 + \theta_1 A_i + \theta_2 D_i))^{1-Y_i} \end{aligned}$$

Non conjugated case, no explicit expression of the posterior $[\boldsymbol{\theta}|\mathbf{Y}]$

Example 2: nlme

Orange dataset

- y_{ij} : circumference of orange tree i at age t_{ij}
- $i = 1, \ldots, 5,\ n_i = 5.$



Example 2: nlme

- Logistic relation between y and t
$$f(t;\phi) = \frac{a}{1 + e^{-\frac{t-b}{c}}}$$
- Gaussian noise
- Individual effect of each tree

Latent variable model

$$\begin{aligned}
Y_{ij} &= \frac{A + a_i}{1 + e^{-\frac{t-(B+b_i)}{C+c_i}}} + \varepsilon_{ij} \\
\varepsilon_{ij} &\sim \mathcal{N}(0, \sigma^2) \\
a_i &\sim_{i.i.d} \mathcal{N}(0, \omega_a^2) \\
b_i &\sim_{i.i.d} \mathcal{N}(0, \omega_b^2) \\
c_i &\sim_{i.i.d} \mathcal{N}(0, \omega_c^2)
\end{aligned}$$

- Latent variables : $\mathbf{a} = (a_1, \ldots, a_5)$,$\mathbf{b} = (b_1, \ldots, b_5)$, $\mathbf{c} = (c_1, \ldots, c_5)$
- Parameters : $\theta = (A, B, C, \omega_a^2, \omega_b^2, \omega_c^2, \sigma^2)$

Example 2: likelihood

$$
\begin{aligned}
p(\mathbf{y}|\mathbf{a},\mathbf{b},\mathbf{c};\theta) &= \prod_{i=1}^{5}\prod_{j=1}^{n_i} \frac{1}{2\pi\sqrt{\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(y_{ij} - f(t_{ij}; A+a_i, B+b_i, C+c_i)\right)^2\right] \\
p(\mathbf{a};\theta) &= \prod_{i=1}^{5} \frac{1}{2\pi\sqrt{\omega_a^2}} \exp\left[-\frac{1}{2\omega_a^2} a_i^2\right] \\
p(\mathbf{b};\theta) &= \prod_{i=1}^{5} \frac{1}{2\pi\sqrt{\omega_b^2}} \exp\left[-\frac{1}{2\omega_b^2} b_i^2\right] \\
p(\mathbf{c};\theta) &= \prod_{i=1}^{5} \frac{1}{2\pi\sqrt{\omega_c^2}} \exp\left[-\frac{1}{2\omega_b^2} c_i^2\right] \\
\ell(\mathbf{y};\theta) &= \int_{\mathbf{a},\mathbf{b},\mathbf{c}} p(\mathbf{y}|\mathbf{a},\mathbf{b},\mathbf{c};\theta)p(\mathbf{a};\theta)p(\mathbf{b};\theta)p(\mathbf{c};\theta)d\mathbf{a}\,d\mathbf{b}\,d\mathbf{c}
\end{aligned}
$$

Not an explicit expression $\Rightarrow$ Impossible to get an expression of the posterior distribution

A few words on MCMC

- Enabled the development of Bayesian inference in the 90's
- Stochastic algorithms

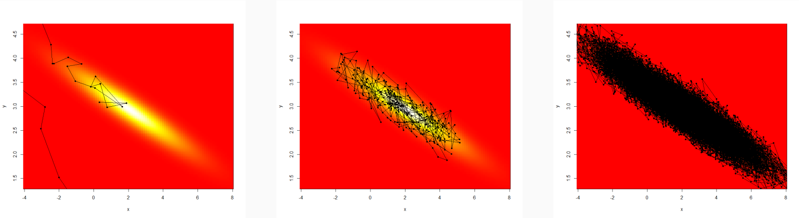
Principle

- Principle: generates a Markov Chain $\boldsymbol{\theta}^{(m)}$ whose ergodic distribution (asymptotic, after a large number of iterations) is the distribution of interest $[\boldsymbol{\theta}|\mathbf{Y}]$
- What it will produce : a sample $(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)})$ from the distribution $[\boldsymbol{\theta}|\mathbf{Y}]$
- What will I do with it? this sample supplies an approximation of the posterior distribution (so : histograms, moments, quantiles...)

$$\widehat{E[\boldsymbol{\theta}|\mathbf{Y}]} = \frac{1}{M}\sum_{m=1}^{M}\boldsymbol{\theta}^{(m)}$$

Metropolis-Hastings algorithm i

- Belongs to the family of Monte Carlo Markov Chains
- Idea: explore the posterior distribution with a random walk using a proposal distribution to move.



- Let's chose an instrumental distribution $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$ which can be easily simulated.

Metropolis-Hastings algorithm ii

A iteration 0

Initialize $\theta^{(0)}$ arbitrarily chosen

At iteration m

1. Propose a candidate $\boldsymbol{\theta}^c \sim q(\boldsymbol{\theta}^c|\boldsymbol{\theta}^{(m-1)})$
2. Calculate an acceptance probability :

$$\rho(\boldsymbol{\theta}^c|\boldsymbol{\theta}^{(m-1)}) = \min\left\{1, \frac{[\boldsymbol{\theta}^c|\mathbf{Y}]}{[\boldsymbol{\theta}^{(m-1)}|\mathbf{Y}]}\frac{q(\boldsymbol{\theta}^{(m-1)}|\boldsymbol{\theta}^c)}{q(\boldsymbol{\theta}^c|\boldsymbol{\theta}^{(m-1)})}\right\}$$

3. Accept the candidate with probability $\rho(\boldsymbol{\theta}^c|\boldsymbol{\theta}^{(m-1)})$, i.e.

$$u \sim \mathcal{U}_{[0,1]} \quad \text{et} \quad \boldsymbol{\theta}^{(m)} = \begin{cases} \boldsymbol{\theta}^c & \text{si } u < \rho(\boldsymbol{\theta}^c|\boldsymbol{\theta}^{(m-1)}) \\ \boldsymbol{\theta}^{(m-1)} & \text{sinon} \end{cases}$$

Why can I apply it?

$$\rho(\boldsymbol{\theta}^c|\boldsymbol{\theta}^{(m-1)}) = \min\left\{1, \frac{[\boldsymbol{\theta}^c|\mathbf{Y}]}{[\boldsymbol{\theta}^{(m-1)}|\mathbf{Y}]}\frac{q(\boldsymbol{\theta}^{(m-1)}|\boldsymbol{\theta}^c)}{q(\boldsymbol{\theta}^c|\boldsymbol{\theta}^{(m-1)})}\right\}$$

$$\begin{aligned}\frac{[\boldsymbol{\theta}^c|\mathbf{Y}]}{[\boldsymbol{\theta}^{(m-1)}|\mathbf{Y}]} &= \frac{[\mathbf{Y}|\boldsymbol{\theta}^c][\boldsymbol{\theta}^c]/\cancel{[\mathbf{Y}]}}{[\mathbf{Y}|\boldsymbol{\theta}^{(m-1)}][\boldsymbol{\theta}^{(m-1)}]/\cancel{[\mathbf{Y}]}}\\ &= \frac{[\mathbf{Y}|\boldsymbol{\theta}^c][\boldsymbol{\theta}^c]}{[\mathbf{Y}|\boldsymbol{\theta}^{(m-1)}][\boldsymbol{\theta}^{(m-1)}]}\end{aligned}$$

- Easy to compute provided I know how to evaluate the likelihood
- Metropolis-Hastings : universal (can be used in a large number of cases = models)

Random walk : particular choice of q

Required qualities on q: easy to propose a candidate: easy to simulate, explicit probability density, with a support larger than the one of the distribution of interest

- $$\theta^c = \theta^{(m-1)} + \xi, \qquad \xi \sim \mathcal{N}_d(0_d, \tau^2 \mathbb{I}_d)$$
- In this case, symmetric kernel: $q(\boldsymbol{\theta}^c|\boldsymbol{\theta}^{(m-1)}) = q(\boldsymbol{\theta}^{(m-1)}|\boldsymbol{\theta}^c)$.

Warning

The choice of the transition kernel $q(\cdot|\cdot)$ strongly influences the theoretical and practical convergence properties.

Visualisation of the principle

We have a look at the wonderful interactive viewer by Chi Feng.

▸ chi-feng interactive MCMC

MH : convergence

- By construction : $[\theta|\mathbf{Y}]$ is stationary
- Explicit transition kernel $K(\theta'|\theta)$
- Prove that for any Borel set A

$$\int_{\theta' \in A} \int_{\theta} K(\theta'|\theta) p(\theta|\mathbf{y}) d\theta d\theta' = \int_{\theta' \in A} p(\theta'|\mathbf{y}) d\theta'$$

MH : kernel transition $K(\theta'|\theta)$

Kernel transition such that

$$\begin{aligned}\boldsymbol{\theta}^c &\sim q(\boldsymbol{\theta}^c|\boldsymbol{\theta})\\ Z &\sim \mathcal{B}ern(\alpha(\theta^c|\theta))\\ \boldsymbol{\theta}' &= Z\boldsymbol{\theta}^c + (1-Z)\boldsymbol{\theta}\end{aligned}$$

Let's prove that

$$K(\theta'|\theta) = \alpha(\theta'|\theta)q(\theta'|\theta) + r(\theta)\delta_\theta(\theta')$$

where

$$r(\theta) = \int_{\theta^c} (1-\alpha(\theta^c|\theta))q(\theta^c|\theta)d\theta^c$$

Kernel transition. Proof

For any measurable function ϕ we need $\mathbb{E}[\phi(\theta')|\theta] = \int \phi(\theta')K(\theta'|\theta)d\theta'$

$$\mathbb{E}[\phi(\theta')|\theta] \quad = \quad \mathbb{E}_{\theta^c,Z}[\phi(Z\boldsymbol{\theta}^c + (1-Z)\boldsymbol{\theta})]$$

Kernel transition. Proof

For any measurable function ϕ we need $\mathbb{E}[\phi(\theta')|\theta] = \int \phi(\theta')K(\theta'|\theta)d\theta'$

$$\begin{aligned}\mathbb{E}[\phi(\theta')|\theta] &= \mathbb{E}_{\theta^c,Z}[\phi(Z\boldsymbol{\theta}^c + (1-Z)\boldsymbol{\theta})] \\ &= \mathbb{E}_{\theta^c,Z}[Z\phi(\boldsymbol{\theta}^c) + (1-Z)\phi(\boldsymbol{\theta})]\end{aligned}$$

