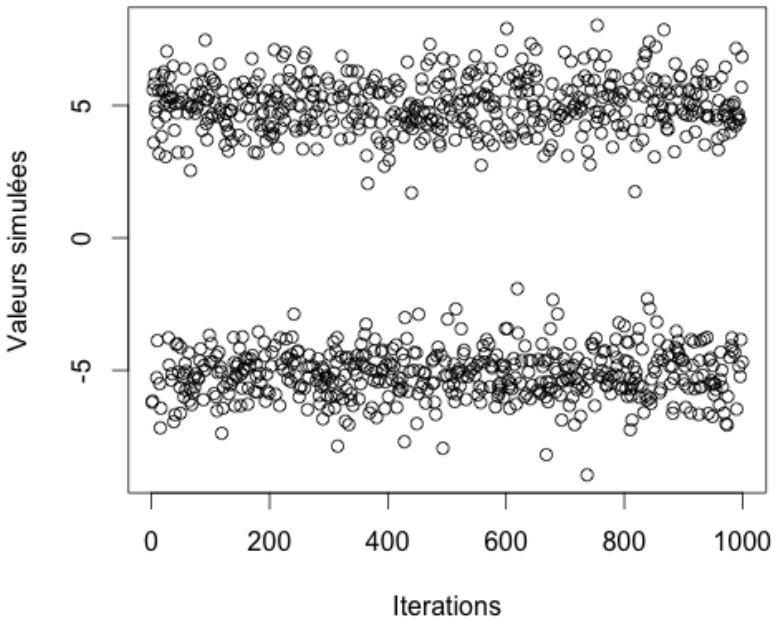
Modèle de mélanges

Gaëtan LE GALL, Joseph LAM et Nicolas HENNETIER

ENSAE Paristech

2016

Simulations d'un mélange de gaussiennes



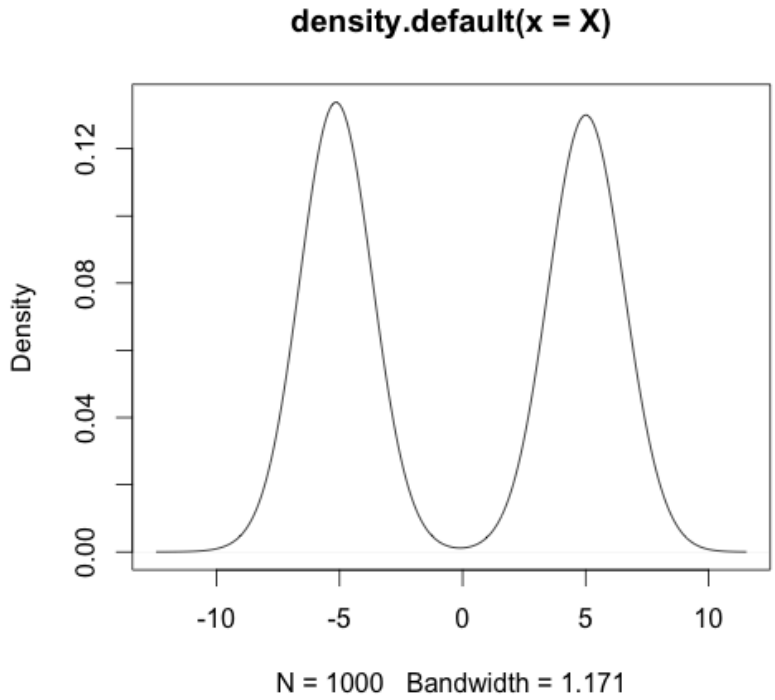


Figure – 1.000 simulations du GMM Figure – Densité estimée du modèle

Le modèle comporte 2 gaussiennes de moyennes -5 et 5 et de variance 1.

Définition des variables latentes

Soit $X_1, X_2, ..., X_n$ i.i.d. tq :

$$\forall 1 \leq i \leq n, f_{X_i}(x) = \frac{1}{K} \sum_{k=1}^{K} \varphi(x; \mu_k, 1)$$

Posons $Z_1, Z_2, ..., Z_n$ i.i.d. tq $X_i = \sum_{k=1}^{K} \mathbb{I}\{Z_i = k\} Y_{ki}$, où
$\forall 1 \leq k \leq K, \forall 1 \leq i \leq n, Y_{ki} \sim \mathcal{N}(\mu_k, 1)$ et
$\forall 1 \leq k \leq K, \mathbb{P}(Z_i = k) = \frac{1}{K}$

Lois *a posteriori*

On a, pour $1 \leq j \leq K$ et $1 \leq i \leq n$:

$$\begin{aligned} \mathbb{P}(Z_i = j|\mu, X_i) &\propto \pi(Z_i = j, X_i|\mu) \\ &\propto \pi(X_i|Z_i = j, \mu)\mathbb{P}(Z_i = j|\mu) \\ &\propto \pi(X_i|Z_i = j, \mu_j)\mathbb{P}(Z_i = j) \\ &\propto \mathcal{N}(X_i; \mu_j, 1) \end{aligned} \tag{1}$$

