Projet 5 : La Méthode EM « Expectation-Maximization »

Habituellement, pour estimer un paramètre on effectue la méthode du maximum de vraisemblance. Cependant elle peut être fastidieuse à mettre en œuvre notamment dans les problèmes de données manquantes. Pour pallier à ce problème, M. Dempster a mis en place une méthode appelée méthode « Expectation-Maximisation » ou plus communément connue sous le nom de la méthode EM. Cette méthode est une méthode d'estimation paramétrique via la vraisemblance d'un modèle probabiliste grâce à un algorithme itératif. Cet algorithme permet de pallier au problème de données manquantes et chaque itération se déroule en 2 étapes : l'étape E : « Expectation » et l'étape M : « Maximisation ».
Etape E :
On note $X^c \coloneqq (X, X^m)$ le vecteur complet, dont X désigne le vecteur des variables observées et X^m le vecteur des données manquantes. Comme on dispose de données inconnues, la log-vraisemblance des « données complètes » va s'écrire $\mathbb{E}_{X^m|X,\theta}\left[\log\left(\mathbb{P}(X, x^m|p)\right)\right]$ ou $\log\left(\mathbb{P}(X, x^m|p)\right)$ est la log-vraisemblance et p le paramètre à estimer.
Etape M :
On va maximiser la log-vraisemblance et ainsi obtenir une nouvelle estimation de p.
Cette méthode est très souvent utilisée dû à sa simplicité et sa robustesse. Elle est notamment utilisée dans plusieurs domaines tels que l'imagerie médicale ou bien encore la classification de données.

Question 1.1

Montrons que $\boldsymbol{f(x^m|x)}$ est la densité d'une Gaussienne sur $\mathbb{R}^{d_2}$ de moyenne $\boldsymbol{\mu_2 + (\Sigma_{12})^T\Sigma_{11}^{-1}(x - \mu_1)}$ et de matrice de covariance $\boldsymbol{\Sigma_{22} - \Sigma_{12}{}^T\,\Sigma_{11}^{-1}\Sigma_{12}}$.

Posons $\mathrm{Z} = \begin{pmatrix} \mathrm{Z}_1 \\ \mathrm{Z}_2 \end{pmatrix} = \mathrm{A}(\mathrm{X^c} - \mu)$ où $\mathrm{A} = \begin{pmatrix} \mathrm{I}_{d_1} & 0 \\ -(\Sigma_{12})^T\Sigma_{11}^{-1} & \mathrm{I}_{d_2} \end{pmatrix}$

Donc $(\mathrm{X^c} - \mu) = \mathrm{A}^{-1}\mathrm{Z} = \begin{pmatrix} \mathrm{I}_{d_1} & 0 \\ (\Sigma_{12})^T\Sigma_{11}^{-1} & \mathrm{I}_{d_2} \end{pmatrix}\begin{pmatrix} \mathrm{Z}_1 \\ \mathrm{Z}_2 \end{pmatrix} = \begin{pmatrix} \mathrm{I}_{d_1}\mathrm{Z}_1 \\ (\Sigma_{12})^T\Sigma_{11}^{-1}\mathrm{Z}_1 + \mathrm{I}_{d_2}\mathrm{Z}_2 \end{pmatrix}$

On pose $x = \mathrm{I}_{d_1}\mathrm{Z}_1$ et $y = (\Sigma_{12})^T\Sigma_{11}^{-1}\mathrm{Z}_1 + \mathrm{I}_{d_2}\mathrm{Z}_2$

$$\begin{cases} Z_1 = xI_{d_1} \\ y = (\Sigma_{12})^T\Sigma_{11}^{-1}Z_1 + I_{d_2}Z_2 \end{cases} \Rightarrow \begin{cases} Z_1 = xI_{d_1} \\ Z_2 = xI_{d_2}(\Sigma_{12})^T\Sigma_{11}^{-1}I_{d_1} - yI_{d_2} \end{cases} \quad \text{et} \quad |\,\mathbf{J}\,| = \begin{vmatrix} \mathrm{I}_{d_1} & 0 \\ \mathrm{I}_{d_2}(\Sigma_{12})^T\Sigma_{11}^{-1}\mathrm{I}_{d_1} & -\mathrm{I}_{d_2} \end{vmatrix} = 1$$

Densité de Z :

$$f(Z_1, Z_2) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\mathrm{e}^{-\frac{1}{2}(\mathrm{X^c}-\mu)^T\Sigma^{-1}(\mathrm{X^c}-\mu)}$$

$$= \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\,\mathrm{e}^{-\frac{1}{2}(\mathrm{A}^{-1}\mathrm{Z})^T\Sigma^{-1}(\mathrm{A}^{-1}\mathrm{Z})}$$

$$= \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\,\mathrm{e}^{-\frac{1}{2}\mathrm{Z}^T(\mathrm{A}^{-1})^T\Sigma^{-1}(\mathrm{A}^{-1}\mathrm{Z})}$$

$$= \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\,e^{-\frac{1}{2}Z^T(A^T)^{-1}\Sigma^{-1}(A^{-1}Z)}$$

