EMs et stochastique

M2 Stats de la SD, CM4, 2025-2026

Hadrien Lorenzo
Aix Marseille Université

MCEM

Monte-Carlo et EM

On utilise l'algorithm EM car pour effectuer une optimisation avec du sens physique… mais aussi car la fonctionelle $Q(\cdot|\theta_0, \boldsymbol{x})$ est, nous l'espérons, simple à écrire.

Que faire si ce n'est pas le cas ?

Monte-Carlo EM (MCEM, Wei and Tanner (1990))

On remplace $Q(\cdot|\theta_k, \boldsymbol{x})$ par une approximation Monte-Carlo. L'étape **E** devient :

$$\forall\, t = 1..T,\ \boldsymbol{Z}_t \sim p(\boldsymbol{Z}|\boldsymbol{x}, \theta_k),$$
$$\widehat{Q}_T(\theta|\theta_k, \boldsymbol{x}) = \frac{1}{T}\sum_{t=1}^{T} \ell_c(\boldsymbol{x}, \boldsymbol{Z}_t, \theta),$$

et alors $\widehat{Q}_T(\cdot|\theta_k, \boldsymbol{x}) \underset{T\to\infty}{\longrightarrow} Q(\cdot|\theta_k, \boldsymbol{x})$ par application de la loi des grands nombres.

Esprit de MCEM

- **Initialisation** : On initialise θ_0 et on fixe T le nombre de simulations
- Pour k=0… jusqu'à convergence :
  - Etape **E** : Calcul de $\widehat{Q}_T(\cdot|\theta_k, \boldsymbol{x})$
  - Etape **M** : On met à jour le paramètre θ qui maximise $\widehat{Q}_T(\cdot|\theta_k, \boldsymbol{x})$

> **Important**
>
> La suite ainsi construite des vraisemblances n'est pas assurée d'être monotone

Exemple des cellules (1)

Voir Dempster, Laird, and Rubin (1977). Soit un échantillon biologique peuplé de 4 types cellulaires différents, $x = (x_1, x_2, x_3, x_4)$, le nombre de chaque type cellulaire dans l'échantillon biologique. On considère le modèle multinomiale suivant :

$$X = (X_1, X_2, X_3, X_4) \sim \mathcal{M}\left(n; \frac{2+\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}, \right),$$

et le paramètre inconnu est $\theta \in [0, 1]$. La vraisemblance de l'échantillon, de taille 1, s'écrit

$$L(x, \theta) = (2+\theta)^{x_1}(1-\theta)^{x_2+x_3}\theta^{x_4},$$

qui n'est pas forcemment simple à maximiser.

Exemple des cellules (2)

On introduit deux variables latentes z_1 et z_2 telles que $x_1 = z_1 + z_2$. z_2 suit une loi binomiale de paramètre $\left(x_1, \frac{\theta}{\theta+2}\right)$. La log-vraisemblance complétée s'écrit alors

$$\ell_c(x, z_2, \theta) = (x_2 + x_3)\log(1-\theta) + (z_2 + x_4)\log(\theta),$$

et là on sait gérer **:)** car l'espérance conditionelle sur θ_k s'écrit :

$$Q(\theta|\theta_k, x) = (x_2 + x_3)\log(1-\theta) + (<z_2>_k + x_4)\log(\theta),$$

avec $<z_2>_k = \frac{\theta_k}{2+\theta_k} x_1$, ça c'est l'EM classique avec :

$$\hat{\theta}_{k+1} = \frac{<z_2>_k + x_4}{<z_2>_k + x_4 + x_2 + x_3}$$

Exemple des cellules (3)

En MCEM, on simule T valeurs pour $Z_2 \sim \mathcal{B}\text{inom}(x_1, \frac{\theta_k}{2+\theta_k})$, notées $(z_{2,k,t})_t$ et on construit la statistique d'intérêt :

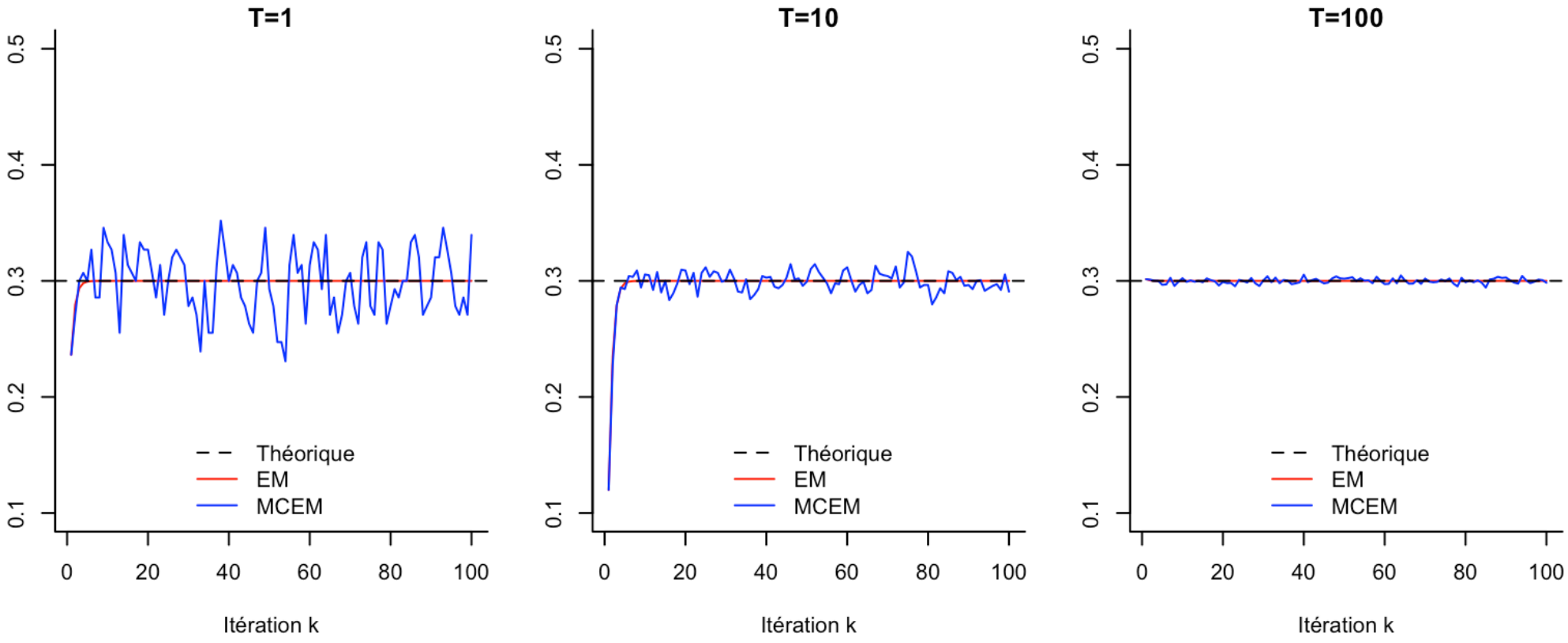
$$\begin{aligned}\widehat{Q}_T(\theta|\theta_k, x) &= \frac{1}{T}\sum_{t=1}^{T} \ell_c(x, z_{2,k,t}, \theta) \\ &= (x_2 + x_3)\log(1-\theta) + \left(\frac{1}{T}\sum_{t=1}^{T} z_{2,t} + x_4\right)\log(\theta).\end{aligned}$$