On déduit donc que $Z_i|\mu, X_i \sim \mathbb{D}((e^{\frac{1}{2}(X_i-\mu_j)^2})_{1\leq j\leq K})$

Lois *a posteriori*

D'autre part, on a :

$$\begin{aligned}
\pi(\mu|X,Z) &\propto \pi(X,Z|\mu)\pi(\mu) \\
&\propto \prod_{i=1}^{n} \pi(X_i,Z_i|\mu)\pi(\mu) \\
&\propto \prod_{i=1}^{n} \pi(X_i|Z_i,\mu) \prod_{k=1}^{K} \mathcal{N}(\mu_k;0,100) \\
&\propto \prod_{i=1}^{n} \mathcal{N}(X_i;\mu_{Z_i},1) \prod_{k=1}^{K} \mathcal{N}(\mu_k;0,100)
\end{aligned} \tag{2}$$

Lois *a posteriori*

Et donc, pour $1 \leq j \leq K$:

$$\begin{aligned}
\pi(\mu_j|X,Z) &\propto \prod_{Z_i=j} \mathcal{N}(X_i;\mu_j,1)\ \mathcal{N}(\mu_j;0,100) \\
&\propto e^{-\frac{1}{2}\frac{\mu_j^2}{100}-\frac{1}{2}\sum_{Z_i=j}(X_i-\mu_j)^2} \\
&\propto e^{-\frac{1}{2}(\frac{1}{100}+\#\{Z_i=j\})\mu_j^2+\mu_j\sum_{Z_i=j}X_i} \\
&\propto \mathcal{N}(\mu_j;\frac{\sum_{Z_i=j}X_i}{\frac{1}{100}+\#\{Z_i=j\}},\frac{1}{\frac{1}{100}+\#\{Z_i=j\}})
\end{aligned} \tag{3}$$

Algorithme de Gibbs-Sampling

Pour simuler selon les lois *a posteriori* $(\mu_1|X,Z),(\mu_2|X,Z),...,(\mu_K|X,Z)$, on peut donc utiliser l'algorithme suivant :

1) Choisir K points de départ $\mu_1^{(0)},\mu_2^{(0)},...,\mu_K^{(0)}$

2) À l'étape (t+1), simuler pour $1 \leq i \leq n$:

$$Z_i^{(t+1)}|\mu^{(t)},X_i \sim \mathbb{D}((e^{-\frac{1}{2}(X_i-\mu_j^{(t)})^2})_{1\leq j\leq K})$$

3) Puis simuler pour $1 \leq k \leq K$:

$$\mu_k^{(t+1)}|X,Z^{(t+1)} \sim \mathcal{N}\left(\frac{\sum_{Z_i^{(t+1)}=k} X_i}{\frac{1}{100}+\#\{Z_i^{(t+1)}=k\}},\frac{1}{\frac{1}{100}+\#\{Z_i^{(t+1)}=k\}}\right)$$

Trace plot des estimateurs

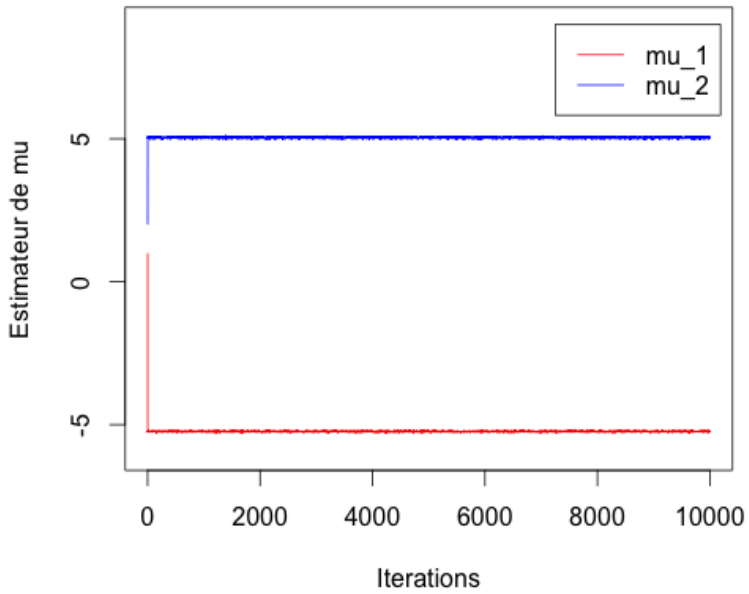


Figure – Simulation selon les lois *a posteriori* des moyennes

On remarque que la période de *burn in* est courte (moins de 100 itérations).