(A est une matrice triangulaire donc $(\mathrm{A}^{-1})^T = (\mathrm{A}^T)^{-1}$)

$$= \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\,e^{-\frac{1}{2}Z^T(A\Sigma A^T)^{-1}Z}$$

$$A\Sigma A^T = \begin{pmatrix} I_{d_1} & 0 \\ -(\Sigma_{12})^T\Sigma_{11}^{-1} & I_{d_2} \end{pmatrix}\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ {\Sigma_{12}}^T & \Sigma_{22} \end{pmatrix}\begin{pmatrix} I_{d_1} & \left(-(\Sigma_{12})^T\Sigma_{11}^{-1}\right)^T \\ 0 & I_{d_2} \end{pmatrix}$$

$$= \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ -(\Sigma_{12})^T\Sigma_{11}^{-1}\Sigma_{11} + {\Sigma_{12}}^T & -(\Sigma_{12})^T\Sigma_{11}^{-1}\Sigma_{12} + \Sigma_{22} \end{pmatrix}\begin{pmatrix} I_{d_1} & \left(-(\Sigma_{12})^T\Sigma_{11}^{-1}\right)^T \\ 0 & I_{d_2} \end{pmatrix}$$

$$= \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ 0 & -(\Sigma_{12})^T\Sigma_{11}^{-1}\Sigma_{12} + \Sigma_{22} \end{pmatrix}\begin{pmatrix} I_{d_1} & -(\Sigma_{11}^{-1})^T\Sigma_{12} \\ 0 & I_{d_2} \end{pmatrix}$$

$$= \begin{pmatrix} \Sigma_{11} & -\Sigma_{11}(\Sigma_{11}^{-1})^T\Sigma_{12} + \Sigma_{12} \\ 0 & -(\Sigma_{12})^T\Sigma_{11}^{-1}\Sigma_{12} + \Sigma_{22} \end{pmatrix}$$

Comme Σ_{11} est une matrice de covariance, elle est symétrique donc ${\Sigma_{11}}^T = \Sigma_{11}$ et on a :
$-\Sigma_{11}\,({\Sigma_{11}}^{-1})^T\,\Sigma_{12} + \Sigma_{12} = -\Sigma_{11}({\Sigma_{11}}^T)^{-1}\Sigma_{12} + \Sigma_{12} = -\Sigma_{11}(\Sigma_{11})^{-1}\Sigma_{12} + \Sigma_{12} = 0$
Finalement, $A\Sigma A^T = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & b \end{pmatrix}$ avec $b = \Sigma_{22} - {\Sigma_{12}}^T\,\Sigma_{11}^{-1}\Sigma_{12}$
$(A\Sigma A^T)^{-1} = \begin{pmatrix} {\Sigma_{11}}^{-1} & 0 \\ 0 & b^{-1} \end{pmatrix} \Rightarrow Z^T(A\Sigma A^T)^{-1}Z = {Z_1}^T{\Sigma_{11}}^{-1}Z_1 + {Z_2}^T\,b^{-1}\,Z_2$ et $|\Sigma| = |\Sigma_{11}||b|$

Donc : $g(Z_1, Z_2) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_{11}|^{\frac{1}{2}}|b|^{\frac{1}{2}}}\, e^{-\frac{1}{2}(Z_1{}^T\Sigma_{11}{}^{-1}Z_1 + Z_2{}^T\,b^{-1}\,Z_2)} = \frac{1}{(2\pi)^{d_1/2}|\Sigma_{11}|^{1/2}}\, e^{-\frac{1}{2}Z_1{}^T\Sigma_{11}{}^{-1}Z_1} \times \frac{1}{(2\pi)^{d_2/2}|b|^{1/2}}\, e^{-\frac{1}{2}Z_2{}^T b^{-1}Z_2}$

La loi marginale de Z_1 est donnée par : $\int_{\mathbb{R}^{d_2}} g\,(Z_1, Z_2)\,dz_2 = \frac{1}{(2\pi)^{d_1/2}|\Sigma_{11}|^{1/2}}\, e^{-\frac{1}{2}Z_1{}^T\Sigma_{11}{}^{-1}Z_1}$

On a donc montré que $Z_1 \sim N\,(0, \Sigma_{11})$
$\Rightarrow (x - \mu_1) \sim N\,(0, \Sigma_{11})$
$\Rightarrow x \sim N\,(\mu_1, \Sigma_{11})$

$$f_{Z_2|Z_1} = \frac{f(Z_1, Z_2)}{f_{Z_1}(z_1)} \sim N(0, \Sigma_{22} - {\Sigma_{12}}^T\,\Sigma_{11}^{-1}\Sigma_{12})$$

$Z_1 \perp\!\!\!\perp Z_2$ donc $Z_2 \sim Z_2|Z_1$ donc $Z_2 = -(\Sigma_{12})^T\Sigma_{11}^{-1}(x - \mu_1) + x^m - \mu_2 \sim N(0, \Sigma_{22} - {\Sigma_{12}}^T\,\Sigma_{11}^{-1}\Sigma_{12})$

$x^m|x \sim N(\mu_2 + (\Sigma_{12})^T\Sigma_{11}^{-1}(x - \mu_1)\,, \Sigma_{22} - {\Sigma_{12}}^T\,\Sigma_{11}^{-1}\Sigma_{12})$

