–Texte–
Estimation d'un mélange avec l'algorithme EM

Mots-clefs : mélange de lois, vraisemblance, lois conditionnelles.

1 Sur les nids de mouettes

En Bretagne, cohabitent quatre espèces de mouettes différentes. Les ornithologues se demandent comment estimer la proportion de mouettes de chaque espèce à partir de l'observation de la taille de nombreux nids. L'avantage de cette méthode (compter les nids plutôt que les oiseaux) c'est que les nids ne bougent pas!!! La difficulté vient du fait que les différentes espèces font des nids assez ressemblants : on ne sait pas par quelles espèces ils ont été construit ("rien ne ressemble plus à un nid de mouette, qu'un autre nid de mouette"). Ce que l'on suppose (sinon la taille des nids n'est pas un outil de mesure très intéressant), c'est que la distribution des nids est caractéristique d'une espèce. La distribution globale de la taille des nids apparaît comme le mélange de quatre lois de probabilité, chacune rendant compte de la répartition de la taille des nids pour chaque espèce d'oiseaux. Modélisons la distribution de la taille des nids construits par l'espèce i par une loi μ_i de densité f_i. La loi μ de la taille des nids de mouettes admet donc pour densité la fonction

$$f(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x) + \alpha_3 f_3(x) + \alpha_4 f_4(x),$$

où les réels $(\alpha_i)_{1 \leqslant i \leqslant 4}$ sont positifs et vérifient $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 1$. Les réels α_i et les densités f_i sont inconnues.

Reste à modéliser la distribution de la taille des nids pour une espèce donnée. Elle sera modélisée par une loi gaussienne de moyenne et variance inconnues (fonctions de l'espèce). La loi de la taille d'un nid apparaît alors comme le mélange de quatre densités gaussiennes dont les coefficients de mélange sont les proportions de chaque espèce. De manière générale, si $\gamma_{m,v}$ désigne la densité de la loi gaussienne $\mathcal{N}(m, v)$ sur $\mathbb{R}$ (v désigne la variance), la densité du mélange de J lois gaussiennes s'écrit

$$\sum_{j=1}^{J} \alpha(j)\gamma_{m(j),v(j)}(x) = \sum_{j=1}^{J} \frac{\alpha(j)}{\sqrt{2\pi v(j)}} \exp\left[-\frac{(x-m(j))^2}{2v(j)}\right],$$

avec,

$$\text{pour tout } j = 1, \ldots, J, \quad \alpha_j \geqslant 0, \quad \text{et} \quad \alpha_1 + \cdots + \alpha_J = 1. \tag{1}$$

Le vecteur α représente les coefficients du mélange tandis que les vecteurs m et v représentent respectivement les vecteurs des moyennes et des variances des lois gaussiennes.

En d'autres termes, si X désigne la taille du nid et Z l'espèce de l'oiseau qui l'a construit, la loi du couple (X, Z) est donnée de la manière suivante : Z suit la loi discrète sur $\{1, \dots, J\}$ de poids $(\alpha(1), \dots, \alpha(J))$ et, conditionnellement à l'événement $\{Z = j\}$, X suit la loi $\mathcal{N}(m(j), v(j))$:

$$\mathcal{L}(Z) = \sum_{j=1}^{J} \alpha(j)\delta_j \quad \text{et, pour } j = 1, \dots, J, \quad \mathcal{L}(X|Z = j) = \mathcal{N}(m(j), v(j)).$$

On ne connaît pas les proportions relatives de chaque espèce de mouettes ni les paramètres caractéristiques de la taille des nids pour chaque espèce. On souhaite estimer toutes ces quantités à la seule vue de la taille de n nids. Pour cela, on suppose qu'il existe θ dans l'ensemble

$$\Theta = \left\{ \theta = (\alpha_j, m_j, \sigma_j^2)_{1 \leqslant j \leqslant J}, (\alpha(j))_{1 \leqslant j \leqslant J} \text{ satisfait } (1) \right\}$$

tel que les mesures $x_1, \dots, x_n$ de la taille de n nids sont des réalisations d'un n-échantillon $X_1, \dots, X_n$ de loi de densité

$$f(x; \theta) = \sum_{j=1}^{J} \alpha(j)\gamma_{m(j),v(j)}(x).$$

Dans la suite du texte, les notations suivantes seront utilisées :