Trace plot des estimateurs

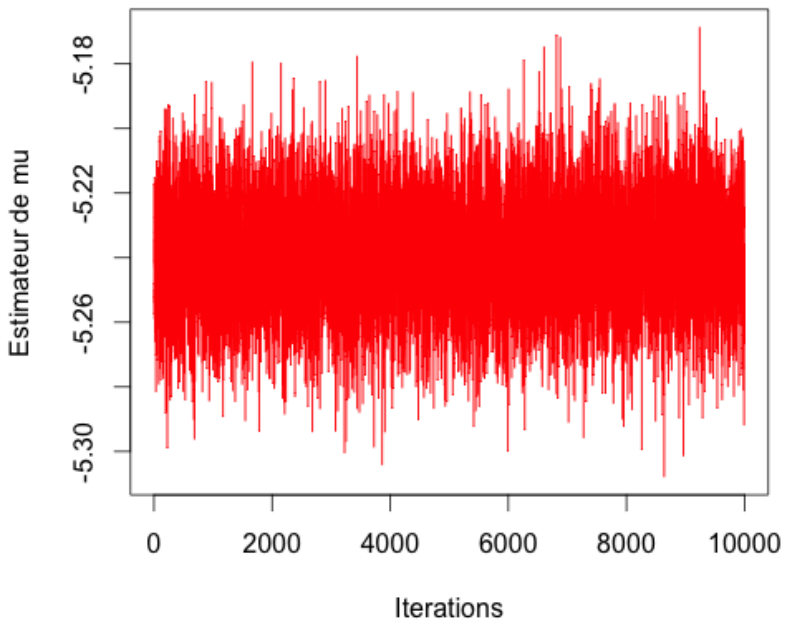


Figure – Estimateur 1

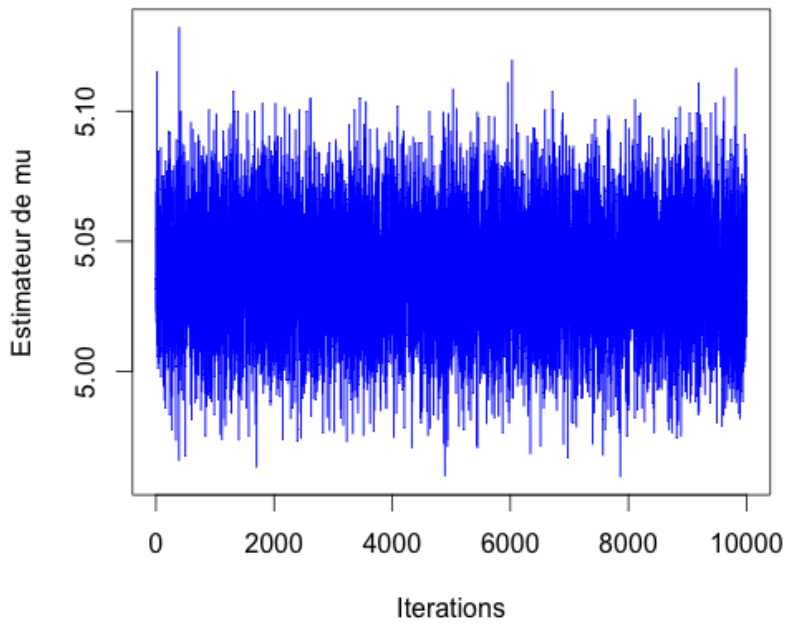


Figure – Estimateur 2

Simulation réalisées sur 10.000 itérations avec un *burn in* de 1.000 itérations.

Boxplot des estimateurs

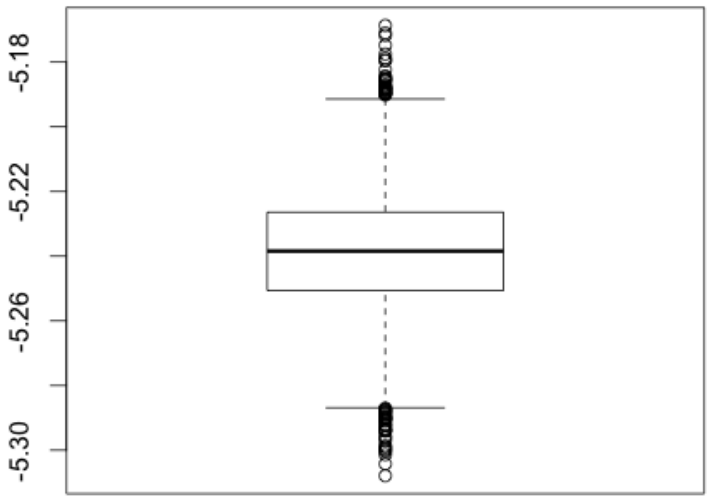


Figure – Estimateur 1

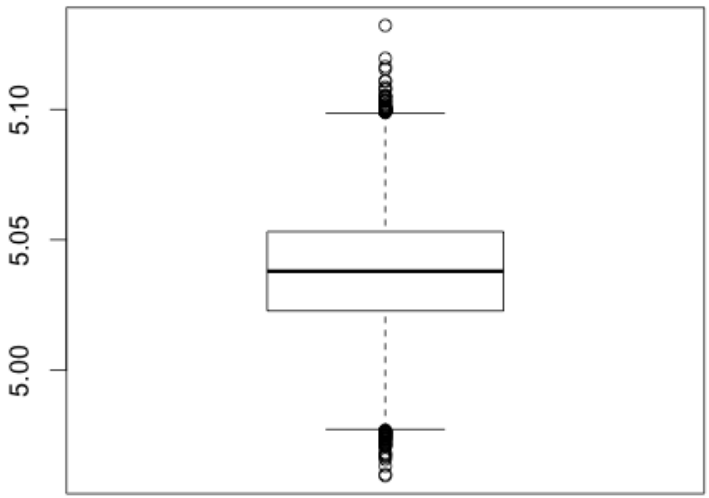


Figure – Estimateur 2

Les chaines simulées semblent suivre la loi *a posteriori* des paramètres.