Question 1.2

Montrons que $\log[f_X(x,p)] = Q(p, p^{old}) - H(p, p^{old})$

$$\begin{aligned} Q(p, p^{old}) - H(p, p^{old}) &= \mathbb{E}[\log f(X^c, p)\,|X = x, p^{old}] - \mathbb{E}[\log f(X^m|x, p)|X = x, p^{old}] \\ &= \int_{\mathbb{R}^{d_2}} \log[f((x, x^m), p)]\, f(x^m|x, p^{old})dx^m - \int_{\mathbb{R}^{d_2}} \log[f(x^m|x, p)] f(x^m|x, p^{old})dx^m \\ &= \int_{\mathbb{R}^{d_2}} \log\left(\frac{f((x, x^m), p)}{f(x^m|x, p)}\right) f(x^m|x, p^{old})dx^m \\ &= \int_{\mathbb{R}^{d_2}} \log\left(\frac{f(x^m|x, p)f_X(x, p)}{f(x^m|x, p)}\right) f(x^m|x, p^{old})dx^m \\ &= \int_{\mathbb{R}^{d_2}} \log[f_X(x, p)]\;\, f(x^m|x, p^{old})dx^m \\ &= \log[f_X(x, p)] \int_{\mathbb{R}^{d_2}} f(x^m|x, p^{old})dx^m = \log[f_X(x, p)] \end{aligned}$$

Donc on a bien $\log[f_X(x,p)] = Q(p,p^{old}) - H(p,p^{old})$

Montrons que $H(p^{old},p^{old}) - H(p,p^{old}) = -\int_{\mathbb{R}^{d_2}} f(x^m|x,p^{old}) log\left(\frac{f(x^m|x,p)}{f(x^m|x,p^{old})}\right) dx^m \geq 0$.

$$\begin{aligned}
&\mathbb{E}[logf(X^m|x,p^{old})|X=x,p^{old}] - \mathbb{E}[logf(X^m|x,p),X=x,p^{old}] \\
&= \int_{\mathbb{R}^{d_2}} log(f(x^m|x,p^{old}))\, f(x^m|x,p^{old})dx^m - \int_{\mathbb{R}^{d_2}} log(f(x^m|x,p))\, f(x^m|x,p^{old})dx^m \\
&= \int_{\mathbb{R}^{d_2}} log\left(\frac{f(x^m|x,p^{old})}{f(x^m|x,p)}\right) f(x^m|x,p^{old})dx^m \\
&= -\int_{\mathbb{R}^{d_2}} log\left(\frac{f(x^m|x,p)}{f(x^m|x,p^{old})}\right) f(x^m|x,p^{old})dx^m
\end{aligned}$$

De plus, $x \mapsto \log(x)$ est strictement concave, donc par l'inégalité de Jensen :

$$\int_{\mathbb{R}^{d_2}} log\left(\frac{f(x^m|x,p)}{f(x^m|x,p^{old})}\right) f(x^m|x,p^{old})dx^m \geq log\left(\int_{\mathbb{R}^{d_2}} \frac{f(x^m|x,p)}{f(x^m|x,p^{old})} f(x^m|x,p^{old})dx^m\right) = log(1) = 0$$

Question 2.1

Montrons que pour tout $s,t > 0$, $\mathbb{P}(T > s+t\,|T > t) = \mathbb{P}(T > s)$.

Soient $s,t > 0$ et $T \sim exp\,(a)$.

$$\mathbb{P}(T > s+t\,|T > t) = \frac{\mathbb{P}(T > s+t, T > t)}{\mathbb{P}(T > t)} = \frac{\mathbb{P}(T > s+t)}{\mathbb{P}(T > t)} = \frac{e^{-(s+t)a}}{e^{-ta}} = e^{-as} = \mathbb{P}(T > s)$$

Donc T est une variable aléatoire « avec perte de mémoire ».

Question 2.2

Montrons que $Q(a,a^{old}) = Nlog(a) - a[\sum_{i=1}^{k} t_i + (N-k)\left(C + \frac{1}{a^{old}}\right)]$

$$\begin{aligned}
&\mathbb{E}[\, log\, f\,(T^c,a)|t_1,\dots,t_k,t_{k+1} \geq C,\dots,t_N \geq C, a^{old}] = \int_{\mathbb{R}^{d_2}} log\,(\prod_{i=1}^{N} ae^{-at_i})\, f(t^m|t,a^{old})dt^m \\
&= \int_{\mathbb{R}^{d_2}} \sum_{i=1}^{N} (log(a) - at_i)\, f(t^m|t,a^{old})dt^m \\
&= N\, log(a) \int_{\mathbb{R}^{d_2}} f(t^m|t,a^{old})dt^m - a \int_{\mathbb{R}^{d_2}} \sum_{i=1}^{N} t_i\; f(t^m|t,a^{old})dt^m \\
&= Nlog(a) - a\left(\sum_{i=1}^{k} t_i \int_{\mathbb{R}^{d_2}} f(t^m|t,a^{old})dt^m + \int_{\mathbb{R}^{d_2}} \sum_{t=k+1}^{N} t_i\, f(t^m|t,a^{old})dt^m\right) \\
&= N\, log(a) - a(\sum_{i=1}^{k} t_i + \int_{\mathbb{R}^{d_2}} \sum_{t=k+1}^{N} t_i\, f(t^m|t,a^{old})dt^m)
\end{aligned}$$

Or $\quad f\left(t^m \middle| t, a^{old}\right) = \dfrac{f((t,t^m),a^{old})}{f_T(t,a^{old})} = \dfrac{a^{old}e^{-a^{old}(t+t^m)}}{\int_{\mathbb{R}^{d_2}} a^{old}e^{-a^{old}(t+t^m)}dt^m} = a^{old}e^{-a^{old}t^m}$