Kernel transition. Proof

For any measurable function ϕ we need $\mathbb{E}[\phi(\theta')|\theta] = \int \phi(\theta')K(\theta'|\theta)d\theta'$

$$\begin{aligned}
\mathbb{E}[\phi(\theta')|\theta] &= \mathbb{E}_{\theta^c,Z}[\phi(Z\boldsymbol{\theta}^c + (1-Z)\boldsymbol{\theta})] \\
&= \mathbb{E}_{\theta^c,Z}[Z\phi(\boldsymbol{\theta}^c) + (1-Z)\phi(\boldsymbol{\theta})] \\
&= \int_{\theta^c} [\phi(\boldsymbol{\theta}^c)\mathbb{P}(Z=1|\theta) + \phi(\boldsymbol{\theta})\mathbb{P}(Z=0|\theta)]\, q(\boldsymbol{\theta}^c|\boldsymbol{\theta})d\theta^c
\end{aligned}$$

Kernel transition. Proof

For any measurable function ϕ we need $\mathbb{E}[\phi(\theta')|\theta] = \int \phi(\theta')K(\theta'|\theta)d\theta'$

$$
\begin{aligned}
\mathbb{E}[\phi(\theta')|\theta] &= \mathbb{E}_{\theta^c,Z}[\phi(Z\boldsymbol{\theta}^c + (1-Z)\boldsymbol{\theta})] \\
&= \mathbb{E}_{\theta^c,Z}[Z\phi(\boldsymbol{\theta}^c) + (1-Z)\phi(\boldsymbol{\theta})] \\
&= \int_{\theta^c} [\phi(\boldsymbol{\theta}^c)\mathbb{P}(Z=1|\theta) + \phi(\boldsymbol{\theta})\mathbb{P}(Z=0|\theta)]\, q(\boldsymbol{\theta}^c|\boldsymbol{\theta})d\theta^c \\
&= \int_{\theta^c} \phi(\boldsymbol{\theta}^c)\alpha(\theta^c|\theta)q(\boldsymbol{\theta}^c|\boldsymbol{\theta})d\theta^c + \phi(\theta)\underbrace{\int_{\theta^c}(1-\alpha(\theta^c|\theta))q(\theta^c|\theta)d\theta^c}_{r(\theta)}
\end{aligned}
$$

Kernel transition. Proof

For any measurable function ϕ we need $\mathbb{E}[\phi(\theta')|\theta] = \int \phi(\theta')K(\theta'|\theta)d\theta'$

$$
\begin{aligned}
\mathbb{E}[\phi(\theta')|\theta] &= \mathbb{E}_{\theta^c,Z}[\phi(Z\boldsymbol{\theta}^c + (1-Z)\boldsymbol{\theta})] \\
&= \mathbb{E}_{\theta^c,Z}[Z\phi(\boldsymbol{\theta}^c) + (1-Z)\phi(\boldsymbol{\theta})] \\
&= \int_{\theta^c} [\phi(\boldsymbol{\theta}^c)\mathbb{P}(Z=1|\theta) + \phi(\boldsymbol{\theta})\mathbb{P}(Z=0|\theta)]\, q(\boldsymbol{\theta}^c|\boldsymbol{\theta})d\theta^c \\
&= \int_{\theta^c} \phi(\boldsymbol{\theta}^c)\alpha(\theta^c|\theta)q(\boldsymbol{\theta}^c|\boldsymbol{\theta})d\theta^c + \phi(\theta)\underbrace{\int_{\theta^c}(1-\alpha(\theta^c|\theta))q(\theta^c|\theta)d\theta^c}_{r(\theta)} \\
&= \int_{\theta'} \phi(\theta')\alpha(\theta'|\theta)q(\theta'|\theta)d\theta' + r(\theta)\phi(\theta)
\end{aligned}
$$

Kernel transition. Proof

For any measurable function ϕ we need $\mathbb{E}[\phi(\theta')|\theta] = \int \phi(\theta')K(\theta'|\theta)d\theta'$

$$
\begin{aligned}
\mathbb{E}[\phi(\theta')|\theta] &= \mathbb{E}_{\theta^c,Z}[\phi(Z\boldsymbol{\theta}^c + (1-Z)\boldsymbol{\theta})] \\
&= \mathbb{E}_{\theta^c,Z}[Z\phi(\boldsymbol{\theta}^c) + (1-Z)\phi(\boldsymbol{\theta})] \\
&= \int_{\theta^c} [\phi(\boldsymbol{\theta}^c)\mathbb{P}(Z=1|\theta) + \phi(\boldsymbol{\theta})\mathbb{P}(Z=0|\theta)]\, q(\boldsymbol{\theta}^c|\boldsymbol{\theta})d\theta^c \\
&= \int_{\theta^c} \phi(\boldsymbol{\theta}^c)\alpha(\theta^c|\theta)q(\boldsymbol{\theta}^c|\boldsymbol{\theta})d\theta^c + \phi(\theta)\underbrace{\int_{\theta^c}(1-\alpha(\theta^c|\theta))q(\theta^c|\theta)d\theta^c}_{r(\theta)} \\
&= \int_{\theta'} \phi(\theta')\alpha(\theta'|\theta)q(\theta'|\theta)d\theta' + r(\theta)\phi(\theta) \\
&= \int_{\theta'} \phi(\theta')\alpha(\theta'|\theta)q(\theta'|\theta) + r(\theta)\int_{\theta'}\phi(\theta')\delta_\theta(\theta')d\theta'
\end{aligned}
$$

Kernel transition. Proof

For any measurable function ϕ we need $\mathbb{E}[\phi(\theta')|\theta] = \int \phi(\theta')K(\theta'|\theta)d\theta'$

$$
\begin{aligned}
\mathbb{E}[\phi(\theta')|\theta] &= \mathbb{E}_{\theta^c,Z}[\phi(Z\boldsymbol{\theta}^c + (1-Z)\boldsymbol{\theta})] \\
&= \mathbb{E}_{\theta^c,Z}[Z\phi(\boldsymbol{\theta}^c) + (1-Z)\phi(\boldsymbol{\theta})] \\
&= \int_{\theta^c} [\phi(\boldsymbol{\theta}^c)\mathbb{P}(Z=1|\theta) + \phi(\boldsymbol{\theta})\mathbb{P}(Z=0|\theta)]\, q(\boldsymbol{\theta}^c|\boldsymbol{\theta})d\theta^c \\
&= \int_{\theta^c} \phi(\boldsymbol{\theta}^c)\alpha(\theta^c|\theta)q(\boldsymbol{\theta}^c|\boldsymbol{\theta})d\theta^c + \phi(\theta)\underbrace{\int_{\theta^c}(1-\alpha(\theta^c|\theta))q(\theta^c|\theta)d\theta^c}_{r(\theta)} \\
&= \int_{\theta'} \phi(\theta')\alpha(\theta'|\theta)q(\theta'|\theta)d\theta' + r(\theta)\phi(\theta) \\
&= \int_{\theta'} \phi(\theta')\alpha(\theta'|\theta)q(\theta'|\theta) + r(\theta)\int_{\theta'}\phi(\theta')\delta_\theta(\theta')d\theta' \\
&= \int_{\theta'} \phi(\theta')\left\{\alpha(\theta'|\theta)q(\theta'|\theta) + r(\theta)\delta_\theta(\theta')\right\}d\theta'
\end{aligned}
$$

MH : stationarity

We have to prove that for any subset A,

$$\int_{\theta' \in A} \int_{\theta} K(\theta'|\theta) p(\theta|y) d\theta d\theta' = \int_{\theta' \in A} p(\theta'|y) d\theta'$$

Proof of stationarity I

$$
\begin{aligned}
& \int_{\theta' \in A} \int_{\theta} K(\theta'|\theta) p(\theta|y) d\theta d\theta' \\
= \quad & \iint_{(\theta,\theta')} \mathbf{1}_A(\theta') \left[\alpha(\theta'|\theta) q(\theta'|\theta) + r(\theta)\delta_\theta(\theta')\right] p(\theta|y) d\theta d\theta'
\end{aligned}
$$

Proof of stationarity I

$$
\begin{aligned}
& \int_{\theta' \in A} \int_{\theta} K(\theta'|\theta) p(\theta|y) d\theta d\theta' \\
= \; & \iint_{(\theta,\theta')} \mathbf{1}_A(\theta') \left[\alpha(\theta'|\theta) q(\theta'|\theta) + r(\theta) \delta_\theta(\theta') \right] p(\theta|y) d\theta d\theta' \\
= \; & \underbrace{\iint_{(\theta,\theta')} \mathbf{1}_A(\theta') \alpha(\theta'|\theta) q(\theta'|\theta) p(\theta|y) d\theta d\theta'}_{=B} \\
& + \underbrace{\iint_{(\theta,\theta')} \mathbf{1}_A(\theta') r(\theta) \delta_\theta(\theta') p(\theta|y) d\theta d\theta'}_{=C}
\end{aligned}
$$

Proof of stationarity I

$$
\begin{aligned}
& \int_{\theta' \in A} \int_{\theta} K(\theta'|\theta) p(\theta|y) d\theta d\theta' \\
= & \iint_{(\theta,\theta')} \mathbf{1}_A(\theta') \left[\alpha(\theta'|\theta) q(\theta'|\theta) + r(\theta)\delta_\theta(\theta')\right] p(\theta|y) d\theta d\theta' \\
= & \underbrace{\iint_{(\theta,\theta')} \mathbf{1}_A(\theta') \alpha(\theta'|\theta) q(\theta'|\theta) p(\theta|y) d\theta d\theta'}_{=B} \\
& + \underbrace{\iint_{(\theta,\theta')} \mathbf{1}_A(\theta') r(\theta) \delta_\theta(\theta') p(\theta|y) d\theta d\theta'}_{=C}
\end{aligned}
$$

Proof of stationarity II i

We set $D = \{(\theta, \theta') | p(\theta'|y)q(\theta|\theta') \leq p(\theta|y)q(\theta'|\theta)\}$ such that

$$\alpha(\theta'|\theta) = \begin{cases} \frac{p(\theta'|y)q(\theta|\theta')}{p(\theta|y)q(\theta'|\theta)} & \forall(\theta, \theta') \in D \\ 1 & \forall(\theta, \theta') \in D^c \end{cases}$$

Note that $(\theta, \theta') \in D \Leftrightarrow (\theta', \theta) \in D^c$.

