Calcul Bayésien

Anne Philippe

Laboratoire de Mathématiques Jean Leray
Université de Nantes

24 mai 2010

Deux grandes étapes

1. Simuler des nombres aléatoires suivant une loi donnée
2. Exploiter ces données :

   $\rightsquigarrow$ **Méthode de Monte Carlo [MC]**

On a besoin de la simulation de variables aléatoires

- Crash de voitures
- Prévisions probabilistes
- Fiabilité : simulation d'événements rares
- Bootstrap
- Problèmes bayésiens ou non d'estimation
  le calcul des estimateurs, les régions de confiance etc
  $\rightsquigarrow$ développement de méthodes numériques basées sur le hasard :
  Méthodes de Monte Carlo

Problèmes spécifiques à l'approche bayésienne

- Calcul de la loi a posteriori $\pi(\theta|x) \propto \pi(\theta)f(x|\theta)$
- Calcul des quantités a posteriori

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

- etc

Techniques usuelles

$\rightsquigarrow$ méthodes d'analyse numérique

- méthode des trapèzes, quadrature par polynômes etc

Commentaires :

1. Difficulté à juger de la valeur de l'approximation.
2. au delà de la dimension 3, mauvaises performances.

Alternative :

Méthode de Monte Carlo

Idée centrale de Monte Carlo

On exprime l'intégrale que l'on cherche à approcher sous la forme

$$\int_{\mathcal{X}} h(x)f(x)\ \mathrm{d}x = \mathbb{E}(h(X)) \quad X \sim f$$

- f probabilité
- h une fonction intégrable [par rapport à f]

Clé : Tirer profit de la représentation de l'intégrale sous la forme d'une moyenne

Description de la méthode

- Simuler

$$X_1, \ldots, X_m \sim f(x)$$

- Utiliser [la loi des grands nombres]

$$I_m = \underbrace{\frac{1}{m}\sum_{i=1}^{m} h(X_i)}_{\textit{Moyenne empirique}} \xrightarrow{m\to\infty} \int_{\mathcal{X}} h(x)f(x)\,\mathrm{d}x$$

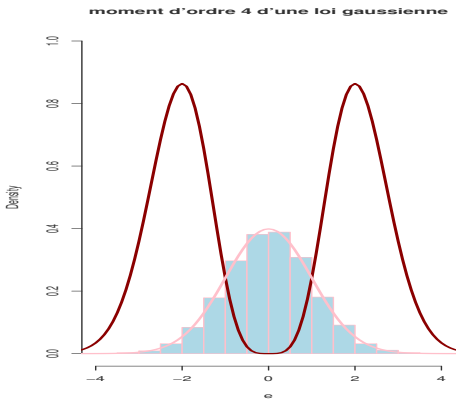
Propriétés : indépendantes de la dimension

- $\mathbb{E}(I_m) = \int_{\mathcal{X}} h(x)f(x)\,\mathrm{d}x$
- $\mathrm{Var}(I_m) = \frac{1}{m}\mathrm{Var}(h(X_1))$

Changement de loi

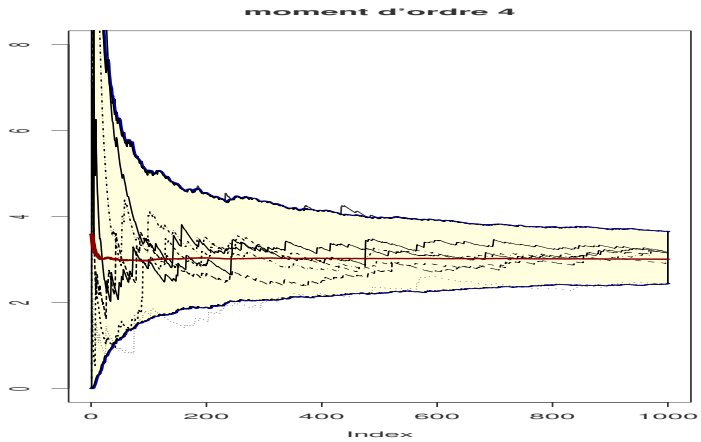
Exemple

On estime le moment d 'ordre 4 d'une loi gaussienne standard



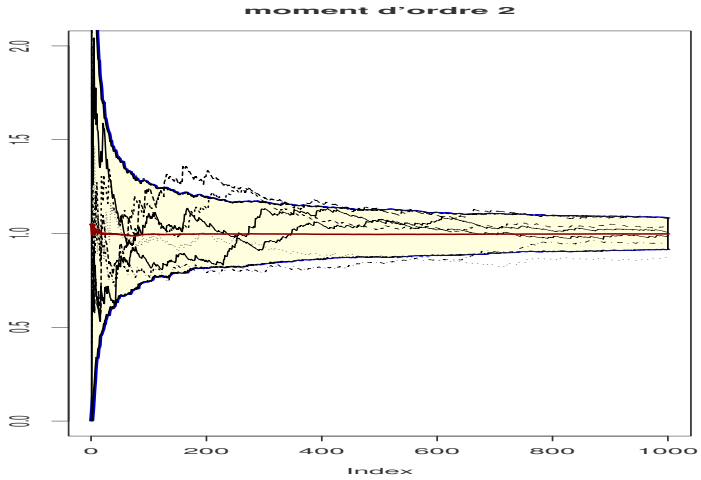
Représentation de $f(x)$ et $x^4 f(x)$

Estimation du moment d'ordre 4 par Monte Carlo



La variance est grande ! !

En effet, pour le moment d'ordre 2, on obtient



Simulation par changement de loi

On utilise la représentation

$$\int_{\mathcal{X}} h(x)f(x)\,\mathrm{d}x = \int_{\mathcal{X}} \frac{h(x)f(x)}{g(x)}g(x)\,\mathrm{d}x$$

g est une probabilité

1. Générer $y_1, \ldots, y_m \sim g$.
2. Utiliser l'approximation

$$\frac{1}{m}\sum_{i=1}^{m} h(y_i)\frac{f(y_i)}{g(y_i)},$$

Comment choisir la densité g

Choix optimal

$$g^*(x) = |h(x)|f(x)\frac{1}{\int_{\mathcal{X}} |h(x)|f(x)\,\mathrm{d}x}$$

Ce résultat est purement théorique car l'estimateur dépend de f/g^* et donc de l'intégrale $\int_{\mathcal{X}} |h(x)|f(x)\,\mathrm{d}x$ qui est inconnue ! !

