Algorithme EM

M2 Stats de la SD, CM3, 2025-2026

Hadrien Lorenzo
Aix Marseille Université

Au menu

Objectifs de l’EM

Double objectif

Dans la suite on considère le cas général d'un modèle statistique

$$(\mathbf{X}, (P_\theta)_{\theta \in \Theta})$$

où $(P_\theta)_{\theta \in \Theta}$ est une famille de probabilités de paramètre $\theta \in \Theta$. Plusieurs objectifs :

- **Estimation d'un paramètre optimal**

Justification de la méthode du maximum de vraisemblance dans le cas de jeux de données avec données manquantes

- **Simulation de données et de paramètres**

Comment construire, conditionellement à des données observées et en présence de données manquantes, des distributions pour :

- des données manquantes,
- des paramètres θ.

Maximisation de vraisemblance

Maximisation de vraisemblance, cas indépendant

Soit une fonction de densité paramétrique $p(x|\theta)$ qui dépend du paramètre $\theta \in \Theta$.

Dans le cas d'un échantillon de n observations indépendantes $\boldsymbol{x} = (x_1, \ldots, x_n)$, la vraisemblance s'écrit :

$$L(\boldsymbol{x}, \theta) = \prod_{i=1}^{n} p(x_i|\theta).$$

Plus $L(\boldsymbol{x}, \theta)$ est importante et plus θ est vraisemblable pour les données $\boldsymbol{x}$.

Ce que signifie $p(x_i|\theta)$

Comprendre ici la dérivée de Radon-Nikodym de la $i^{\text{ème}}$ mesure $P_{i,\theta}$ par rapport à une mesure de référence μ_i pour l'observation i En vertu du théorème de Radon-Nikodym pour une mesure produit, voir CM2. Nous sommes en présence d'un produit de tribus : $P_\theta = \bigotimes_{i=1}^{n} P_{i,\theta}$.

Maximisation de vraisemblance, données manquantes

Le meilleur paramètre pour nos données $\boldsymbol{x}$ sera donc celui qui maximise cette vraisemblance.

$$\max_{\theta \in \Theta} \; L(\boldsymbol{x}, \theta).$$

Une solution correspond à l'estimateur **du maximum de vraisemblance** (EMV) qui peut ne pas être unique.

Il nous intéresse aujourd'hui de regarder une approche de maximisation de cette vraisemblance lorsque des données sont manquantes.

A retenir tout de suite

L'algorithme EM est une méthode itérative qui permet de maximiser la vraisemblance.

C'est un algorithme d'optimisation, une méthode numérique.

Il est utilisé dans le cadre de la gestion de données manquantes car il permet d'approximer l'**EMV**.

Donc

L'algorithme EM **n'est pas** toujours utilisé pour gérer des données manquantes, il a pour principale utilité d'approximer l'EMV.

Un algorithme alterné

Deux étapes à itérer

EM (pour Expectation-Maximization) permet d'estimer des paramètres θ dans un modèle à vraisemblance, mais aussi les données manquantes.

C'est un algorithme itératif à deux étapes. On part d'une initialisation et pour une itération $k > 0$, il vient :

- Expectation : Ecrire l'espérance conditionelle de la log-vraisemblance complétée en fonction du jeu de donnée observé et du paramètre estimé à l'itération précédente, θ_{k-1}.

- Maximisation : Maximiser l'espérance conditionelle précédente sous $\theta \in \Theta$.

Outils utilisés

Soit $\boldsymbol{x} \in \mathcal{X}$ les données observées et $\boldsymbol{z} \in \mathcal{Z}$ les données manquantes, alors la fonction de densité jointe de $(\boldsymbol{x}, \boldsymbol{z})$ peut s'écrire

$$p\left(\boldsymbol{x}, \boldsymbol{z} \middle| \theta\right) = p\left(\boldsymbol{x} \middle| \boldsymbol{z}, \theta\right) p\left(\boldsymbol{z} \middle| \theta\right).$$

Ainsi, par intégration sur $\mathcal{Z}$ l'ensemble formé par les évènements possibles pour les données manquantes il vient

$$p\left(\boldsymbol{x} \middle| \theta\right) = \int_{\boldsymbol{z} \in \mathcal{Z}} p\left(\boldsymbol{x}, \boldsymbol{z} \middle| \theta\right) d\boldsymbol{z} = \int_{\boldsymbol{z} \in \mathcal{Z}} p\left(\boldsymbol{x} \middle| \boldsymbol{z}, \theta\right) p\left(\boldsymbol{z} \middle| \theta\right) d\boldsymbol{z}$$