Autocorrélations empiriques des estimateurs

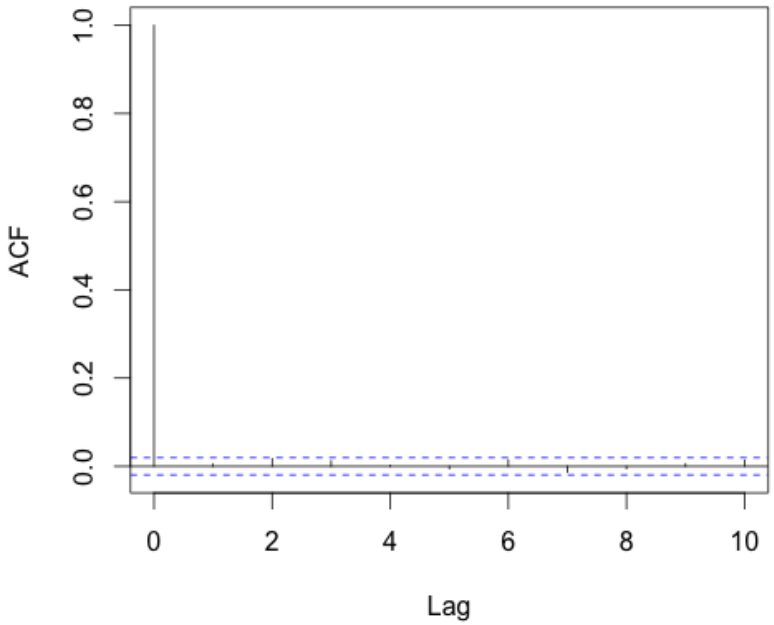


Figure – Estimateur 1

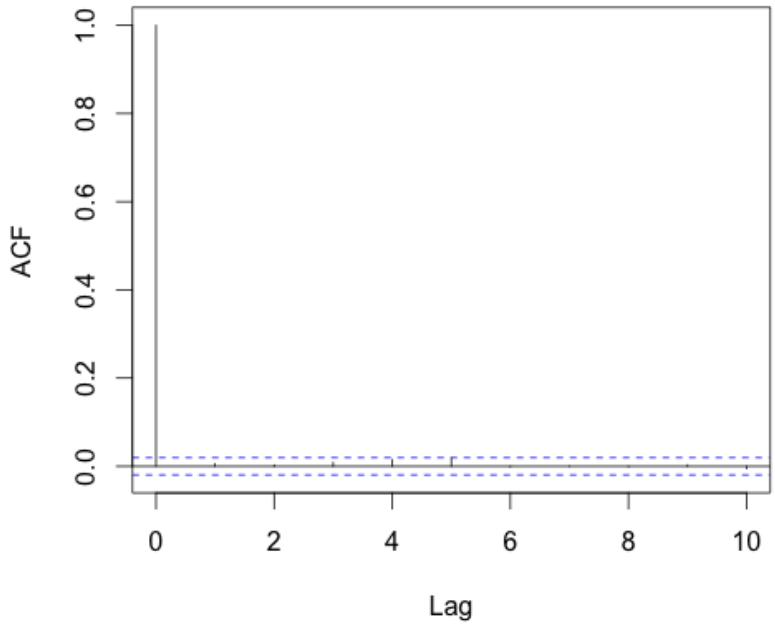


Figure – Estimateur 2

Les autocorrélations empiriques de nos chaines sont rapidement très faibles.

Gibbs-Sampling et le Label Switching

- Nos simulations semblent avoir de bonnes propriétés asymptotiques (*good mixing*, convergence vers des vraies valeurs des paramètres...)
- Remarquons que le modèle initial est invariant par permutation des $\mu_1, ..., \mu_K$:

$$\forall 1 \leq i \leq n, f_{X_i}(x) = \frac{1}{K} \sum_{k=1}^{K} \phi(x; \mu_k, 1)$$

- De plus, la loi *a priori* des $(\mu_k)_{1 \leq k \leq K}$ est indépendante de k (ce sont des $\mathcal{N}(0, 100)$)

$\Rightarrow$ La loi *a posteriori* des $(\mu_k)_{1 \leq k \leq K}$ doit donc être également indépendante de k.

Gibbs-Sampling et le Label Switching

- Le modèle possède donc $K!$ jeux de paramètres possibles qui maximisent la vraisemblance, toutes les permutations de $(\mu_k)_{1 \leq k \leq K}$.
- Notre simulation semble converger vers un unique jeu de paramètres sans permettre aux estimateurs de « switcher » entre les différentes valeurs possibles.
- La convergence vers un jeu de paramètres dépend très fortement des valeurs intiales de l'algorithme.

