

Chapitre 2 : Méthodes de Monte Carlo et algorithme EM

- Introduction
- Intégration par la méthode de Monte Carlo
- Fonctions d'importance
- Méthodes d'accélération
- Algorithme EM

Utilisations de la simulation

① intégration

$$\mathfrak{J} = \mathbb{E}_f[h(X)] = \int h(x)f(x)dx$$

② comportement limite/stationnaire de systèmes complexes

③ optimisation

$$\arg \min_x h(x) = \arg \max_x \exp\{-\beta h(x)\} \quad \beta > 0$$

Exemple (Propagation d'une épidémie)

Sur un territoire quadrillé, on représente par x, y les coordonnées d'un point.

La probabilité d'attraper la maladie est

$$P_{x,y} = \frac{\exp(\alpha + \beta \cdot n_{x,y})}{1 + \exp(\alpha + \beta \cdot n_{x,y})} \mathbb{I}_{n_{x,y} > 0}$$

si $n_{x,y}$ dénote le nombre de voisins de (x, y) ayant déjà cette maladie.

La probabilité de guérir de la maladie est

$$Q_{x,y} = \frac{\exp(\delta + \gamma \cdot n_{x,y})}{1 + \exp(\delta + \gamma \cdot n_{x,y})}$$

Exemple (Propagation d'une épidémie (2))

Question

En fonction de $(\alpha, \beta, \gamma, \delta)$, quelle est la vitesse de propagation de cette épidémie ? la durée moyenne ? le nombre de personnes infectées ?

Intégration par Monte Carlo

Loi des grands nombres

Si X_1, \dots, X_n simulés suivant f ,

$$\hat{\mathfrak{J}}_n = \frac{1}{n} \sum_{i=1}^n h(X_i) \longrightarrow \mathfrak{J}$$

Théorème Central Limit

Evaluation de l'erreur par

$$\hat{\sigma}_n^2 = \frac{1}{n^2} \sum_{i=1}^n (h(X_i) - \hat{\mathfrak{J}})^2$$

et

$$\hat{\mathfrak{J}}_n \approx \mathcal{N}(\mathfrak{J}, \hat{\sigma}_n^2)$$

Exemple (Normale)

Pour une loi normale, $\mathbb{E}[X^4] = 3$. Par la méthode de Monte Carlo,

n	5	50	500	5000	50,000	500,000
$\hat{\mathcal{J}}_n$	1.65	5.69	3.24	3.13	3.038	3.029



Exemple (Cauchy / Normale)

On considère le modèle joint

$$X|\theta \sim \mathcal{N}(\theta, 1), \quad \theta \sim \mathcal{C}(0, 1)$$

Après observation de X , on estime θ par

$$\delta^\pi(x) = \frac{\int_{-\infty}^{\infty} \frac{\theta}{1 + \theta^2} e^{-(x-\theta)^2/2} d\theta}{\int_{-\infty}^{\infty} \frac{1}{1 + \theta^2} e^{-(x-\theta)^2/2} d\theta}$$

Exemple (Cauchy / Normale (2))

Cette forme δ^π suggère de simuler des variables iid

$$\theta_1, \dots, \theta_m \sim \mathcal{N}(x, 1)$$

et de calculer

$$\hat{\delta}_m^\pi(x) = \frac{\sum_{i=1}^m \frac{\theta_i}{1 + \theta_i^2}}{\sum_{i=1}^m \frac{1}{1 + \theta_i^2}}.$$

Par la **Loi des Grands Nombres**,

$$\hat{\delta}_m^\pi(x) \longrightarrow \delta^\pi(x) \quad \text{quand } m \longrightarrow \infty.$$

Exemple (FdR normale)

Approximation de la fonction de répartition de la loi normale

$$\Phi(t) = \int_{-\infty}^t \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$$

par

$$\hat{\Phi}(t) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{X_i \leq t},$$

ayant généré un échantillon de taille n , (X_1, \dots, X_n) , via l'algorithme de Box-Muller.

Exemple (FdR normale (2))

- Variance

$$\Phi(t)(1 - \Phi(t))/n,$$

car les variables $\mathbb{I}_{X_i \leq t}$ sont iid Bernoulli($\Phi(t)$).