Donc $\displaystyle\int_{\mathbb{R}^{d_2}} \sum_{t=k+1}^{N} t_i\, f\left(t^m \middle| t, a^{old}\right)dt^m = \int_{\mathbb{R}^{d_2}} \sum_{t=k+1}^{N} t_i\, a^{old}e^{-a^{old}t^m}dt^m$

$$= \int_{\mathbb{R}^{d_2}} \sum_{t=k+1}^{N} t_i\, a^{old}e^{-a^{old}\sum_{i=k+1}^{N}t_i}dt^m = \mathrm{E}\left[T^m \middle| T = t, a^{old}\right]$$

On a $t_{k+1}, \dots, t_N \ \geq C$ et l'espérance d'une loi exponentielle de paramètre a^{old} vaut $\frac{1}{\boldsymbol{a^{old}}}$

$\Rightarrow \mathbb{E}\left[t_i \middle| T = t, a^{old}\right] = \left(C + \frac{1}{a^{old}}\right)$ pour tout i = k+1,…,N

$\Rightarrow \mathbb{E}\left[T^m \middle| T = t, a^{old}\right] = (N-k)(C + \frac{1}{a^{old}})$

Et donc on a bien :

$Q\left(a, a^{old}\right) = \mathbb{E}\left[\, log\ f\ (T^c, a) \middle| t_1, \dots\ , t_{k,} t_{k+1} \geq C, \dots, t_N\ \geq C, a^{old}\right]$

$\qquad = Nlog(a) - a[\, \sum_{i=1}^{k} t_i +\ (N-k)\left(C + \frac{1}{a^{old}}\right)]$

En déduire que si $\boldsymbol{a^{new}}$ maximise $\boldsymbol{a \mapsto Q\left(a, a^{old}\right)}$ alors $\boldsymbol{a^{new} = \frac{N}{\sum_{i=1}^{k} t_i + (N-k)\left(C + \frac{1}{a^{old}}\right)}}$

$\frac{\partial Q(a,a^{old})}{\partial a} = \frac{N}{a} - \left(\sum_{i=1}^{k} t_i +\ (N-k)\left(C + \frac{1}{a^{old}}\right)\right)$

$\frac{\partial Q(a,a^{old})}{\partial a} = 0 \ \Leftrightarrow\ \frac{N}{a} = \left(\sum_{i=1}^{k} t_i +\ (N-k)\left(C + \frac{1}{a^{old}}\right)\right) \Leftrightarrow a^{new} = \frac{N}{\sum_{i=1}^{k} t_i + (N-k)\left(C + \frac{1}{a^{old}}\right)}$

$\frac{\partial^2 Q(a,a^{old})}{\partial a^2} = -\frac{N}{a^2} < 0$. Donc $a^{new} = \frac{N}{\sum_{i=1}^{k} t_i + (N-k)\left(C + \frac{1}{a^{old}}\right)}$ maximise $a \mapsto Q\left(a, a^{old}\right)$.

En déduire que l'algo EM converge vers $\hat{a} = \frac{k}{\sum_{i=1}^{k} t_i + (N-k)C}$

$$a^{new} = \frac{N}{\sum_{i=1}^{k} t_i +\ (N-k)\left(C + \frac{1}{a^{old}}\right)}$$

Si l'algorithme converge vers une limite â alors $a^{old} \to$ â et $a^{new} \to$ â, on a donc :

$$a^{new} \to\ \hat{a} = \frac{N}{\sum_{i=1}^{k} t_i + (N-k)\left(C + \frac{1}{\hat{a}}\right)} \Leftrightarrow \hat{a}\left(\sum_{i=1}^{k} t_i +\ (N-k)\left(C + \frac{1}{\hat{a}}\right)\right) = N$$

$$\Leftrightarrow \hat{a}\sum_{i=1}^{k} t_i + \hat{a}(N-k)C + N - k = N$$

$$\Leftrightarrow \hat{a}\left(\sum_{i=1}^{k} t_i + (N-k)C\right) = N - N + k$$

$$\Leftrightarrow \hat{a} = \frac{k}{\sum_{i=1}^{k} t_i + (N-k)C}$$

Donc si l'algorithme EM converge, il converge vers $\hat{a} = \frac{k}{\sum_{i=1}^{k} t_i + (N-k)C}$

Montrons que â est l'EMV d'une expérience où les données observées sont N variables aléatoires i.i.d $\boldsymbol{X_1, \dots, X_N}$ distribuées comme $\boldsymbol{X = T\mathbb{I}_{(T \leq C)} + C\mathbb{I}_{(T > C)}}$ et $\boldsymbol{k = |\{i{:}\, T_i \leq C\}|}$.

$$\mathbb{P}(X \le x) = \mathbb{P}\big(T\mathbb{I}_{(\boldsymbol{T\le C})} + C\mathbb{I}_{(\boldsymbol{T>C})} \le x\big) = \mathbb{P}(T \le x, T \le C) + \mathbb{P}(C \le x, T > C)$$
$$= \begin{cases} \mathbb{P}(T \le x, T \le C) \text{ pour tout T} \in \{t_1, \dots, t_k\} \\ \mathbb{P}(C \le x, T > C) \text{ pour tout T} \in \{t_{k+1}, \dots, t_N\} \end{cases}$$