En pratique ...

On prend g qui ressemble à la fonction $|h|f$ et $|h|f/g$ bornée

[finitude de la variance]

Estimation d'évènements rares

Exemple

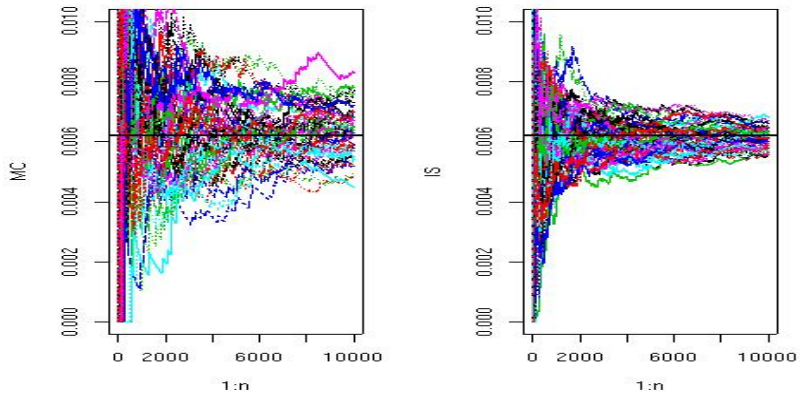
Queue de la distribution d'une va gaussienne

$$P(X > x) = \mathbb{E}(\mathbb{I}_{[x\infty[}(X))$$

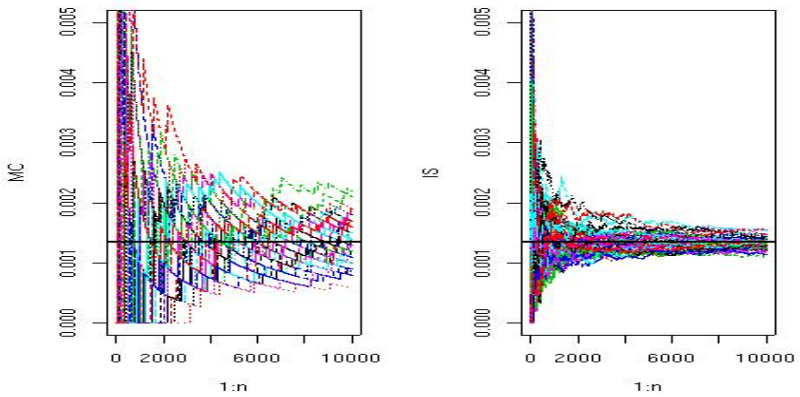
On va choisir une loi qui donne plus de poids que la gaussienne aux queues de la distribution
par exemple la loi de Cauchy.

Résultats pour $x = 2.5$

- A gauche : échantillon simulé suivant la loi gaussienne
- A droite : échantillon simulé suivant la loi de Cauchy



Résultats pour $x = 3$



Existence

On peut toujours simuler un échantillon suivant une loi donnée à partir

- d'un échantillon iid suivant la loi uniforme sur $[0, 1]$
- F^- le pseudo inverse de la fonction de répartition

Exemple : pour la loi exponentielle on a $F^-(U) = -\log(1 - U)$
En pratique : souvent F^- n'est pas connue explicitement ....

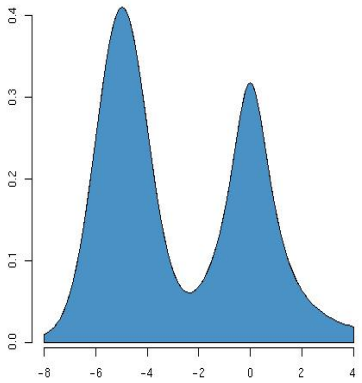
Théorème fondamental

Théorème

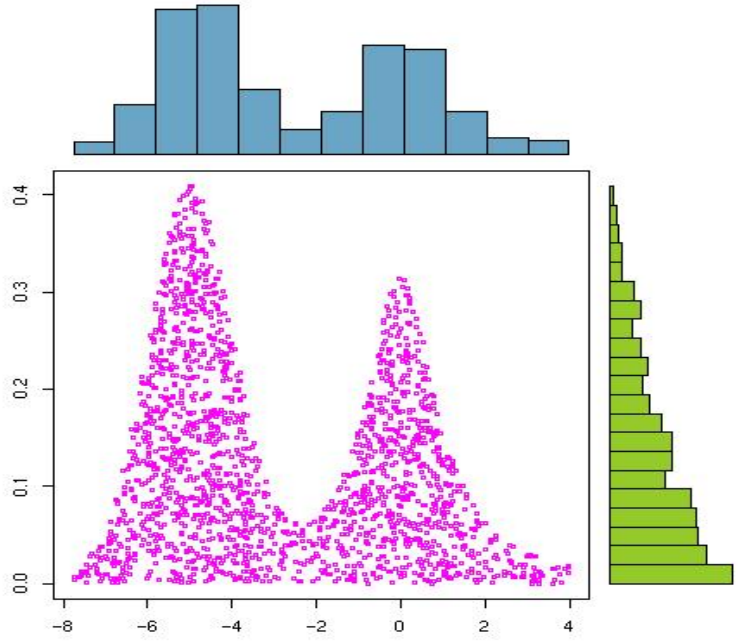
Soit h une fonction positive.
Si

$$(X, Y) \sim \mathcal{U}niforme\{(x, y) \ : \ 0 \leq y \leq h(x)\}$$

alors la loi marginale de X admet pour densité
$\frac{h}{\int h(t)\ \mathrm{d}t}$