Ainsi, en notant $\widehat{<z_2>}_k = \frac{1}{T}\sum_{t=1}^{T} z_{2,k,t}$, la valeur mise à jour de θ devient

$$\check{\theta}_{k+1} = \frac{\widehat{<z_2>}_k + x_4}{\widehat{<z_2>}_k + x_4 + x_2 + x_3}$$

Exemple des cellules (4)

$$(x_1, x_2, x_3, x_4) = (115, 40, 40, 10) \quad \text{et} \quad \theta = 0.3$$



Remarques

Remarque sur l'exemple

Dans cet exemple, il est inutile d'avoir recours au MCEM, mais ce n'est pas toujours le cas

Remarque générale

Si $T = 1$ on parle de **stochastic EM** (SEM)

Modèle logit à effet aléatoire (1)

On considère le modèle logit à effet aléatoire suivant (voir Booth and Hobert 1999) :

$$[Y_{i,j}|x_{i,j}, z_i, \beta] \sim \mathcal{B}\text{ernoulli}\left(\frac{\exp(\beta x_{i,j} + z_i)}{1 + \exp(\beta x_{i,j} + z_i)}\right),$$
$$Z_i \sim \mathcal{N}(0, \sigma^2),$$

où Z_i est l'effet aléatoire de l'individu i et la modalité j allant de 1 à m. On écrit la vraisemblance complétée de ce modèle

$$L_c\left((x, y), z, (\beta, \sigma^2)\right) = \frac{1}{\sqrt{\sigma^2}^n} \mathrm{e}^{-\sum_i \frac{z_i^2}{2\sigma^2}} \prod_{i,j} \frac{\exp(y_{i,j}(\beta x_{i,j} + z_i))}{1 + \exp(\beta x_{i,j} + z_i)}.$$

et son pendant logarithmique est alors :

$$\begin{aligned}\ell_c\left((x, y), z, (\beta, \sigma^2)\right) &= -n \log \sigma - \frac{\sum_i z_i^2}{2\sigma^2} + \sum_{i,j} y_{i,j}(\beta x_{i,j} + z_i) \\ &\quad - \sum_{i,j} \log(1 + \exp(\beta x_{i,j} + z_i)).\end{aligned}$$

Modèle logit à effet aléatoire (2)

L'espérance conditionelle de son logarithme, pour un paramètre $\theta_0 = (\beta_0, \sigma_0^2)$:

$$\begin{aligned} Q(\theta|\theta_0,(x,y)) &= -n\log\sigma - \tfrac{1}{2\sigma^2}\textstyle\sum_i \mathbb{E}_{\theta_0}[Z_i^2] + \sum_{i,j} y_{i,j}(\beta x_{i,j} + \mathbb{E}_{\theta_0}[Z_i]) \\ &\quad - \textstyle\sum_{i,j} \mathbb{E}_{\theta_0}[\log(1+\exp(\beta x_{i,j} + Z_i))]. \end{aligned}$$

La maximisation en σ est possible :

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_i \mathbb{E}_{\theta_0}[Z_i^2],$$

mais la maximisation en β est plus complexe…

Modèle logit à effet aléatoire (3)

D'après la forme de la vraisemblance complétée, on peut déduire la forme de la loi instrumentale des Z_i :

$$p(z_i|(x_{i,j},y_{i,j})_j,\beta,\sigma^2) \propto \frac{\mathrm{e}^{-\frac{1}{2\sigma^2}\left(z_i-\sigma^2\sum_j y_{i,j}\right)^2}}{\prod_j\left(1+\exp\left(\beta x_{i,j}+z_i\right)\right)}.$$

En *oubliant* le dénominateur on pourrait croire que

$$[Z_i|(x_{i,j},y_{i,j})_j,\beta,\sigma^2] \sim \mathcal{N}(\sigma^2\sum_j y_{i,j},\sigma^2),$$

mais ce n'est pas le cas…

Il est nécessaire d'avoir recours à des méthodes d'échantillonnage de type MCMC.

Modèle logit à effet aléatoire (4)

Note

On va utiliser un algorithme MCMC pour simuler chacun des Z_i, indépendemment. A l'instant t, on utilise la loi instrumentale $\mathcal{N}(z_{i,t-1}, \sigma^2)$ pour simuler un candidat $\check{z}_i$. On calcule la probabilité d'acceptation

$$\alpha_i = \min\left[1, \frac{p(\check{z}_i|(x_{i,j}, y_{i,j})_j, \beta, \sigma^2)}{p(z_{i,t-1}|(x_{i,j}, y_{i,j})_j, \beta, \sigma^2)}\right].$$

