L'algorithme EM et ses variantes :

On va se concentrer maintenant sur un autre algorithme itératif aussi qui possède de nombreux avantages : sa forme initiale est très utilisée et possède, sous des conditions réalistes, des propriétés de convergence. De plus, quand le problème se complexifie un peu, plusieurs variantes sont disponibles permettant de garder des propriétés de convergence intéressantes et d'avoir une implémentation facile et efficace.

Cet algorithme est l'algorithme EM pour Espectation - Maximisation. On va d'abord voir comment il est apparu puis ses propriétés et enfin ses variantes.

I Rappel du cadre : (Dempster, Laird, Rubin ; 77)

On cherche à maximiser une quantité se mettant sous la forme d'une espérance :

$$\theta^* = \underset{\theta \in \Theta}{\operatorname{argmax}}\ q(y;\theta) = \underset{\theta \in \Theta}{\operatorname{argmax}}\ \mathbb{E}_z[q(y,z;\theta)] \quad (*)$$

où y est une donnée que l'on observe, typiquement un n-uplet de scalaires $(y_1, \dots, y_n)$ et z est une variable aléatoire de loi $q(z;\theta)$. Notez que cette loi peut dépendre de θ.

Rq : • Si $n=1$, on se retrouve dans le cas où on a à maximiser $\int q(y_1, z;\theta)\, d\mu(z)$ ce qui se ramène au cas considéré précédemment.

• Si $n \neq 1$ mais (y_i) i.i.d, on se retrouve aussi avec $\int (\sum q(y_i, z;\theta))\, d\mu(z)$ à maximiser. C'est souvent ce qui se passe en pratique.

• Pourquoi EM plutôt que GS ? Pour éviter des calculs d'espérance du gradient ; on va voir comment tout de suite.

Terminologie : $y_1, \dots, y_n$: données observées tq $\forall i\ \exists z_i$ tq $y_i | z_i \sim q(y_i | z_i;\theta)$
$z_1, \dots, z_n$: données manquantes
$(\vec{y}, \vec{z})$: données complètes.

Rq : • écrit sous la forme $(*)$, c'est un problème de maximum de vraisemblance. Si θ est considéré comme une v.a sur laquelle on a un apriori, on peut aussi utiliser ce cadre de l'EM pour un maximum a posteriori

• Si on doit maximiser une intégrale DÉTERMINISTE, on réécrivant $\int$ en $\mathbb{E}_\mu \to$ EM.

1) D'où vient l'EM ?

* Pour des raisons de commodité, on se place dans le cadre éventuellement Bayesien $(;\to)$
* On commence avec un θ_0 initial. L'état courant de la suite est noté θ^t.
* Trouver $\underset{\theta\in\Theta}{\operatorname{argmax}}\ q(y,\theta)$ est équivalent à trouver $\theta^* = \operatorname{argmax} \log q(y,\theta)$

  Le problème est qu'il est difficile de trouver cet argmax du fait de l'$\int$.
* idée : Trouver $B(\theta,\theta^t) \le \log q(\theta|y)$ et de maximiser cette fonction B au lieu de $q(.|y)$ $\to$ en itérant, on veut qu'on peut espérer une convergence vers un max local pourvu que B soit "optimale" à chaque étape.
* On réécrit le problème pour trouver cette fonction B optimale.

$$\log q(y,\theta) = \log \int_{\mathcal{Z}} q(y,z,\theta)\, d(z) = \log \int_{\mathcal{Z}} f^t(z)\, \frac{q(y,z,\theta)}{f^t(z)}\, d(z)$$

