Bayesian calculus

Marie Etienne, Etienne Rivot

November 19, 2017

Our goal today

Getting the posterior distribution

A posteriori distribution is defined by

$$[\theta|y] = \frac{[y|\theta][\theta]}{[y]}$$

with $[y] = \int_\theta [y|\theta][\theta] d\theta$.

Getting the posterior distribution

A posteriori distribution is defined by

$$[\theta|y] = \frac{[y|\theta][\theta]}{[y]}$$

with $[y] = \int_\theta [y|\theta][\theta] d\theta$.

$[y]$ is mostly unavailable.

Getting the posterior distribution

A posteriori distribution is defined by

$$[\theta|y] = \frac{[y|\theta][\theta]}{[y]}$$

with $[y] = \int_\theta [y|\theta][\theta] d\theta$.

$[y]$ is mostly unavailable.

But nothing can stop us !!

Getting the posterior distribution

A posteriori distribution is defined by

$$[\theta|y] = \frac{[y|\theta][\theta]}{[y]}$$

with $[y] = \int_\theta [y|\theta][\theta] d\theta$.

$[y]$ is mostly unavailable.

But nothing can stop us !!

$$[\theta|y] = ?$$

Analytical posterior determination

Binomial example

- Data model :
$$Y \sim \mathcal{B}(n, p), \quad n \text{ known}$$
- Prior uniform
$$p \sim \mathcal{U}(0, 1)$$
-
$$[p|y] = ?$$

Normal example

- Model : $Y_k = \beta_0 + \beta_1 x_k + E_k, \quad E_k \overset{ind}{\sim} \mathcal{N}(0, \sigma^2)$
- Normal prior on $\theta = (\beta_0, \beta_1)$, ($\sigma^2$ assumed to be known)

$$[\beta_0, \beta_1] = \mathcal{N}(\mu_{prior}, \Lambda_{prior}),$$

  with Λ_{prior} denoting the precision matrix.
- Posterior distribution

Normal example

- Model : $Y_k = \beta_0 + \beta_1 x_k + E_k, \quad E_k \overset{ind}{\sim} \mathcal{N}(0, \sigma^2)$
- Normal prior on $\theta = (\beta_0, \beta_1)$, ($\sigma^2$ assumed to be known)

$$[\beta_0, \beta_1] = \mathcal{N}(\mu_{prior}, \Lambda_{prior}),$$

with Λ_{prior} denoting the precision matrix.

- Posterior distribution

$$[\beta_0, \beta_1 | y] \sim \mathcal{N}(\mu_{post}, \Lambda_{post})$$

with

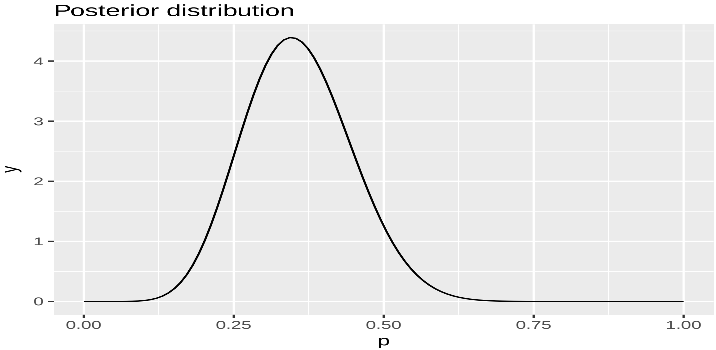
$$\Lambda_{post} = \left(\frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{\sigma^2} + \boldsymbol{\Lambda}_{prior} \right)$$

$$\mu_{post} = \left(\frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{\sigma^2} + \boldsymbol{\Lambda}_{prior} \right)^{-1} \left(\frac{\mathbf{X}^{\mathrm{T}}\mathbf{Y}}{\sigma^2} + \boldsymbol{\Lambda}_{prior} \boldsymbol{\mu}_{prior} \right)$$

Vérifier le calcul pour la cas des beta non indépendats

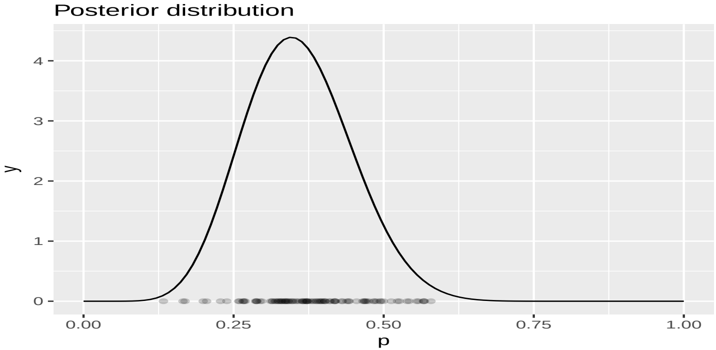
Sampling from posterior distribution

Why a sample is mostly enough ?