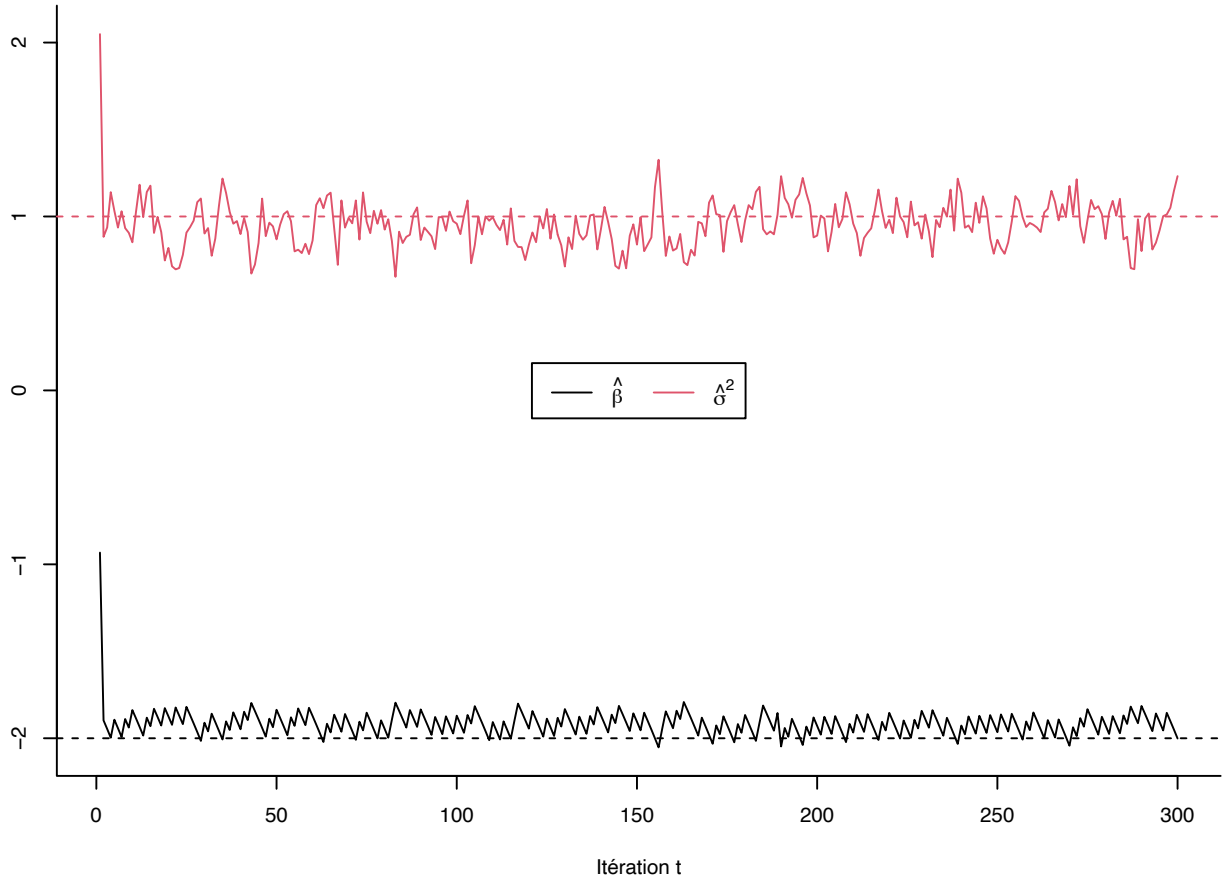
On tire alors un U uniformemment sur $[0,1]$: $z_{i,t} \leftarrow \check{z}_i$ si $U \leq \alpha_i$ et $z_{i,t} \leftarrow z_{i,t-1}$ sinon.

Il vient alors l'étape **M** de MCEM :

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{nT}\sum_{i,t} z_{i,t}^2, \\
\hat{\beta} &= \arg\max_{\beta} \sum_{i,j,t} y_{i,j}(\beta x_{i,j} + z_{i,t}) - \log(1 + \exp(\beta x_{i,j} + z_{i,t})).
\end{aligned}$$

Modèle logit à effet aléatoire (5)

Il ne reste plus qu'à implémenter, exercice laissé au lecteur



Conclusion sur MCEM

- **Avantages** : permet de contourner des problèmes de calculs de vraisemblances conditionelles complexes

- **Inconvénients** :
  - nécessité de simuler des données, ce qui peut être coûteux en temps de calcul
  - nécessité d'un critère de convergence adapté
  - pas de prise en compte de la variabilité associée au mécnaisme de données manquantes

Variational Bayes EM (VBEM)

VBEM : plus loin avec l'EM (1)

La méthode **Variational Bayes EM** (VBEM) est une extension de l'EM qui permet de contourner les problèmes de calculs de vraisemblances conditionelles complexes en supposant que la distribution postérieure des données est formée par le produit d'une densité en Z, les données manquantes, et d'une densité en θ, le paramètre :

$$p(\boldsymbol{z}, \theta|\boldsymbol{x}) \approx q_{\boldsymbol{z}}(\boldsymbol{z}|\boldsymbol{x}) q_\theta(\theta|\boldsymbol{x}) = q(\boldsymbol{z}, \theta|\boldsymbol{x}).$$

$\boldsymbol{z}$ et θ sont considérées comme indépendantes conditionellement à $\boldsymbol{x}$.

On associe Attias (1999) à la découverte de cette méthode, mais d'autres travaux (beaucoup) existent. Voir les travaux de Hinton, récemment primés par un prix Nobel de physique.

VBEM : plus loin avec l'EM (2)

L'idée est d'estimer les lois en $\boldsymbol{z}$ et en θ alternativement.

Ecrire les équations nécessite d'introduire un nouvel objet : **la vraisemblance marginale**. Cette distribution marginalise sur les paramètres θ en supposant un modèle paramétrique $\mathcal{M}$, en substance :

$$
\begin{aligned}
\log p(\boldsymbol{x}|\mathcal{M}) &= \log \textstyle\int_{\Theta\times \boldsymbol{z}} p(\boldsymbol{x},\boldsymbol{z},\theta|\mathcal{M}) d\theta d\boldsymbol{z}, \\
&= \log \textstyle\int_{\Theta\times \boldsymbol{z}} q(\boldsymbol{z},\theta|\boldsymbol{x}) \dfrac{p(\boldsymbol{x},\boldsymbol{z},\theta|\mathcal{M})}{q(\boldsymbol{z},\theta|\boldsymbol{x})} d\theta d\boldsymbol{z}, \\
&\geq \textstyle\int_{\Theta\times \boldsymbol{z}} q(\boldsymbol{z},\theta|\boldsymbol{x}) \log\left(\dfrac{p(\boldsymbol{x},\boldsymbol{z},\theta|\mathcal{M})}{q(\boldsymbol{z},\theta|\boldsymbol{x})}\right) d\theta d\boldsymbol{z}, \\
&\geq \textstyle\int_{\Theta\times \boldsymbol{z}} q_{\boldsymbol{z}}(\boldsymbol{z}|\boldsymbol{x}) q_\theta(\theta|\boldsymbol{x}) \log\left(\dfrac{p(\boldsymbol{x},\boldsymbol{z},\theta|\mathcal{M})}{q_{\boldsymbol{z}}(\boldsymbol{z}|\boldsymbol{x}) q_\theta(\theta|\boldsymbol{x})}\right) d\theta d\boldsymbol{z} \;=:\; F_m(q_{\boldsymbol{z}}, q_\theta, \boldsymbol{x}),
\end{aligned}
$$

d'après l'inégalité de Jensen appliquée à la fonction $\log$ car $q(\boldsymbol{z},\theta|\boldsymbol{x})$ somme à 1.
$F_{\mathcal{M}}(q_Z, q_\theta, \boldsymbol{x})$ est donc un minorant de la log-**vraisemblance marginale**.