- Pour t près de $t = 0$ la variance vaut approximativement $1/4n$:
une précision de quatre décimales demande en moyenne

$$\sqrt{n} = \sqrt{2} 10^4$$

simulations, donc, **200 millions d'itérations.**

- Plus grande précision [absolue] dans les queues

Exemple (FdR normale (3))

n	0.0	0.67	0.84	1.28	1.65	2.32	2.58	3.09	3.72
10^2	0.485	0.74	0.77	0.9	0.945	0.985	0.995	1	1
10^3	0.4925	0.7455	0.801	0.902	0.9425	0.9885	0.9955	0.9985	1
10^4	0.4962	0.7425	0.7941	0.9	0.9498	0.9896	0.995	0.999	0.9999
10^5	0.4995	0.7489	0.7993	0.9003	0.9498	0.9898	0.995	0.9989	0.9999
10^6	0.5001	0.7497	0.8	0.9002	0.9502	0.99	0.995	0.999	0.9999
10^7	0.5002	0.7499	0.8	0.9001	0.9501	0.99	0.995	0.999	0.9999
10^8	0.5	0.75	0.8	0.9	0.95	0.99	0.995	0.999	0.9999

Evaluation de quantiles normaux par Monte Carlo fondée sur n générations normales.

Fonctions d'importance

Représentation alternative :

$$\mathfrak{J} = \int h(x)f(x)dx = \int h(x)\frac{f(x)}{g(x)}g(x)dx$$

Donc, si Y_1, \dots, Y_n simulés suivant g ,

$$\tilde{\mathfrak{J}}_n = \frac{1}{n} \sum_{i=1}^n h(Y_i) \frac{f(Y_i)}{g(Y_i)} \longrightarrow \mathfrak{J}$$

Intérêt

- Fonctionne pour tout choix de g tel que

$$\text{supp}(g) \supset \text{supp}(f)$$

- Amélioration possible de la variance
- Recyclage de simulations $Y_i \sim g$ pour d'autres densités f
- Utilisation de lois simples g

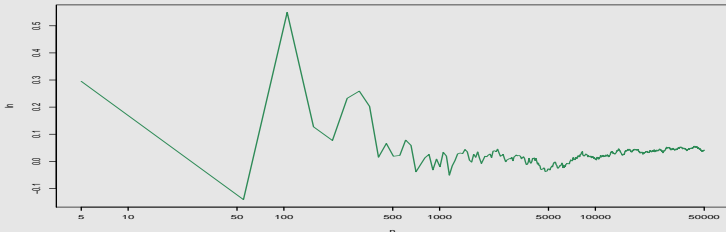
Exemple (Normale)

Pour la loi normale et l'approximation de $\mathbb{E}[X^4]$,

$$\int_{-\infty}^{\infty} x^4 e^{-x^2/2} dx \stackrel{[y=x^2]}{=} 2 \int_0^{\infty} y^{3/2} \frac{1}{2} e^{-y/2} dy$$

suggère d'utiliser $g(y) = \exp(-y/2)/2$

n	5	50	500	5000	50000
\tilde{J}_n	3.29	2.89	3.032	2.97	3.041



Choix de la fonction d'importance

La “bonne” fonction g dépend de la densité f *et* de la fonction h

Théorème (Importance optimale)

Le choix de g minimisant la variance de $\tilde{\mathcal{J}}_n$ est

$$g^*(x) = \frac{|h(x)|f(x)}{\mathfrak{J}}$$

Remarques

- Variance finie seulement si

$$\mathbb{E}_f \left[h^2(X) \frac{f(X)}{g(X)} \right] = \int_{\mathcal{X}} h^2(x) \frac{f(X)}{g(X)} dx < \infty .$$

- Variance **nulle** pour g^* si h positive (!!)
- g^* dépend de \mathfrak{J} que l'on cherche à estimer (??)
- Remplacement de $\tilde{\mathfrak{J}}_n$ par **moyenne harmonique**

$$\check{\mathfrak{J}}_n = \frac{\sum_{i=1}^n h(y_i)/|h(y_i)|}{\sum_{i=1}^n 1/|h(y_i)|}$$

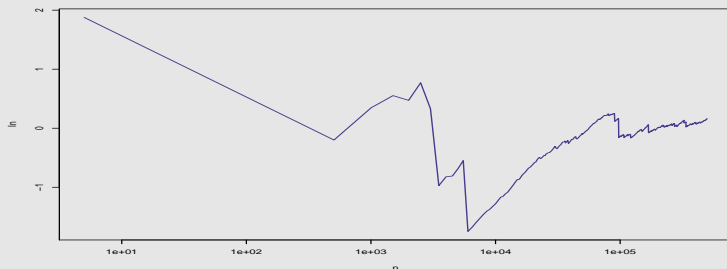
(numérateur et dénominateur sont convergents)
souvent **mauvais** (variance infinie)