Proof of stationarity II ii

We divide the $B = \iint_{(\theta,\theta')} \mathbf{1}_A(\theta')\alpha(\theta'|\theta)q(\theta'|\theta)p(\theta|y)d\theta d\theta'$ term into two parts:

$$\begin{aligned} B &= \iint_{(\theta',\theta)\in D} \mathbf{1}_A(\theta')\alpha(\theta'|\theta)q(\theta'|\theta)p(\theta|y)d\theta d\theta' \\ &\quad + \iint_{(\theta',\theta)\in D^c} \mathbf{1}_A(\theta')\alpha(\theta'|\theta)q(\theta'|\theta)p(\theta|y)d\theta d\theta' \end{aligned}$$

Proof of stationarity III

Using the fact that $(\theta, \theta') \in D \Leftrightarrow (\theta, \theta') \in D^c$. we make a variable change in B_2 : $(\theta, \theta') \rightarrow (\theta', \theta)$

$$
\begin{aligned}
B &= \underbrace{\iint_{(\theta',\theta)\in D} \mathbf{1}_A(\theta') p(\theta'|y) q(\theta|\theta') d\theta d\theta'}_{B_1} \\
&\quad + \underbrace{\iint_{(\theta',\theta)\in D} \mathbf{1}_A(\theta) p(\theta'|y) q(\theta|\theta') d\theta d\theta'}_{B_2}
\end{aligned}
$$

Proof of stationarity IV : about C

$$
\begin{aligned}
C &= \iint_{(\theta,\theta')} \mathbf{1}_A(\theta') r(\theta) \delta_\theta(\theta') p(\theta|y) d\theta d\theta' \\
&= \int_\theta r(\theta) \mathbf{1}_A(\theta) p(\theta|y) d\theta \\
&= \int_\theta \left[\int_{\theta'} \underbrace{(1-\alpha(\theta'|\theta))}_{=0, \forall (\theta,\theta') \in D^c} q(\theta'|\theta) d\theta' \right] \mathbf{1}_A(\theta) p(\theta|y) d\theta \\
&= \iint_{(\theta,\theta') \in D} (1-\alpha(\theta'|\theta))\, q(\theta'|\theta) \mathbf{1}_A(\theta) p(\theta|y) d\theta d\theta' \\
&= \underbrace{\iint_{(\theta,\theta') \in D} q(\theta'|\theta) \mathbf{1}_A(\theta) p(\theta|y) d\theta d\theta'}_{C_1} \\
&\quad - \iint_{(\theta,\theta') \in D} \alpha(\theta'|\theta) q(\theta'|\theta) \mathbf{1}_A(\theta) p(\theta|y) d\theta d\theta' \qquad (= B_2)
\end{aligned}
$$

Proof of stationarity IV : conclusion

$$C = C_1 - B_2$$

$$C_1 = \iint_D q(\theta'|\theta)\mathbf{1}_A(\theta)p(\theta|y)d\theta d\theta' = \iint_{D^c} q(\theta|\theta')\mathbf{1}_A(\theta')p(\theta'|y)d\theta d\theta'$$

So

$$\begin{aligned}
& \int_{\theta'\in A}\int_\theta K(\theta'|\theta)p(\theta|y)d\theta d\theta' = B + C = B_1 + \cancel{B_2} + C_1 - \cancel{B_2} \\
= & \iint_D \mathbf{1}_A(\theta')p(\theta'|y)q(\theta|\theta')d\theta d\theta' + \iint_{D^c} q(\theta|\theta')\mathbf{1}_A(\theta')p(\theta'|y)d\theta d\theta' \\
= & \iint \mathbf{1}_A(\theta')p(\theta'|y)q(\theta|\theta')d\theta d\theta' \\
= & \int_{\theta'} \mathbf{1}_A(\theta') \underbrace{\int_\theta q(\theta|\theta')d\theta}_{=1} p(\theta'|y)d\theta' = \int_A p(\theta'|y)d\theta'
\end{aligned}$$

cqfd

Convergence

Theoretical convergence

- By construction : $[\theta|\mathbf{Y}]$ is stationary
- The theoretical convergence depends on the distribution of interest and the instrumental distribution . [Robert and Casella, 1999]

Practical convergence

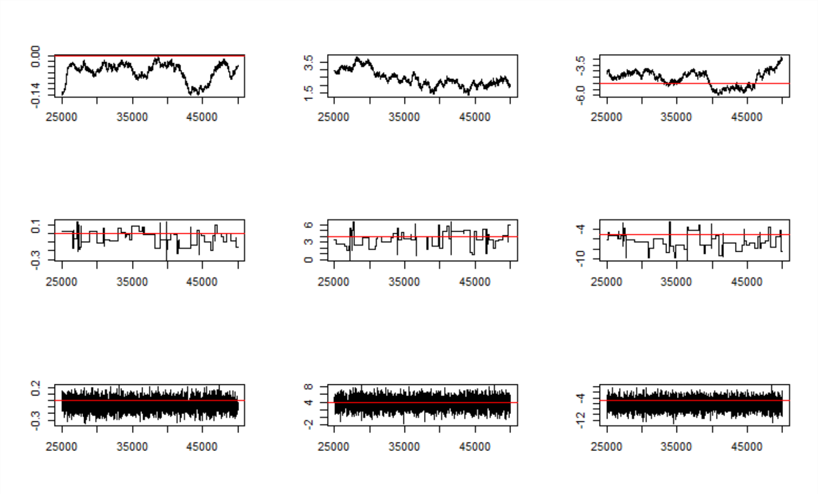
About the acceptance rate

For the random walk

$$\theta^c = \theta^{(m-1)} + \xi, \qquad \xi \sim \mathcal{N}_d(0_d, \tau^2 \mathbb{I}_d)$$

- τ small : we are moving very slowly in the parameters space because the steps are small. I accept a lot but I won't visit all the parameter space
- τ big : we are moving slowly in the parameter space because the steps are big. The algorithm does not accept a lot, we are not moving enough
- τ medium' : we reach quickly the stationary distribution

Trajectories $(\theta^{(m)})_{m\geq 0}$



Chains obtained for 3 values of τ (resp. 0.01,1.5,10). We remove a burn-in period (25000 iterations over the total 50000 iterations)

Remarks

- Target an acceptance rate of 25 % in problems of small dimension, 50% in large dimension problems.
- Can also consider mixtures of kernels $\rho_1 < \rho_2 < \rho_3$

$$\xi \sim p_1 \mathcal{N}(0, \rho_1) + p_2 \mathcal{N}(0, \rho_2) + (1 - p_1 - p_2) \mathcal{N}(0, \rho_3)$$

- Be careful if the parameter leaves in a constrained set.

Exercice

Let us consider the Poisson regression :

$$\begin{array}{rcl} y_i & \sim & \mathcal{P}(\mu_i) \\ \log \mu_i & = & x_i \beta \\ \beta & \sim & \mathcal{N}(0, \sigma^2 I_p) \end{array}$$

- Write (in R) a MCMC such that its asymptotic distribution is $p(\theta|y)$.
- Tune the size of the random walk to observe changes in the behavior
- See codes in `BayesRegressionPoisson_MH.R`

General Gibbs algorithm

If we want to sample a distribution $p(\theta_1, \dots, \theta_d|\mathbf{y})$ such that all the conditional distributions $g_j(\theta_1|\theta_{\{-j\}}, \mathbf{y})$ are explicit, then the Gibbs algorithm is:

Iteration 0: Initialize $\theta_1^{(0)} \dots, \theta_d^{(0)}$

Iteration m $(m = 1 \dots M)$: Given the current values of $\theta_1^{(m-1)}, \dots, \theta_d^{(m-1)}$,

- Simulate $\theta_1^{(m)} \sim g_1(\theta_1|\theta_2^{(m-1)}, \dots, x_d^{(m-1)}, \mathbf{y})$
- Simulate $\theta_2^{(m)} \sim g_2(\theta_2|\theta_1^{(m)}, \theta_3^{(m-1)} \dots, \theta_d^{(m-1)}, \mathbf{y})$
- Simulate $\theta_3^{(m)} \sim g_3(\theta_3|\theta_1^{(m)}, \theta_2^{(m)}, \theta_4^{(m-1)} \dots, \theta_d^{(m-1)}, \mathbf{y})$
- $\cdots$
- Simulate $\theta_d^{(m)} \sim g_d(\theta_d|\theta_1^{(m)}, \dots, \theta_{d-1}^{(m)}, \mathbf{y})$

The stationary distribution is the joint one $p(\theta_1, \dots, \theta_p|\mathbf{y})$

Gibbs for latent variables

Assume that we introduce latent variables $\mathbf{Z}$ in the model such that $[\mathbf{Z}|\mathbf{Y},\theta]$ and $[\theta|\mathbf{Y},\mathbf{Z}]$ have an explicit form and can be easily simulated.

Iteration 0: Initialise $\boldsymbol{\theta}^{(0)}$ et $\mathbf{Z}^{(0)}$

Iteration m $(m = 1 \ldots M)$: Given the current values of $\mathbf{Z}^{(m-1)}$, $\boldsymbol{\theta}^{(m-1)}$

- Simulate $\mathbf{Z}^{(m)} \sim [\mathbf{Z}|\boldsymbol{\theta}^{(m-1)}, \mathbf{Y}]$
- Simulate $\boldsymbol{\theta}^{(m)} \sim [\boldsymbol{\theta}|\mathbf{Z}^{(m)}, \mathbf{Y}]$

We will get a sample of $(\mathbf{Z}^{(m)}, \boldsymbol{\theta}^{(m)})_{m\geq 1}$ under the posterior distribution $[\boldsymbol{\theta}, \mathbf{Z}|\mathbf{Y}]$ and so marginally $\boldsymbol{\theta}^{(m)} \sim [\boldsymbol{\theta}|\mathbf{Y}]$

Exercice : Stationarity of $p(\theta, Z|Y)$

1. Explicit the kernel transition of the chain.
2. Prove that $p(\theta, Z|Y)$ is stationary.

Convergence

Ergodicity and convergence studied in [Robert and Casella, 1999].