VBEM : plus loin avec l'EM (3)

La fonctionelle $F_{\mathcal{M}}(q_Z, q_\theta, \boldsymbol{x})$ est maximisée par rapport à q_Z et q_θ (voir Neal and Hinton 1998) pour donner l'algorithme **VBEM** :

- Etape **E** : Choisir $q_Z^{(t)}$ qui maximise $q_Z \mapsto F_{\mathcal{M}}(q_Z, q_\theta^{(t-1)}, \boldsymbol{x})$,
- Etape **M** : Choisir $q_\theta^{(t)}$ qui maximise $q_\theta \mapsto F_{\mathcal{M}}(q_Z^{(t)}, q_\theta, \boldsymbol{x})$.

Pour aller plus loin, il faudra lire les dits papiers

Monte Carlo par chaînes de Markov (MCMC)

Les méthodes MCMC permettent d'échantillonner selon une loi de probabilité complexe. Elles sont basées sur la construction d'une chaîne de Markov dont la distribution stationnaire est la distribution cible.

Nous allons étudier deux approches MCMC :

- Algorithme de **Metropolis-Hastings**
- Algorithme de **Gibbs**

Algorithme de Métropolis-Hastings

Définition

On souhaite simuler X qui suit une loi de densité $x \to p(x)$, connue à une constante près grâce à $x \to f(x)$.

On fixe une loi instrumentale $g(\cdot|y)$ qui est connue à une constante (indépendante de y) près.

A l'instant t, on utilise g pour simuler un candidat $y \sim g(x|x_{t-1})$

On calcule la probabilité d'acceptation

$$\alpha(x_{t-1}, y) = \min\left[1, \frac{f(y)g(x_{t-1}|y)}{f(x_{t-1})g(y|x_{t-1})}\right].$$

On tire alors un U uniformemment sur $[0,1]$: $x_t \leftarrow y$ si $U \leq \alpha(x_{t-1}, y)$ et $x_t \leftarrow x_{t-1}$ sinon.

On lance plusieurs fois cet algorithme (on parle de **chaînes**) que l'on laisse tourner un certain temps (les **itérations**)

Algorithme de Métropolis-Hastings (2)

Remarques

- la probabilité d'acceptation fait intervenir le rapport f/f qui est égale au rapport p/p : il est donc inutlie de connaître la constante de normalisation
- La valeur de x_t est indépendante des valeurs antérieurs à l'instant $t-1$ conditionellement à la valeur de x_{t-1} via la loi $g(\cdot|x_{t-1})$. La suite des $(x_t)_{t>0}$ forme donc une chaîne de Markov

Noyau de transition

On apelle **noyau de transition** une distribution conditionelle $P(\cdot, \cdot) : (x, A) \mapsto P(x, A)$ qui fixe la probabilité de passer de $x \in \mathbb{R}^p$ à tout élément de l'ensemble $A \in \mathcal{B}(\mathbb{R}^p)$. Il vient $P(x, \mathbb{R}^p) = 1$. On considère que la transition de x à x est possible, donc $P(x, \{x\}) > 0$. Ce noyau va guider la transition de la chaîne de Markov.

Densité invariante

On dit qu'une chaîne, de noyau $P(\cdot, \cdot)$ converge vers la densité **invariante** $p(\cdot)$ si :

$$p^\star(dy) = \int_{\mathbb{R}^p} P(x, dy) p(x) dx, \quad \text{où} \quad p^\star(dx) = p(x)dx,$$

où $p^\star(\cdot)$ est la distribution associée à $p(\cdot) : \forall A \in \mathcal{B}(\mathbb{R}^p), p^\star(A) = \int_A p(x)dx$.

Dans le cas des méthodes MCMC, la densité $p(\cdot)$ est connue à une constante près, via $f(\cdot)$, mais le noyau de transition $P(\cdot, \cdot)$ est inconnu. On cherche à le construire.

Probabilité de non déplacement $r(x)$

On suppose une fonction $h(x,y)$ telle que $h(x,x)=0$, avec un noyau de cette forme

$$P(x,dy) = h(x,y)dy + \delta_x(dy)r(x),$$

où $\delta_x(dy) = 1$ si $x \in dy$ et 0 sinon. Par intégration il vient, car $P(\cdot, dy)$ est une probabilité :

$$\int_{\mathbb{R}^p} P(x,dy) = 1 = \int_{\mathbb{R}^p} h(x,y)dy + r(x)\int_{\mathbb{R}^p}\delta_x(dy) = \int_{\mathbb{R}^p} h(x,y)dy + r(x)\cdot 1,$$

et donc $r(x) = 1 - \int_{\mathbb{R}^p} h(x,y)dy$, qui est la **probabilité de rester en x**.

Remarque

En général $r(x) \neq 0$ et donc $y \mapsto h(x,y)$ n'est pas une probabilité : ne somme pas à 1.

Hypothèse de réversibilité de la chaîne

On suppose que $h(x, y)$ vérifie la **contrainte de réversibilité** : $p(x)h(x, y) = p(y)h(y, x)$.

Interprétation de l'hypothèse de réversibilité

Intuitivement, $p(x)h(x, y)$ donne la probabilité de passer de x à y et $p(y)h(y, x)$ donne la probabilité de passer de y à x.

L'hypothèse de réversibilité assure que ces deux probabilités sont égales : il y a autant de chance de passer de x à y que de y à x.

Théorème : Densité invariante sous l'hypothèse de réversibilité

Soit une densité de probabilité $p(\cdot)$ et un noyau de transition $P(x, dy) = h(x, y)dy + \delta_x(dy)r(x)$, où $h(x, x) = 0$, vérifiant la contrainte de réversibilité $p(x)h(x, y) = p(y)h(y, x)$. Alors $p(\cdot)$ est la densité invariante par $P(\cdot, \cdot)$.