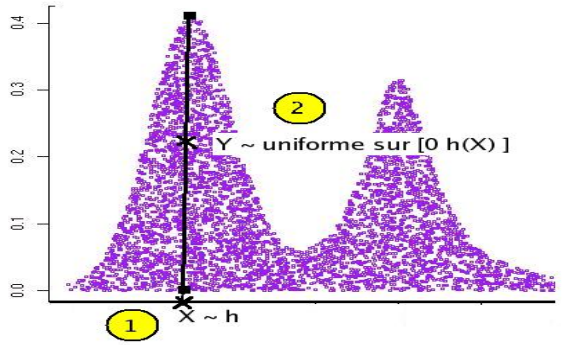
Le résultat



Algorithme

Pour simuler x suivant la loi uniforme sur

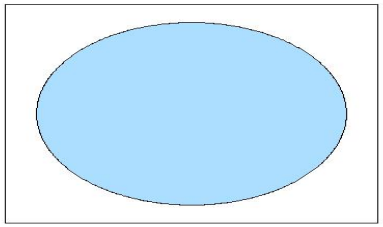
$$\{(x, y) \ : \ 0 \leq y \leq h(x)\}$$



1. On simule x suivant la loi de densité $\frac{h}{\int h(t) \ \mathrm{d}t}$
2. On simule y suivant la loi uniforme sur $[0, h(x)]$

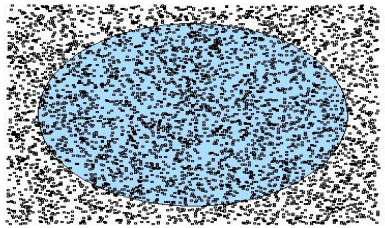
Cas particulier de la loi uniforme

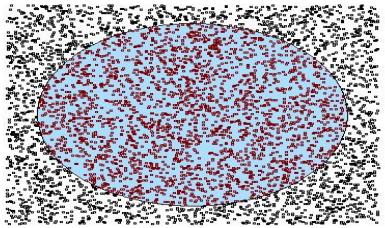
Pour simuler x suivant la loi uniforme sur A (région en bleue)

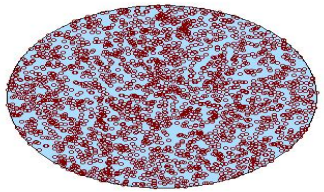


1. On simule y suivant la loi uniforme sur $[0, 1]^2$
2. Si $y \in A$ alors on pose $x = y$ sinon on recommence

Résultats





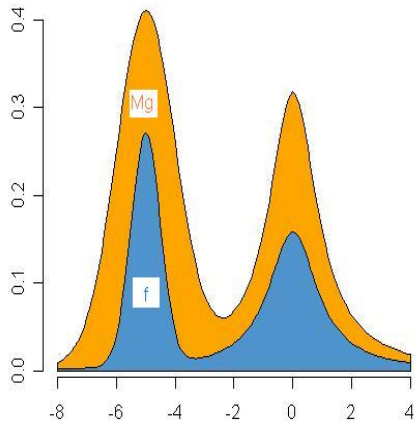


Algorithme d'acceptation rejet

On veut simuler $x \sim f$.

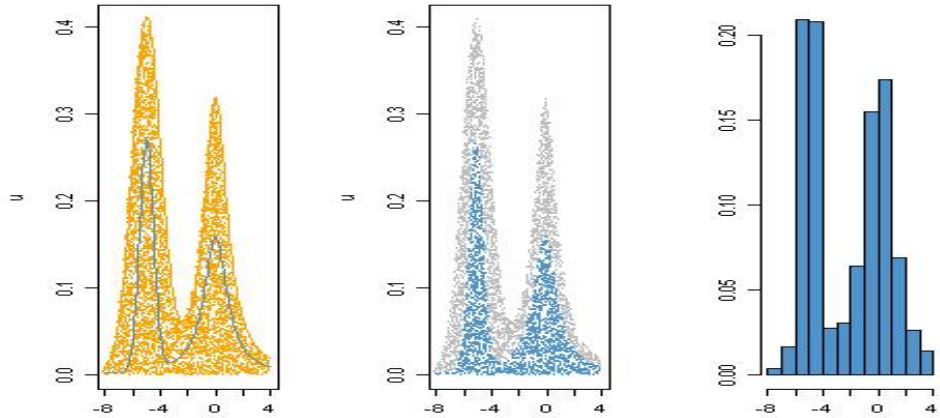
Ingrédient : g une probabilité telle que

- f/g est bornée par M
- on peut facilement simuler suivant g



L'algorithme

1. Générer $z \sim g(z)$ et $u \sim \text{Uniforme}[0, Mg(z)]$.
2. Si $u \leq f(z)$, on prend $x = z$;
   Sinon on répète 1.



Comment choisir g

- Le nombre moyen de variables simulées suivant g pour obtenir une variable suivant f est égal M

  $\rightsquigarrow$ M proche de 1 : l'algorithme est efficace

- En pratique on prend g qui ressemble à f
- Il n'est pas nécessaire de connaître la constante de normalisation de f
  Intérêt en bayésien il n'est pas utile de connaître la loi marginale des observations

Idée : Il est souvent plus facile de simuler

une chaîne de Markov de loi invariante f

que

des variables indépendantes de loi f

Chaînes de Markov -

Définition

Une chaîne de Markov est une suite de variables (X_t) telle que pour tout n

la loi conditionnelle de X_{n+1} sachant $(X_n, \ldots, X_0)$ et la loi conditionnelle de X_{n+1} sachant X_n coïncident.

Exemple

Soit (u_t) une suite de variables aléatoires iid

- on pose $X_o \sim \pi_0$ et $X_t = \rho X_{t-1} + u_t$
  (X_t) est une chaîne de Markov.
- Plus généralement : on peut prendre $X_t = g(X_{t-1}, u_t)$ où g est une application mesurable

Construction

Une chaîne de Markov est donc définie par

- la loi de X_0
- la loi de X_t sachant X_{t-1} (transition de la chaîne)
  - l'espace d'état E est fini : la transition est une matrice P
    $P(i,j) = P(X_1 = j | X_O = i)$
  - l'espace d'état E est continu : la transition est une famille de densités $\{K(x,\cdot),\ x \in E\}$ où $K(X_n,\cdot)$ est la densité de la loi conditionnelle de X_{n+1} sachant X_n

$$P(X_{n+1} \in A | X_n) = \int_A K(X_n, y)\, \mathrm{d}y$$

  - dans le cas général la transition est $P(x, A) = P(X_1 \in A | X_O = x)$
- $P^n(x, A)$ probabilité que $x_n \in A$ sachant que $x_0 = 0$

Propriétés des chaînes de Markov

chaîne irréductible : toute région d'intérêt de l'espace d'états peut être visitée.