Illustration : Gibbs sampler for a Poisson mixture model

- Mixture distribution

$$Y_i \sim_{i.i.d.} \sum_{k=1}^{K} \pi_k \mathcal{P}(\mu_k)$$

- Prior distribution

$$\begin{aligned} \mu_k &\sim \Gamma(\alpha, \beta) \\ \pi &\sim \mathcal{D}ir(\nu, \dots, \nu) \end{aligned}$$

- Posterior distribution

$$[\pi, \mu_1, \dots \mu_K | Y] \propto \prod_{i=1}^{n} \left(\sum_{k=1}^{K} \pi_k e^{-\mu_k} \frac{\mu_k^{Y_i}}{Y_i!} \right) \prod_{k=1}^{K} \pi_k^{\nu-1} \prod_{k=1}^{K} \mu_k^{\alpha-1} e^{-\beta \mu_k}$$

Not explicit

Gibbs sampler for a Poisson mixture : latent variable version

- Latent variables version

$$
\begin{array}{rcl}
Y_i|Z_i = k & \sim & {}_{i.i.d.}\mathcal{P}(\mu_k) \\
P(Z_i = k) & = & \pi_k \\
(Z_{i1}, \ldots, Z_{iK}) & \sim & \mathcal{M}(1, \pi)
\end{array}
$$

with $Z_{ik} = \mathbf{1}_{Z_i = k}$

- Conditional posterior distributions

$$
\begin{array}{rcccl}
p(\mu, \pi|Y, Z) & \propto & p(Y, Z, \mu, \pi) & = & p(Y|Z, \mu)p(Z|\pi)p(\mu)p(\pi) \\
p(Z|Y, \mu, \pi) & \propto & p(Y, Z, \mu, \pi) & = & p(Y|Z, \mu)p(Z|\pi)\cancel{p(\theta)}
\end{array}
$$

Gibbs sampler for a Poisson mixture: $p(\mu|Y, Z)$

$$p(\mu|Y, Z) \propto p(Y|Z; \mu)p(\mu)$$

- $p(Y|Z, \mu)$

$$p(Y|Z, \mu) = \prod_{i=1}^{n} \frac{1}{Y_i!} e^{-\mu_{Z_i}} \mu_{Z_i}^{Y_i}$$

Gibbs sampler for a Poisson mixture: $p(\mu|Y,Z)$

$$p(\mu|Y,Z) \propto p(Y|Z;\mu)p(\mu)$$

- $p(Y|Z,\mu)$

$$p(Y|Z,\mu) \quad = \quad \prod_{i=1}^{n} \frac{1}{Y_i!} e^{-\mu_{Z_i}} \mu_{Z_i}^{Y_i} \quad \propto \quad \prod_{k=1}^{K} \prod_{i=1, Z_i k=1}^{n} e^{-\mu_k} \mu_k^{Y_i}$$

Gibbs sampler for a Poisson mixture: $p(\mu|Y, Z)$

$$p(\mu|Y,Z) \propto p(Y|Z;\mu)p(\mu)$$

- $p(Y|Z,\mu)$

$$\begin{aligned} p(Y|Z,\mu) &= \prod_{i=1}^{n} \frac{1}{Y_i!} e^{-\mu_{Z_i}} \mu_{Z_i}^{Y_i} \quad \propto \quad \prod_{k=1}^{K} \prod_{i=1, Z_i k=1}^{n} e^{-\mu_k} \mu_k^{Y_i} \\ &\propto \prod_{k=1}^{K} e^{-\mu_k N_k} \mu_k^{S_k} \end{aligned}$$

  with $N_k = \sum_{i=1}^{n} Z_{ik}$, $S_k = \sum_{i=1}^{n} Z_{ik} Y_i$
- $p(\mu)$

$$p(\mu) \propto \prod_{k=1}^{K} \mu_k^{\alpha-1} e^{-\beta \mu_k}$$

Gibbs sampler for a Poisson mixture: $p(\mu|Y, Z)$

$$p(\mu|Y, Z) \propto p(Y|Z; \mu)p(\mu)$$

$$\begin{aligned}
p(\mu|Y, Z) &\propto \prod_{k=1}^{K} e^{-\mu_k N_k} \mu_k^{S_k} \prod_{k=1}^{K} \mu_k^{\alpha-1} e^{-\beta\mu_k} \\
&\propto \prod_{k=1}^{K} e^{-\mu_k(N_k+\beta)} \mu_k^{\alpha+S_k-1} \\
\mu_k|Z, Y &\sim {}_{i.i.d.}\Gamma(\alpha + S_k - 1, N_k + \beta)
\end{aligned}$$

Gibbs sampler for a Poisson mixture: $p(\pi|Y,Z)$

$$p(\pi|Y,Z) \propto p(Z|\pi)p(\pi)$$

- $p(Z|\pi)$

$$p(Z|\pi) = \prod_{i=1}^{n} \pi_{Z_i} \prod_{k=1}^{K} \prod_{i=1|Z_{ik}=1}^{n} \pi_k \propto \prod_{k=1}^{K} \pi_k^{N_k}$$

- $p(\pi)$

$$p(\pi) \propto \prod_{k=1}^{K} \pi_k^{\nu-1}$$

- $p(\pi|Y,Z)$

$$\begin{aligned} p(\pi|Y,Z) &\propto \prod_{k=1}^{K} \pi_k^{N_k+\nu-1} \\ \pi|Y,Z &\sim \mathcal{D}ir(\nu+N_1,\ldots,\nu+N_K) \end{aligned}$$

Gibbs sampler for a Poisson mixture: $p(Z|Y,\theta)$

- $$\begin{aligned} p(Z|Y,\theta) &\propto p(Y|Z,\mu)p(Z|\pi) \\ &\propto \prod_{i=1}^{n} e^{-\mu_{Z_i}} \mu_{Z_i}^{Y_i} \pi_{Z_i} \end{aligned}$$
- Z_i independent conditionnally to Y and $Z_i \in \{1, \ldots, K\}$
- $$\begin{aligned} P(Z_i = k|Y,\theta) &\propto e^{-\mu_k} \mu_k^{Y_i} \pi_k \\ &= \frac{e^{-\mu_k} \mu_k^{Y_i} \pi_k}{\sum_{k'=1}^{K} e^{-\mu_{k'}} \mu_{k'}^{Y_i} \pi_{k'}} \end{aligned}$$

Gibbs sampler for a Poisson mixture

Iteration 0: Initialize $\theta^{(0)}$ et $Z^{(0)}$

Iteration m $(m = 1 \dots M)$: Given current values of $Z^{(m-1)}$, $\theta^{(m-1)}$

- Simulate $Z^{(m)} \sim [Z|\theta^{(m-1)}, Y]$ $\forall i = 1, \dots, n$, $\forall g = 1, \dots, G$

$$P(Z_i = k|Y, \theta^{(m-1)}) \quad \propto \quad e^{-\mu_k^{(m-1)}}(\mu_k^{(m-1)})^{Y_i}\pi_k^{(m-1)}$$

- Simulate $\theta^{(m)} \sim [\theta|Z^{(m)}, Y]$
  - $N_k^{(m)} = \sum_{i=1}^n \mathbf{1}_{Z_i^{(m)}=k}$ et $S_k^{(m)} = \sum_{i=1}^n \mathbf{1}_{Z_i^{(m)}=k} Y_i$
  - $\mu_k^{(m)}|Z^{(m)}, Y \sim \Gamma\left(\alpha + S_k^{(m)}, b + N_k^{(m)}\right)$
  - $\pi^{(m)}|Z, Y \sim \mathcal{D}ir(N_1^{(m)} + \nu, \dots, N_K^{(m)} + \nu)$

Exercice

Write the Gibbs corresponding to the SBM model

$$Y_{ij}|Z_i = k, Z_j = l \sim \mathcal{P}(\mu_{kl}) \quad , \quad P(Z_i = k) = \pi_k$$

1. Write the complete likelihood
2. Propose prior distributions
3. Calculate $P(\mu_{kl}|Y, Z)$
4. Calculate $P(\pi|Y, Z)$
5. Are the Z_i's independant conditionnally to Y? How will you proceed?

Remarks on the Gibbs sampler

- For multidimensional distributions
- Does not work if the number of parameters is variable
- Constraining on the conditional distributions (have to be explicit)
- No tuning of the algorithm: + and -

Visualization ▸ chi-feng interactive MCMC (Gibbs)

Metropolis-Hastings within Gibbs

Convenient for latent variable models. Gibbs and Metropolis-Hastings combined

Iteration 0: Initialise $\theta^{(0)}$ et $\mathbf{Z}^{(0)}$
Iteration m ($m = 1 \dots M$): Given the current values of $\mathbf{Z}^{(m-1)}$, $\theta^{(m-1)}$

- On the latent variables $\mathbf{Z}$
  - Propose $\mathbf{Z}^{(c)} \sim q(\mathbf{Z}|\mathbf{Z}^{(m-1)}, \theta^{(m-1)})$
  - Accept with probability such that $[\mathbf{Z}|\theta, \mathbf{Y}]$ is the stationary distribution
- For each component of θ
  - Propose $\theta_k^{(c)} \sim q(\theta_k|\theta_{-\{k\}}^{(m-1)}, \mathbf{Z}^{(m)})$
  - Accept with probability such that $[\theta_k|\theta_{-\{k\}}, \mathbf{Z}, \mathbf{Y}]$ is the stationary distribution

We will get a sample of $(\mathbf{Z}^{(m)}, \theta^{(m)})_{m\geq 1}$ under the posterior distribution $[\boldsymbol{\theta}, \mathbf{Z}|\mathbf{Y}]$ and so marginally $\theta^{(m)} \sim [\theta|\mathbf{Y}]$

Great but and now...

- Many packages to automatically construct the MCMC from your model.
- Very flexible and adapted to latent variable models
- Based on the writing of the model : automatically designed proposals

Software

- WinBUGS : Bayesian inference Using Gibbs Sampling for Windows. 'Point-and-click' windows interface version. May also be called from R.
- With R : package R2WinBUGS
- OpenBUGS
- JAGS : Just An Other Gibbs Sampler. More recent.From R: r2JAGS or rJAGS...
- STAN : developed by Andrew Gelman, coding more complex but more powerful.

In pratice

We will have a look at the file

exempleLinearModelIrispresentation.html

Tunning

- As we saw : step-size will have a non-neglectable influence on the convergence.
- Solution : run the algorithm for a few iterations, check the acceptance rate
  - If the acceptance rate is too low, decrease the step-size.
  - If the acceptance rate is too high, increase the step-size.
- **Be careful**: not possible to adapt the acceptance rate along the iterations, because in that case, it would not be a Markov Chain anymore (theoretical convergence conditions do not hold anymore)

Burn-in

- Period where the chain will reach the stationary distribution
- Need to remove the first iterations (check the traces to calibrate)

Thinning

- With our sample $\theta^{(1)}, \ldots, \theta^{(M)}$ we want to compute expectations, kernel density estimates of the posterior, etc...

$$\frac{1}{M} \sum_{m=1}^{M} \phi(\theta^{(m)})$$

- The convergence of such estimates is ensured (LGN) if the $\theta^{(m)}$ are independent and identically distributed.
- In our case : $\theta^{(m)}$ realisations of a Markov Chain, so not independent.
- To break the dependence, **thin**: take one realization over ... (to be set).

Number of iterations

Must take into account

- The complexity of the model (number of parameters to sample)
- The burn-in period you need
- The thinning parameter you need
- The time you have

From 10000 to . . . millions?