Exemple (Normale)

Pour la loi normale et l'approximation de $\mathbb{E}[X^4]$,
 $g^*(x) \propto x^4 \exp(-x^2/2)$, loi de la racine d'une $\mathcal{G}a(5/2, 1/2)$

[Exercice]

n	5	50	500	5,000	50,000	500,000
\tilde{J}_n	4.877	2.566	2.776	2.317	2.897	3.160



Exemple (Loi de Student)

$X \sim \mathcal{T}(\nu, \theta, \sigma^2)$, de densité

$$f(x) = \frac{\Gamma((\nu + 1)/2)}{\sigma\sqrt{\nu\pi} \Gamma(\nu/2)} \left(1 + \frac{(x - \theta)^2}{\nu\sigma^2}\right)^{-(\nu+1)/2}.$$

Soient $\theta = 0$, $\sigma = 1$ et

$$\mathfrak{J} = \int_{2.1}^{\infty} x^5 f(x) dx.$$

à calculer

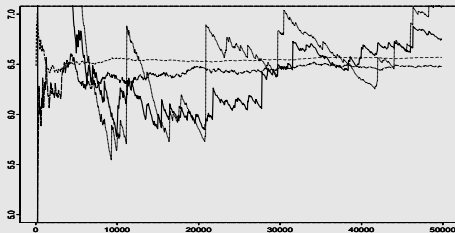
Exemple (Loi de Student (2))

- Choix de fonctions d'importance

- f , car $f = \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_\nu/\nu}}$
- Cauchy $\mathcal{C}(0,1)$
- Normale $\mathcal{N}(0,1)$
- $\mathcal{U}([0, 1/2.1])$

Résultats:

- Uniforme optimale
- Cauchy OK
- f et Normale mauvaises



Simulations corrélées

La corrélation négative...

Deux échantillons (X_1, \dots, X_m) et (Y_1, \dots, Y_m) suivant f pour estimer

$$\mathfrak{J} = \int_{\mathbb{R}} h(x)f(x)dx .$$

Soient

$$\hat{\mathfrak{J}}_1 = \frac{1}{m} \sum_{i=1}^m h(X_i) \quad \text{et} \quad \hat{\mathfrak{J}}_2 = \frac{1}{m} \sum_{i=1}^m h(Y_i)$$

de moyenne \mathfrak{J} et variance σ^2

Simulations corrélées (2)

...réduit la variance

La variance de la moyenne vaut

$$\text{var} \left(\frac{\hat{\mathcal{J}}_1 + \hat{\mathcal{J}}_2}{2} \right) = \frac{\sigma^2}{2} + \frac{1}{2} \text{cov}(\hat{\mathcal{J}}_1, \hat{\mathcal{J}}_2).$$

Par conséquent, si les deux échantillons sont **néativement corrélés**,

$$\text{cov}(\hat{\mathcal{J}}_1, \hat{\mathcal{J}}_2) \leq 0,$$

ils font mieux que deux échantillons indépendants de même taille

Variables antithétiques

Construction de variables négativement corrélées

- ① Si f symétrique autour de μ , prendre $Y_i = 2\mu - X_i$
- ② Si $X_i = F^{-1}(U_i)$, prendre $Y_i = F^{-1}(1 - U_i)$
- ③ Si $(A_i)_i$ est une partition de \mathcal{X} , échantillonnage partitionné en prenant des X_j dans chaque A_i (nécessite de connaître $\Pr(A_i)$)

Variables de contrôle

Soit

$$\mathfrak{I} = \int h(x)f(x)dx$$

à évaluer et

$$\mathfrak{I}_0 = \int h_0(x)f(x)dx$$

connue

On estime quand même \mathfrak{I}_0 par $\hat{\mathfrak{I}}_0$ (et \mathfrak{I} par $\hat{\mathfrak{I}}$)

Variables de contrôle (2)