L'objectif de l'algorithme est de maximiser la log vraisemblance $\ell(\boldsymbol{x}, \theta)$:

$$\ell(\boldsymbol{x}, \theta) = \ln\left(p(\boldsymbol{x}|\theta)\right).$$

D_{KL}

Si l'on choisit $\boldsymbol{z} \mapsto q(\boldsymbol{z})$, une densité de probabilité sur $\mathcal{Z}$ qui est nulle seulement là où $\boldsymbol{z} \mapsto p\left(\boldsymbol{x}, \boldsymbol{z} | \theta\right)$ est aussi nulle, alors :

$$\begin{aligned}
\ell(\boldsymbol{x}, \theta) &= \int_{\mathcal{Z}} q(\boldsymbol{z}) \ell(\boldsymbol{x}, \theta) d\boldsymbol{z} = \int_{\mathcal{Z}} q(\boldsymbol{z}) \ln\left(\frac{p(\boldsymbol{x}|\theta) p(\boldsymbol{z}|\boldsymbol{x}, \theta)}{p(\boldsymbol{z}|\boldsymbol{x}, \theta)}\right) d\boldsymbol{z}, \\
&= \int_{\mathcal{Z}} q(\boldsymbol{z}) \ln\left(\frac{p(\boldsymbol{x}, \boldsymbol{z}|\theta)}{p(\boldsymbol{z}|\boldsymbol{x}, \theta)}\right) d\boldsymbol{z} = \int_{\mathcal{Z}} q(\boldsymbol{z}) \ln\left(\frac{p(\boldsymbol{x}, \boldsymbol{z}|\theta) q(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x}, \theta) q(\boldsymbol{z})}\right) d\boldsymbol{z}, \\
&= \int_{\mathcal{Z}} q(\boldsymbol{z}) \ln\left(\frac{p(\boldsymbol{x}, \boldsymbol{z}|\theta)}{q(\boldsymbol{z})}\right) d\boldsymbol{z} + \int_{\mathcal{Z}} q(\boldsymbol{z}) \ln\left(\frac{q(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x}, \theta)}\right) d\boldsymbol{z}, \\
&= F(q, \theta) + KL\left\{q || p(\cdot|\boldsymbol{x}, \theta)\right\}.
\end{aligned}$$

On apelle divergence de Kullblack-Leibler : $(p, q) \mapsto D_{KL}\left\{p || q\right\} = \int_{\mathcal{Z}} p \log\left(p/q\right)$ et on peut montrer que cette divergence est positive ou nulle. Nulle si et seulement si $p = q$.

ELBO

Il vient que la fonctionelle $(q, \theta) \mapsto F(q, \theta)$ agit comme un minorant de la log-vraisemblance, on l'apelle souvant **Evidence Lower Bound (ELBO)** et ainsi :

$$\ell(\boldsymbol{x}, \theta) \geq F(q, \theta).$$

L'objectif de l'**EM** est de maximiser cette fonctionnelle qui va ainsi assurer, après convergence d'un algorithme itératif, à la log-vraisemblance d'être maximale.

$\mathbb{E}_{\theta_0}\left[\ln p\left(\boldsymbol{X}, \boldsymbol{Z}|\theta\right)|\boldsymbol{X} = \boldsymbol{x}\right]$

Si l'on connaît la forme de la loi instrumentale de Z sachant un paramètre $\theta_0 \in \Theta$ et nos données $\boldsymbol{x}$, on peut écrire $q(\boldsymbol{z}) = p(\boldsymbol{z}|\boldsymbol{x}, \theta_0)$ et il vient :