- *transient* : nombre moyen de passages est fini
- *recurrent* : garantie de retour

Loi invariante : On dit que la chaîne admet une loi invariante s'il existe f telle que

si $x_n \sim f$ **alors** $x_{n+1} \sim f$

Les chaînes construites par les algorithmes MCMC admettent une loi invariante et elle est unique.

Notions de convergence

Étant donné une loi de probabilité de densité f
On cherche des transitions telles que

1. la loi invariante est unique et de densité f,
2. la loi de x_n est "proche" de la loi invariante f lorsque n assez grand et $x_0 \sim P_0 \neq f$
$$|| \int P^n(x, \ ) \ \mathrm{d}P_0(x) - f||_{TV} \to 0$$
3. Théorème ergodique
$$\frac{1}{T}\sum_{t=1}^{T} h(x^{(t)}) \xrightarrow[T\to\infty]{} \int h(x)f(x) \ \mathrm{d}x$$

Algorithme de Métropolis–Hasting

On veut simuler $x \sim f$.
Ingrédient : q une probabilité $q(\cdot|x^{(t)})$ telle que

$$\text{supp } q(\cdot|x) \supset \text{supp } f,$$

L'algorithme : partant de $x^{(t)}$, la transition de la chaîne est

$$x^{(t+1)} = \begin{cases} \xi \sim q(\xi|x^{(t)}) & \texttt{avec probabilité } \rho, \\ x^{(t)} & \texttt{sinon,} \end{cases}$$

où

$$\rho = \min\left\{1, \frac{f(\xi)q(\mathbf{x}^{(t)}|\xi)}{f(x^{(t)})q(\xi|x^{(t)})}\right\}$$

propriétés

Lorsque $\text{supp } q(\cdot|x) \supset \text{supp } f$,
$(x^{(t)})$ est

1. irréductible
2. ergodique,
3. de distribution invariante f.

Exemples de lois instrumentales

1. Lois indépendantes :
$$q(\xi|x) = g(\xi)$$
On "généralise" l'algorithme d'acceptation-rejet dans le sens où il y a moins de contraintes sur la loi instrumentale g
2. Lois symétriques :
$$q(\xi|x^{(t)}) = h(|\xi - x^{(t)}|)$$
Le taux d'acceptation ne dépend pas de la loi q,
$$\rho = \min\left(\frac{f(\xi)}{f(\theta^{(t)})}, 1\right).$$

les marches aléatoires

Un exemple de lois symétriques : partant de $x^{(t)}$, la transition s'écrit

$$\xi = x^{(t)} + \varepsilon$$

où $\varepsilon^{(t)}$ est une suite de variables aléatoires iid et indépendantes de $(X_0, \ldots, X_t)$
la loi de ε est symétrique par rapport à zéro,
par exemple

- les lois gaussiennes $\mathcal{N}(0, \sigma^2)$,
- les lois uniformes sur $[-a, a,]$
- etc...

Exemple originel de Hastings (1970)

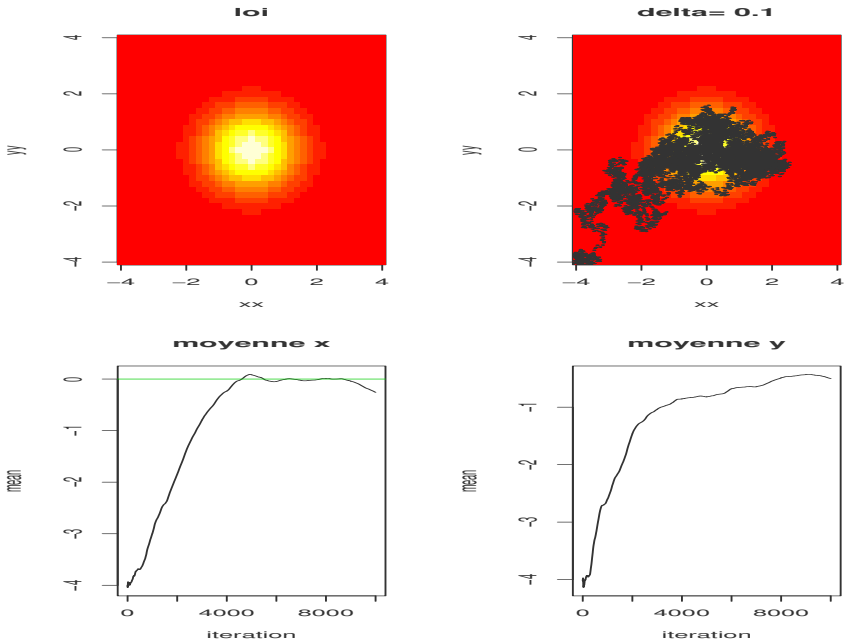
On simule un vecteur gaussien dans le plan $\mathcal{N}\left(0, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$