Preuve

$$
\begin{aligned}
\int P(x, A)p(x)dx &= \int \left[\int_A p(x)h(x, y)dy\right]dx + \int r(x)\left[\int_A \delta_x(y)dy\right]p(x)dx, \\
&= \int_A \left[\int p(x)h(x, y)dx\right]dy + \int r(x)\delta_x(A)p(x)dx, \qquad (\circ) \\
&= \int_A \left[\int p(y)h(y, x)dx\right]dy + \int_A r(x)p(x)dx, \qquad (\bullet) \\
&= \int_A p(y)\left[\int h(y, x)dx\right]dy + \int_A r(x)p(x)dx, \\
&= \int_A p(y)\left[1 - r(y)\right]dy + \int_A r(x)p(x)dx, \\
&= \int_A p(x)dx = p^\star(A),
\end{aligned}
$$

et donc $p(\cdot)$ est la densité invariante par $P(\cdot, \cdot)$. Où $(\circ) \to (\bullet)$ découle de la réversibilité.

Une densité candidate qui ne vérifie pas la contrainte de réversibilité

Mettons que nous ayons à disposition une densité candidate $g(\cdot|\cdot)$. Cette densité peut vérifier la condition de réversibilité, mais pas forcemment. Supposons (x, y) tel que la contrainte de réversibilité ne soit pas satisfaite dans un sens :

$$p(x)g(y|x) > p(y)g(x|y).$$

Donc l'algorithme passera plus facilement de x à y que de y à x. Il faut donc réduire la probabilité de passer de x à y en utilisant une probabilité $\alpha(\cdot, \cdot)$ telle que, dans notre cas, $\alpha(x, y) < 1$ et $\alpha(y, x) = 1$. La condition de réversibilité s'écrit alors :

$$p(x)g(y|x)\alpha(x, y) = p(y)g(x|y)\alpha(y, x) = p(y)g(x|y),$$

et alors $\alpha(\cdot, \cdot)$ prend la forme :

$$\alpha(x, y) = \begin{cases} \dfrac{p(y)g(x|y)}{p(x)g(y|x)} & \text{si } p(x)g(y|x) > p(y)g(x|y), \\ 1 & \text{sinon.} \end{cases}$$

Forme générale du noyau de l'algorithme MH

Il vient la forme générale

$$P_{\mathrm{MH}}(x, dy) = g(y|x)\alpha(x, y)dy + \delta_x(dy)\left(1 - \int_{\mathbb{R}^p} h(x, y)dy\right),$$
$$\text{où} \quad \alpha(x, y) = \min\left[1, \frac{p(y)g(x|y)}{p(x)g(y|x)}\right].$$

dont $p(\cdot)$ est la densité invariante.

Probabilité d'acceptation

On apelle **probabilité d'acceptation** de y depuis x la fonction $(x, y) \to \alpha(x, y)$.

Bien comprendre ce noyau de transition : c'est bien l'algorithme MH

- On part d'un x courant,
- on simule un y grâce à $g(\cdot|x)$,
- on accepte y avec une probabilité $\alpha(x, y)$.

Algorithme de Métropolis-Hastings par marche aléatoire (RWMH)

RWMH (1)

Reprendre l'algorithme précédent et choisir une loi instrumentale de la forme :

$$g(x|y) = g_s(x - y) = g_s(y - x),$$

c'est le cas de la loi normale. Donc $g(\cdot|y)$ vérifie $g(x|y) = g(y|x)$. Ce qui implique que la probabilité d'acceptation est simplifiée en

$$\alpha = \frac{f(y)}{f(x_{t-1})}.$$

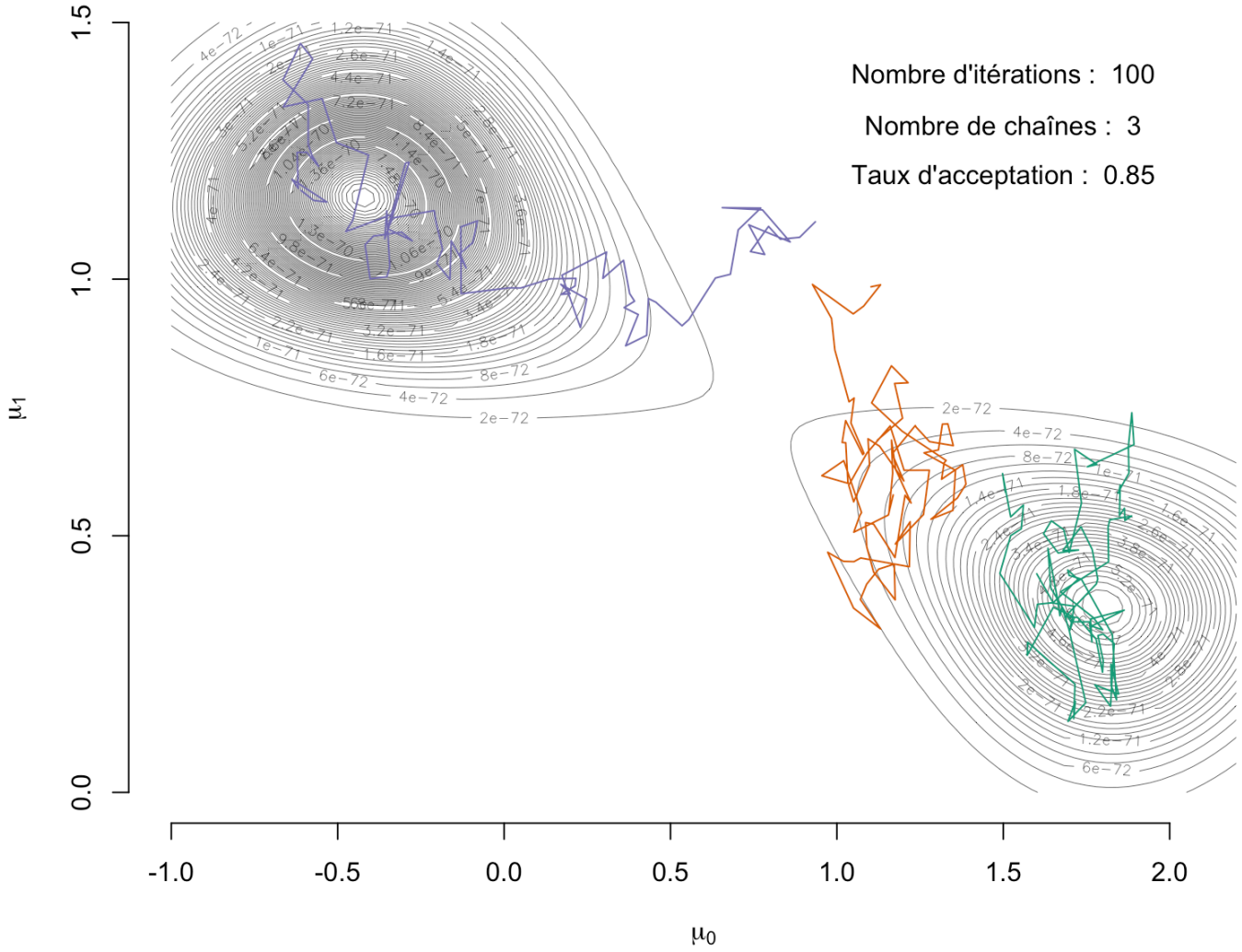
RWMH (2)