$$
\begin{aligned}
F(q, \theta) &= \int_{\mathcal{Z}} q(\boldsymbol{z}) \ln\left(\frac{p(\boldsymbol{x}, \boldsymbol{z}|\theta)}{q(\boldsymbol{z})}\right) d\boldsymbol{z}, \\
&= \int_{\mathcal{Z}} p(\boldsymbol{z}|\boldsymbol{x}, \theta_0) \ln\left(\frac{p(\boldsymbol{x}, \boldsymbol{z}|\theta)}{p(\boldsymbol{z}|\boldsymbol{x}, \theta_0)}\right) d\boldsymbol{z}, \\
&= \int_{\mathcal{Z}} p(\boldsymbol{z}|\boldsymbol{x}, \theta_0) \ln\left(p(\boldsymbol{x}, \boldsymbol{z}|\theta)\right) d\boldsymbol{z} - \int_{\mathcal{Z}} p(\boldsymbol{z}|\boldsymbol{x}, \theta_0) \ln\left(p(\boldsymbol{z}|\boldsymbol{x}, \theta_0)\right) d\boldsymbol{z}, \\
&= Q(\theta|\theta_0, \boldsymbol{x}) + cste(\theta_0, q), \\
&= \mathbb{E}_{\theta_0}\left[\ln p\left(\boldsymbol{X}, \boldsymbol{Z}|\theta\right)|\boldsymbol{X} = \boldsymbol{x}\right] + cste(\theta_0, q).
\end{aligned}
$$

Il vient l'objectif, finalement, de maximiser $Q(\theta|\theta_0, \boldsymbol{x})$, qui est

l'espérance conditionelle de la log-vraisemblance complétée.

Suite croissante et majorée

L'algorithme EM construit une suite de paramètres $(\theta_k)_{k \geq 0}$ tels que $\forall k > 0$:

$$\theta_{k+1} = \arg\max_{\theta \in \Theta} \; Q(\theta | \theta_k, \boldsymbol{x}).$$

Ces éléments vérifient $\ell(\theta_{k+1}|\boldsymbol{x}) \geq \ell(\theta_k|\boldsymbol{x}) \geq \ldots \geq \ell(\theta_0|\boldsymbol{x})$. La suite $(\ell(\theta_k|\boldsymbol{x}))_{k \geq 0}$ est donc croissante. Cette suite est aussi majorée par le maximum de la function maximisée. Cette suite converge donc.

Important

Dans les faits, cette convergence est lente et se fait souvent vers un maximum local, ce qui nécessite plusieurs initialisations.

L'espérance conditionelle

$$\begin{aligned} Q\left(\theta|\theta_0, \boldsymbol{x}\right) &= \mathbb{E}_{\theta_0}\left[\ln p\left(\boldsymbol{X}, \boldsymbol{Z}|\theta\right)|\boldsymbol{X} = \boldsymbol{x}\right] \\ \ell_c(\boldsymbol{x}, \boldsymbol{Z}, \theta) &= \ln p\left(\boldsymbol{x}, \boldsymbol{Z}|\theta\right). \end{aligned}$$

La log-vraisemblance complétée, $\ell_c(\boldsymbol{x}, \boldsymbol{Z}, \theta)$, est qualifiée ainsi car les données manquantes, $\boldsymbol{Z}$, sont ajoutées, ce qui permet d'avoir accès à un jeu de données complet $(\boldsymbol{x}, \boldsymbol{Z})$.

La log-vraisemblance complétée

L'objectif est donc de trouver une variable aléatoire $\boldsymbol{Z}$ telle que la vraisemblance *complétée* $\ell_c(\boldsymbol{x}, \boldsymbol{Z}, \theta)$, ou plus précisément son espérance sous un paramètre θ_0 particulier, soit plus simple à maximiser que la vraisemblance *observée* $\ell(\boldsymbol{x}, \theta)$.

Remarque importante

Dans certains cas, le calcul de l'espérance conditionelle revient au calcul des espérances conditionelles des variables latentes z_i.

Important

On voit souvent la notation, pour une variable Z et un paramètre θ_k donnés :

$$\mathbb{E}_{\theta_k}[Z|X = x] = <z>_k.$$

Exemple des cellules

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

$$\begin{aligned} L(x, \theta) &= (2+\theta)^{x_1}(1-\theta)^{x_2+x_3}\theta^{x_4}, \\ \ell(x, \theta) &= (x_2 + x_3)\log(1-\theta) + x_4 \log(\theta) + x_1 \log(2+\theta), \end{aligned}$$

qui n'est pas forcément simple à maximiser.