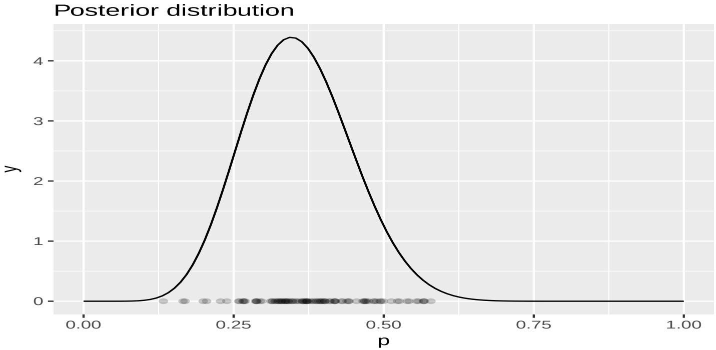
- $E[p|y] = ?$
- $CI_{0.95}(p) = ?$

Why a sample is mostly enough ?



- $E[p|y] \approx ?$
- $CI_{0.95}(p) \approx ?$

Why a sample is mostly enough ?



- $E[p|y] \approx ?$
- $CI_{0.95}(p) \approx ?$

```
##          Sum   Theory     MC100    MC1000
##         Mean 0.3571429 0.3803773 0.3552553
## 5%     CIInf 0.2166169 0.2270933 0.2133460
## 95%    CISup 0.5094782 0.5550791 0.5037494
```

Importance sampling algorithm

Importance sampling approach

Main idea :

$$E_{d_X}(h(X)) = \int_u h(u)\, d_X(u)du = \int_u h(u)\, \frac{d_X(u)}{d_Z(u)} d_Z(u)du$$
$$= \int_u h(u)\, \frac{d_X(u)}{d_Z(u)} d_Z(u)du = E_{d_Z}\left(h(Z)\frac{d_X(Z)}{d_Z(Z)}\right)$$

IS algorithm

1. Sample from proposal distribution : $(z_i)_{i=1,\ldots N}$.

IS algorithm

1. Sample from proposal distribution : $(z_i)_{i=1,\dots N}$.
2. For every particle i, Compute weight

$$w_i = d_X z_i / d_Z(z_i)$$

IS algorithm

1. Sample from proposal distribution : $(z_i)_{i=1,...N}$.
2. For every particle i, Compute weight

$$w_i = d_X z_i / d_Z(z_i)$$

3. Normalize weight

$$\tilde{w}_i = \frac{w_i}{\sum_{i=1}^{N} w_i}$$

IS algorithm

1. Sample from proposal distribution : $(z_i)_{i=1,\ldots N}$.
2. For every particle i, Compute weight

$$w_i = d_X z_i / d_Z(z_i)$$

3. Normalize weight

$$\tilde{w}_i = \frac{w_i}{\sum_{i=1}^{N} w_i}$$

$(z_i, \tilde{w}_i)$ is a weighted sample from d_X.

IS algorithm

1. Sample from proposal distribution : $(z_i)_{i=1,...N}$.
2. For every particle i, Compute weight

$$w_i = d_X z_i / d_Z(z_i)$$

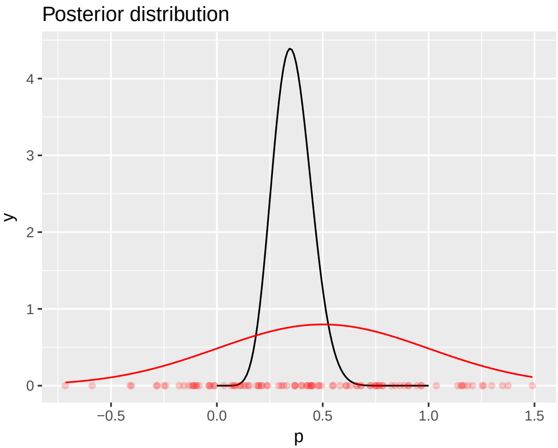
3. Normalize weight

$$\tilde{w}_i = \frac{w_i}{\sum_{i=1}^{N} w_i}$$

   $(z_i, \tilde{w}_i)$ is a weighted sample from d_X.
4. Resample to get unweighted sample. \ Sample in (z_i) with replacement with a probability $\tilde{w}_i$ to draw z_i.