- $f(x; \theta)$ désigne la densité au point x de la loi de X (par rapport à la mesure de Lebesgue sur $\mathbb{R}$),
- $f(x|Z; \theta)$ désigne la densité au point x de la loi de X sachant Z,
- $g(z; \theta)$ désigne la densité au point z de la loi de Z (par rapport à la mesure de comptage dN sur $\mathbb{N}$),
- $g(z|X; \theta)$ désigne la densité au point z de la loi de Z sachant X (par rapport à la mesure de comptage sur $\mathbb{N}$),
- $h(x, z; \theta)$ désigne la densité au point (x, z) de la loi de (X, Z) par rapport à la mesure $d\lambda \otimes dN$.
- on notera $\overline{X}$ le vecteur $(X_1, \dots, X_n)$ et $\overline{Z}$ le vecteur $(Z_1, \dots, Z_n)$.

2 Une situation favorable mais irréaliste

Supposons dans un premier temps que l'on observe à la fois X et Z (la taille du nid et l'espèce qui a fait le nid). Estimer les paramètres inconnus est alors aisé grâce à la méthode du maximum de vraisemblance. La loi du couple (X, Z) admet la densité

suivante par rapport à la mesure $d\lambda \otimes dN$ (où $d\lambda$ désigne la mesure de Lebesgue sur $\mathbb{R}$ et dN la mesure de comptage sur $\mathbb{N}$) :

$$h(x, j; \theta) = \alpha(j)\gamma_{m(j),v(j)}(x)_{\{1,\ldots,J\}}(j).$$

La log-vraisemblance du modèle complet (i.e. le logarithme de la densité de la loi de l'échantillon $(X_1, Z_1, \ldots, X_n, Z_n)$ par rapport à la mesure $(d\lambda \otimes dN)^{\otimes n}$ s'écrit donc

$$\begin{aligned} L(\overline{X}, \overline{Z}, \theta) &= \ln \prod_{i=1}^{n} h(X_i, Z_i; \theta) \\ &= \sum_{i=1}^{n} \left[\ln \alpha(Z_i) + \ln \gamma_{m(Z_i),v(Z_i)}(X_i)\right] \\ &= -\frac{n}{2}\ln(2\pi) + \frac{1}{2}\sum_{i=1}^{n}\left[2\ln\alpha(Z_i) - \ln(v(Z_i)) - \frac{(X_i - m(Z_i))^2}{v(Z_i)}\right]. \end{aligned}$$

Pour un échantillon de taille n, notons, pour $j = 1, \ldots, J$,

$$A_j = \{i = 1, \ldots, n,\ Z_i = j\} \quad \text{et} \quad C_j = \text{card}(A_j).$$

La log-vraisemblance complète s'écrit alors, en regroupant les termes en fonction des valeurs de Z,

$$L(\overline{X}, \overline{Z}, \theta) = \sum_{j=1}^{J} C_j \ln \alpha(j) + \sum_{j=1}^{J}\sum_{i\in A_j} \ln \gamma_{m(j),v(j)}(X_i).$$

Proposition 2.1. *La vraisemblance complète est maximale, sous la contrainte que le vecteur $(\alpha(1), \ldots, \alpha(J))$ définisse une mesure de probabilité sur $\{1, \ldots, J\}$, pour le choix suivant des paramètres :*

$$\hat{\alpha}(j) = \frac{C_j}{n}, \quad \hat{m}(j) = \frac{1}{C_j}\sum_{i\in A_j} X_i \quad \textit{et} \quad \hat{v}(j) = \frac{1}{C_j}\sum_{i\in A_j}(X_i - m(j))^2.$$

Ce résultat ne répond pas à la question initiale puisque, en pratique, ne sont observées que les tailles des nids $x_1, \ldots, x_n$. Il ne faudrait donc pas raisonner sur la log-vraisemblance totale mais sur la log-vraisemblance de l'échantillon $X_1, \ldots, X_n$, notée L pour *log-vraisemblance des observations*, qui s'écrit :

$$L(\overline{X}; \theta) = \ln \prod_{i=1}^{n} f(X_i; \theta) = \sum_{i=1}^{n} \ln\left[\sum_{j=1}^{J} \alpha(j)\gamma_{m(j),v(j)}(X_i)\right]. \tag{2}$$

Trouver le jeu de paramètres θ rendant cette quantité maximale tourne à la mission impossible...

