Données manquantes : l'algorithme EM

Résumé

A travers deux exemples (l'un discret, l'autre continu) de problèmes à données manquantes, on introduit l'algorithme EM et on esquisse son étude de façon théorique et par des simulations.

1 Deux problèmes à données manquantes

1.1 Génétique

Un modèle génétique conduit à la considération de cinq génotypes possibles, ayant pour fréquences dans la population respectivement

$$p_1 = \left(\frac{\theta}{4}, p_2 = \frac{1}{4}(1-\theta), p_3 = \frac{1}{4}(1-\theta), p_4 = \frac{\theta}{4}, p_5 = \frac{1}{2}\right),$$

où θ est un paramètre inconnu qu'il convient d'estimer. Pour cela, un expériementateur inspecte un échantillon de n animaux et transmet ses résultats pour analyse à un statisticien.

Les cinq génotypes conduisent à des phénotypes différents, sauf les numéros 4 et 5 qui produisent des animaux indistingables. Il transmet donc comme données un quadruplet $\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4)$, où Y_i correspond au nombre d'animaux de génotype i pour $1 \leqslant i \leqslant 3$, et Y_4 au nombre d'animaux de génotype 4 ou 5.

Si l'expérimentateur observait directement les nombres d'occurrences $\mathbf{X} = (X_1, \ldots, X_5)$ de chaque génotype, le problème de l'estimation de θ serait très simplement résolu par un calcul direct du maximum de vraisemblance. Mais puisque que l'on n'observe que $\mathbf{Y} = f(\mathbf{X})$ (avec $f(x_1, x_2, x_3, x_4, x_5) = (x_1, x_2, x_3, x_4 + x_5)$) les choses sont un peu plus compliquées.

1.2 Mélange gaussien

Dans une population d'animaux que l'on étudie, les mâles et les femelles ont des poids significativement différents. On considère que le poids des mâles est distribué selon une loi normale de moyenne μ_1 et de variance 1, tandis que celui des femelles est distribué selon une loi normale de moyenne μ_2 et de variance 1.

Une expérience a été menée pour estimer μ_1 et μ_2 : un échantillon de n animaux a été pesé, et les résultats $\mathbf{Y} = (Y_1, \ldots, Y_n)$ est transmis par les biologistes aux statisticiens pour être analysé. Malheureusement, les biologistes ont oublié de noter le sexe de chaque animal ! Soit Z_i le sexe du i-ème animal de l'échantillon. La donnée complète $X_i = (Y_i, Z_i)$ permettrait de traiter très facilement le problème. Mais en ne disposant que des données partielles $Y_i = f(X_i)$ (f désignant ici la projection sur la première coordonnées), il est beaucoup plus difficile de calculer l'estimateur du maximum de vraisemblance !

2 L'algorithme EM

Nous traitons ici l'exemple de génétique, le candidat est invité à justifier les éléments avancés ici et à adapter cette modélisation au cas du mélange gaussien.

Si la valeur du paramètre est $\theta \in [0, 1]$, la variable aléatoire X a une loi P_θ multinomiale sur l'ensemble $\{1, \ldots, 5\}$: pour $\mathbf{x} \in \mathbb{N}^5$,

$$P_\theta(\mathbf{X} = \mathbf{x}) = \binom{n}{x_1\ x_2\ x_3\ x_4\ x_5} \left(\frac{\theta}{4}\right)^{x_1+x_4} \left(\frac{1-\theta}{4}\right)^{x_2+x_3} \left(\frac{1}{2}\right)^{x_5}.$$

Si l'on observait $\mathbf{X} = \mathbf{x}$, on utiliserait l'estimateur du maximum de vraisemblance $\hat{\hat{\theta}}(\mathbf{x}) = \frac{x_1+x_4}{x_1+x_2+x_3+x_4}$ dont il est facile de voir qu'il est consistant.