Gibbs-Sampling et le Label Switching

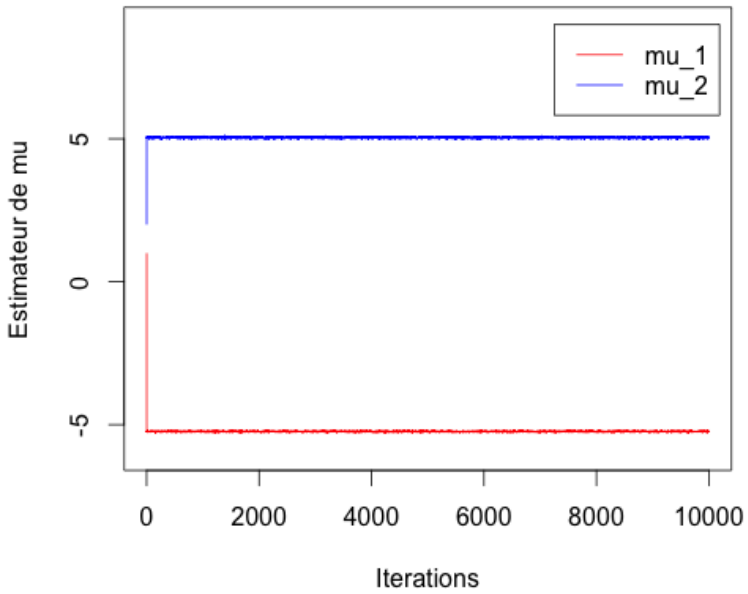


Figure – Simulations pour $(\mu_1^{(0)}, \mu_2^{(0)}) = (1, 2)$

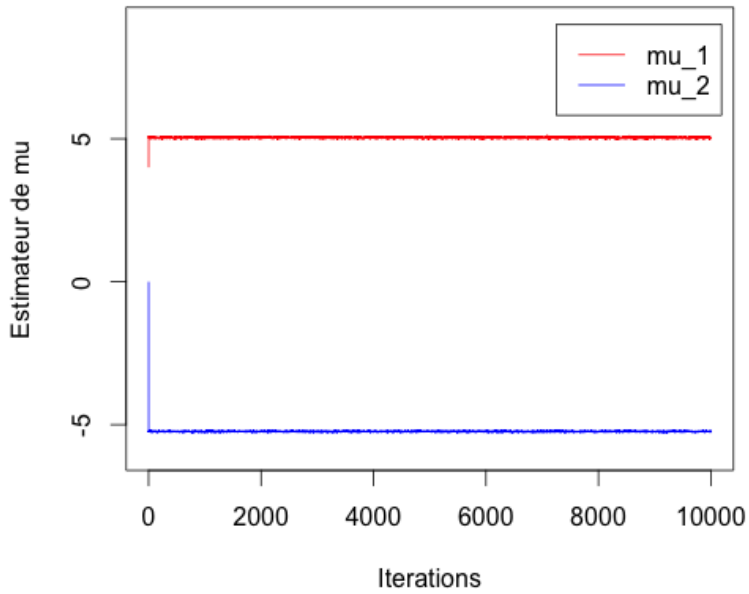


Figure – Simulations pour $(\mu_1^{(0)}, \mu_2^{(0)}) = (4, 0)$

Les simulations dépendent des valeurs initiales, ce qui n'est pas évident en regardant les lois *a posteriori* !

Loi *a posteriori* de μ

Toujours grâce à la formule de Bayes, on peut écrire :

$$\begin{aligned}\pi(\mu|X) &\propto \pi(\mu)\pi(X|\mu) \\ &\propto [\prod_{k=1}^{K} \mathcal{N}(\mu_k; 0, 100)][\prod_{i=1}^{n} \frac{1}{K} \sum_{k=1}^{K} \mathcal{N}(X_i; \mu_k, 1)] \end{aligned} \tag{4}$$

Algorithme de Metropolis-Hastings

Pour simuler selon la loi à posteriori $(\mu|X)$, on peut donc utiliser l'algorithme suivant :

1) Choisir K points de départ $\mu_1^{(0)}, \mu_2^{(0)}, ..., \mu_K^{(0)}$

2) À l'étape (t+1), on simule :

$$Z|\mu^{(t)} \sim \mathcal{N}(\mu^{(t)}, \sigma_0)$$

3) Puis on pose $\mu^{(t+1)} = Z$ avec la probabilité $min(1, \frac{\pi(Z|X)}{\pi(\mu^{(t)}|X)})$; sinon on pose $\mu^{(t+1)} = \mu^{(t)}$

Trace plot des estimateurs

Véritable problème de l'algorithme de Métropolis : deux effets antagonistes concernant le choix de σ_0 :

- $\sigma_0 >> 1$ pour inciter au *Label Switching* : la probabilité d'accepter en étape 3 est souvent très faible

$\Rightarrow$ Algorithme très lent à converger (de l'ordre de 123.600 itérations réelles pour accepter 500 fois)