On introduit deux variables latentes Z_1 et Z_2 telles que $X_1 = Z_1 + Z_2$ et $Z_2|X = x$ suit une loi binomiale de paramètre $(x_1, \theta/(\theta+2))$. Il vient que $Z_1|X = x, Z_2 = z_2$ est une masse de Dirac centrée en $x_1 - z_2$:

$$[Z_1|X = x, Z_2 = z_2] \sim \delta_{x_1 - z_2}.$$

Soit $x = (x_1, x_2, x_3, x_4) \in \mathbb{N}^4$ tel que $n = x_1 + x_2 + x_3 + x_4$ et $z = (z_1, z_2) \in \mathbb{N}^2$ tel que $x_1 = z_1 + z_2$. Alors $\delta_{x_1 - z_2}(z_1) = 1$ et la vraisemblance complétée $L_c(x, z, \theta) = \mathbb{P}(X = x, Z = z|\theta)$, vérifie :

$$\begin{aligned}
L_c(x, z, \theta) &= \mathbb{P}(Z_1 = z_1, Z_2 = z_2|X = x, \theta)\mathbb{P}(X = x|\theta) \\
&= \mathbb{P}(Z_1 = z_1|X = x, Z_2 = z_2, \theta)\mathbb{P}(Z_2 = z_2|X = x, \theta)\mathbb{P}(X = x|\theta) \\
&= \delta_{x_1 - z_2}(z_1)\mathbb{P}(Z_2 = z_2|X = x, \theta)L(x, \theta) \\
&= \left(\frac{\theta}{\theta + 2}\right)^{z_2}\left(\frac{2}{\theta + 2}\right)^{x_1 - z_2} L(x, \theta) \\
&\propto \left(\frac{\theta}{\theta + 2}\right)^{z_2}\left(\frac{2}{\theta + 2}\right)^{x_1 - z_2} (2 + \theta)^{x_1}(1 - \theta)^{x_2 + x_3}\theta^{x_4} \\
&\propto \theta^{z_2 + x_4}(1 - \theta)^{x_2 + x_3},
\end{aligned}$$

La log-vraisemblance complétée, en repassant aux variables aléatoires Z_1 et Z_2 vérifiant $x_1 = Z_1 + Z_2$, s'écrit donc :

$$l_c(x, Z, \theta) = \mathrm{cste} + (Z_2 + x_4)\log(\theta) + (x_2 + x_3)\log(1-\theta),$$

Sur laquelle on peut appliquer l'opérateur espérance conditionelle sachant $(X = x, \theta_0)$:

$$\begin{aligned} Q(\theta|\theta_0, x) &= \mathbb{E}_{\theta_0}[l_c(x, Z, \theta)|X = x] \\ &= \mathrm{cste} + (\mathbb{E}_{\theta_0}[Z_2|X = x] + x_4)\log(\theta) + (x_2 + x_3)\log(1-\theta). \end{aligned}$$

On se souvient que $[Z_2|X = x]$ suit une binomiale de paramètre $(x_1, \theta_0/(\theta_0 + 2))$ et ainsi $\mathbb{E}_{\theta_0}[Z_2|X = x] = x_1\theta_0/(\theta_0 + 2)$. Il vient :

$$Q(\theta|\theta_0, x) = \mathrm{cste} + \left(x_1\frac{\theta_0}{\theta_0 + 2} + x_4\right)\log(\theta) + (x_2 + x_3)\log(1-\theta).$$

$$Q(\theta|\theta_0, x) = \text{cste} + \left(x_1 \frac{\theta_0}{\theta_0 + 2} + x_4\right) \log(\theta) + (x_2 + x_3) \log(1 - \theta),$$

qui est relativement simple à maximiser :

$$\theta_1 = \arg\max_{\theta \in [0,1]} Q(\theta|\theta_0, x) = \frac{x_1 \frac{\theta_0}{\theta_0 + 2} + x_4}{x_1 \frac{\theta_0}{\theta_0 + 2} + x_2 + x_3 + x_4}.$$

A retenir

On complexifie l'espace de mesure pour avoir accès à une vraisemblance complétée plus simple à maximiser.

On maximise ensuite l'espérance conditionelle de cette log-vraisemblance complétée pour obtenir un paramètre plus vraisemblable.

On réitère ces deux étapes jusqu'à convergence.

Critère(s) de convergence

L'algrotihme ainsi construit maximise la vraisemblance, il paraît ainsi raisonnable d'utiliser la stabilisation de la vraisemblance commme critère de convergence.

Attention 1 : La vraisemblance s'exprime souvent dans une échelle très faible : lui préférer la log-vraisemblance.

Attention 2 : La vraisemblance peut être complexe/couteuse à évaluer, lui préférer parfois un critère sur la convergence du paramètre θ_k.