On suppose toutefois que seul est observé le vecteur $\mathbf{y} = (y_1, \ldots, y_4)$ des nombre d'occurrences de chaque phénotype dans l'échantillon. On a

$$P_\theta(\mathbf{Y} = \mathbf{y}) = \binom{n}{y_1\ y_2\ y_3\ y_4} \left(\frac{\theta}{4}\right)^{y_1} \left(\frac{1-\theta}{4}\right)^{y_2+y_3} \left(\frac{1}{2} + \frac{\theta}{4}\right)^{y_4}.$$

Notons $P_\theta(\cdot|y)$ la loi conditionnelle de X sachant $\{Y = y\}$, c'est-à-dire $P_\theta(x|y) = \frac{P_\theta(X=x \cap Y=y)}{P_\theta(Y=y)}$, et θ_0 la vraie valeur du paramètre.

L'idée de l'algorithme EM est, à partir d'une estimation $\hat{\theta}^0 \in [0,1]$ (idée que l'on a a priori du paramètre θ inconnu), de calculer un estimateur $\tilde{\mathbf{x}} = \mathbb{E}_{P_{\hat{\theta}^0}}[\mathbf{X}|\mathbf{Y}=\mathbf{y}]$ des données non observées (ou plutôt de leurs statistiques suffisantes pour l'estimation de θ), puis d'utiliser l'estimateur classique $\hat{\theta}^1 = \hat{\hat{\theta}}(\tilde{\mathbf{x}})$ comme raffinement de $\hat{\theta}^0$. On peut ensuite itérer le processus, en espérant qu'il converge vers le maximum de vraisemblance.

Plus formellement, on construit par récurrence une suite d'estimateurs $(\hat{\theta}^n)_n$ par une suite d'itérations se décomposant en deux phases :

1. **Phase E** (Expectation)

$$Q(\theta|\hat{\theta}^n) = \mathbb{E}_{P_{\hat{\theta}^n}(\cdot|\mathbf{y})}[\log P_\theta(\mathbf{X})].$$

2. **Phase M** (Maximisation)

$$\hat{\theta}^{n+1} = \arg\max_{\theta\in[0,1]} Q(\theta|\hat{\theta}^n).$$

Pour notre problème, l'espérance $\tilde{\mathbf{x}}$ sous $\mathbb{P}_{\hat{\theta}^n}$ des données non observées conditionnellement à $\mathbf{Y}=\mathbf{y}$ s'écrit

$$\tilde{x} = \left(\tilde{x}_1 = y_1, \tilde{x}_2 = y_2, \tilde{x}_3 = y_3, \tilde{x}_4 = y_4\frac{\hat{\theta}^n}{2+\hat{\theta}^n}, \tilde{x}_5 = y_4\frac{2}{2+\hat{\theta}^n}\right).$$

Or la log-vraisemblance de $\theta \in [0,1]$ pour $\tilde{x}$ sous $P_{\hat{\theta}^n}(\cdot|\mathbf{y})$ s'écrit :

$$\mathbb{E}_{P_{\hat{\theta}^n}(\cdot|\mathbf{y})}[\log P_\theta(\mathbf{X})] = C(\tilde{\mathbf{x}}) + \tilde{x}_1 \log\frac{\theta}{4} + (\tilde{x}_2+\tilde{x}_3)\log\frac{1-\theta}{4} + \tilde{x}_4\log\frac{\theta}{4}$$

où $C(\tilde{\mathbf{x}})$ ne dépend pas de θ. On aboutit donc à l'itération :

$$\hat{\theta}^{n+1} = \frac{y_1 + y_4\frac{\hat{\theta}^n}{2+\hat{\theta}^n}}{y_1+y_2+y_3+y_4\frac{\hat{\theta}^n}{2+\hat{\theta}^n}}.$$

Dans ce cas d'école, un calcul permet de trouver explicitement la limite $\hat{\theta}$ de la suite. On peut ainsi vérifier que la suite d'estimateurs $(\hat{\theta}^n)_n$ converge vers $\hat{\theta}$, et même étudier à quelle vitesse la convergence se fait.