$\mathbb{P}(T \le x, T \le C) = \mathbb{P}(T \le x)\ car\ X = T\mathbb{I}_{(T\le C)} + C\mathbb{I}_{(T>C)} \le C \quad \forall T$

$\mathbb{P}(C \le x) = \begin{cases} 1\ si\ C \le x \\ 0\ si\ C > x \end{cases}$ donc :

$\mathbb{P}(C \le x, T > C) = \mathbb{P}(T > C)\mathbb{P}(C \le x) = \mathbb{P}(T > C)\mathbb{I}_{(C\le x)} = \big(1 - \mathbb{P}(T \le C)\big)\mathbb{I}_{(C\le x)} =$
$(1 - (1 - e^{-ac}))\mathbb{I}_{(C\le x)} = e^{-ac}\mathbb{I}_{(C\le x)}$

Finalement :

$$\mathbb{P}(X \le x) = \begin{cases} \mathbb{P}(T \le x) & \text{T} \in \{t_1, \dots, t_k\} \\ e^{-ac}\mathbb{I}_{(C\le x)} & \text{T} \in \{t_{k+1}, \dots, t_N\} \end{cases} = \begin{cases} 1 - e^{-ax} & \text{T} \in \{t_1, \dots, t_k\} \\ e^{-ac}\mathbb{I}_{(C\le x)} & \text{T} \in \{t_{k+1}, \dots, t_N\} \end{cases}$$

Donc la vraisemblance est donnée par :

$\prod_{i=1}^{N}\left(ae^{-at_i}\mathbb{I}_{(0\le x_i<\boldsymbol{C})} + e^{-aC}\mathbb{I}_{(C\le x_i)}\right)$ où $f_T(t_i) = ae^{-at_i}\ \mathbb{I}_{(\boldsymbol{t_i\ge 0})}$ et $\mathbb{P}(T > C)\mathbb{I}_{(C\le x_i)} = e^{-aC}\mathbb{I}_{(C\le x_I)}$

$\prod_{i=1}^{N}\left(ae^{-at_i}\mathbb{I}_{(0\le x_i<\boldsymbol{C})} + e^{-aC}\mathbb{I}_{(C\le x_i)}\right) = a^k e^{-a\sum_{i=1}^{k} x_i}\ \ e^{-aC(N-k)}$ car $k = |\{i: T_i \le C\}|$

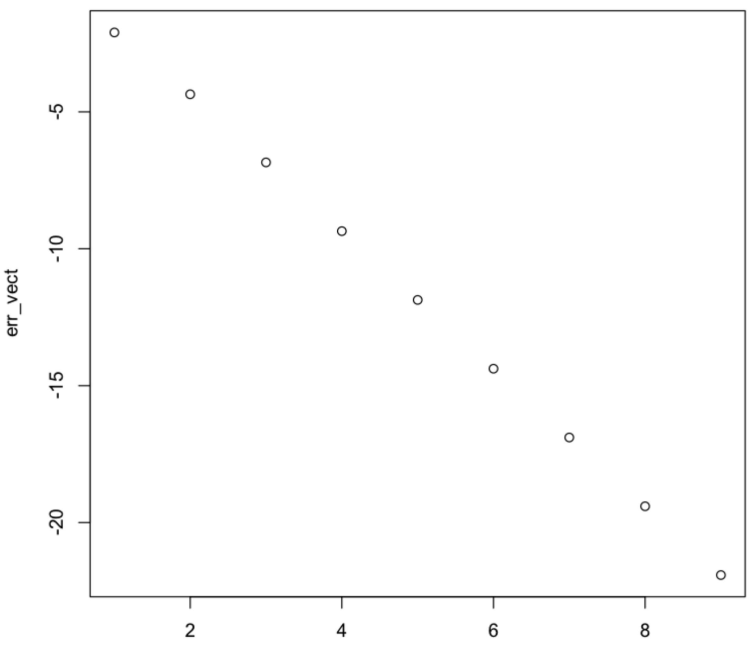
$\mathcal{L}(a) = k\ log(a) - a\sum_{i=1}^{k} t_i - aC(N-k)$

$\mathcal{L}'(a) = \frac{\mathrm{k}}{a} - \sum_{i=1}^{k} t_i - C(N-k)$

$\mathcal{L}'^{(a)} = 0 \Leftrightarrow \frac{\mathrm{k}}{a} - \sum_{i=1}^{k} t_i - C(N-k) = 0 \Leftrightarrow \frac{\mathrm{k}}{a} = \sum_{i=1}^{k} t_i + C(N-k) \Leftrightarrow \hat{a} = \frac{k}{\sum_{i=1}^{k} t_i + C(N-k)}$