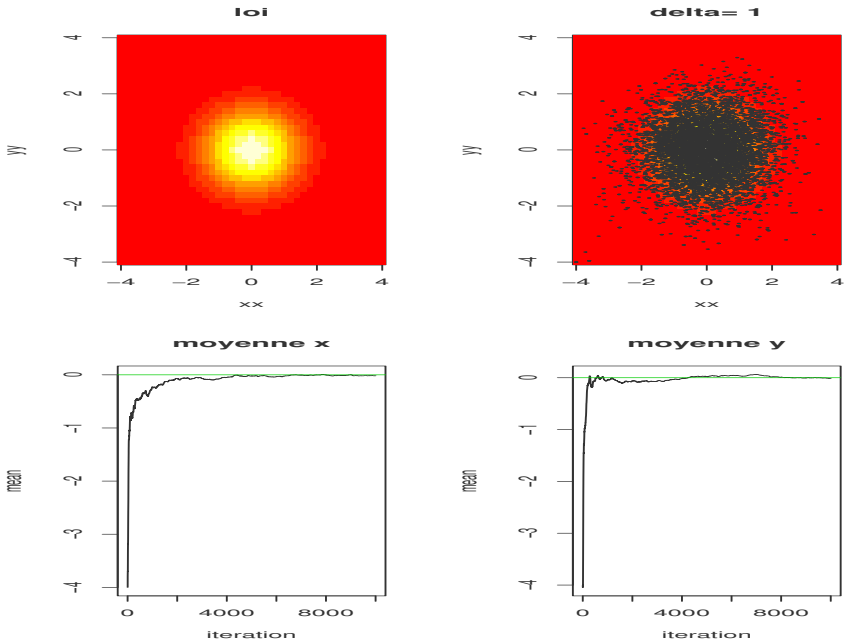
La loi instrumentale est symétrique, on prend une marche aléatoire construite à partir de la loi uniforme $\mathcal{U}([-\alpha, \alpha])^2$,

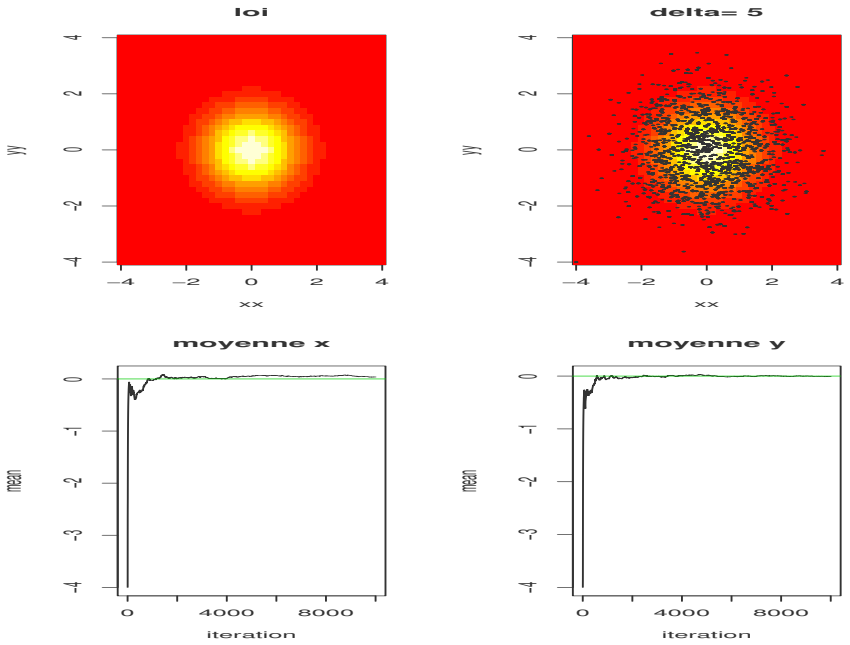
$$\xi = x^{(t)} + \alpha\varepsilon_t \ , \quad \varepsilon_t \sim \mathcal{U}_{[-1,1]^2}$$

On a

$$\rho = \min(\exp\{(x^{(t)^2} - \xi^2)/2\}, 1)$$



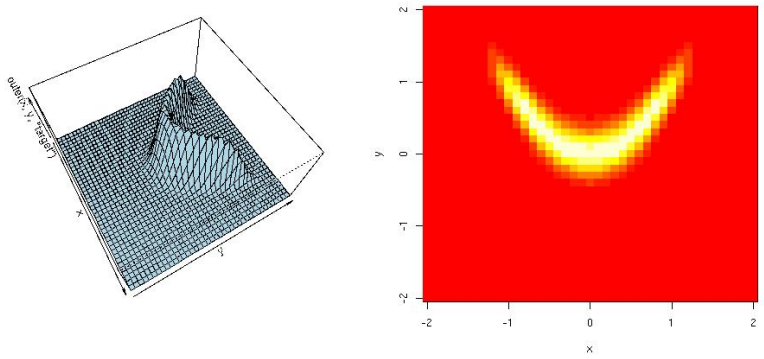




Un second exemple

On simule un échantillon suivant une loi de densité dans le plan.

$$f(x,y) \propto \exp\{-10(x^2-y)^2-(y-1/4)^4\}$$



L'algorithme

On utilise un algorithme d'Hasting Métropolis dont la loi instrumentale est une marche aléatoire gaussienne
Partant de x_t, la valeur proposée ξ est distribuée suivant

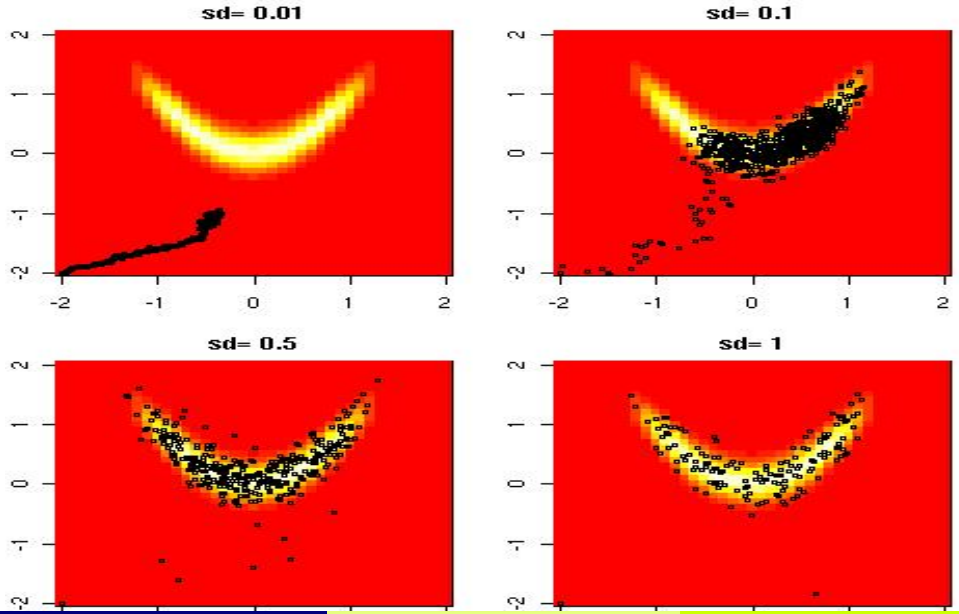
$$\xi \sim \mathcal{N}\left(x_t, \begin{pmatrix}\sigma^2 & 0\\ 0 & \sigma^2\end{pmatrix}\right)$$

c'est à dire

$$\xi = x_t + \sigma\varepsilon_t \qquad \varepsilon_t \sim \mathcal{N}\left(0, \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right)$$