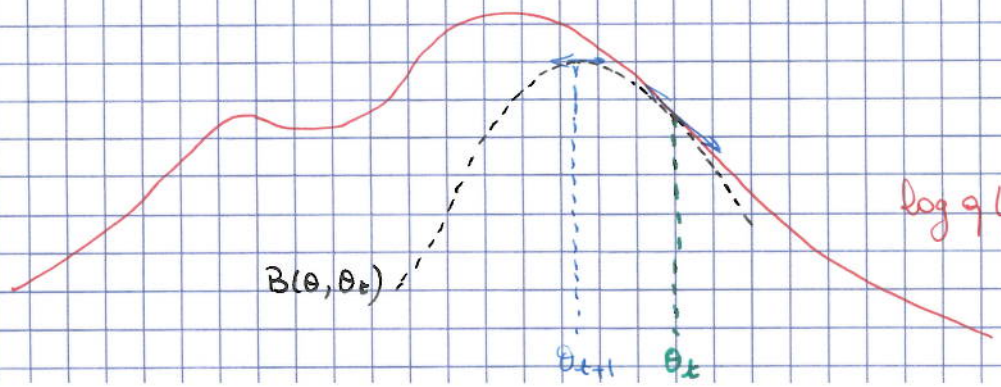
où $f_t(z)$ est une densité non nulle sur $\mathcal{Z}$ dépendant de la valeur courante de θ.

Par l'inégalité de Jensen : $\log \int_{\mathcal{Z}} f^t(z) \frac{q(y,z,\theta)}{f^t(z)}\, dz \geqslant \int f^t(z) \log\left[\frac{q(y,z,\theta)}{f^t(z)}\right] dz$

Cette inégalité de convexité est optimale (car $\exists$ cas d'égalité) donc on pose $B(\theta,\theta^t) = \int f^t(z) \log\left[\frac{q(y,z,\theta)}{f^t(z)}\right] dz$

* Maintenant que l'on a la forme générale de B, il faut trouver la meilleure densité $f_t(z)$ qui permet que l'inégalité soit la meilleure possible.

  L'idée est de prendre $f_t(z)$ tq B touche la fonction objectif $\log q(y,\theta)$ en l'état courant θ_t :



En équation : $B(\theta, \theta_t) = \int f_t(z) \log\left[\frac{q(y,z,\theta)}{f_t(z)}\right] dz$

On veut que $B(\theta_t, \theta_t)$ touche $q(y, \theta_t)$ donc comme B est un minorant on optimise en f_t en maximisant $B(\theta_t, \theta_t)$ par dérivation en f_t.

On cherche donc $f_t^* = \underset{f_t}{\operatorname{argmax}} \int f_t(z) \log\left[\frac{q(y,z,\theta)}{f_t(z)}\right] dz$ sous contrainte de densité : $f_t(z) \geq 0 \ \forall z$ et $\int f_t(z)\, dz = 1$

On introduit un multiplicateur de Lagrange :

$$G(f_t) = \lambda\left[1 - \int f_t(z)\, dz\right] + \int f_t(z) \log\left[\frac{q(y,z,\theta_t)}{f_t(z)}\right] dz$$

$$= \lambda\left[1 - \int f_t(z)\, dz\right] + \int f_t(z) \log q(y,z,\theta_t)\, dz - \int f_t(z) \log f_t(z)\, dz$$

$$\begin{cases} \frac{\partial G}{\partial f_t(z)} = -\lambda + \log q(y,z,\theta_t) - \log f_t(z) - 1 = 0 & (1) \\ \frac{\partial G}{\partial \lambda} = 1 - \int f_t(z)\, dz & (2) \end{cases}$$

$(1) \Leftrightarrow \log f_t(z) = \log q(y,z,\theta_t) - \lambda - 1$

$(2) \Leftrightarrow 1 = \int \exp\left[(-\lambda - 1)\, q(y,z,\theta_t)\right] dz = e^{-\lambda-1} \int q(y,z,\theta_t)\, dz$

$= e^{-\lambda-1} \int q(y,\theta_t)$ (marginale)

D'où $\boxed{f_t(z) = q(y,z,\theta_t) \times \frac{1}{q(y,\theta_t)} = q(z|y,\theta_t)}$