Estimateur combiné

$$\hat{\mathcal{J}}^* = \hat{\mathcal{J}} + \beta(\hat{\mathcal{J}}_0 - I_0)$$

$\hat{\mathcal{J}}^*$ est sans biais pour \mathcal{J} et

$$\text{var}(\hat{\mathcal{J}}^*) = \text{var}(\hat{\mathcal{J}}) + \beta^2 \text{var}(\hat{\mathcal{J}}_0) + 2\beta \text{cov}(\hat{\mathcal{J}}, \hat{\mathcal{J}}_0)$$

Variables de contrôle (3)

Choix optimal de β

$$\beta^* = -\frac{\text{cov}(\hat{\mathcal{J}}, \hat{\mathcal{J}}_0)}{\text{var}(\hat{\mathcal{J}}_0)},$$

avec

$$\text{var}(\hat{\mathcal{J}}^*) = (1 - \rho^2) \text{var}(\hat{\mathcal{J}}),$$

où ρ corrélation entre $\hat{\mathcal{J}}$ et $\hat{\mathcal{J}}_0$

Exemple (Approximation de quantiles)

Soit à évaluer

$$\varrho = \Pr(X > a) = \int_a^{\infty} f(x)dx$$

par

$$\hat{\varrho} = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(X_i > a), \quad X_i \stackrel{\text{iid}}{\sim} f$$

avec $\Pr(X > \mu) = \frac{1}{2}$

Exemple (Approximation de quantiles (2))

La variable de contrôle

$$\frac{1}{n} \sum_{i=1}^n \mathbb{I}(X_i > a) + \beta \left(\frac{1}{n} \sum_{i=1}^n \mathbb{I}(X_i > \mu) - \Pr(X > \mu) \right)$$

améliore $\hat{\theta}$ si

$$\beta < 0 \quad \text{et} \quad |\beta| < 2 \frac{\text{cov}(\delta_1, \delta_3)}{\text{var}(\delta_3)} = 2 \frac{\Pr(X > a)}{\Pr(X > \mu)}.$$

Intégration par conditionnement

Tirer parti de l'inégalité

$$\text{var}(\mathbb{E}[\delta(\mathbf{X})|\mathbf{Y}]) \leq \text{var}(\delta(\mathbf{X}))$$

appelée aussi **Théorème de Rao-Blackwell**

Conséquence :

Si $\hat{\mathcal{J}}$ est un estimateur sans biais de $\mathcal{J} = \mathbb{E}_f[h(X)]$, avec X simulé à partir de la densité jointe $\tilde{f}(x, y)$, où

$$\int \tilde{f}(x, y) dy = f(x),$$

l'estimateur

$$\hat{\mathcal{J}}^* = \mathbb{E}_{\tilde{f}}[\hat{\mathcal{J}}|Y_1, \dots, Y_n]$$

domine $\hat{\mathcal{J}}(X_1, \dots, X_n)$ en variance (et est aussi sans biais)

Exemple (Espérance de loi de Student)

Soit à calculer

$$\mathbb{E}[h(x)] = \mathbb{E}[\exp(-x^2)] \quad \text{avec} \quad X \sim \mathcal{T}(\nu, 0, \sigma^2)$$

La loi de Student peut être simulée par

$$X|y \sim \mathcal{N}(\mu, \sigma^2 y) \quad \text{et} \quad Y^{-1} \sim \chi_{\nu}^2.$$

Exemple (Espérance de loi de Student (2))

La moyenne empirique

$$\frac{1}{m} \sum_{j=1}^m \exp(-X_j^2),$$

peut être améliorée à partir de l'échantillon joint

$$((X_1, Y_1), \dots, (X_m, Y_m))$$

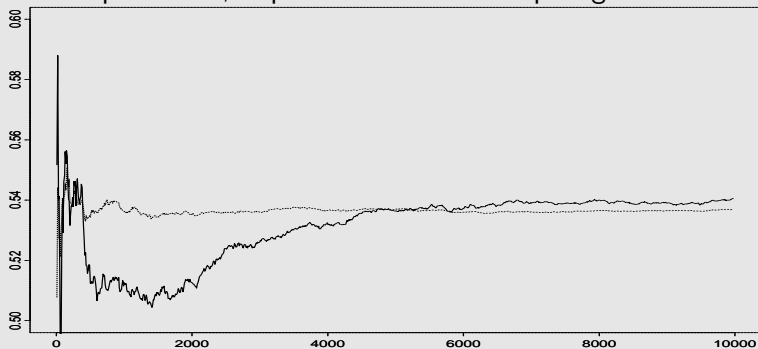
puisque

$$\frac{1}{m} \sum_{j=1}^m \mathbb{E}[\exp(-X^2)|Y_j] = \frac{1}{m} \sum_{j=1}^m \frac{1}{\sqrt{2\sigma^2 Y_j + 1}}$$

est l'espérance conditionnelle

Exemple (Espérance de loi de Student (3))