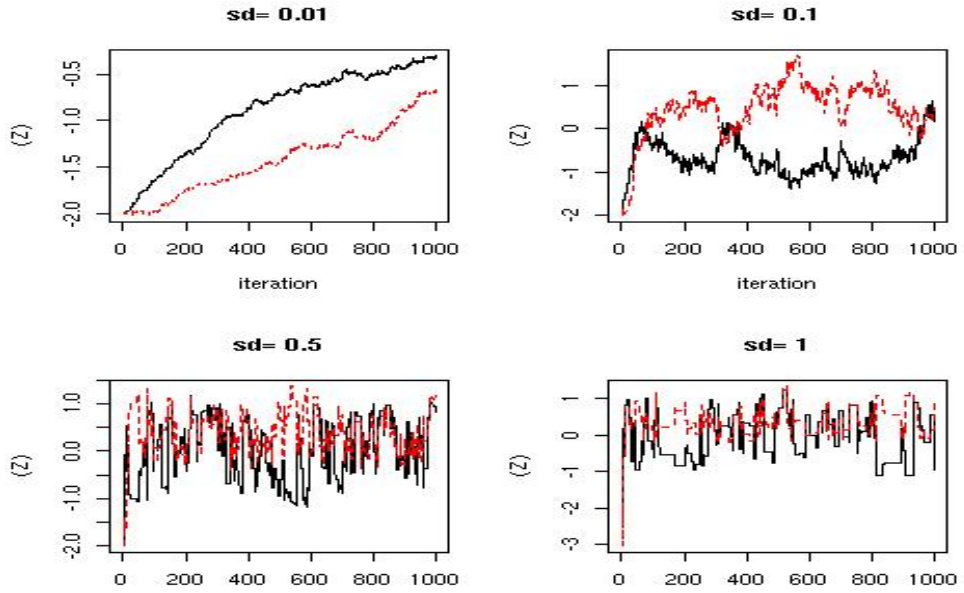
sites visités par la chaîne

Nombre d'itérations = 1000.



Trajectoires des deux coordonnées

Nombre d'itérations = 1000.



L'algorithme : la transition de la chaîne est

1. $x_1^{(t+1)} \sim f_1(x|x_2^{(t)}, \ldots, x_m^{(t)})$

$\ldots$

i. $x_i^{(t+1)} \sim f_i(x|x_1^{(t+1)}, \ldots, x_{i-1}^{(t+1)}, x_i^{(t)} x_m^{(t)})$

$\ldots$

m. $x_m^{(t+1)} \sim f_m(x|x_1^{(t+1)}, \ldots, x_{m-1}^{(t+1)})$

Algorithme de Gibbs

On veut simuler $x \sim f$ en dimension $p > 1$! ! !
Outil central : une décomposition de $x \in \mathbb{R}^p$ en m blocs

$$x = (x_1, \ldots, x_m)$$

telle que

les lois conditionnelles

$$f_j(x_j|x_\ell, \ell \neq j)\,, \qquad j = 1, \ldots, m$$

sont simulables.

Exemple

La loi à simuler admet pour densité

$$f(x, y) \propto e^{-x-y-xy}\, \mathbb{I}_{R^+ \times \mathbb{R}^+}(x, y)$$

On calcule les deux lois conditionnelles

- la loi de X sachant Y est la loi exponentielle de paramètre $Y + 1$
- la loi de Y sachant X est la loi exponentielle de paramètre $X + 1$

L'algorithme :

Initialisation $Y_0 \sim \mathcal{E}(1)$

Itération 1. $X_1 \sim \mathcal{E}(1 + Y_0)$ et $Y_1 \sim \mathcal{E}(1 + X_1)$

$\vdots$

Intération i. $X_i \sim \mathcal{E}(1 + Y_{i-1})$ et $Y_i \sim \mathcal{E}(1 + X_i)$

Propriétés

Si $f(x) > 0$ sur E alors $(x^{(t)})$ est une chaîne de Markov

1. ergodique
2. de loi invariante $f(x)$
3. Harris récurrentes *[lorsque les densités sont continues]*.

Augmentation de la dimension

Lorsque

f est une probabilité sur $\mathbb{R}$

ou/et les lois conditionnelles sont difficiles à simuler

on peut augmenter la dimension pour simplifier la simulation

$$f(x) = \int g(x, y)\, \mathrm{d}y$$

On simule alors $(x^{(t)}, y^{(t)})$ une chaîne de Markov de loi stationnaire g via un algorithme de Gibbs.
La suite $(x^{(t)})$ est aussi une chaîne de Markov de loi stationnaire f.

un second intérêt à l'augmentation de la dimension

On peut améliorer l'approximation de l'intégrale $\mathbb{E}(h(X))$ avec $X \sim f$ en prenant

$$\mathbb{E}[h(X)] \simeq \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[h(X)|y^{(k)}]$$

En effet

$$\mathbb{E}(h(X)) = \mathbb{E}(\underbrace{\mathbb{E}(h(X)|Y)}_{\text{fonction de } Y})$$

Il y a réduction de la variance

[théorème de Rao-Blackwell]