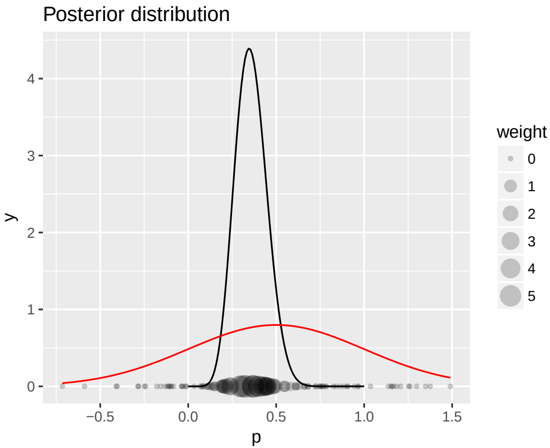
IS algorithm : graphical point of view

1. Step 1 : sample from proposal ($N = 100$)



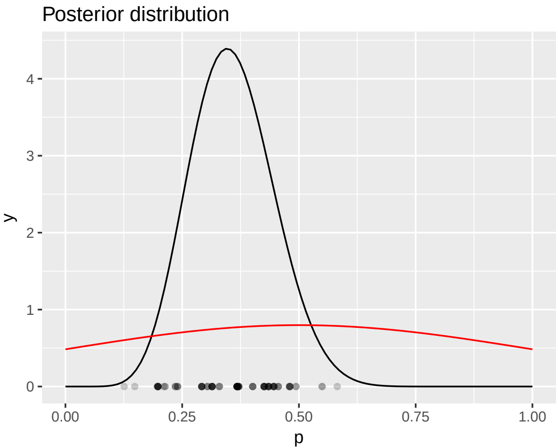
IS algorithm : graphical point of view

2. Step 2 : compute weight ($N = 100$)



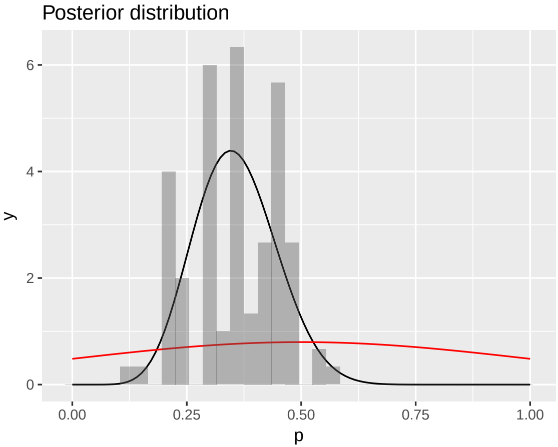
IS algorithm : graphical point of view

3. Step 3 : Resample to get unweighted sample ($N = 100$)



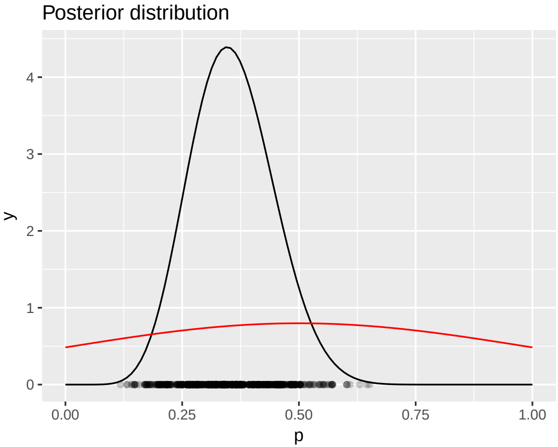
IS algorithm : graphical point of view

3. Step 3 : Resample to get unweighted sample ($N = 100$)



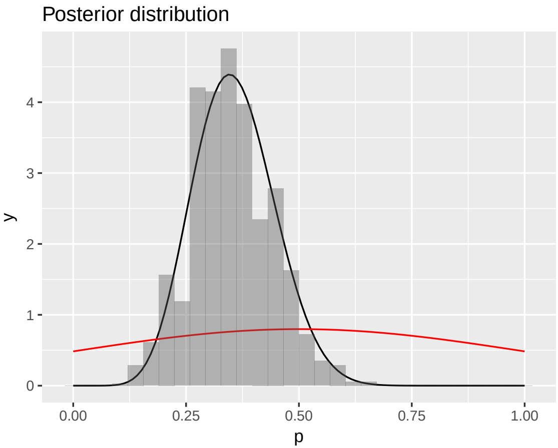
IS algorithm : graphical point of view

3. Step 3 : Resample to get unweighted sample ($N = 1000$)



IS algorithm : graphical point of view

3. Step 3 : Resample to get unweighted sample ($N = 1000$)