Remarques

- Cet algorithme (Metropolis et al. (1953)) n'est pas récent mais est toujours énormemment utilisé
- Si y augmente la densité de f vis à vis de x_{t-1} alors $\alpha = \dfrac{f(y)}{f(x_{t-1})} > 1$ et y est systématiquement accepté.

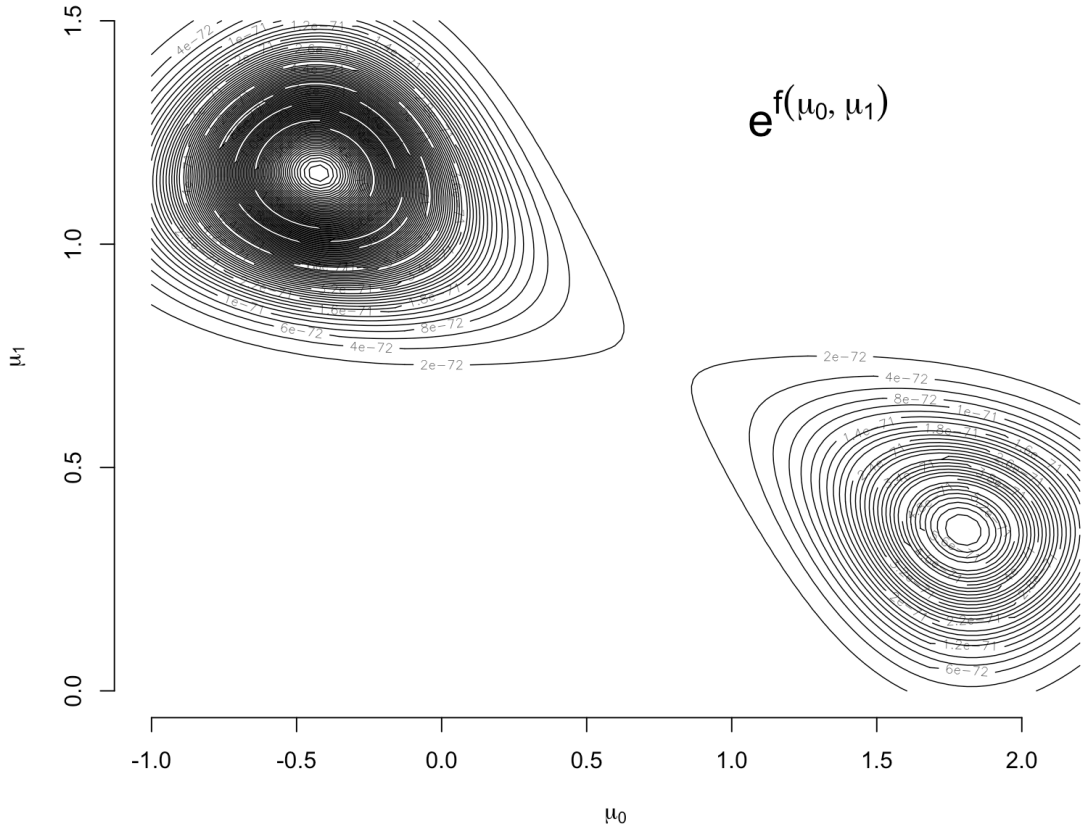
RWMH (3)

- Code


Nombre d'itérations : 100
Nombre de chaînes : 3
Taux d'acceptation : 0.85
μ1
μ0
0.0
0.5
1.0
1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0

Choix de la loi instrumentale g

Soit la fonction $f(\mu_0, \mu_1) = \sum_{i=1}^{n} \log\left(\varphi(x_i - \mu_0) + 3\varphi(x_i - \mu_1)\right)$ où $x_i \sim 1/4\mathcal{N}(0,1) + 3/4\mathcal{N}(1,1)$.