Question 2.3

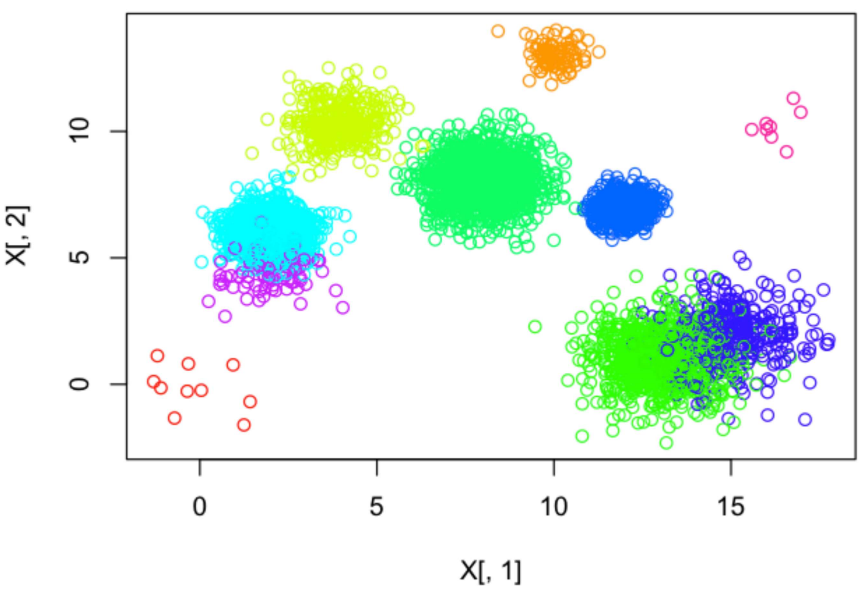
Montrons la convergence de l'algo EM et estimons la vitesse de convergence en fonction du nombre d'itérations.



```r
### mise en place des variables
a<-0.5
N<-10000 #Nombre d'observations
C<-5 #seuil

### mise en place des conditions de convergence (nb d'itération et tolérance)
a_new<-0.1
iter<-0
err<-1 #erreur
err_vect=NULL
tol<-0.000000001


### Algorithme EM

t<-rexp(N,a) ###simulation de N variables de loi exp(a)
t<-sort(t)
K=t<C #mesure de comptage, nombre d'élément de t<C
t<-t[K] #on prend les k premiers éléments du vecteur t
k<-length(t)
a_chapeau<-k/(sum(t)+(N-k)*C) ###Valeur de convergence théorique



while (iter<100000 && err>tol){
  a_new<-N/(sum(t)+(N-k)*(C+1/a_new))
  err<-abs(a_new-a_chapeau)
  err_vect<-c(err_vect,log(err))
  iter<-iter+1
}

iter
err
a
a_chapeau
a_new
plot(err_vect)
```

Question 3.1

Représenter graphiquement un mélange de 10 gaussiennes dans $\mathbb{R}^2$



code sur R pour obtenir le graphe ci-dessus :

```
N<-5000 #nombre de données connues
k<-10 #nombre de loi gaussienne
X<-matrix(ncol = 3,nrow=N)
couleur = palette (rainbow(10))

ALPHA<-c(0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1)
mu1<-c(0,10,4,13,8,2,12,15,2,16)
mu2<-c(0,13,10.3,1,8,6,7,2,4.5,10)

sigma<-c(0.75,0.5,0.75,1,0.86,0.6,0.4,1,0.75,0.625)


for(i in 1:N){
  j<-rbinom(1,k-1,0.5) #Tirage des loi normales selon une loi
 binomiale
  X[i,1]<-rnorm(1,mu1[j+1],sigma[j+1])
  X[i,2]<-rnorm(1,mu2[j+1],sigma[j+1])
  X[i,3]<-j+1
}

plot(X[,1],X[,2], col=couleur[X[,3]])
```

Question 3.2

Montrons que $\boldsymbol{Q(p,p^{old}) = \sum_{n=1}^{N}\sum_{k=1}^{K} f(k|\, x_n\,,p^{old})\, log\, \alpha_k + \sum_{n=1}^{N}\sum_{k=1}^{K} f(k|x_n\,,p^{old})\log f_k(x_n,p_k)}$

$$Q(p,p^{old}) = \mathbb{E}[\log f(X^c,p)\,|X = x,p^{old}] = \sum_{k=1}^{K} f(k|x,p^{old})\log\left(f(x^c,p)\right)$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} f(k|x_n,p^{old})\left(\log(\alpha_k) + \log\ f_k(x_n,p_k)\right)$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} f(k|x_n,p^{old})\log(\alpha_k) + \sum_{n=1}^{N}\sum_{k=1}^{K} f(k|x_n\,,p^{old})\log f_k(x_n,p_k)$$

Question 3.3

On veut résoudre le problème d'optimisation suivant :

$\boldsymbol{\mathrm{argmax}_{\alpha_1\ldots\alpha_k}\, Q(p,p^{old})}$ ***sous la contrainte*** $\boldsymbol{\sum_{k=1}^{K}\alpha_k = 1}$

On va utiliser la méthode des multiplicateurs de Lagrange.

$L(\alpha, \lambda) = Q(p, p^{old}) - \lambda(\sum_{k=1}^{K} \alpha_k - 1)$

On annule la dérivée partielle de α, λ.

$\frac{dL(\alpha,\lambda)}{\partial \alpha_k} = 0 \quad \Leftrightarrow \qquad \frac{dL(\alpha,\lambda)}{\partial \alpha_k} = \sum_{n=1}^{N} \frac{f(k|x_n, p^{old})}{\alpha_k} - \lambda = 0 \qquad \Leftrightarrow \qquad \alpha_k = \frac{1}{\lambda} \sum_{n=1}^{N} f(k|x_n, p^{old})$

$\frac{dL(\alpha,\lambda)}{\partial \lambda} = 0 \quad \Leftrightarrow \qquad \frac{dL(\alpha,\lambda)}{\partial \lambda} = \sum_{n=1}^{N} \alpha_k - 1 = 0 \qquad \Leftrightarrow \qquad \sum_{k=1}^{K} \alpha_k = 1$

Comme $\sum_{k=1}^{K} \alpha_k = 1$ et $\alpha_k = \frac{1}{\lambda} \sum_{n=1}^{N} f(k|x_n, p^{old})$ alors $\sum_{k=1}^{K} \alpha_k = \sum_{k=1}^{K} \frac{1}{\lambda} \sum_{n=1}^{N} f\left(k \middle| x_n, p^{old}\right) = 1$

$\Leftrightarrow \lambda = \sum_{n=1}^{N} \sum_{k=1}^{K} f(k|x_n, p^{old})$ or $\sum_{k=1}^{K} f(k|x_n, p^{old}) = 1$ car $f(k|x_n, p^{old})$ est une densité.