Remarque importante

Dans l'exemple précédent nous avons explicitement écrit le conditionnement en X pour les calculs d'espérances $\mathbb{E}_{\theta_0}[l_c(x, Z, \theta)|X = x]$ et $\mathbb{E}_{\theta_0}[Z_2|X = x]$.

Ceci car le contexte était très clairement coloré probabiliste. Dans un contexte purement statistique, ce conditionnement est inutil puisque l'on travaille toujours à données acquises.

Pour cette raison le conditionnement en X est souvent omis dans la littérature.

Capteur saturé

On accède aux données d'un capteur de température qui sature à partir d'une température a. On sait que la température suit une loi normale de moyenne θ et de variance 1. Parmi les $n = n_1 + n_2$ données, les n_1 premières ne sont pas saturées ($x_i < a$) alors que les n_2 suivantes sont saturées ($x_i \geq a$). On cherche tout de même à estimer θ, la vraie valeur de la température.

Dans ce cas, la vraisemblance du problème s'écrit

$$L(\boldsymbol{x}, \theta) = (1 - \Phi(a - \theta))^{n_2} \prod_{i=1}^{n_1} \varphi(x_i - \theta),$$

où φ et Φ sont les fonctions de densité et de répartition de la loi normale centrée réduite.

On introduit des variables latentes z_i et on écrit alors la vraisemblance complétée

$$L_c(\boldsymbol{x}, \boldsymbol{z}, \theta) = \prod_{i=1}^{n_1} \varphi(x_i - \theta) \prod_{i=1}^{n_2} \varphi(z_i - \theta),$$

qui est simple à optimiser. La log-vraisemblance complétée $\ell_c(\boldsymbol{x}, \boldsymbol{z}, \theta)$ devient :

$$\ell_c(\boldsymbol{x}, \boldsymbol{z}, \theta) = -\frac{1}{2} \sum_{i=1}^{n_1} (x_i - \theta)^2 - \frac{1}{2} \sum_{i=1}^{n_2} (z_i - \theta)^2 + \boldsymbol{cste}.$$

Son espérance conditionelle est, à une constante additive près :

$$\begin{aligned} \mathbb{E}_{\theta_k}\left[\ell_c(\boldsymbol{X}, \boldsymbol{Z}, \theta) | \boldsymbol{X} = \boldsymbol{x}\right] &= -\frac{1}{2} \sum_{i=1}^{n_1} (x_i - \theta)^2 - \frac{1}{2} \sum_{i=1}^{n_2} \mathbb{E}_{\theta_k}[(Z_i - \theta)^2 | \boldsymbol{X} = \boldsymbol{x}] \\ &= -\frac{1}{2} \sum_{i=1}^{n_1} (x_i - \theta)^2 - \frac{1}{2} \sum_{i=1}^{n_2} \mathbb{E}_{\theta_k}[(Z_i - \theta)^2 | X_i \geq a] \end{aligned}$$

$$
\begin{aligned}
\mathbb{E}_{\theta_k}[(Z_i-\theta)^2|X_i\geq a] &= \mathbb{E}_{\theta_k}[Z_i^2-2Z_i\theta+\theta^2|X_i\geq a] \\
&= \mathbb{E}_{\theta_k}[Z_i^2|X_i\geq a]-2\theta\mathbb{E}_{\theta_k}[Z_i|X_i\geq a]+\theta^2 \\
&= <z_i^2>_k -2\theta <z_i>_k +\theta^2
\end{aligned}
$$

$$
\frac{d\,\mathbb{E}_{\theta_k}\left[\ell_c(\boldsymbol{X},\boldsymbol{Z},\cdot)|\boldsymbol{X}=\boldsymbol{x}\right]}{d\,\theta}(\theta)=\sum_{i=1}^{n_1}(x_i-\theta)+\sum_{i=1}^{n_2}(<z_i>_k -\theta)
$$

La fonctionelle $Q(\theta|\theta_k,\boldsymbol{x})=\mathbb{E}_{\theta_k}\left[\ell_c(\boldsymbol{X},\boldsymbol{Z},\theta)|\boldsymbol{X}=\boldsymbol{x}\right]$ est alors maximale en

$$
\theta_{k+1}=\frac{\sum_{i=1}^{n_1}x_i+\sum_{i=1}^{n_2}<z_i>_k}{n}
$$

et le terme en n_2 a tendance à tirer la moyenne vers lui.