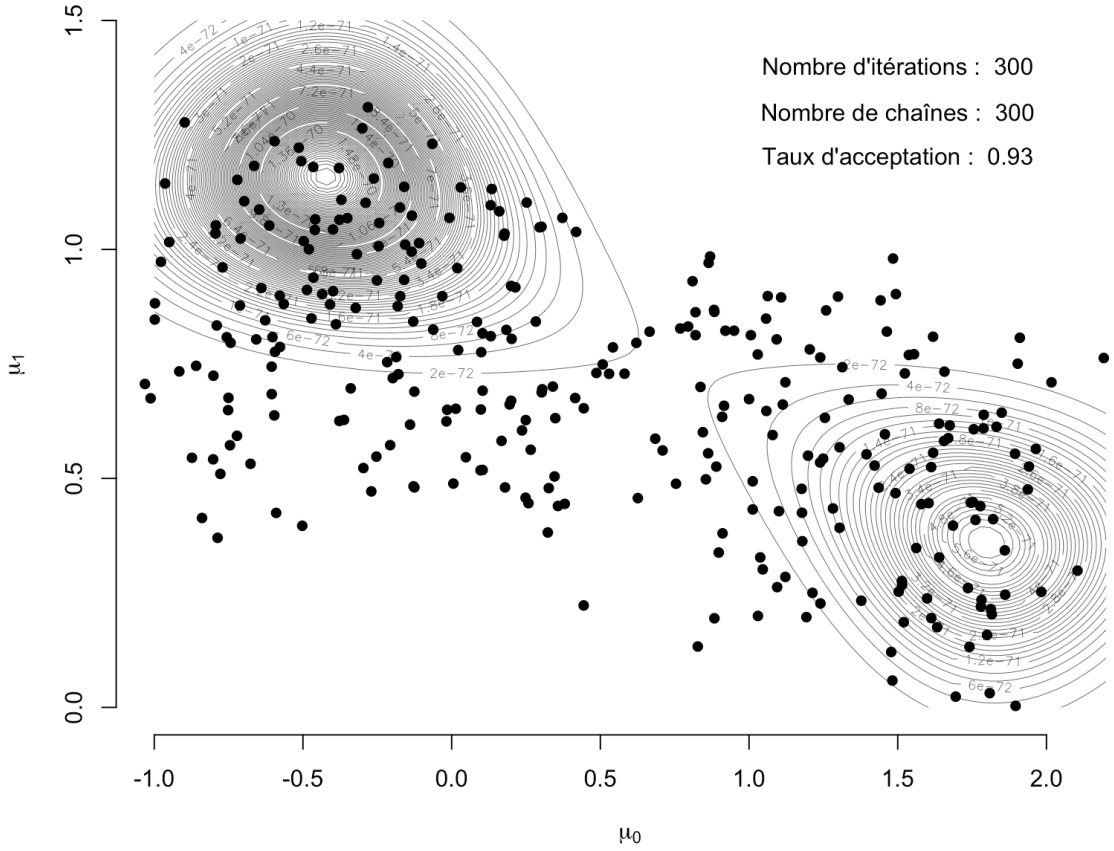


On veut générer un échantillon de densité $\propto e^{f(\mu_0,\mu_1)}$. On conserve le dernier élément de chaque chaîne. Ici $g(x|x_0) = \mathcal{N}(x_0, \sigma^2\mathbb{I}_2)$, mais quid de σ^2 ?

Choix de la loi instrumentale (2)

Si l'on fait des trop petits pas… $\sigma^2 = 10^{-2}$…

▸ Code

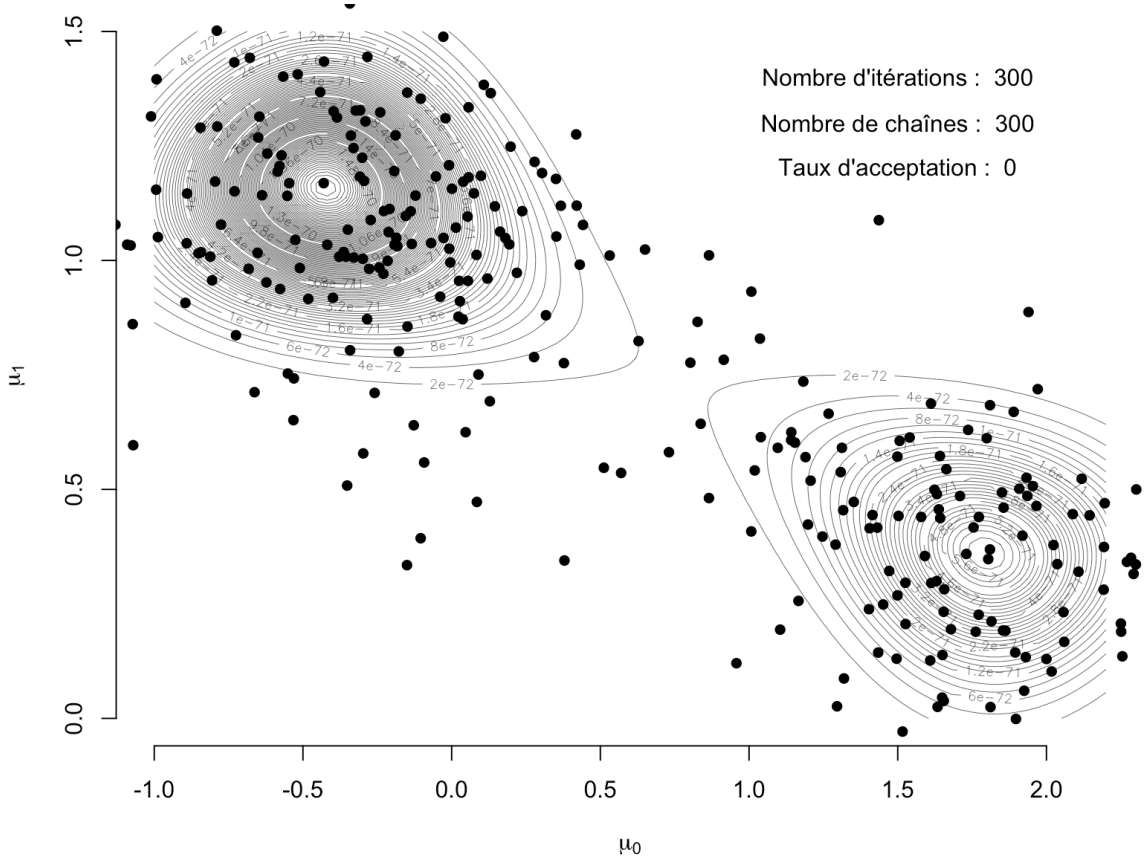


…le taux d'acceptation est excellent mais **on ne couvre pas correctement l'espace**

Choix de la loi instrumentale (3)