$\Leftrightarrow \lambda = \sum_{n=1}^{N} 1 = N$

Donc on a bien $\alpha_k = \frac{1}{N} \sum_{n=1}^{N} f(k|x_n, p^{old})$

Vérifions que α_k est un maximum :

$\frac{d^2 L(\alpha,\lambda)}{\partial {\alpha_k}^2} = \sum_{n=1}^{N} \left(-\frac{f(k|x_n, p^{old})}{{\alpha_k}^2} \right) \leq 0$ Donc *L(α,λ)* est concave et α_k est bien un maximum.

Question 3.4

Montrons que $\boldsymbol{f\left(k \middle| x_n, p^{old}\right) = \frac{\frac{\alpha_k^{old}}{\sigma_k^{old}} exp\left(-(x_n - \mu_k^{old})^2 / \left(2(\sigma_k^{old})^2\right)\right)}{\sum_{k=1}^{K} \frac{\alpha_k^{old}}{\sigma_k^{old}} exp\left(-(x_n - \mu_k^{old})^2 / \left(2(\sigma_k^{old})^2\right)\right)}}$

$$f\left(k \middle| x_n, p^{old}\right) = \frac{f\left(k, x_n, p^{old}\right)}{f\left(x_n, p^{old}\right)} = \frac{f(k)\, f\left(x_n, p^{old} \middle| k\right)}{f\left(x_n, p^{old}\right)} = \frac{f(k)\, f_k\left(x_n, p^{old}\right)}{\sum_{k=1}^{K} \alpha_k^{old} f_k(x_n, p^{old})} = \frac{\frac{\alpha_k^{old}}{\sqrt{2\pi}\,\sigma_k^{old}} \exp\left(-(x_n - \mu_k^{old})^2 / \left(2(\sigma_k^{old})^2\right)\right)}{\sum_{k=1}^{K} \frac{\alpha_k^{old}}{\sigma_k^{old}} \exp\left(-(x_n - \mu_k^{old})^2 / \left(2(\sigma_k^{old})^2\right)\right)}$$

$$= \frac{\frac{\alpha_k^{old}}{\sigma_k^{old}} \exp\left(-(x_n - \mu_k^{old})^2 / \left(2(\sigma_k^{old})^2\right)\right)}{\sum_{k=1}^{K} \frac{\alpha_k^{old}}{\sigma_k^{old}} \exp\left(-(x_n - \mu_k^{old})^2 / \left(2(\sigma_k^{old})^2\right)\right)}$$

Montrons que $\boldsymbol{\mu_k^{new} = \frac{\sum_{n=1}^{N} x_n\, f(k|x_n, p^{old})}{\sum_{n=1}^{N} f(k|x_n, p^{old})}}$

Pour cela nous allons dériver *Q(p,p^old)*

$f(k|x_n, p^{old})$ et α_k sont des constantes vu qu'on connaît déjà p^{old}. On calcule la dérivée partielle de $log(f_k(x_n, p_k))$.

$$\frac{df_k(x_n, p^{old})}{\partial \mu_k} = \frac{1}{\sqrt{2\pi}\,\sigma_k^{old}} \frac{(x_n - \mu_k^{old})}{2(\sigma_k^{old})^2} \exp\left(-\frac{\left(x_n - \mu_k^{old}\right)^2}{2(\sigma_k^{old})^2}\right) = f_k(x_n, p^{old}) \frac{(x_n - \mu_k^{old})}{2(\sigma_k^{old})^2}$$

$$\frac{dlog(f_k(x_n, p^{old}))}{\partial \mu_k} = \frac{f_k(x_n, p^{old}) \frac{(x_n - \mu_k^{old})}{2(\sigma_k^{old})^2}}{f_k(x_n, p^{old})} = \frac{(x_n - \mu_k^{old})}{2(\sigma_k^{old})^2}$$

Posons $A = \sum_{k=1}^{K} f\left(k \middle| x_n, p^{old}\right) \log(\alpha_k)$ et $B = \sum_{k=1}^{K} f\left(k \middle| x_n, p^{old}\right) \log\left(f_k(x_n, p_k)\right)$