- $\sigma_0 < 1$ pour maintenir un taux d'acceptation correct en étape 3

$\Rightarrow$ Plus de *Label Switching*

Simulations pour $\sigma_0 = 0.5$

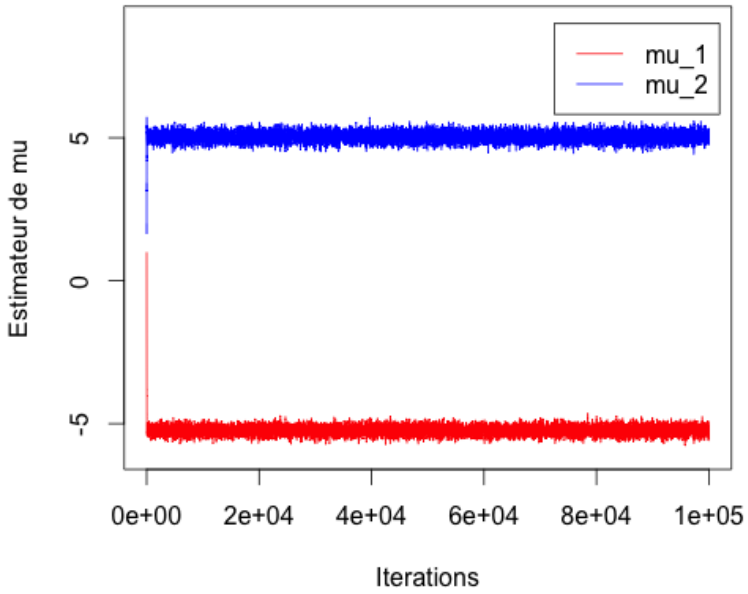


Figure – Simulations selon les lois *a posteriori*

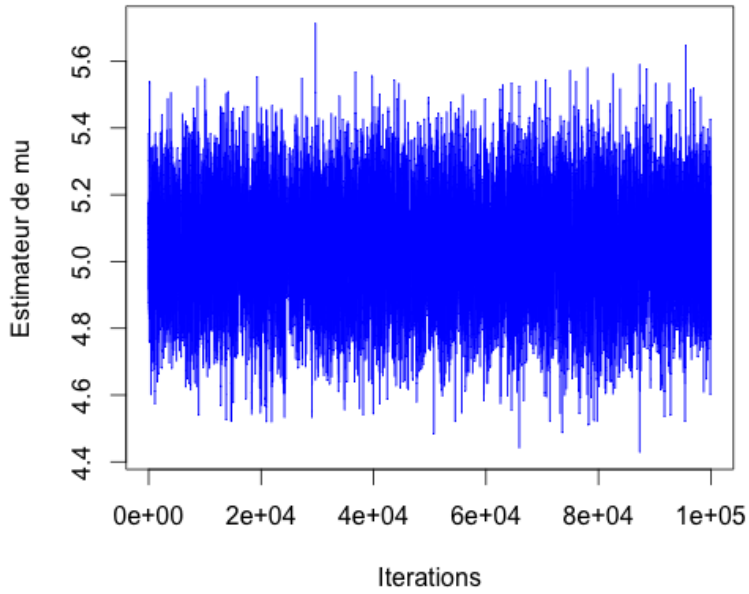


Figure – Simulation de μ_2 après la phase de *burn in*

Simulations réalisées sur 100.000 itérations avec un *burn in* de 10.000 itérations. Pas de *Label Switching* !