On vérifie facilement que $B(\theta_t, \theta_t) = q(y, \theta_t)$:

$$B(\theta_t, \theta_t) = \int q(z|y,\theta_t) \log\left[\frac{q(y,z,\theta_t)}{q(z|y,\theta_t)}\right] dz$$

$$= \int q(y,z,\theta_t) \frac{1}{q(y,\theta_t)} \log\left[\frac{q(y,z,\theta_t)\, q(y,\theta_t)}{q(y,z,\theta_t)}\right] dz$$

$$= \frac{\log q(y,\theta_t)}{q(y,\theta_t)} \int q(y,z,\theta_t)\, dz = \log q(y,\theta_t) \times \frac{q(y,\theta_t)}{q(y,\theta_t)}$$

$$= \log q(y,\theta_t)$$

* Dernière étape : il faut maximiser $B(\theta, \theta_t)$ en θ.

$$B(\theta,\theta_t) = \mathbb{E}_{f_t(z)}\left[\log q(y,z,\theta_t) - \log f^t(z)\right]$$

$$= \mathbb{E}_{f_t}\left[\log q(y,z|\theta) + \log q(\theta)\right] + \mathcal{H}(f_t)$$

↳ dans le cas Bayesien ↳ entropie de f

$$= Q(\theta|\theta_t) + \log q(\theta) + \underbrace{\mathcal{H}(f_t)}_{\text{indépendante de } \theta \text{ à optimiser}}$$

Donc $\theta_{t+1} = \underset{\theta\in\Theta}{\operatorname{argmax}}\ Q(\theta|\theta_t)$ (+ $\log q(\theta)$ dans le cas Bayesien)

2) L'algorithme en deux étapes :

(E) : Calculer $Q(\theta|\theta_t) = \mathbb{E}\left[\log q(y,z,\theta) \mid y, \theta_t\right]$

où l'espérance est prise contre la loi a posteriori des données manquantes en le paramètre courant.

(M) $\theta_{t+1} = \underset{\theta\in\Theta}{\operatorname{argmax}}\ Q(\theta|\theta_t)$ $\left(+\log q(\theta)\right]$

Remarque : Ne nécessite plus de calculer qu'une $\mathbb{E}$ et non J et ∇J du GS.

Cet avantage reste un inconvénient : il faut calculer cette $\int$ sous la loi a posteriori ...

3) Exemple :

L'exemple le plus courant est celui d'un modèle de mélange :

La densité des observations est $f_\theta(y) = \sum_{j=1}^{m} \alpha_j f_{\beta_j}(y)$

Le modèle sous-jacent est celui-ci : $(**)$ $\begin{cases} \mathbb{P}(Z_k = i) = \alpha_i : Z_k \sim \sum_{j=1}^{m} \alpha_j \delta_j \\ Y_k \mid Z_k = j \sim f_{\beta_j}(y) \end{cases}$

On cherche $\hat{\theta}_n = \underset{\theta\in\Theta}{\operatorname{argmax}} \sum_{j=1}^{m} \log f_\theta(Y_j)$ où θ sont les paramètres de f_i et aussi les poids α_i

Supposons ici que $f_\theta(Y) = \sum_{j=1}^{m} \alpha_j \frac{\lambda_j^Y}{Y!} e^{-\lambda_j}$: mélange de lois de Poisson

La vraisemblance complète (en log) est donc :

$$\log q_\theta(Y,Z) = -\log(Y!) + \sum_{j=1}^{m}\left[\log\alpha_j - \lambda_j\right] \mathbb{1}_{\{Z=j\}}$$

$$+ \sum_{j=1}^{m} \log(\lambda_j)\, Y\, \mathbb{1}_{\{Z=j\}}$$

$$Q(\theta|\theta_t) = \mathbb{E}\left[\log q_\theta(Y,Z) \mid Y, \theta_t\right]$$

où $q(Z|Y,\theta_t)$ est une loi discrète.