Assessing convergence

- Plot of the chains, parameter by parameter
- Plot the autocorrelations plots
- Compute numerical indicators

Gelman-Rubin convergence diagnostic

- Relies on several chains run in parallel
- Let c be the index for the chain.
- Must be initialized from *over dispersed initial values* $\theta^{c(0)}$ with respect to the targeted distribution.
- Formulae compare the variances intra and inter chains
  - Within-chain variance averaged over the chains:

$$s_c^2 = \frac{1}{M-1}\sum_{m=1}^{M}(\theta^{c(m)} - \overline{\theta^c})^2 \quad W = \frac{1}{C}\sum_{c=1}^{C} s_c^2$$

  - Between-chain variance:

$$B = \frac{M}{C-1}\sum_{c=1}^{C}(\overline{\theta^c} - \overline{\overline{\theta}})^2$$

  - Variance of $\theta|y$ is estimated as a weighted mean of these two quantities

$$\widehat{\text{var}}(\theta|y) = \frac{M-1}{M}\,W + \frac{1}{M}\,B.$$

  - *Potential scale reduction statistic* is defined by

Geweke convergence diagnostic

- Perform a test on two parts of the chain.
- Assume that the chain is of M iterations
- Take $M\alpha_1$ first iterations and $M\alpha_2$ last iterations (such that $\alpha_1 + \alpha_2 < 1$
- Compute the mean of θ on the two parts
- If we are at the stationary distribution, then the two means should be equal
- Correction by the variances (taking into account the dependance between the realisations
- Geweke is the Z-statistic of the test.
- A z-score higher than the absolute value of 1.96 is associated with a p-value of $<$ .05 (two-tailed). The absolute value of Z should therefore be lower than 1.96.

Approximating the posterior : variational Bayes

In a latent variable model, one wants to approximate $p(\mathbf{Z}, \theta|\mathbf{y})$.

- Denote $\tilde{q}(\mathbf{Z}, \theta)$ the approximation of $p(\mathbf{Z}, \theta|\mathbf{y})$.
- We want to minimize

$$KL\left(\tilde{q}(\mathbf{Z}, \theta), p(\mathbf{Z}, \theta|\mathbf{y})\right)$$

  where KL is the Kullback Leibler divergence
- Essential identity

$$\underbrace{\log p(\mathbf{y})}_{Cste} = KL\left(\tilde{q}(\mathbf{Z}, \theta), p(\mathbf{Z}, \theta|\mathbf{y})\right) + \int \tilde{q}(\mathbf{Z}, \theta) \log \frac{p(\mathbf{y}, \mathbf{Z}, \theta)}{\tilde{q}(\mathbf{Z}, \theta)} d\theta d\mathbf{Z}$$

- Minimizing KL is equivalent to maximizing

$$J(\mathbf{y}, \tilde{q}(\mathbf{Z}, \theta)) = \int \tilde{q}(\mathbf{Z}, \theta) \log \frac{p(\mathbf{y}, \mathbf{Z}, \theta)}{\tilde{q}(\mathbf{Z}, \theta)} d\theta d\mathbf{Z}$$

Approximating the posterior : variational Bayes

$$J(\mathbf{y}, \tilde{q}(\mathbf{Z}, \theta)) = \int \tilde{q}(\mathbf{Z}, \theta) \log \frac{p(\mathbf{y}, \mathbf{Z}, \theta)}{\tilde{q}(\mathbf{Z}, \theta)} d\theta d\mathbf{Z}$$

- $\log p(\mathbf{y}, \mathbf{Z}|\theta)$ is explicit in a latent model
- Key point Choose $\tilde{q}(\mathbf{Z}, \theta)$ such that $J(\mathbf{y}, \tilde{q}(\mathbf{Z}, \theta))$ can be computed explicitely.

Approximating the posterior

- Assume that
$$\tilde{q}(\mathbf{Z}, \theta) = \tilde{q}(\mathbf{Z})\tilde{q}(\theta)$$
(simplification)
- Alternatively maximize in $\tilde{q}(\mathbf{Z})$ and $\tilde{q}(\theta)$
- Minimizing a functional with respect to a function $\rightarrow$ Calculus of variations
- Equivalent to iterate
  1. $\tilde{q}(\mathbf{Z}) \propto \exp\left[\int \log p(\mathbf{y}, \mathbf{Z}|\theta)\tilde{q}(\theta)d\theta\right]$
  2. $\tilde{q}(\theta) \propto \pi(\theta)\exp\left[\int \log p(\mathbf{y}, \mathbf{Z}|\theta)\tilde{q}(\mathbf{Z})d\mathbf{Z}\right]$

Application to the Poisson mixture

We consider the following Poisson mixture model

$$\begin{array}{rcl} Y_i|Z_i = k & \sim & \mathcal{P}(\mu_k) \\ P(Z_i = k) & = & \pi_k \\ Z_{ik} & = & \mathbf{1}_{Z_i=k} \end{array}$$

with the prior distributions:

$$\begin{array}{rcl} \mu_k & \sim & \Gamma(a_k, b_k) \\ (\pi_1, \ldots, \pi_K) & \sim & \mathcal{D}ir(e_1, \ldots, e_K) \end{array}$$

About $\tilde{q}(\theta)$

$\tilde{q}(\mathbf{Z}) = \prod_{i=1}^{n} \tilde{q}_i(Z_i)$ with $\tilde{q}_i(Z_i = k) = \tau_{ik}$

$$\begin{aligned}
\mathbb{E}_{\tilde{q}(Z)}\left[\log p(\mathbf{y}, \mathbf{Z}|\theta)\right] &= \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}(-\mu_k + y_i \log \mu_k) + \sum_{i,k} \tau_{ik} \log \pi_k + Cste \\
&= \sum_{k}^{K} \left(-\mu_k \sum_{i=1}^{n} \tau_{ik} + \log \mu_k \sum_{i=1}^{n} \tau_{ik} y_i \right) + \sum_{k=1}^{K} \log \pi_k \sum_{i=1}^{n} \tau_{ik}
\end{aligned}$$

So

$$\begin{aligned}
\tilde{q}(\theta) &\propto \pi(\theta) \exp\left[\mathbb{E}_{\tilde{q}(Z)}\left[\log p(\mathbf{y}, \mathbf{Z}|\theta)\right]\right] \\
&\propto \prod_{k=1}^{K} e^{-\mu_k \sum_{i=1}^{n} \tau_{ik}} \mu_k^{\sum_i y_i \tau_{ik}} \prod_{k=1}^{K} e^{-\mu_k b_k} \mu_k^{a_k - 1} \pi_k^{e_k - 1} \\
&\propto \underbrace{\prod_{k=1}^{K} \pi_k^{\tilde{e}_k - 1}}_{\mathcal{D}ir(\tilde{e}_1, \dots, \tilde{e}_K)} \prod_{k=1}^{K} \underbrace{e^{-\tilde{b}_k \mu_k} \mu_k^{\tilde{a}_k - 1}}_{\Gamma(\tilde{a}_k, \tilde{b}_k)}
\end{aligned}$$

with

$$\tilde{e}_k = e_k + \sum^{n} \tau_{ik}, \quad \tilde{a}_k = a_k + \sum^{n} y_i \tau_{ik} \quad \tilde{b}_k = b_k + \sum^{n} \tau_{ik}$$

About $\tilde{q}(\mathbf{Z})$

$$\begin{aligned}
&\mathbb{E}_{\tilde{q}(\theta)}\left[\log p(\mathbf{y}, \mathbf{Z}|\theta)\right] = \\
= \quad & \sum_{i=1}^{n}\sum_{k=1}^{K} -Z_{ik}\mathbb{E}_{\tilde{q}(\theta)}[\mu_k] + Z_{ik}y_i\mathbb{E}_{\tilde{q}(\theta)}[\log\mu_k] + Z_{ik}\mathbb{E}_{\tilde{q}(\theta)}[\log\pi_k] + Cste \\
= \quad & \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik}\underbrace{\left[-\frac{\tilde{a}_k}{\tilde{b}_k} + y_i\left[\Psi(\tilde{a}_k) - \log(\tilde{b}_k) + \Psi(\tilde{e}_k) - \Psi(\bar{\tilde{e}})\right]\right]}_{\rho_{ik}}
\end{aligned}$$

where Ψ is the digamma function.

$$\begin{aligned}
\tilde{q}(\mathbf{Z}) \quad & \propto \quad \exp\left[\int \log p(\mathbf{y}, \mathbf{Z}|\theta)\tilde{q}(\theta)d\theta\right] \\
& \propto \quad e^{\sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik}\rho_{ik}} \\
& \propto \quad \prod_{i=1}^{n}\prod_{k=1}^{K}(e^{\rho_{ik}})^{Z_{ik}} \\
\tau_{ik} = P_{\tilde{q}(\mathbf{Z})}(Z_{ik} = 1) \quad & \propto \quad e^{\rho_{ik}}
\end{aligned}$$

Remarks on the methods

- Algorithm (VBEM) iterates the two previously described steps.
- Optimization algorithm provides an approximation of the posterior distribution.
- Quick but wrong
- Under-estimate the posterior variance
- If considering minimizing

$$KL\left(p(\mathbf{Z},\theta|\mathbf{y}),\tilde{q}(\mathbf{Z},\theta)\right)$$

$\Rightarrow$ Expectation Propagation EP on wikipedia

Remarks in the implementation

- Calculus adapted to each model. Less universal than MCMC.
- Variational bayes R Packages : LaplacesDemon by Henrik Singmann
  $\Rightarrow$ Not working on our example

In a few words

- Uses the Laplace approximation (Taylor extension around the MAP)
- OK for Gaussian Latent Model:

$$
\begin{aligned}
Y_i|x,\theta_2 &\sim p(\cdot|x_i,\theta_2)\\
x|\theta_1 &\sim p(x|\theta_1) = \mathcal{N}(0,\Sigma)\\
\theta = (\theta_1,\theta_2) &\sim p(\theta)
\end{aligned}
$$

- Many models are included
  Exemple : generalized linear model

$$
\begin{aligned}
Y_i &\sim \mathcal{N}(\phi(\mu_i),\sigma^2) \quad \mu_i = \alpha + \sum_{k=1}^{K}\beta_k z_{ki}\\
x = (\alpha,\beta_1,\ldots,\beta_K) &\sim \mathcal{N}(0,\Sigma)\\
\theta_2 = \frac{1}{\sigma^2} &\sim \Gamma(a,b)
\end{aligned}
$$

- Particularly adapted to spatial models

Importance sampling i

- For any function φ (...),

$$E_{\theta|\mathbf{Y}}[\varphi(\theta)] = \int_{\Theta} \varphi(\theta)\pi(\theta|\mathbf{Y})d\theta = \int_{\Theta} \varphi(\theta)\frac{\pi(\theta|\mathbf{Y})}{\eta(\theta)}\eta(\theta)d\theta$$

- with η easily simulable distribution, such that its support contains the one of $\pi(\theta|\mathbf{Y})$, whose density can be computed.