Si l'on fait des trop grands pas… $\sigma^2 = 10$…

▸ Code

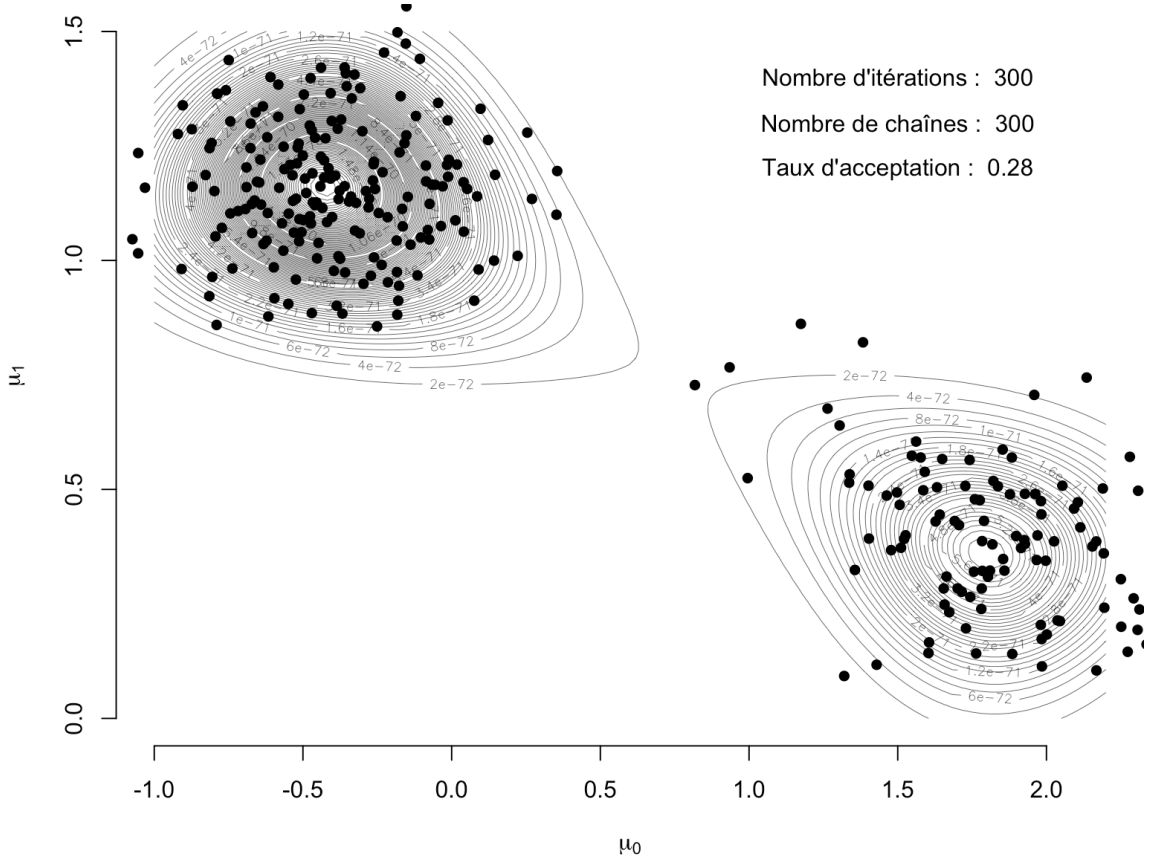


…le taux d'acceptation est **mauvais**

Choix de la loi instrumentale (4)

Si l'on fait des pas raisonnables… $\sigma^2 = 1/2$…

▸ Code



…le taux d'acceptation est acceptable et on couvre correctement l'espace

Conclusion sur RWHM

- Méthode simple à mettre en place
- Si la variance de g est trop faible, on converge lentement
- Sensibilité du choix de la loi instrumentale g
- Un fort taux d'acceptation peut signifier peu de déplacements
- D'après Roberts, Gelman, and Gilks (1997) : viser un taux d'acceptation dans $[0.25, 0.5]$
- En pratique on lance peu de chaînes et on conserve un petit nombre d'itérations, en bout de chaîne. La partie jetée *à la poubelle* est appelé le **burning**

Gibbs

Echantilloneur de Gibbs

L'objectif est ici de simuler une loi sur un paramètre multivarié $\boldsymbol{\theta} = (\theta^{(1)}, \ldots, \theta^{(p)})$.
Souvent on est dans le cas de lois postérieures : $[\boldsymbol{\theta}|\boldsymbol{x}]$. Voici l'algorithme :

- Partir d'un point $\boldsymbol{\theta}_0 = (\theta_0^{(1)}, \ldots, \theta_0^{(p)})$
- Pour chaque itération $k > 0$ faire :
  - $\theta_{t+1}^{(1)} \sim p(\theta^{(1)}|\theta_t^{(2)}, \theta_t^{(3)}, \ldots, \theta_t^{(p)}, \boldsymbol{x})$
  - $\theta_{t+1}^{(2)} \sim p(\theta^{(2)}|\theta_{t+1}^{(1)}, \theta_t^{(3)}, \ldots, \theta_t^{(p)}, \boldsymbol{x})$
  - $\ldots$
  - $\theta_{t+1}^{(p)} \sim p(\theta^{(p)}|\theta_{t+1}^{(1)}, \theta_{t+1}^{(2)}, \ldots, \theta_{t+1}^{(p-1)}, \boldsymbol{x})$

Sous certaines conditions, qui dépassent le cadre de ce cours, pour t grand, $\boldsymbol{\theta}_t$ converge vers une réalisation de la loi cible de la loi $[\boldsymbol{\theta}|\boldsymbol{x}]$.

Echantilloneur de Gibbs (2)

Important

Les lois conditionelles peuvent être définies et simulables mais pas la loi jointe… 😟

Remarques

- On trouve souvent la notation $\boldsymbol{\theta}_{t+1}^{(-k)}$ qui désigne tout $\boldsymbol{\theta}$ sauf sa $k^{\text{ième}}$ coordonnée.
- Les coordonnées $\theta^{(k)}$ sont unidimensionelles mais cet algorithme peut être utilisé en regroupant par blocs de coordonnées de $\boldsymbol{\theta}$.
- L'algorithme de Gibbs est en fait un algorithme de Metropolis-Hastings où la loi instrumentale est la loi conditionnelle, qui change donc pour chaque coordonnée et où l'on accepte tout nouveau candidat.

Références

Attias, Hagai. 1999. “A Variational Baysian Framework for Graphical Models.” *Advances in Neural Information Processing Systems* 12.

Booth, J. G., and J. P. Hobert. 1999. “Maximizing Generalized Linear Mixed Model Likelihoods with an Automated Monte Carlo EM Algorithm.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61 (1): 265–85. https://doi.org/https://doi.org/10.1111/1467-9868.00176.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. “Maximum Likelihood from Incomplete Data via the EM Algorithm.” *Journal of the Royal Statistical Society: Series B (Methodological)* 39 (1): 1–22. https://doi.org/https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

Metropolis, Nicholas, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. 1953. “Equation of State Calculations by Fast Computing Machines.” *The Journal of Chemical Physics* 21 (6): 1087–92.

Neal, Radford M, and Geoffrey E Hinton. 1998. “A View of the EM Algorithm That Justifies Incremental, Sparse, and Other Variants.” In *Learning in Graphical Models*, 355–68. Springer.

Roberts, G. O., A. Gelman, and W. R. Gilks. 1997. “Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms.” *The Annals of Applied Probability* 7 (1): 110–20. http://www.jstor.org/stable/2245134.

Wei, Greg C. G., and Martin A. Tanner. 1990. “A Monte Carlo Implementation of the EM Algorithm and the Poor Man’s Data Augmentation Algorithms.” *Journal of the American Statistical Association* 85 (411): 699–704. https://doi.org/10.1080/01621459.1990.10474930.