$$Q(\theta|\theta_t) = \int \sum_{i=1}^{n} \left[\log(Y_i!) + \sum_{j=1}^{m} \left[\log \alpha_j - \lambda_j\right] \mathbb{1}_{\{Z_i=j\}} + \sum_{j=1}^{m} \log(\lambda_j) Y_i \mathbb{1}_{\{Z=j\}}\right] \underbrace{\prod_{i=1}^{n} q(Z_i \mid Y_i, \theta_t)}_{\text{i.i.d}} \, dZ_i$$

$$= -\sum_{i=1}^{n} \log(Y_i!) + \sum_{i=1}^{n}\sum_{j=1}^{m} (\log \alpha_j - \lambda_j)\, \mathbb{P}(Z_i=j \mid Y_i, \theta_t) + \sum_{i=1}^{n}\sum_{j=1}^{m} \log(\lambda_j)\, Y_i\, \mathbb{P}(Z_i=j \mid Y_i, \theta_t)$$

$$= \underbrace{C}_{\text{indépendante de } Z,\theta} + \left[\sum_{j=1}^{m} (\log \alpha_j - \lambda_j) \times \left[\sum_{i=1}^{n} \mathbb{P}(Z_i=j \mid Y_i, \theta_t)\right]\right] + \left[\sum_{j=1}^{m} \log(\lambda_j) \times \sum_{i=1}^{n} Y_i\, \mathbb{P}(Z_i=j \mid Y_i, \theta_t)\right]$$

Maintenant la mise à jour de λ_j^{t+1} et α_j^{t+1} $\forall i$: par dérivation :

- α_j^{t+1} est solution d'une minimisation sous contrainte ; comme précédemment avec un multiplicateur de Lagrange $\mu \times \left(\sum_{j=1}^{m} \alpha_j - 1\right)$ on trouve :

$$\alpha_j^{t+1} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}(Z_i=j \mid Y_i, \theta_t)$$

- $$\lambda_j^{t+1} = \frac{\sum_{i=1}^{n} Y_i\, \mathbb{P}(Z_i=j \mid Y_i, \theta_t)}{\sum_{i=1}^{n} \mathbb{P}(Z_i=j \mid Y_i, \theta_t)}$$

Le calcul de $\mathbb{P}(Z_i=j \mid Y_i, \theta_t)$ se fait à partir du modèle génératif (**).

4) Convergence :

Une remarque tout d'abord : la simplicité apparente de cet algorithme dans le cas de notre exemple n'est pas le fait de la simplicité de ce modèle en particulier. Elle vient du fait que l'on peut écrire simplement la vraisemblance complète $q(Y,Z;\theta_t)$ et que son log donne une fonction simple qui sépare des termes dépendant de Z et ceux de θ.

C'est une propriété que l'on retrouve comme condition de convergence.

Convergence montrée (puis corrigée) par Dempster, Laird, Rubin et Wu (1983) puis reformulée avec des hypothèses plus simples par Delyon, Lavielle, Moulines (1999)

(M1) L'espace des paramètres Θ est un ouvert de $\mathbb{R}^p$ et la vraisemblance complète du modèle est donnée par : $q(y,z;\theta) = \exp\left[-\psi(\theta) + \langle S(y,z), \varphi(\theta)\rangle\right]$

où $\langle ; \rangle$ est le produit scalaire euclidien dans $\mathbb{R}^m$, S est une fonction borélienne en z sur $\mathbb{R}^q$ à valeurs dans $\mathcal{S} \subset \mathbb{R}^m$. S est appelée statistique exhaustive du modèle (unique à $\subset^q$ x près)

On suppose de plus que l'enveloppe convexe de $S(\mathbb{R}^q) \subset \mathcal{S}$ et que $\forall \theta \in \Theta$

$$\int_{\mathbb{R}^q} |S(y,z)|\, q(z|y,\theta)\, dz < \infty$$

(M2) φ et ψ sont C^2 sur Θ

(M3) $\bar{s} : \Theta \to \mathcal{S}$ tq $\bar{s}(\theta) \triangleq \int_{\mathbb{R}^q} S(y,z)\, q(z|y,\theta)\, dz$ est C^1 sur Θ

(M4) $\ell(\theta) \triangleq \log q(y,\theta)$ est C^1 sur Θ et

$$\partial_\theta \int q(y,z,\theta)\, dz = \int \partial_\theta q(y,z,\theta)\, dz$$