Importance sampling ii

Monte Carlo estimator : $\theta^{(1)}, \ldots, \theta^{(M)} \sim_{i.i.d.} \eta$

$$\begin{aligned}
\widehat{E}_n[\varphi(\theta)] &= \frac{1}{M} \sum_{m=1}^{M} \frac{\pi(\theta^{(m)}|\mathbf{Y})}{\eta(\theta^{(m)})} \varphi(\theta^{(m)}) \\
&= \frac{1}{M} \frac{1}{p(\mathbf{Y})} \sum_{m=1}^{M} \underbrace{\frac{\ell(\mathbf{Y}|\theta^{(m)})\pi(\theta^{(m)})}{\eta(\theta^{(m)})}}_{w^{(m)}} \varphi(\theta^{(m)})
\end{aligned}$$

But $p(\mathbf{Y})$ without explicit expression:

$$p(\mathbf{Y}) = \int \ell(\mathbf{Y}|\theta)\pi(\theta)d\theta = \int \frac{\ell(\mathbf{Y}|\theta)\pi(\theta)}{\eta(\theta)} \eta(\theta) d\theta$$

Importance sampling iii

$$\widehat{p(\mathbf{Y})} = \frac{1}{M}\sum_{m=1}^{n}\frac{\ell(\mathbf{Y}|\theta^{(m)})\pi(\theta^{(m)})}{\eta(\theta^{(m)})} = \frac{1}{M}\sum_{m=1}^{N} w^{(m)}$$

$$\begin{aligned}
\widehat{E}_{\theta|\mathbf{Y}}[\varphi(\theta)] &= \frac{1}{M}\sum_{m=1}^{M}\frac{w^{(m)}}{p(\mathbf{Y})}\varphi(\theta^{(m)}) \\
\widehat{\widehat{E}}_{\theta|\mathbf{Y}}[\varphi(\theta)] &= \frac{\cancel{\frac{1}{M}}\sum_{m=1}^{M} w_n^{(m)}\varphi(\theta^{(m)})}{\cancel{\frac{1}{M}}\sum_{m=1}^{M} w^{(m)}} \\
&= \sum_{m=1}^{M} W^{(m)}\varphi(\theta^{(m)}) \text{ avec } W^{(m)} = \frac{w^{(m)}}{\sum_{m=1}^{M} w^{(m)}}
\end{aligned}$$

Consistant Estimator

Importance Sampling

Summary

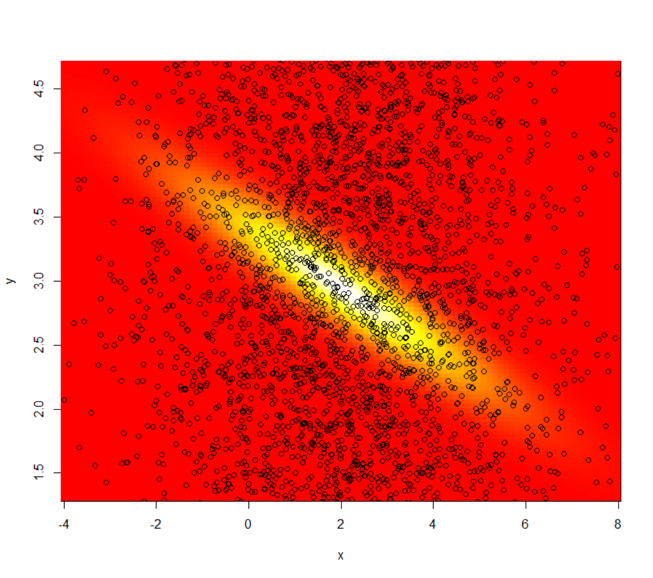
Approximate $\pi(\theta|\mathbf{Y})$ by a weighted sample $(\theta^{(m)}, W^{(m)})_{m=1\ldots M}$ such that

$$\theta^{(m)} \sim_{i.i.d.} \eta(\cdot)$$

$$W^{(m)} = \frac{w^{(m)}}{\sum_{m=1}^{M} w^{(m)}}$$

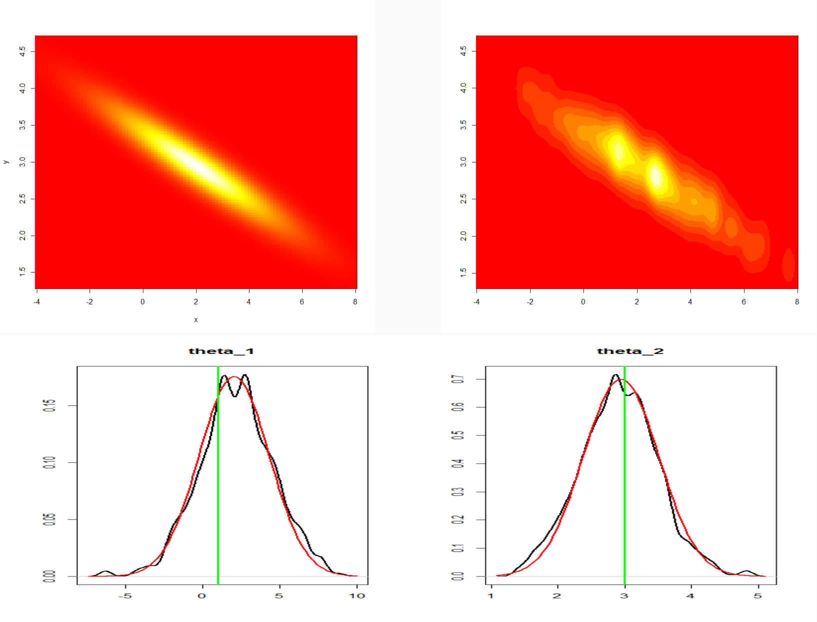
$$w^{(m)} = \frac{\ell(\mathbf{Y}|\theta^{(m)})\pi(\theta^{(m)})}{\eta(\theta^{(m)})}$$

Importance Sampling: example

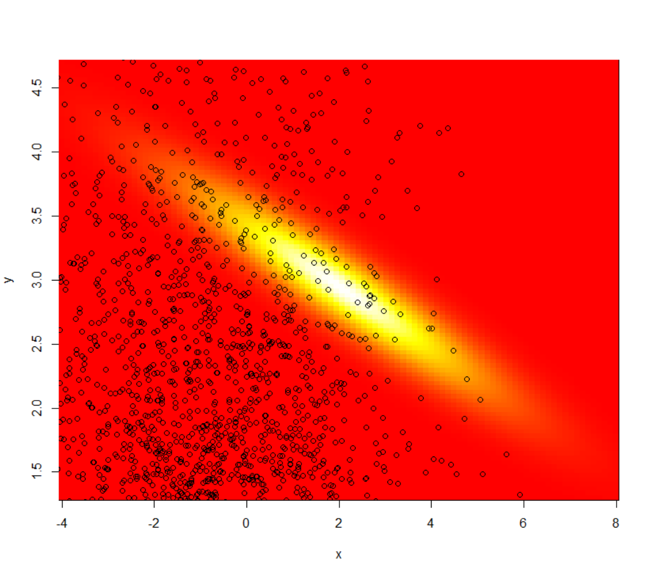


Simulated particles, without their weights

Importance Sampling: posterior

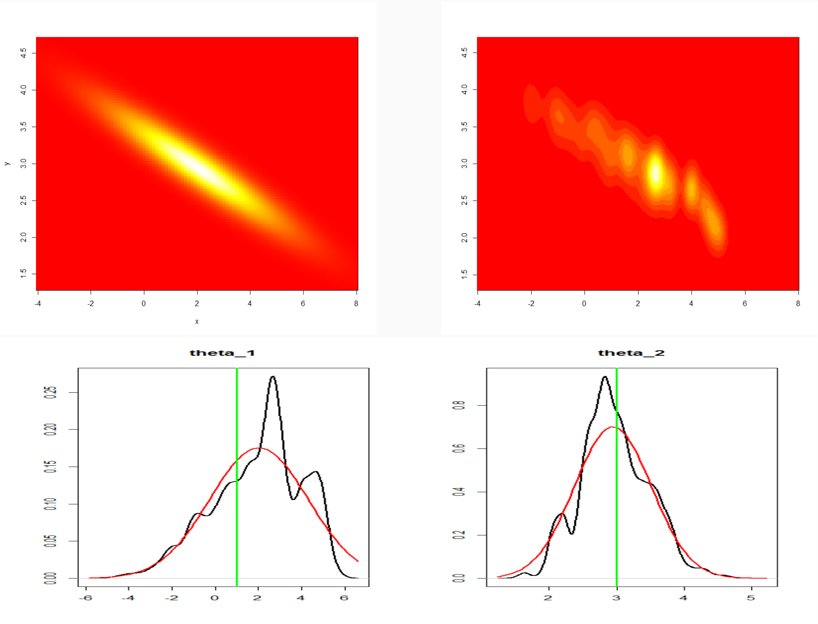


Importance Sampling: an example that does not work



Simulated particles, without their weights

Importance Sampling: un example that does not work



Importance sampling methods: comments i

- Convergence ensured by the large numbers law.
- But the quality of the estimator (variance) for a given M ?
- Problem if some weights are very large while others are very small.
  - Calculus of the Effective Sample Size:

$$ESS = \frac{1}{\sum_{m=M}^{N} \left(W^{(m)}\right)^2}$$

  - $EES \in [1, M]$.
  - The weighted sample $(W^{(m)}, \theta^{(m)})$ corresponds to a no-weighted sample of size ESS

Importance sampling methods: comments ii

- Essential to chose η carefully such that $\ell(\mathbf{Y}|\theta^{(m)})\pi(\theta^{(m)})$ not two small.
- Not possible in large dimension problems: need to sequentially build η Sequential Monte Carlo
- Advantage: easy estimation of $p(\mathbf{Y})$ par $\sum_{m=1}^{M} w^{(m)}$

Problematic

Can we use the variational approximation of the posterior distribution in a IS procedure. Can we correct its tendancy to under-estimate the posterior variance?

Let η^{VB} be the VB posterior approximation of $\pi(\theta|\mathbf{Y})$.