Dans ce cas particulier, la précision est **dix fois** plus grande



Estimateurs de $\mathbb{E}[\exp(-X^2)]$: moyenne empirique (traits pleins) contre espérance conditionnelle (pointillés) pour $(\nu, \mu, \sigma) = (4.6, 0, 1)$.

Optimisation par l'algorithme EM

Maximum de vraisemblance

On observe $\mathbf{X} = (X_1, \dots, X_n)$, iid $g(x|\theta)$ et on cherche

$$\theta^* = \arg \max_{\theta} L(\theta|\mathbf{x}) = \prod_{i=1}^n g(X_i|\theta).$$

Utilisation de la représentation marginale

$$g(x|\theta) = \int_{\mathbf{z}} f(x, \mathbf{z}|\theta) d\mathbf{z}$$

de la vraisemblance pour obtenir le maximum de vraisemblance

Exemples

- Données censurées :

$$X = \min(X^*, a) \quad X^* \sim \mathcal{N}(\theta, 1)$$

- Données de mélange

$$X \sim .3 \mathcal{N}_1(\mu_0, 1) + .7 \mathcal{N}_1(\mu_1, 1),$$

- Modèles de déséquilibre

$$X = \min(X^*, Y^*) \quad X^* \sim f_1(x|\theta) \quad Y^* \sim f_2(x|\theta)$$

Complétion

La méthode consiste à ajouter aux données un complément \mathbf{z} , où

$$(\mathbf{X}, \mathbf{Z}) \sim f(\mathbf{x}, \mathbf{z}|\theta)$$

\mathbf{Z} est appelé **vecteur des données manquantes** et le couple (\mathbf{X}, \mathbf{Z}) **vecteur des données complétées**

Notons que la densité conditionnelle de \mathbf{Z} sachant les données \mathbf{x} vaut

$$k(\mathbf{z}|\theta, \mathbf{x}) = \frac{f(\mathbf{x}, \mathbf{z}|\theta)}{g(\mathbf{x}|\theta)}$$

Décomposition de la vraisemblance

La vraisemblance associée aux données complétées

$$L^c(\theta|\mathbf{x}, \mathbf{z}) = f(\mathbf{x}, \mathbf{z}|\theta)$$

et la vraisemblance associée aux données observées

$$L(\theta|\mathbf{x})$$

sont liées par

$$\begin{aligned} \log L(\theta|\mathbf{x}) &= \mathbb{E}[\log L^c(\theta|\mathbf{x}, \mathbf{Z})|\theta_0, \mathbf{x}] \\ &\quad - \mathbb{E}[\log k(\mathbf{Z}|\theta, \mathbf{x})|\theta_0, \mathbf{x}], \end{aligned} \quad (1)$$

pour tout θ_0 , où l'intégration se fait suivant la loi conditionnelle de \mathbf{Z} sachant les données, $k(\mathbf{z}|\theta_0, \mathbf{x})$

Remarque

Il y a “deux θ ” ! : dans (1), θ_0 est une valeur fixe (mais arbitraire) qui sert à faire l'intégration, tandis que θ est libre (et variable)

Maximiser la vraisemblance **observée**

$$L(\theta|\mathbf{x})$$

revient à maximiser le terme de droite dans (1)

Intuition

Au lieu de maximiser (en θ) le terme de droite dans (1), on ne maximise que la première partie

$$\mathbb{E}[\log L^c(\theta|\mathbf{x}, \mathbf{Z})|\theta_0, \mathbf{x}]$$

même si la seconde partie dépend de θ :

On cherche donc à maximiser la log-vraisemblance complétée ou plutôt son espérance puisque \mathbf{Z} est inconnu, et on corrige l'omission du second terme de (1) ainsi que la dépendance à θ_0 en itérant les maximisations.