Monte Carlo Markov Chain algorithm (MCMC)

Markov chain definition

A Markov chain is a sequence of random variables $X_1, \ldots, X_n)$ verifying the Markov property.

Markov chain definition

A Markov chain is a sequence of random variables $X_1, \ldots, X_n)$ verifying the Markov property.

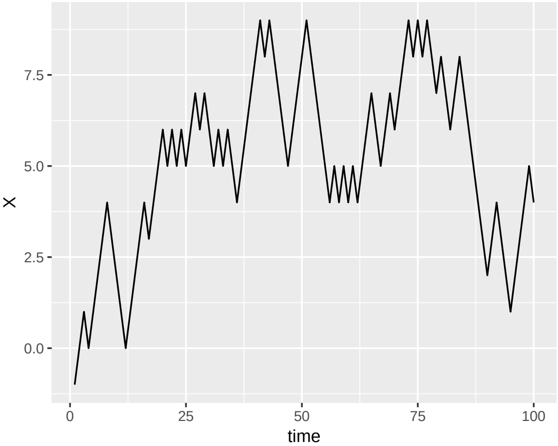
$$[X_{i+1} | X_{1:i}] = [X_{i+1} | X_i].$$

Markov chain example

Random walk

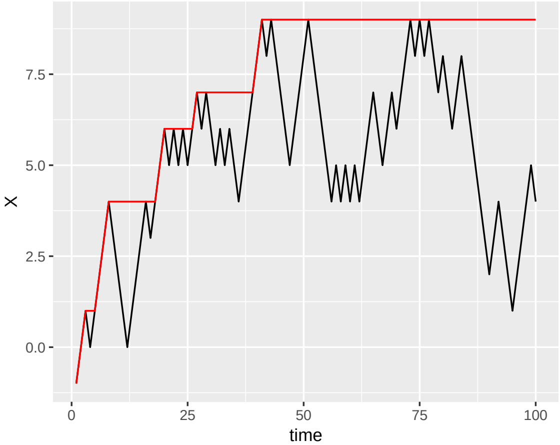
$$X_{i+1} = X_i + E_{i+1}, \quad E_{i+1} \overset{ind}{\sim} \mathcal{U}(\{-1, 1\})$$

(X_i) is a Markov chain.



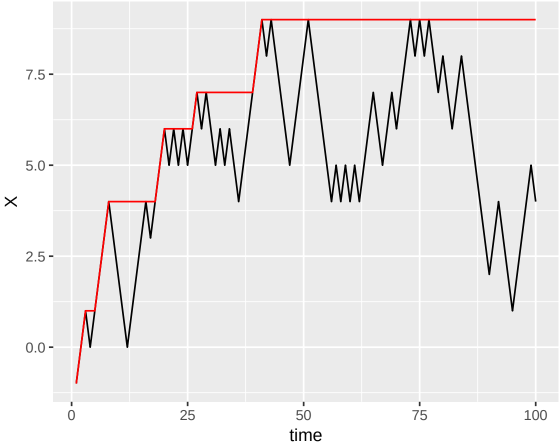
Markov chain example

Supremum of a random walk $Z_i = max_{k=1}^{i} max(X_k)$,



Markov chain example

Supremum of a random walk $Z_i = max_{k=1}^{i} max(X_k)$,



(Z_i) is not a Markov chain.