3 Etude abstraite de la convergence

Pour comprendre ce phénomène au delà de l'exemple génétique, introduisons l'*information de Küllback-Leibler* $K(q|r)$ entre les deux lois de probabilités q et r définies sur l'ensemble $\{1,\dots,5\}$:

$$K(q|r) = \mathbb{E}_q\left[\log\frac{q(X)}{r(X)}\right] = \sum_{j=1}^{5} q(j)\log\frac{q(j)}{r(j)}.$$

On prouve aisément que $K(q|r) \geqslant 0$, et que l'on a égalité si et seulement si $q=r$.

Remarquons que, quel que soit $\theta' \in [0,1]$, la log-vraisemblance du paramètre θ pour l'observation $\mathbf{y}$ s'écrit :

$$\log P_\theta(Y=y) = \mathbb{E}_{P_{\theta'}(\cdot|y)}[\log P_\theta(X)] - \mathbb{E}_{P_{\theta'}(\cdot|y)}[\log P_\theta(X|y)].$$

Ainsi,

$$\begin{aligned}
\log P_{\hat{\theta}^{n+1}}(Y=y) - \log P_{\hat{\theta}^n}(Y=y) &= \mathbb{E}_{P_{\hat{\theta}^n}(\cdot|y)}\left[\log P_{\hat{\theta}^{n+1}}(X)\right] - \mathbb{E}_{P_{\hat{\theta}^n}(\cdot|y)}\left[\log P_{\hat{\theta}^{n+1}}(X|y)\right] \\
&\quad -\mathbb{E}_{P_{\hat{\theta}^n}(\cdot|y)}\left[\log P_{\hat{\theta}^n}(X)\right] + \mathbb{E}_{P_{\hat{\theta}^n}(\cdot|y)}\left[\log P_{\hat{\theta}^n}(X|y)\right] \\
&= \mathbb{E}_{P_{\hat{\theta}^n}(\cdot|y)}\left[\log P_{\hat{\theta}^{n+1}}(X)\right] - \mathbb{E}_{P_{\hat{\theta}^n}(\cdot|y)}\left[\log P_{\hat{\theta}^n}(X)\right] \quad (1) \\
&\quad +\mathbb{E}_{P_{\hat{\theta}^n}(\cdot|y)}\left[\log P_{\hat{\theta}^n}(X|y)\right] - \mathbb{E}_{P_{\hat{\theta}^n}(\cdot|y)}\left[\log P_{\hat{\theta}^{n+1}}(X|y)\right]. \quad (2) \\
&\geqslant 0.
\end{aligned}$$

En effet, la première différence (1) est positive par définition de $\hat{\theta}^{n+1}$. D'autre part, la différence (2) est égale à $K\left(P_{\hat{\theta}^n}(\cdot|y)|P_{\hat{\theta}^{n+1}}(\cdot|y)\right)$. Cela suffit à montrer que la suite des vraisemblances est croissante. On voit aisément que la limite est nécessairement un maximum local de la fonction de vraisemblance. Mais comme celle-ci est concave, on peut conclure que $\hat{\theta}^n$ converge vers le maximum de vraisemblance.

4 Suggestions

1. Simuler des variables aléatoires $\mathbf{X}$ et $\mathbf{Y}$ pour les deux problèmes présentés dans la section 1.
2. Programmer et illustrer le fonctionnement de l'algorithme EM présenté dans la section 2.
3. Prouver les affirmations contenues dans les sections 2 et 3.
4. Adapter la discussion proposée dans la section 2 au problème du mélange gaussien. Programmer l'algorithme EM dans ce cas et illustrer son fonctionnement.
5. Adapter la preuve de la section 3 au cas gaussien.