Exemple : loi normale tronquée

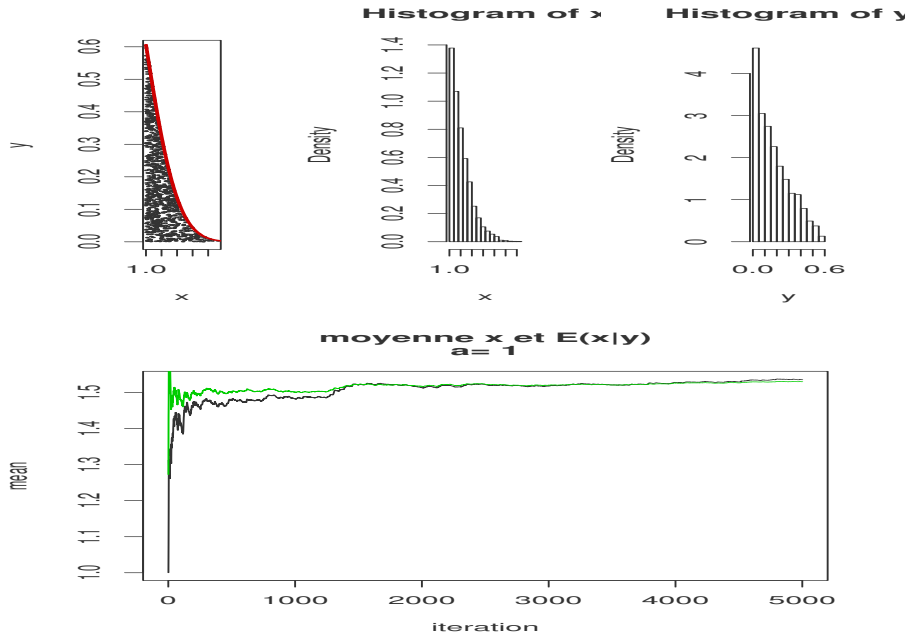
$$f(x) \propto e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \mathbb{I}_{[a,\infty]}(x) = \int g(x, y)\mathrm{d}y$$

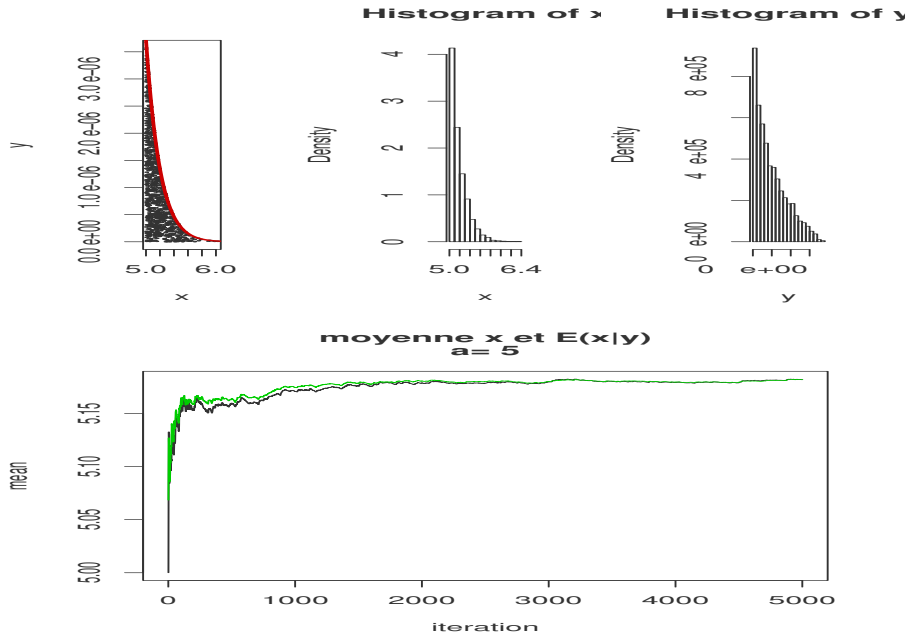
où

$$g(x, y) \propto \mathbb{I}_{[a,\infty]}(x) \mathbb{I}_{[0, e^{-\frac{1}{2\sigma^2}(x-\mu)^2}]}(y)$$

Les lois conditionnelles sont des lois uniformes

$$g(x|y) \sim \mathcal{U}(a, \mu + \sqrt{-2\sigma^2 \log(y)})$$
$$g(y|x) \sim \mathcal{U}(0, e^{-\frac{1}{2\sigma^2}(x-\mu)^2})$$





Application aux mélanges de lois exponentielles

$$x \sim p\mathcal{E}(\ell 1) + (1-p)\mathcal{E}(\ell_2)$$

On cherche à approcher la loi a posteriori des paramètres

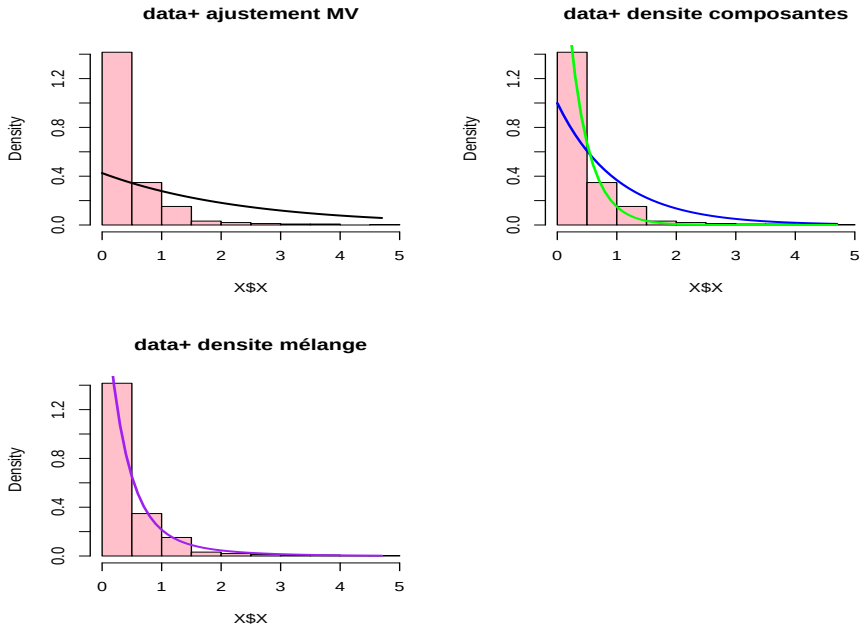
$$p,\ \ell_1, \ell_2 = \tau\ell_1,\ z_i,\ i = 1...n$$