Espérance/Maximisation

On note l'**E**spérance de la log-vraisemblance complète

$$Q(\theta|\theta_0, \mathbf{x}) = \mathbb{E}[\log L^c(\theta|\mathbf{x}, \mathbf{Z})|\theta_0, \mathbf{x}]$$

pour insister sur la dépendance en θ_0 et à l'échantillon \mathbf{x}

Principe

EM construit une suite d'estimateurs $\hat{\theta}_{(j)}$, $j = 1, 2, \dots$, par itération des étapes **E**spérance et **M**aximisation :

$$Q(\hat{\theta}_{(j)}|\hat{\theta}_{(j-1)}, \mathbf{x}) = \max_{\theta} Q(\theta|\hat{\theta}_{(j-1)}, \mathbf{x}).$$

Algorithme EM

Itérer (en m)

1. (*étape E*) Calculer

$$Q(\theta|\hat{\theta}_{(m)}, \mathbf{x}) = \mathbb{E}[\log L^c(\theta|\mathbf{x}, \mathbf{Z})|\hat{\theta}_{(m)}, \mathbf{x}],$$

2. (*étape M*) Maximiser $Q(\theta|\hat{\theta}_{(m)}, \mathbf{x})$ en θ et prendre

$$\hat{\theta}_{(m+1)} = \arg \max_{\theta} Q(\theta|\hat{\theta}_{(m)}, \mathbf{x}).$$

tant qu'un point fixe [de Q] n'est pas obtenu

Justification

On démontre que la vraisemblance observée

$$L(\theta|\mathbf{x})$$

augmente à chaque étape de EM

$$L(\hat{\theta}_{(m+1)}|\mathbf{x}) \geq L(\hat{\theta}_{(m)}|\mathbf{x})$$

[Exercice]

Exemple (Données censurées)

Soit un échantillon $\mathcal{N}(\theta, 1)$ censuré à droite en a , de vraisemblance :

$$L(\theta|\mathbf{x}) = \frac{1}{(2\pi)^{m/2}} \exp \left\{ -\frac{1}{2} \sum_{i=1}^m (x_i - \theta)^2 \right\} [1 - \Phi(a - \theta)]^{n-m}$$

La logvraisemblance complétée vaut

$$\log L^c(\theta|\mathbf{x}, \mathbf{z}) \propto -\frac{1}{2} \sum_{i=1}^m (x_i - \theta)^2 - \frac{1}{2} \sum_{i=m+1}^n (z_i - \theta)^2,$$

où les z_i correspondent aux observations censurées, de densité

$$k(z|\theta, \mathbf{x}) = \frac{\exp\{-\frac{1}{2}(z - \theta)^2\}}{\sqrt{2\pi}[1 - \Phi(a - \theta)]} = \frac{\varphi(z - \theta)}{1 - \Phi(a - \theta)}, \quad a < z.$$

Exemple (Données censurées (2))

A la j -ième itération de EM,

$$\begin{aligned} Q(\theta | \hat{\theta}_{(j)}, \mathbf{x}) &\propto -\frac{1}{2} \sum_{i=1}^m (x_i - \theta)^2 - \frac{1}{2} \mathbb{E} \left[\sum_{i=m+1}^n (Z_i - \theta)^2 \middle| \hat{\theta}_{(j)}, \mathbf{x} \right] \\ &\propto -\frac{1}{2} \sum_{i=1}^m (x_i - \theta)^2 \\ &\quad - \frac{1}{2} \sum_{i=m+1}^n \int_a^{\infty} (z_i - \theta)^2 k(z | \hat{\theta}_{(j)}, \mathbf{x}) dz_i \end{aligned}$$

Exemple (Données censurées (3))

En différenciant en θ , on obtient

$$n \hat{\theta}_{(j+1)} = m \bar{x} + (n - m) \mathbb{E}[Z | \hat{\theta}_{(j)}],$$

où

$$\mathbb{E}[Z | \hat{\theta}_{(j)}] = \int_a^\infty z k(z | \hat{\theta}_{(j)}, \mathbf{x}) dz = \hat{\theta}_{(j)} + \frac{\varphi(a - \hat{\theta}_{(j)})}{1 - \Phi(a - \hat{\theta}_{(j)})}.$$

Par conséquent, la suite EM est donnée par

$$\hat{\theta}_{(j+1)} = \frac{m}{n} \bar{x} + \frac{n - m}{n} \left[\hat{\theta}_{(j)} + \frac{\varphi(a - \hat{\theta}_{(j)})}{1 - \Phi(a - \hat{\theta}_{(j)})} \right],$$

et elle converge vers le maximum de vraisemblance $\hat{\theta}$.