Pour ce qui est de $<z_i>_k$:

$$\begin{aligned}<z_i>_k = \mathbb{E}_{\theta_k}[Z_i|X_i \geq a] = \mathbb{E}_{\theta_k}[X_i|X_i \geq a] &= \frac{\mathbb{E}_{\theta_k}[X_i \mathbb{1}_{X_i \geq a}]}{\mathbb{P}_{\theta_k}(X_i \geq a)} \\ &= \frac{\int_a^\infty x\varphi(x-\theta_k)dx}{\int_a^\infty \varphi(x-\theta_k)dx} = \frac{\int_a^\infty z\varphi(z-\theta_k)dz}{1-\Phi(a-\theta_k)} \\ &= \frac{\int_a^\infty (t+\theta_k)\varphi(t)dt}{1-\Phi(a-\theta_k)} = \frac{\theta_k \int_{a-\theta_k}^\infty \varphi(t)dt + \int_{a-\theta_k}^\infty t\varphi(t)dt}{1-\Phi(a-\theta_k)} \\ &= \frac{\theta_k(1-\Phi(a-\theta_k)) + \int_{a-\theta_k}^\infty t\varphi(t)dt}{1-\Phi(a-\theta_k)}\end{aligned}$$

Il faut alors se souvenir que $\varphi'(t) = -t\varphi(t)$ et donc $\int_{a-\theta_k}^\infty t\varphi(t)dt = -\int_{a-\theta_k}^\infty \varphi'(t)dt = -[\varphi(t)]_{a-\theta_k}^\infty = \varphi(a-\theta_k)$ et ainsi :

$$<z_i>_k = \theta_k + \frac{\varphi(a-\theta_k)}{1-\Phi(a-\theta_k)}.$$

L'algorithme **EM** recherché réalise donc les deux étapes suivantes, après initialisation de θ via θ_0, et pour toute itération $k > 0$:

- **Espérance** : Calcul de $Q(\theta|\theta_k, \boldsymbol{x})$, celà revient ici à calculer:

$$\forall i \in \{1, \ldots, n_2\}, \quad <z_i>_k = \theta_k + \frac{\varphi(a - \theta_k)}{1 - \Phi(a - \theta_k)} = <z_1>_k.$$

- **Maximisation** : on calcule $\theta_{k+1} = \dfrac{\sum_{i=1}^{n_1} x_i + n_2 <z_1>_k}{n}$.

> **Note**
>
> L'espérance conditionelle $\mathbb{E}_{\theta_k}[Z_i|\boldsymbol{X} = \boldsymbol{x}]$ est indépendante de l'individu considéré : on arrive à la conclusion que l'hypothèse **MNAR**, non discutée précédemment, est bien vérifiée.

ⓘ Qui de l'imputation ?