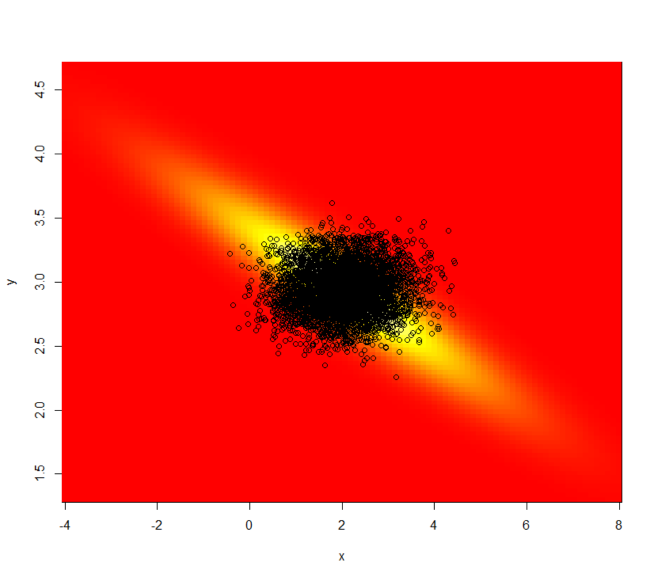
Naive idea

- IS using η_{VB} as a sampling distribution
- But: η_{VB} has a support smaller than the one of $\pi(\theta|\mathbf{Y})$

Naive idea 2

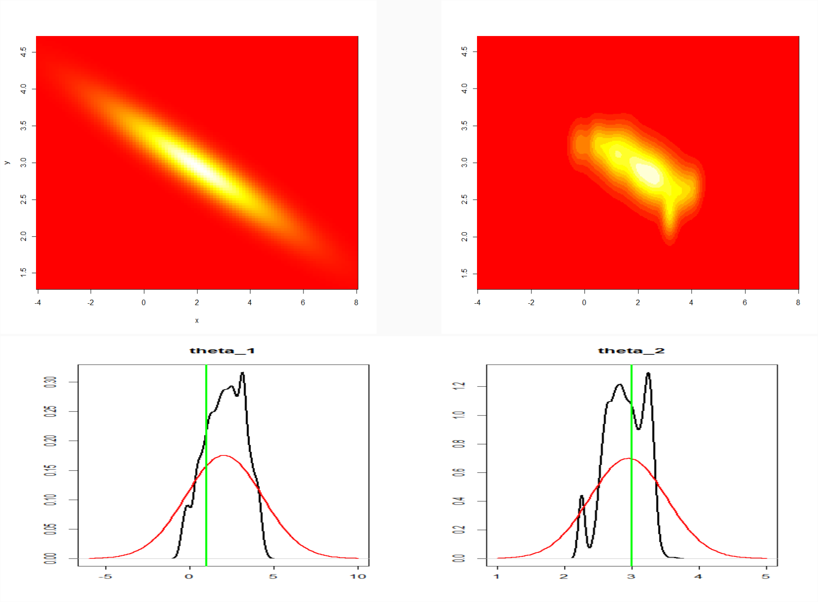
- Using a dilated version of η_{VB}
- Problems : how? how much?
- The problems of neglected dependencies remains

IS with η_{VB}



Particles simulated not weighted

IS with η_{VB}



One solution

Sequential sampling of a sequence of distributions

Let α_n be a increasing sequence such that $\alpha_0 = 0$ et $\alpha_N = 1$.
Sample sequentially

$$\pi_n(\theta) \propto \eta_{VB}(\theta)^{1-\alpha_n}(\ell(\mathbf{Y}|\theta)\pi(\theta))^{\alpha_n} = \frac{\gamma_n(\theta)}{Z_n}$$

using at each step n a sampling distribution η_n "judicious".

Remarks:

- $\log \pi_n(\theta) = Cste + (1-\alpha_n)\log \eta_{VB}(\theta) + \alpha_n \log(\ell(\mathbf{Y}|\theta)\pi(\theta))$
- $n = 0 : \pi_n(\theta) = \eta_{VB}(\theta)$. Easy to simulate.
- $n = N : \pi_n(\theta) \propto \ell(\mathbf{Y}|\theta)\pi(\theta) = \pi(\theta|\mathbf{Y})$: goal reached
- If α_n does not increase too fast $\pi_n(\theta) \approx \pi_{n+1}(\theta)$.

From η_n to η_{n+1}

Assume that at itération n, we have built η_n efficient for π_n :

$$\theta_n^{(1)}, \ldots, \theta_n^{(M)} \sim \eta_n(\theta)$$

At iteration $n+1$, we want to simulate $\pi_{n+1}(\theta)$ using that previous sample

- Intuition : if $\pi_n \approx \pi_{n+1}$ simulate $\theta_n^{(m)}$ in a neighbourood of $\theta_{n+1}^{(m)}$, i.e. using a Markovian kernel

  $$\theta_{n+1}^{(m)}|\theta_n^{(m)} \sim K_{n+1}(\theta_n^{(m)}, \theta_{n+1}^{(m)})$$

  Example : $\theta_{n+1}^{(m)} = \theta_n^{(m)} + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0, \rho^2)$
- New weights:

  $$w_{n+1}(\theta) = \frac{\gamma_{n+1}(\theta)}{\eta_{n+1}(\theta)}$$

But

After N iterations

- At iteration 0, simulate $\theta_0^{(1)}, \ldots, \theta_0^{(M)} \sim \eta_{VB}(\cdot) = \pi_0(\cdot)$
- At iteration 1, use the instrumental distribution:

$$\eta_1(\theta_1) = \int_\Theta \pi(\theta_0) K_1(\theta_0, \theta_1) d\theta_0$$

- At iteration N, use:

$$\eta_N(\theta_N) = \int_{\Theta^{N-1}} \pi(\theta_0) \prod_{n=1}^{N} K_n(\theta_{n-1}, \theta_n) d\theta_{0:N-1}$$

SMC by [Del Moral et al., 2006]

- Prove that one can apply the previous algorithm without having to comptue $\eta_n(\theta_n)$
- Main idea : introduce an atifical backward Markovian kernel $L_{n-1}(\theta_n, \theta_{n-1})$ such that $\int_\Theta L_{n-1}(\theta_n, \theta_{n-1}) d\theta_{n-1} = 1$
- Sample

$$\widetilde{\pi}_n(\theta_0, \ldots, \theta_{n-1}, \theta_n) = \pi_n(\theta_n) \prod_{k=0}^{n-1} L_k(\theta_{k+1}, \theta_k)$$

- By properties of the ‘backward" kernel, the marginal version of $\widetilde{\pi}_n(\theta_0, \ldots, \theta_n)$ is π_n

$$\begin{aligned}
\int_{\Theta^{n-1}} \widetilde{\pi}_n(\theta_0, \ldots, \theta_{n-1}, \theta_n) d\theta_{0:n-1} &= \int_{\Theta^{n-1}} \pi_n(\theta_n) \prod_{k=0}^{n-1} L_k(\theta_{k+1}, \theta_k) d\theta_{0:n-1} \\
&= \pi_n(\theta_n) \underbrace{\int_{\Theta^{n-1}} \prod_{k=0}^{n-1} L_k(\theta_{k+1}, \theta_k) d\theta_{0:n-1}}_{=1}
\end{aligned}$$

SMC by [Del Moral et al., 2006]

$$\widetilde{\pi}_n(\theta_0, \ldots, \theta_{n-1}, \theta_n) = \pi_n(\theta_n) \prod_{k=0}^{n-1} L_k(\theta_{k+1}, \theta_k) = \frac{\widetilde{\gamma}_n(\theta)}{\widetilde{Z}_n}$$

- Assume that at iteration $n-1$ we have $\left\{W_{n-1}^{(m)}, \theta_{1:n-1}^{(m)}\right\}$ approximating $\widetilde{\pi}_{n-1}$
- At time n, we propose

$$\theta_n^{(m)} \sim K_n(\theta_{n-1}^{(m)}, \theta_n^{(m)})$$

- $\eta_n(\theta_0, \ldots, \theta_n) = K_n(\theta_{n-1}, \theta_n)\eta_{n-1}(\theta_0, \ldots, \theta_{n-1})$
- Un-normalized weights:

$$\begin{aligned} w_n(\theta_{0:n}) &= \frac{\widetilde{\gamma}_n(\theta_{0:n})}{\eta_n(\theta_{0:n})} \\ &= \frac{\widetilde{\gamma}_{n-1}(\theta_{0:n-1})}{\eta_{n-1}(\theta_{0:n-1})} \frac{L_{n-1}(\theta_n, \theta_{n-1})}{K_n(\theta_{n-1}, \theta_n)} \frac{\gamma_n(\theta_n)}{\gamma_{n-1}(\theta_{n-1})} \\ &= w_{n-1}(\theta_{0:n-1})\widetilde{w}_{n-1}(\theta_{n-1}, \theta_n) \end{aligned}$$

Choosing K_n

- Independant kernels :

$$K_n(\theta_{n-1}, \theta_n) = K_n(\theta_{n-1})$$

$\Rightarrow$ Poorly efficient for complicated distributions : no learning.

- Local random network :

$$\theta_n = \theta_{n-1} + \mathcal{N}(0, \rho^2)$$

$\Rightarrow$ Choice of ρ^2? Does not use π_n

- MCMC type kernel : K_n such that π_n is invariant.
  - If $\pi_{n-1} \approx \pi_n$ and the chain moves fastly then we can hope that $\eta_n \approx \pi_n$.
  - But, anyway, the divergence between η_n and π_n is corrected.
  - Allows to use practical knowledge and theory from MCMC

Choosing L_{n-1}

- Purely artificial, but used to avoid the inegration against η_n when calculating the weights
- Price to pay: increase of the domain $\Theta \to \Theta^n$ and increasing of the weight variance
- Possibility to give the expression of the optimal L_{n-1}^{opt} minimizing the weigth variance $w_n(\theta_{0:n})$ (without explicit expression)
- In practice, look for $L_{n-1} \approx L_{n-1}^{opt}$ or the one simplifying the calculus

Choosing L_{n-1} for a MCMC type kernel i

- For K_n MCMC-kernel of stationnary ditribution π_n, on choose

$$L_{n-1}(\theta_n, \theta_{n-1}) = \frac{\pi_n(\theta_{n-1}) K_n(\theta_{n-1}, \theta_n)}{\pi_n(\theta_n)} = \frac{\gamma_n(\theta_{n-1}) K_n(\theta_{n-1}, \theta_n)}{\gamma_n(\theta_n)}$$

- Then

$$\begin{aligned}
\int_{\theta_{n-1}} L_{n-1}(\theta_n, \theta_{n-1}) d\theta_{n-1} &= \frac{\int_{\theta_{n-1}} \pi_n(\theta_{n-1}) K_n(\theta_{n-1}, \theta_n) d\theta_{n-1}}{\pi_n(\theta_n)} \\
&= \frac{\pi_n(\theta_n)}{\pi_n(\theta_n)} \text{ by stationarity of } \pi_n \ / \ K_n \\
&= 1
\end{aligned}$$

Choosing L_{n-1} for a MCMC type kernel ii

- Consequences on the non-normalized weights

$$
\begin{aligned}
w_n(\theta_{0:n}) &= w_{n-1}(\theta_{0:n-1}) \frac{L_{n-1}(\theta_n, \theta_{n-1})}{K_n(\theta_{n-1}, \theta_n)} \frac{\gamma_n(\theta_n)}{\gamma_{n-1}(\theta_{n-1})} \\
&= w_{n-1}(\theta_{0:n-1}) \frac{\gamma_n(\theta_{n-1})}{\cancel{\gamma_n(\theta_n)}} \frac{\cancel{\gamma_n(\theta_n)}}{\gamma_{n-1}(\theta_{n-1})} \\
&= w_{n-1}(\theta_{0:n-1}) \frac{\eta_{VB}(\theta_{n-1})^{1-\alpha_n} (\ell(\mathbf{Y}|\theta_{n-1})\pi(\theta_{n-1}))^{\alpha_n}}{\eta_{VB}(\theta_{n-1})^{1-\alpha_{n-1}} (\ell(\mathbf{Y}|\theta_{n-1})\pi(\theta_{n-1}))^{\alpha_{n-1}}} \\
&= w_{n-1}(\theta_{0:n-1}) \left[\frac{\ell(\mathbf{Y}|\theta_{n-1})\pi(\theta_{n-1})}{\eta_{VB}(\theta_{n-1})} \right]^{\alpha_n - \alpha_{n-1}}
\end{aligned}
$$

- Do not depend on θ_n : can be computed before the move.