Exemple (Mélanges)

Dans le cadre d'un mélange de deux distributions normales

$$.3 \mathcal{N}_1(\mu_0, 1) + .7 \mathcal{N}_1(\mu_1, 1),$$

échantillon X_1, \dots, X_n et paramètre $\theta = (\mu_0, \mu_1)$

Donnée manquante : $Z_i \in \{0, 1\}$, variable indicatrice de la composante associée à X_i ,

$$X_i | z_i \sim \mathcal{N}(\mu_{z_i}, 1) \quad Z_i \sim \mathcal{B}(.7)$$

La vraisemblance complétée vaut

$$\begin{aligned} \log L^c(\theta | \mathbf{x}, \mathbf{z}) &\propto -\frac{1}{2} \sum_{i=1}^n z_i (x_i - \mu_1)^2 - \frac{1}{2} \sum_{i=1}^n (1 - z_i) (x_i - \mu_0)^2 \\ &= -\frac{1}{2} n_1 (\hat{\mu}_1 - \mu_1)^2 - \frac{1}{2} (n - n_1) (\hat{\mu}_0 - \mu_0)^2 \end{aligned}$$

Exemple (Mélanges (2))

A la j -ième itération de EM,

$$Q(\theta|\hat{\theta}_{(j)}, \mathbf{x}) = \frac{1}{2} \mathbb{E} \left[n_1(\hat{\mu}_1 - \mu_1)^2 + (n - n_1)(\hat{\mu}_0 - \mu_0)^2 \mid \hat{\theta}_{(j)}, \mathbf{x} \right]$$

Différenciant,

$$\hat{\theta}_{(j+1)} = \begin{pmatrix} \mathbb{E} \left[n_1 \hat{\mu}_1 \mid \hat{\theta}_{(j)}, \mathbf{x} \right] / \mathbb{E} \left[n_1 \mid \hat{\theta}_{(j)}, \mathbf{x} \right] \\ \mathbb{E} \left[(n - n_1) \hat{\mu}_0 \mid \hat{\theta}_{(j)}, \mathbf{x} \right] / \mathbb{E} \left[(n - n_1) \mid \hat{\theta}_{(j)}, \mathbf{x} \right] \end{pmatrix}$$

Exemple (Mélanges (3))

Soit $\hat{\theta}_{(j+1)}$ égal à

$$\left(\begin{array}{l} \sum_{i=1}^n \mathbb{E} \left[Z_i \mid \hat{\theta}_{(j)}, x_i \right] x_i / \sum_{i=1}^n \mathbb{E} \left[Z_i \mid \hat{\theta}_{(j)}, x_i \right] \\ \sum_{i=1}^n \mathbb{E} \left[(1 - Z_i) \mid \hat{\theta}_{(j)}, x_i \right] x_i / \sum_{i=1}^n \mathbb{E} \left[(1 - Z_i) \mid \hat{\theta}_{(j)}, x_i \right] \end{array} \right)$$

Conclusion

L'étape (E) de EM consiste à remplacer les données manquantes Z_i par leur espérance conditionnellement à \mathbf{x} (espérance qui dépend de $\hat{\theta}_{(m)}$).

Propriétés

EM est un algorithme qui

- converge vers un maximum local ou un point-selle de la vraisemblance
- dépend de la condition initiale $\theta_{(0)}$
- nécessite plusieurs initialisations si la vraisemblance n'est pas unimodale

MCEM

Une difficulté supplémentaire avec EM est que le calcul de $Q(\theta|\theta_0, \mathbf{x})$ n'est pas toujours possible

On peut remplacer l'espérance par une approximation de Monte Carlo

$$\hat{Q}(\theta|\theta_0, \mathbf{x}) = \frac{1}{m} \sum_{i=1}^m \log L^c(\theta|\mathbf{x}, \mathbf{Z}_i),$$

où $\mathbf{Z}_1, \dots, \mathbf{Z}_m \sim k(\mathbf{z}|\mathbf{x}, \theta_0)$.

Quand $m \rightarrow \infty$, cette approximation converge vers

$$Q(\theta|\theta_0, \mathbf{x})$$

Inconvénient :

A moins de bénéficier de conditions spéciales, il faut recommencer la simulation des \mathbf{Z}_i à chaque itération de MCEM. Besoin de grandes valeurs de m pour obtenir la stabilité.