Markov chain properties

Definition : ν is a stationnay distribution if and only if

$$X_i \sim \nu \Longrightarrow X_{i+1} \sim \nu$$

Example :

$$X_1 \sim \mathcal{B}(p_{init}), \quad X_{i+1}|X_i \sim \mathcal{B}(p_{X_i})$$

Markov chain properties

Definition : ν is a stationnay distribution if and only if

$$X_i \sim \nu \Longrightarrow X_{i+1} \sim \nu$$

Example :

$$X_1 \sim \mathcal{B}(p_{init}), \quad X_{i+1}|X_i \sim \mathcal{B}(p_{X_i})$$

```
n     <- 100
pinit <- 1/3
pr    <- c(0.2, 0.6)
X     <- rep(NA,n)

X[1] <- sample(c(0,1), size=1, prob = c(1-pinit, pinit))
for( i in 1:(n-1)){
 X[i+1] <-   sample(x=c(0,1), size = 1,
                    prob = c(1-pr[X[i]+1], pr[X[i]+1 ]))
}
```

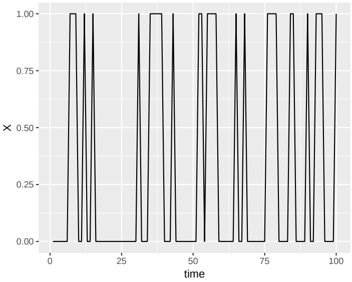
Markov chain properties

Definition : ν is a stationnay distribution if and only if

$$X_i \sim \nu \Longrightarrow X_{i+1} \sim \nu$$

Example :

$$X_1 \sim \mathcal{B}(p_{init}), \quad X_{i+1}|X_i \sim \mathcal{B}(p_{X_i})$$



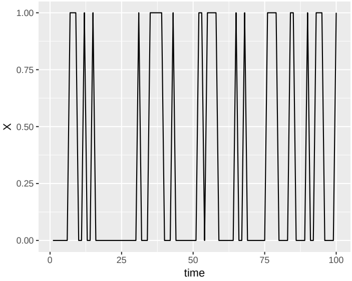
Markov chain properties

Definition : ν is a stationnay distribution if and only if

$$X_i \sim \nu \Longrightarrow X_{i+1} \sim \nu$$

Example :

$$X_1 \sim \mathcal{B}(p_{init}), \quad X_{i+1}|X_i \sim \mathcal{B}(p_{X_i})$$



Distribution of X_1, X_2, ... ?

Markov chain properties

Ergodic property :

If a Markov chain (X_i) is irreducible, aperiodic and recurrent then there is exists a unique stationnary distribution π and

$$[X_n] \underset{n \to \infty}{\longrightarrow} \pi.$$

If a Markov chain (X_i) is reversible $([X_i][X_{i+1}|X_i] = [X_{i+1}][X_i|X_{i+1}])$ then this markov chain has a stationnary distribution.

Consequences of the ergodic theorem

If (X_n) is a Markov chain with stationnary distribution, for any initial distribu $[X_1]$, $[X_n]$ is close to the stationnary distribution.

Consequences of the ergodic theorem

If (X_n) is a Markov chain with stationnary distribution, for any initial distribu $[X_1]$, $[X_n]$ is close to the stationnary distribution.

Back to the example : stationnary distribution is $\pi = (0.7, 0.3)$

```
freq = table(X)/n
print(freq)
```

```
## X
##    0    1
## 0.69 0.31
```

Metropolis Hastings algorithm

Key idea : building a reversible Markov chain with $[\theta|y]$ as stationnary distribution

Metropolis Hastings algorithm

Key idea : building a reversible Markov chain with $[\theta|y]$ as stationnary distribution

1. Initialization $\theta^{(0)}$ an admissible initial value
2. For i in 1:nIter

- Propose a new candidate value $\theta_c^{(i)}$ sampled from a proposal distribution $g(.|\theta^{(i-1)})$
- Compute Metropolis Hastings ratio

$$r_i = \frac{[y|\theta_c^{(i)}][\theta_c^{(i)}]}{[y|\theta^{(i-1)}][\theta^{(i-1)}]} \frac{g(\theta^{(i-1)}|\theta^{(i)})}{g(\theta_c^{(i)}|\theta^{(i-1)})}$$

  - Define

$$\theta^{(i)} = \begin{cases} \theta_c^{(i)} \text{ with probablity } min(r_i, 1) \\ \theta_c^{(i-1)} \text{ with probablity } 1 - min(r_i, 1) \end{cases}$$