(M5) Soit $L : \mathcal{S} \times \Theta \longrightarrow \mathbb{R}$ tq $L(s,\theta) \triangleq -\psi(\theta) + \langle S(y,z), \varphi(\theta)\rangle$

Alors $\exists\, \hat{\theta} : \mathcal{S} \to \Theta$ C^1 sur $\mathcal{S}$ tq

$$\forall s \in \mathcal{S}\ \forall \theta \in \Theta \quad L(s, \hat{\theta}(s)) \geq L(s,\theta)$$

On définit $\mathcal{L} = \{\theta \in \Theta \text{ tq } \partial_\theta \ell(\theta) = 0\}$ et $d(x,A)$ la distance euclidienne du point x à l'ensemble fermé A.

Théorème (D,L,R,'99) On suppose que le modèle vérifie (M1-5).

Alors $\forall \theta_0 \in \Theta$, la suite $(\ell(\theta_k))_k$ donnée par l'algorithme EM est croissante et la suite $(\theta_k)_k$ vérifie $\lim_{k\to\infty} d(\theta_k, \mathcal{L}) = 0$

On ne va pas prouver ici ce théorème. Juste commenter les hypothèses. Je vous renvoie au DLM pour les détails.

Rq: • $(\theta_k)_k$ est une suite déterministe donc elle dépend de θ_0. Il se peut que pour un θ_0 on obtienne un max local ⇒ tester pour plusieurs θ_0 et choisir celui pour lequel $l(\theta_k)$ maximal.

• L'écriture sous forme exponentielle: beaucoup de modèles entre dans cette catégorie même des très complexes. ⇝ pas tellement une contrainte.

Dans notre exemple:

$$S_{1,j}(y_i, z_i) = \mathbb{1}_{\{z_i = j\}} \qquad \varphi_j(\theta) = \log(\alpha_j) - \lambda_j$$

$$S_{2,j}(y_i, z_i) = y_i \mathbb{1}_{\{z_i = j\}} \qquad \varphi_{2,j}(\theta) = \log \lambda_j$$

→ S de dimension $2m$ → en exo

• (H4) garantie une certaine régularité de log vraisemblance

• (H5) permet une mise à jour facile de θ connaissant S.

Dans notre exemple: en exo

$$\hat{\alpha}_j^{t+1} = \frac{1}{n} \sum_{i=1}^{n} S_{1,j}(y_i, z_i)$$

$$\hat{\lambda}_j^{t+1} = \frac{\sum_{i=1} S_{2,j}(y_i, z_i)}{\sum_{i=1} S_{1,j}(y_i, z_i)}$$

II Variantes de l'EM:

1) SEM: (Simulation EM) Celeux, Diebolt (1986)

• Cherche à éviter le calcul de l'espérance selon la loi a posteriori

• On simule $z_i^k \sim q(\cdot \mid y_i, \theta_k)$

• On maximise $\sum_{i=1} \log q(y_i, z_i^k, \theta)$

⊕ → évite le calcul de $\mathbb{E}$

→ l'aléat introduit réduit la dépendance en θ_0 car exploration plus générale des modes de $q(z \mid y, \theta)$

⊖ → Convergence prouvée en moyenne seulement (ps si mélange à la recuit)

2) MCEM : (Monte Carlo EM) Wei, Tanner (1990)

- Consiste à approcher l'espérance conditionnelle par une somme de Monte Carlo.
- On simule un (grand) nombre T de variables $(z_i^t)_{t \in \{1,\dots,T\}} \sim q(z | y_i, \theta_k)$ i.i.d
- On approche $Q(\theta | \theta_k) \simeq \frac{1}{T} \sum_{i=1}^{n} \sum_{t=1}^{T} \log q(y_i, z_i^t, \theta_k)$

⊖ Allongement du temps de calcul

⊖ Pas de résultats théoriques de convergence connus.

3) SAEM : Stochastic Approximation-EM: DLM(99)

L'idée est de profiter du passé des simulations avec $\theta_0 \dots \theta_{k+1}$ pour mettre à jour notre approximation de $Q(\theta | \theta_k)$. Cela est fait à l'aide des approximations stochastiques.