$\frac{dA}{\partial \mu_k} = 0$

$\frac{dB}{\partial \mu_k} = f\left(k \middle| x_n, p^{old}\right) \frac{dlog(f_k(x_n, p_k))}{\partial \mu_k} = f\left(k \middle| x_n, p^{old}\right) \frac{(x_n - \mu_k^{old})}{2(\sigma_k^{old})^2}$

$$\frac{dQ(p,p^{old})}{\partial\mu_k} = \sum_{n=1}^{N}\frac{dA}{\partial\mu_k} + \frac{dB}{\partial\mu_k} = \sum_{n=1}^{N} f\left(k\middle|x_n,p^{old}\right)\frac{(x_n-\mu_k^{old})}{2(\sigma_k^{old})^2}$$

$$\frac{dQ(p,p^{old})}{\partial\mu_k} = 0 \quad\Leftrightarrow\quad \sum_{n=1}^{N} f\left(k\middle|x_n,p^{old}\right)(x_n-\mu_k^{new}) = 0$$

$$\Leftrightarrow\quad \mu_k^{new} = \frac{\sum_{n=1}^{N} x_n\, f(k|x_n,p^{old})}{\sum_{n=1}^{N} f(k|x_n,p^{old})}$$

Vérifions que μ_k est un maximum :

$\frac{d^2Q(p,p^{old})}{\partial{\mu_k}^2} = \sum_{n=1}^{N} f\left(k\middle|x_n,p^{old}\right)\frac{-1}{2(\sigma_k^{old})^2}$ or $f\left(k\middle|x_n,p^{old}\right)\frac{-1}{2(\sigma_k^{old})^2} < 0$ Donc $Q\left(p,p^{old}\right)$ est concave et μ_k est bien un maximum.

Démontrons $(\boldsymbol{\sigma_k^{new}})^2 = \frac{\sum_{n=1}^{N}(\boldsymbol{x_n}-\boldsymbol{\mu_k^{new}})^2 \boldsymbol{f(k|x_n,p^{old})}}{\sum_{n=1}^{N}\boldsymbol{f(k|x_n,p^{old})}}$

Calculons la dérivée partielle de $log(f_k(x_n,p_k))$:

$$\frac{df_k(x_n,p_k)}{\partial\sigma_k} = \frac{-1}{\sqrt{2\pi}\,(\sigma_k^{old})^2}\exp\left(-\frac{(x_n-\mu_k^{new})^2}{2(\sigma_k^{old})^2}\right) + \frac{1}{\sqrt{2\pi}\,(\sigma_k^{old})}\,\frac{(x_n-\mu_k^{new})^2}{2(\sigma_k^{old})^4}\,2\sigma_k^{old}\exp\left(-\frac{(x_n-\mu_k^{new})^2}{2(\sigma_k^{old})^2}\right)$$

$$= f_k\left(x_n,p^{old}\right)\left(-\frac{1}{\sigma_k^{old}} + \frac{(x_n-\mu_k^{new})^2}{(\sigma_k^{old})^3}\right)$$

$$\frac{dlog(f_k(x_n,p^{old}))}{\partial\sigma_k} = \frac{f_k(x_n,p^{old})\left(-\frac{1}{\sigma_k^{old}}+\frac{(x_n-\mu_k^{new})^2}{(\sigma_k^{old})^3}\right)}{f_k(x_n,p^{old})} = \left(-\frac{1}{\sigma_k^{old}} + \frac{(x_n-\mu_k^{new})^2}{(\sigma_k^{old})^3}\right)$$

Calculons $\frac{dA}{\partial\sigma_k}$ et $\frac{dB}{\partial\sigma_k}$:

$$\frac{dA}{\partial\sigma_k} = 0$$

$$\frac{dB}{\partial\sigma_k} = f\left(k\middle|x_n,p^{old}\right)\frac{dlog(f_k(x_n,p_k))}{\partial\sigma_k} = f\left(k\middle|x_n,p^{old}\right)\left(-\frac{1}{\sigma_k^{old}} + \frac{(x_n-\mu_k^{new})^2}{(\sigma_k^{old})^3}\right)$$

$$\frac{dQ(p,p^{old})}{\partial\sigma_k} = \sum_{n=1}^{N}\frac{dA}{\partial\sigma_k} + \frac{dB}{\partial\sigma_k} = \sum_{n=1}^{N} f\left(k\middle|x_n,p^{old}\right)\left(-\frac{1}{\sigma_k^{old}} + \frac{(x_n-\mu_k^{new})^2}{(\sigma_k^{old})^3}\right)$$

$$\frac{dQ(p,p^{old})}{\partial\sigma_k} = 0 \quad\Leftrightarrow\quad \sum_{n=1}^{N} f\left(k\middle|x_n,p^{old}\right)\left(-\frac{1}{\sigma_k^{new}} + \frac{(x_n-\mu_k^{new})^2}{(\sigma_k^{new})^3}\right) = 0$$

$$\Leftrightarrow\quad \sum_{n=1}^{N} f\left(k\middle|x_n,p^{old}\right)\frac{1}{\sigma_k^{new}} = \sum_{n=1}^{N} f\left(k\middle|x_n,p^{old}\right)\frac{(x_n-\mu_k^{new})^2}{(\sigma_k^{new})^3}$$

$$\Leftrightarrow\quad \frac{(\sigma_k^{new})^3}{\sigma_k^{new}} = \frac{\sum_{n=1}^{N}(x_n-\mu_k^{new})^2 f\left(k\middle|x_n,p^{old}\right)}{\sum_{n=1}^{N} f\left(k\middle|x_n,p^{old}\right)}$$

$$\Leftrightarrow\quad (\sigma_k^{new})^2 = \frac{\sum_{n=1}^{N}(x_n-\mu_k^{new})^2 f(k|x_n,p^{old})}{\sum_{n=1}^{N} f(k|x_n,p^{old})}$$

Question 3.5

Proposer un algorithme de « Clustering »