Simulations pour $\sigma_0 = 0.5$

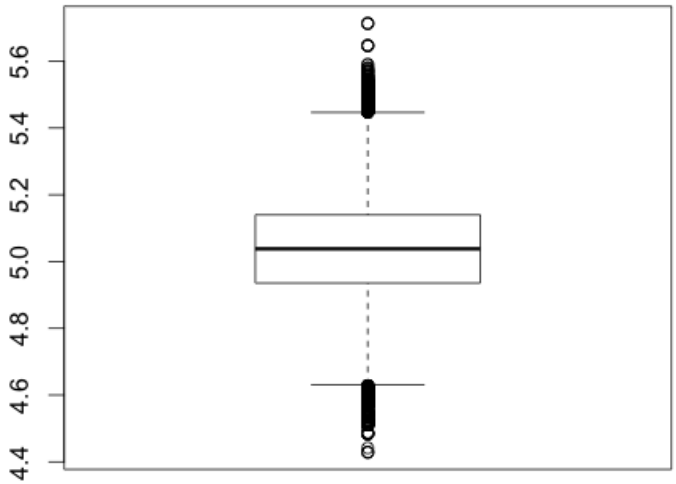


Figure – Boxplot de μ_2 après la phase de *burn in*

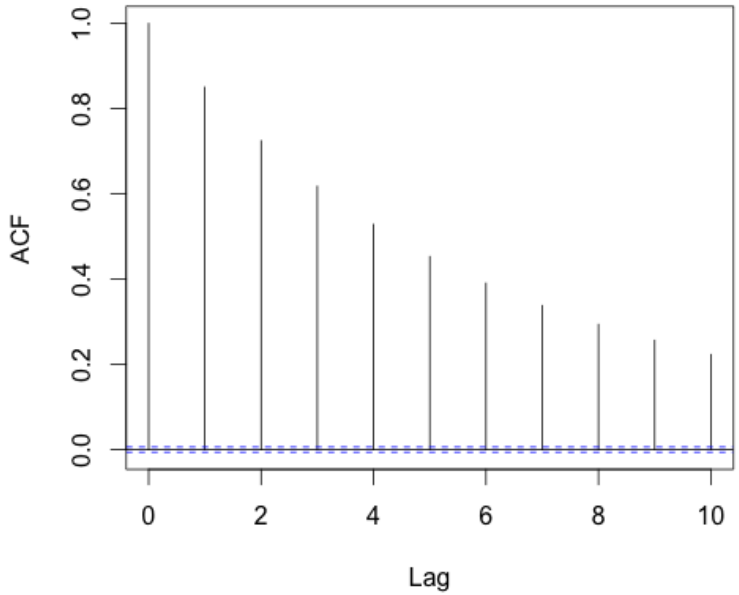


Figure – ACF de μ_2 après la phase de *burn in*

Les simulations semblent pourtant posséder de bonnes propriétés de *mixing* et de convergence.

Simulations pour $\sigma_0 = 10$

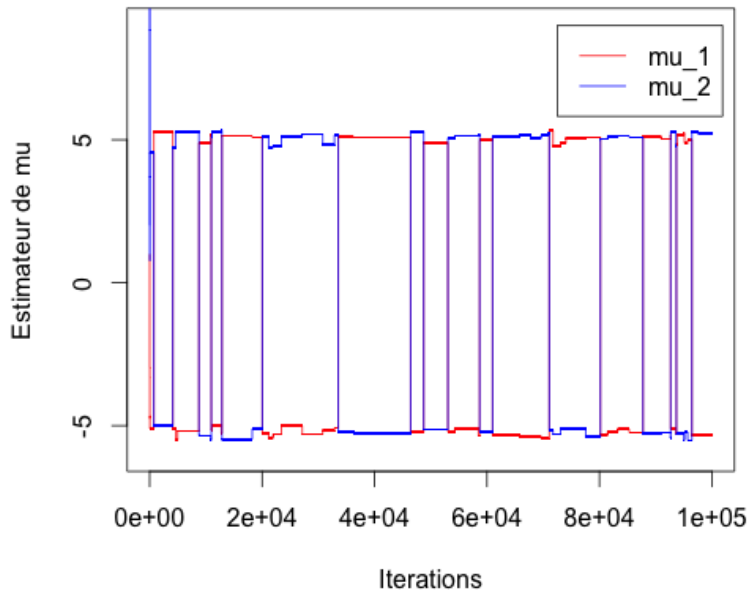


Figure – Simulations selon les lois *a posteriori*

Simulations réalisées sur 100.000 itérations avec un *burn in* de 10.000 itérations. On constate un fort *Label Switching*.

Simulations pour $\sigma_0 = 10$

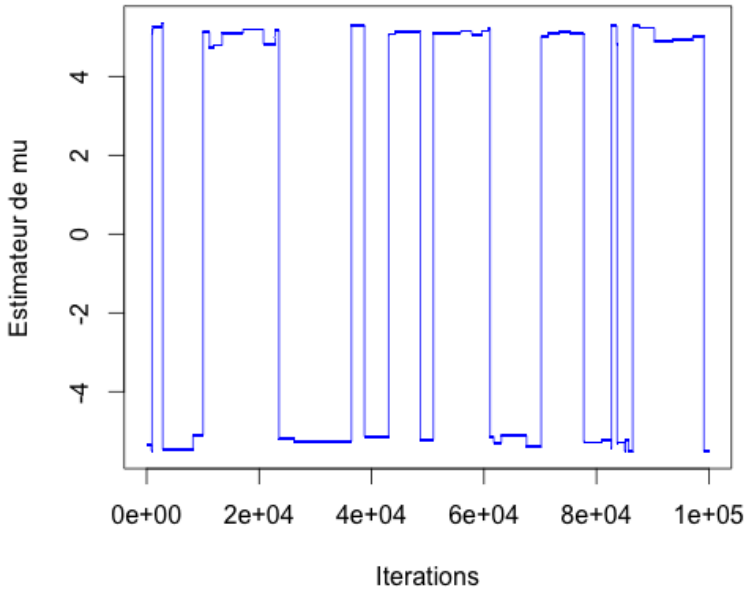


Figure – Simulation selon la loi *a posteriori* de μ_2

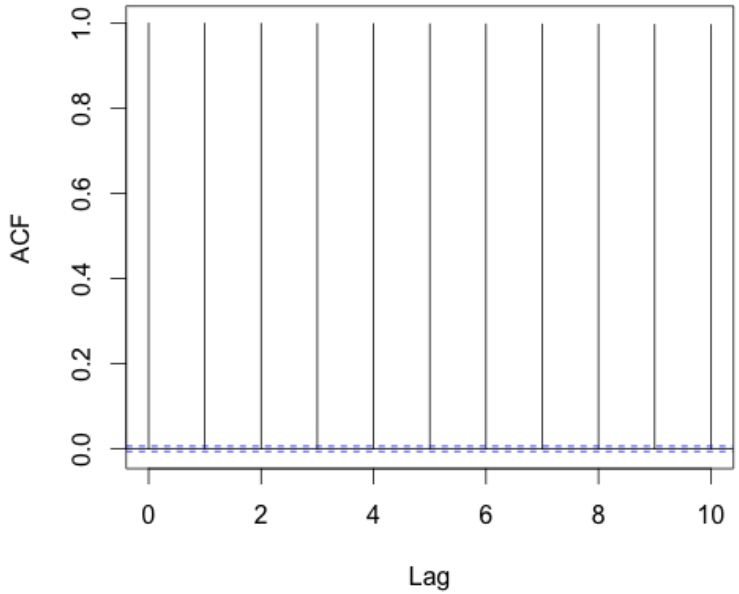


Figure – ACF de μ_2 après la phase de *burn in*

Les simulations n'ont plus du tout de bonnes propriétés de *mixing* (seulement 69 acceptations sur 100.000 simulations).