Algorithm SMC : initialization

Initialization : $n = 0$

- Pour $m = 1 \ldots N$, $\theta_0^{(m)} \sim_{i.i.d} \eta_{VB}(\cdot)$
- Calculer $w_0^{(m)} = 1$ et $W_0^{(m)} = \frac{1}{M}$.

Algorithm SMC

At iteration n

- $\forall m = 1 \dots M$, calculate

$$w_n^{(m)} = w_{n-1}(\theta_{0:n-1}^{(m)}) \left[\frac{\ell(\mathbf{Y}|\theta_{n-1}^{(m)})\pi(\theta_{n-1}^{(m)})}{\eta_{VB}(\theta_{n-1}^{(m)})} \right]^{\alpha_n - \alpha_{n-1}}$$

- Deduce $W_n^{(i)}$ and compute the effective sample size: $ESS(W_n^{(i)})$.
- If $ESS > seuil$: $\theta_n^{(m)} = \theta_{n-1}^{(m)}$
- If $ESS < seuil$:
  - Resample: $\widetilde{\theta}_n^{(m)} \sim \sum_{i=1}^M W_n^{(i)} \delta_{\{\theta_{n-1}^{(i)}\}}$ and $w_n^{(m)} = 1 \ \forall m = 1 \dots M$.
  - Propagation: $\theta_n^{(m)} \sim K_n(\widetilde{\theta}_n^{(m)}, \cdot)$ where K_n is made of a kew iterations of a MH of stationnary distribution π_n.

Adaptative version

At each iteration, push α_n until the ESS falls under a threshold $ESS < seuil$.

At iteration n

- Find α_n such that: $\alpha_n = \inf_{\alpha > \alpha_{n-1}} \{ESS_n(\alpha) < seuil\}$ with

$$w_{n,\alpha}^{(m)} = \left[\frac{\ell(\mathbf{Y}|\theta_{n-1}^{(m)})\pi(\theta_{n-1}^{(m)})}{\eta_{VB}(\theta_{n-1}^{(m)})} \right]^{\alpha - \alpha_{n-1}}, \quad W_{n,\alpha}^{(m)} = \frac{w_{n,\alpha}^{(m)}}{\sum_{m=1}^{M} w_{n,\alpha}^{(m)}}$$

$$ESS_n(\alpha) = \frac{1}{\sum_{m=M}^{N} \left(W_{n,\alpha}^{(m)} \right)^2}$$

- Re-sample: $\widetilde{\theta}_n^{(m)} \sim \sum_{i=1}^{M} W_{n,\alpha_n}^{(i)} \delta_{\{\theta_{n-1}^{(i)}\}}$ and $w_n^{(m)} = 1 \ \forall m = 1 \dots M$.
- Propagate: $\theta_n^{(m)} \sim K_n(\widetilde{\theta}_n^{(m)}, \cdot)$ where K_n is made of a few iterations of a MH of stationnary distribution $\pi_n(\theta) = \eta_{VB}(\theta)^{1-\alpha_n} (\ell(\mathbf{Y}|\theta)\pi(\theta))^{\alpha_n}$.

Simulated data

- Mixture of 4-dimensional Bernoulli distributions
- n = number of individuals
- K = number of mixture components
- Y_{ij}: observation of individual i of component j.
- Z_{ik} : equal 1 if i belongs to group k. $Z_{i\,\bullet} = (Z_{i1}, \dots, Z_{iK})$
- $\forall i = 1, \dots, n, \forall j = 1 \dots 4$

$$Y_{ij} | Z_{i\bullet} \sim_{i.i.d} \mathcal{B}ern(Z_{i\bullet}\gamma_{\bullet\, j})$$

$$P(Z_i = k) = \pi_k$$

- $\theta = (\boldsymbol{\pi}, \gamma)$ with $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$ and γ probability matrix of size $K \times 4$

Prior and variational posterior

Prior

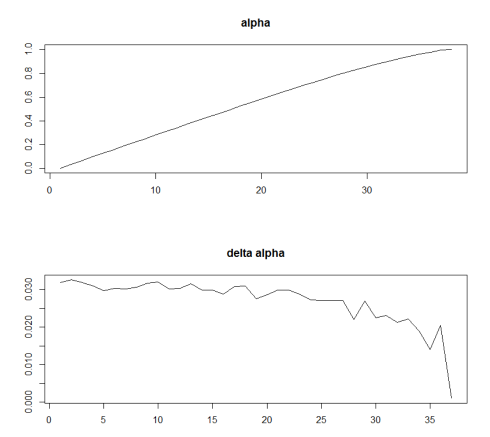
$$
\begin{aligned}
(\pi_1 \ldots, \pi_K) &\sim \mathcal{D}(1, \ldots, 1), \quad d_k \in \mathbb{R}^{+*} \\
\gamma_{kj} &\sim_{i.i.d.} \mathcal{B}(1, 1), \quad (j, k) \in \{1, \ldots, J\} \times \{1, \ldots, K\}
\end{aligned}
$$

Posterior distribution given by VBEM $\eta_{VB}(\theta, \mathbf{Z}|\mathbf{Y})$

$$
\begin{aligned}
(\pi_1 \ldots, \pi_K) &\sim \mathcal{D}ir(\widetilde{d}_1, \ldots, \widetilde{d}_K), \quad \widetilde{d}_k \in \mathbb{R}^{+*} \\
\gamma_{kj} &\sim_{i.i.d.} \mathcal{B}eta(\widetilde{a}_{kj}, \widetilde{b}_{kj}), \quad (j, k) \in \{1, \ldots, J\} \times \{1, \ldots, K\} \\
\mathbf{Z}_{i,\bullet} &\sim \mathcal{M}ult(\widetilde{\tau}_{i\bullet}), \quad \sum_{i=1}^{n} \widetilde{\tau}_{ik} = 1
\end{aligned}
$$

Tuning of the algorithm

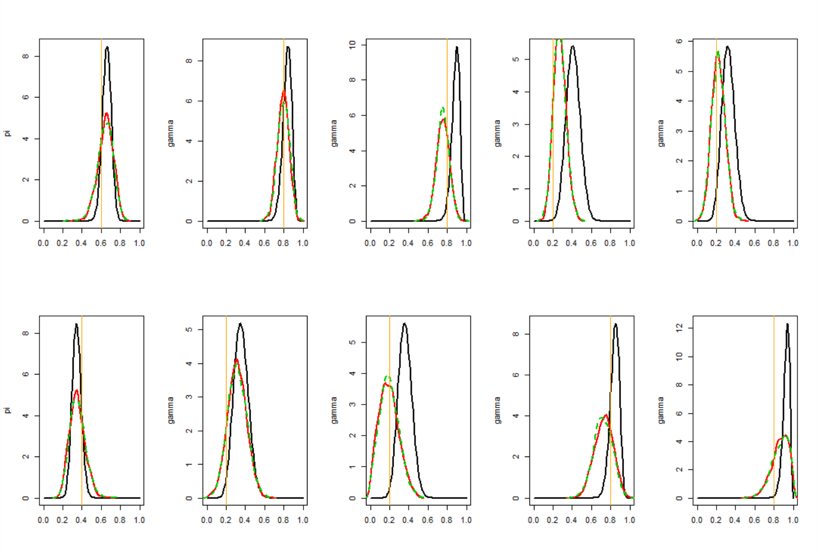
- $N = 2000$ particles
- Kernel K_n : 5 iterations of a standard Gibbs (explicit conditional distributions)
- ESS threshold: 1000 $\Rightarrow$ 39 iterations.
- Less than 5 minutes



Comparison with a standard Gibbs

- 5 chains, 39×2000 iterations to respect the tame computational budget
- Chains initialized on $\theta^{(0)} \sim \eta_{VB}(\cdot)$
- Convergence checked empirically
- In the end : thining $= 5$. Sample of size 2000.

Posterior



Black : VB, Green SMC, red MCMC

See [Donnet and Robin, 2017] for more examples

About latent variable models

- Latent variables naturally arise is many models
- Require specific inference methods because
  - the likelihood is not explicit anymore (NLME)
  - the likelihood can not be computed in a reasonnable time (SBM)
  - we are interested in the posterior distribution of the latent variables $p(\mathbf{Z}|\mathbf{Y})$ (mixture models)

About the Bayesian inference

- MCMC are VERY flexible tools to infer latent variable models
- Universal package for ANY model
- However
    - Reach their limit for models with large latent space.
    - For a complicated model the MCMC will require tunnings to make it converge, SMC may be more efficient
- People trying to propose universal tools for other methods to get the posterior distribution (INLA for gaussian latent variable models for instance...)
- New tools gathering all the possibilities : Stan, LaplaceDemon...

Références i

Del Moral, P., Doucet, A., and Jasra, A. (2006).
Sequential monte carlo samplers.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(3):411–436.

Donnet, S. and Robin, S. (2017).
Using deterministic approximations to accelerate smc for posterior sampling.

Moran, M., Walsh, C., Lynch, A., Coen, R. F., Coakley, D., and Lawlor, B. A. (2004).
Syndromes of behavioural and psychological symptoms in mild alzheimer's disease.
International Journal of Geriatric Psychiatry, 19(4):359–364.

Robert, C. and Casella, G. (1999).
Monte Carlo Statistical Methods.
Springer texts in statistics. Springer.