L'étape E est divisée ici en deux sous étapes :

S : simulation de $z_k \sim q(z | y, \theta_k)$ (pour tout i, on l'omet pour des raisons de commodité)

A : Approximation stochastique : Soit $(\gamma_k)_k$ une suite de pas :

$$\begin{cases} Q_{k+1}(\theta) = Q_k(\theta) + \gamma_k (\log q(y, z_k, \theta) - Q_k(\theta)) \\ Q_0 \text{ donné} \end{cases}$$

M : Maximisation : $\theta_{k+1} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, Q_{k+1}(\theta)$

⊕ . Ne demande qu'une simulation à chaque étape

. Garde en mémoire le passé via la suite $(\gamma_k)_k$

. Si le modèle est exponentiel et $\exists \hat{\theta}$ comme dans M1 et M5 alors l'AS se fait uniquement sur S ce qui produit une suite $(s_k)_k$ tq

$$s_{k+1} = s_k + \gamma_k (S(y, z_k) - s_k)$$

et $\theta_{k+1} = \hat{\theta}(s_{k+1})$

Rq : $Q_{k+1} = (1 - \gamma_k) Q_k + \gamma_k \log q(y, z_k, \theta_k)$

$= (1 - \gamma_k)(1 - \gamma_{k-1}) Q_{k-1} + \gamma_{k-1} \log q(y, z_{k-1}, \theta_{k-1}) + \gamma_k \log q(y, z_k, \theta_k)$

$= \sum_{p=1}^{k} \mu_p \log q(y, z_p, \theta_p)$ avec $\sum_{p=1}^{k} \mu_p = 1$ → Approx de type MC

- Résultats de convergence :

En plus de (H1 - 5) il faut des conditions supplémentaires qu'on devine en regardant la somme précédente. Elles portent sur $(\gamma_k)_k$ et la continuité de ℓ et $\tilde{\theta}$.

Soit $(\mathcal{F}_k)_k$ la filtration engendrée par la suite $(z_k)_k$ des simulations.

(SAEM 1) $\forall k \in \mathbb{N}$ $\gamma_k \in [0,1]$, $\sum_{k\geq 0} \gamma_k = +\infty$, $\sum_{k\geq 0} \gamma_k^2 < \infty$

(SAEM 2) $\ell : \Theta \to \mathbb{R}$ et $\tilde{\theta} : \mathcal{S} \to \Theta$ sont m fois continuement différentiables où m est tel que $\mathcal{S}$ ouvert de $\mathbb{R}^m$.

(SAEM 3) . $\forall \phi$ borélienne positive

$$\mathbb{E}\left[\phi(z_{k+1}) \mid \mathcal{F}_k\right] = \int \phi(z) q(z \mid y, \theta_k) dz$$

. $\forall \theta \in \Theta$ $\int \| S(y,z) \|^2 q(z \mid y, \theta) dz < \infty$ et

$$\Gamma(\theta) \triangleq \mathrm{Var}\left(S(y,z) \mid y, \theta\right) = \int_{\mathbb{R}^e} S(y,z)^2 q(z \mid y, \theta) dz - \left[\int_{\mathbb{R}^e} S(y,z) q(z \mid y, \theta) dz\right]^2$$

est C^0 en θ

Théorème (DLM) On suppose (H1-5) et (SAEM 1-3) vérifiés ainsi que

(C) $(\theta_k)_{k\geq 0}$ est à valeur dans un ss ensemble compact de Θ

Alors $\lim_{k\to\infty} d(\theta_k, \mathcal{L}) = 0$ ps.

Rq : ⊕ vitesse de convergence

⊖ vers un point stationnaire de ℓ ... mais un point selle ;

il faut des conditions type convexité au moins locale pour conclure à un maximum.

Sous des conditions un peu plus restrictives qui ressemblent à celles vues pour les A.S, on obtient aussi un TCL.

→ détails DLM 99.