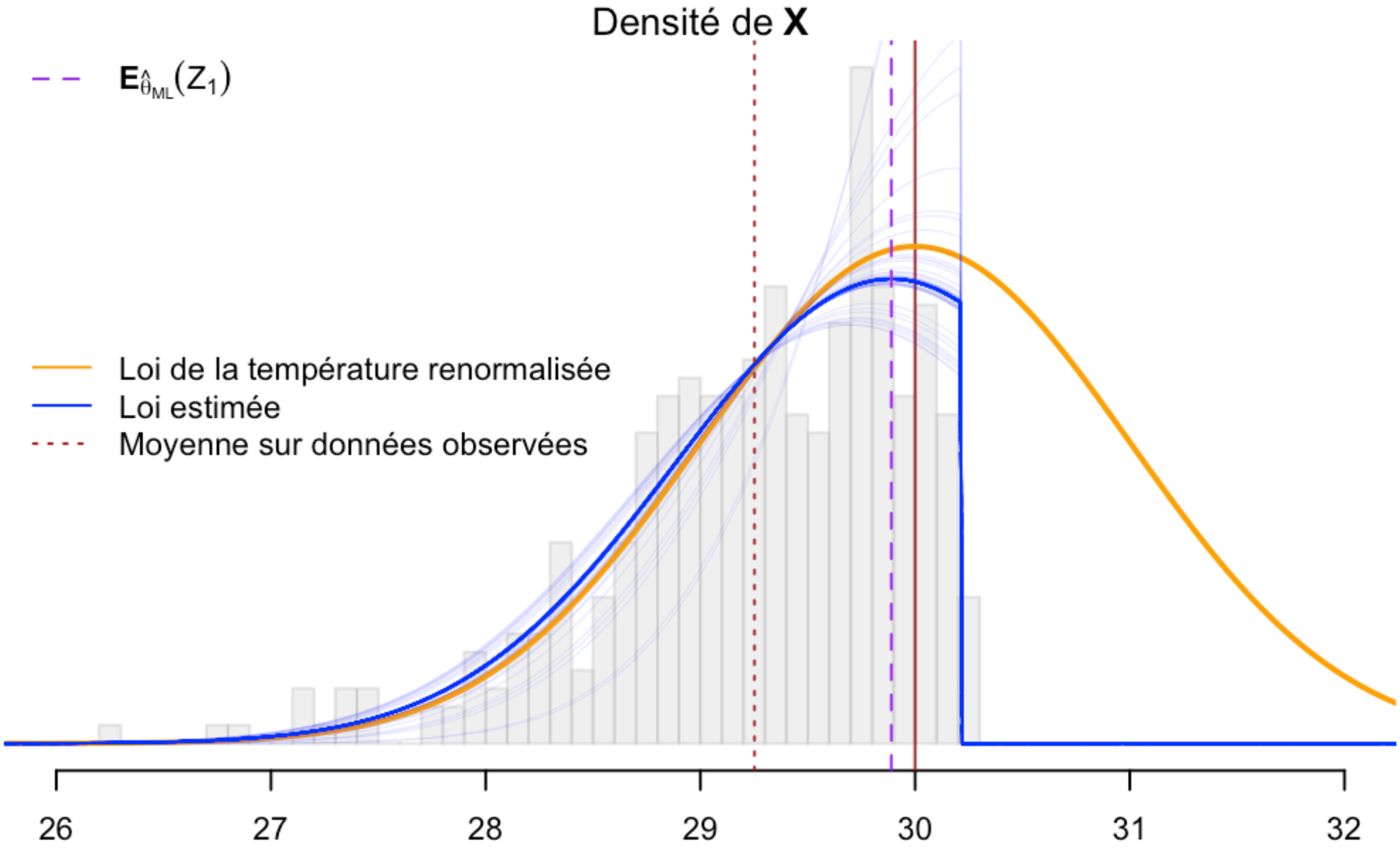
Dans cet exemple, l'estimateur du maximum de vraissemblance, noté $\hat{\theta}_{ML}$, est la limite de la suite

$$\hat{\theta}_{ML} = \lim_{k \to \infty} \frac{\sum_{i=1}^{n_1} x_i + n_2 <z_1>_k}{n}.$$

Sous le modèle choisi, tel que $\theta = \hat{\theta}_{ML}$, la **meilleure** valeur imputée pour chaque z_i (comprendre **la plus vraisemblable**) est

$$\forall i \in \{1, \ldots, n_2\}, \quad <z_i>_{ML} = \hat{\theta}_{ML} + \frac{\varphi(a - \hat{\theta}_{ML})}{1 - \Phi(a - \hat{\theta}_{ML})}.$$

Application numérique avec $\theta = 30, n_1 = 400, n_2 = 250$ et $a = 30.22$:

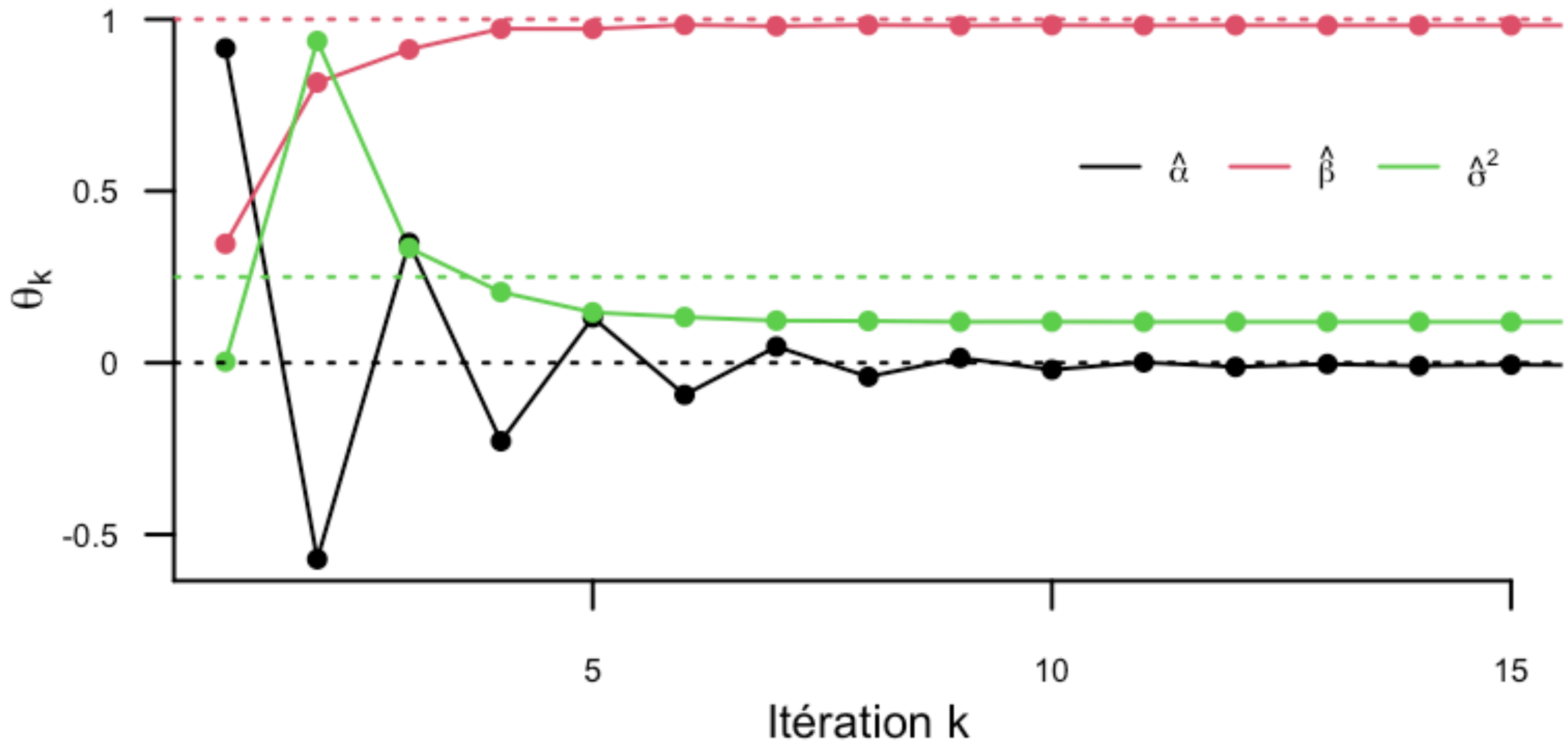


Important

On voit deux choses importantes :

- L'estimation du paramètre θ sur les données seulement observées (29.25) est plus mauvaise que l'estimation sur les données observées et sur l'information de saturation, $\hat{\theta}_{ML} = 29.89$.
- L'imputation par la valeur la plus vraissemblable, $< z_i >_{ML}$, est *mauvaise*. Ceci est principalement dû au fait que le processus est **MNAR** : l'imputation ne peut pas particulariser sur l'individu considéré.

Retour à notre exemple



Les valeurs estimées sont alors

$$\hat{\alpha}_{\text{EM}} \approx -0.007 \pm 0.04,\ \hat{\beta}_{\text{EM}} \approx 0.982 \pm 0.038,\ \hat{\sigma}^2_{\text{EM}} \approx 0.119$$

Conclusion

Nous avons discuté de l'approche EM pour la maximisation de la vraisemblance en présence de données manquantes.

Nous avons vu que cette méthode itérative permettait d'approximer l'estimateur du maximum de vraisemblance.

Aussi, nous avons vu que l'algorithme EM pouvait être relié à une imputation des données manquantes, conditionnellement au modèle statistique estimé à la convergence de l'algorithme.

Il est à noter que l'imputation n'est pas nécessaire pour le focntionnement de l'algorithme EM.

Dans la suite, nous verrons comment l'algorithme EM peut être réalisé dans le cadre de modèles plus complexes, où l'espérance conditionelle est difficile calculable. Dans de nombreux cas, il est nécessaire de simuler selon des lois non tabulées et les approches par chaînes de Markov seront alors utiles.

Dans la suite du cours, nous étudierons les approches MCMC (Monte Carlo Markov Chain).

Références

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.