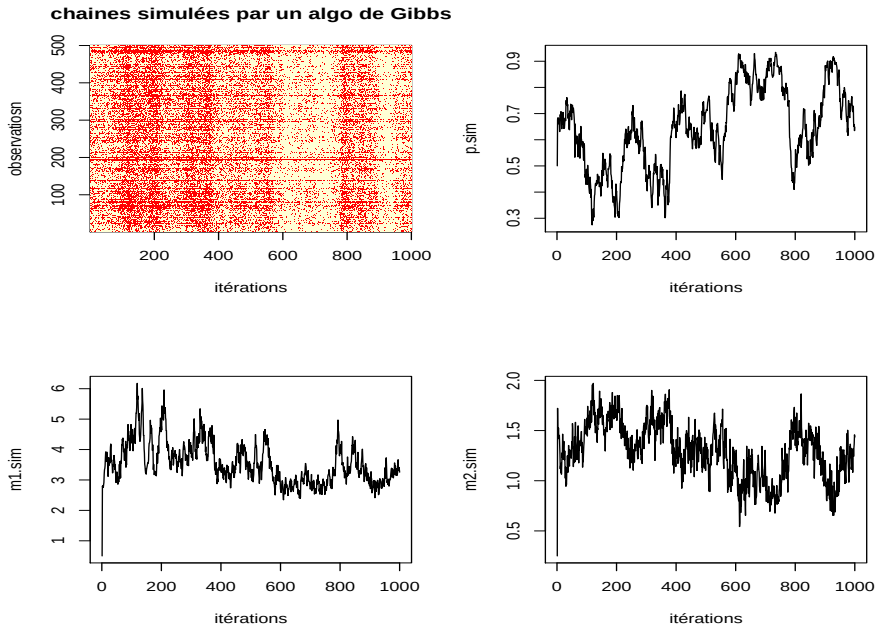
Les lois conditionnelles sont facilement simulables

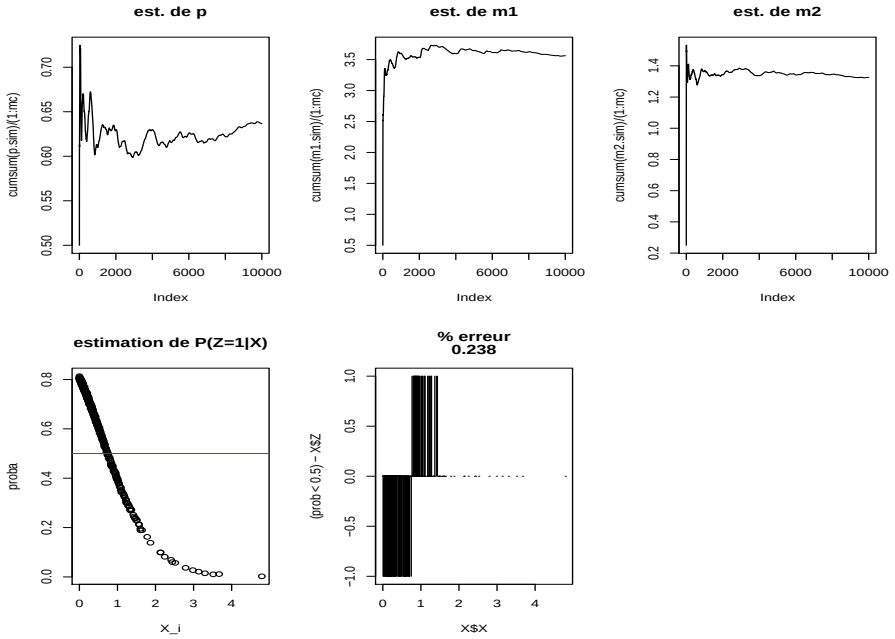
$$z_i|x_1,...x_n,p,\theta_1,\theta_2 \sim \text{bernoulli}\left(\frac{pf_1(x_i)}{pf_1(x_i)+(1-p)f_2(x_i)}\right)$$

$$p|x_1,...x_n,z_1,...z_n,\ell_1,\ell_2 \sim \text{beta}(1+\sum_i z_i, 1+n-\sum_i z_i)$$

etc







Contrôle de la convergence des algorithmes

Il est plus simple de simuler une chaîne de Markov de loi stationnaire f que des variables aléatoires suivant f
MAIS
Il faut contrôler

- la convergence vers la loi stationnaire de la chaîne de Markov
- la qualité de l'approximation de $\int h(x)f(x)\,\mathrm{d}x$ par la moyenne empirique $T^{-1}\sum_{t=1}^{T} h(x^{(t)})$

Deux critères d'arrêt :

- T_1 le temps de chauffe : on élimine les premières valeurs $x_1, \ldots, x_{T_1}$
- T_2 le temps de simulation

$$T_2^{-1} \sum_{t=T_1+1}^{T_1+T_2} h(x^{(t)}) \approx \int h(x)f(x)\,\mathrm{d}x$$

Méthodes graphiques

- Tracé de la série brute,
- Comparaison de différentes estimations pour une même quantité [Moyenne, Rao-Blackwell]
- comparaison des estimations d'une même quantité effectuées sur des chaînes indépendantes

Méthode basée sur plusieurs chaînes Gelman & Rubin

On suppose que l'on a simulé M chaînes de Markov indépendantes

$$\overline{x}_m = \frac{1}{T}\sum_{t=1}^{T} x_m^{(t)}, \qquad \overline{x} = \frac{1}{M}\sum_{m=1}^{M} \overline{x}_m \,,$$

Estimateurs des variances *inter* et *intra*

$$B_T = \frac{1}{M}\sum_{m=1}^{M} (\overline{x}_m - \overline{x})^2 \,,$$

$$W_T = \frac{1}{M}\sum_{m=1}^{M} s_m^2 = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{T}\sum_{t=1}^{T} (x_m^{(t)} - \overline{x}_m)^2$$

Critère

Pour T assez grand

$$W_T \sim \frac{T-1}{T} W_T + B_T$$