Introduction au Parallel Tempering

- Simuler N_{sweep} copies du système initialisé de façon aléatoire à différentes températures.
  La distribution *a posteriori* utilisée est :

$$\pi_T(x) = l(x)^{\frac{1}{T}} p(x)$$

  Si $T \to \infty$, on se ramène à la loi *a priori*.
- Appliquer Metropolis-Hastings sur les N_{sweep} systèmes de façon indépendante.
- Suivant le critère de Métropolis, échanger les configurations à différentes températures.

Echange des configurations

Pour 2 chaines indépendantes, de distributions cibles $f(x)$ et $g(y)$. La distribution est jointe est donnée par :

$$\pi(x, y) = f(x)g(y)$$

On prend $x^* = y$, $y^* = x$. La probabilité d'acceptation de cette transformation est : $min(1, A)$ avec

$$A = \frac{\pi(x^*, y^*)}{\pi(x, y)} = \frac{\pi(y, x)}{\pi(x, y)} = \frac{f(y)g(x)}{\pi(x, y)}$$

Simulations par Parallel Tempering

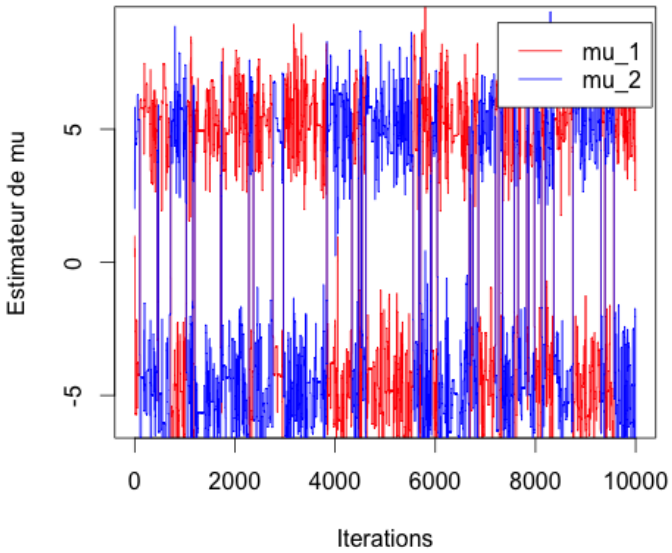


Figure – Simulations

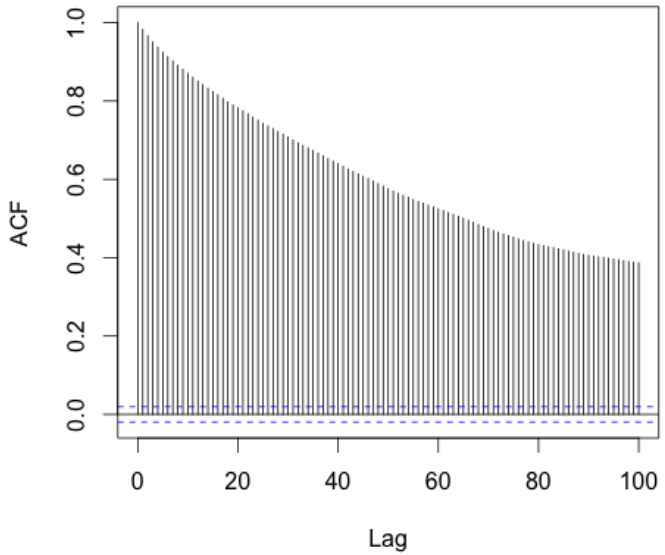


Figure – ACF de μ_1

Cet algorithme améliore les propriétés de *mixing* des chaines simulées (seulement 1.373 acceptations sur 10.000 simulations). Simulations pour $N_{sweep} = 100$.

Références

- Liang L., *On Simulation Methods for Two Component Normal Mixture Models under Bayesian Approach*, U.U.D.M. Project Report 2009 :17, septembre 2009
- A. Jasra, C. C. Holmes and D. A. Stephens, *Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modelling*
- Charles J. Geyer and Elizabeth A. Thompson, *Annealing Markov Chain Monte Carlo with Applications to Ancestral Inference*, Journal of the American Statistical Association Vol. 90, No. 431 (Sep., 1995), pp. 909-920
- Darren Wilkinson, *Parallel tempering and Metropolis coupled MCMC*, https ://darrenjw.wordpress.com/2013/09/29/parallel-tempering-and-metropolis-coupled-mcmc/, septembre 2013