3 Une solution au vrai problème

On ne dispose pas de la log-vraisemblance complète $\ln L$ puisque l'on n'observe pas les variables $(Z_i)_i$. On va la remplacer par son espérance conditionnelle sachant les observations : on définit la *log-vraisemblance conditionnelle des observations* (que nous noterons $L_c(\overline{X};\theta,\theta_k)$), par :

$$L_c(\overline{X};\theta,\theta_k) = \mathbb{E}(L(\overline{X},\overline{Z};\theta)|\overline{X};\theta_k) = \sum_{i=1}^{n}\int g(z|X=X_i;\theta_k)\ln h(X_i,z;\theta)\,dz.$$

Cette quantité est l'espérance de la log-vraisemblance totale conditionnellement aux observations sous la loi de paramètre θ_k.

L'algorithme EM consiste à répéter successivement deux étapes consécutives de la manière suivante :

- étape E(xpectation) : étant donnée une valeur θ_k du paramètre, on calcule la log-vraisemblance conditionnelle des observations $L_c(\overline{X};\theta,\theta_k)$.
- étape M(aximization) : on choisit θ_{k+1} pour que la fonction qui à θ associe $L_c(\overline{X};\theta,\theta_k)$ soit maximale au point θ_{k+1}.

(Au moins) deux questions se posent :

- Peut-on effectivement executer simplement ces deux opérations ?
- En quoi ceci nous aide à trouver une valeur de θ qui donne une « grande vraisemblance » ?

3.1 Comment ça marche ?

Le calcul de la fonction L_c nécessite la connaissance de la loi de Z sachant X. La proposition suivante assure que c'est le cas.

Proposition 3.1. *La densité de la loi de Z sachant que $X = x$ par rapport à la mesure de comptage sur $\mathbb{N}$ est donnée par :*

$$g(z|X=x;\theta) = \frac{h(x,z;\theta)}{f(x;\theta)} = \frac{\alpha(z)\gamma_{m(z),v(z)}(x)}{\sum_{j=1}^{J}\alpha(j)\gamma_{m(j),v(j)}(x)}_{\{1,\dots,J\}}(z).$$

Dans la section précédente, on a déterminé la vraisemblance totale ce qui donne l'expression de L_c.

Proposition 3.2. *La fonction L_c est de la forme suivante :*

$$\begin{aligned} L_c(\overline{X};\theta,\theta_k) &= -\frac{n}{2}\ln(2\pi) + \sum_{j=1}^{J}\left(\sum_{i=1}^{n} g(j|X=X_i;\theta_k)\right)\ln\alpha(j) \\ &\quad -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{J}\left[\ln v(j) + \frac{(X_i-m(j))^2}{v(j)}\right]g(j|X=X_i;\theta_k). \end{aligned}$$

Enfin, la fonction L_c admet un unique maximum comme le montre la proposition suivante.

Proposition 3.3. *L'étape M consiste à choisir θ_{k+1} de la façon suivante :*

$$\begin{aligned}
\alpha_{k+1}(j) &= \frac{1}{n}\sum_{i=1}^{n} g(j|X=X_i;\theta_k) \\
m_{k+1}(j) &= \frac{\sum_{i=1}^{n} X_i g(j|X=X_i;\theta_k)}{\sum_{i=1}^{n} g(j|X=X_i;\theta_k)} \\
\sigma_{k+1}(j)^2 &= \frac{\sum_{i=1}^{n} (X_i-m_{k+1})^2 g(j|X=X_i;\theta_k)}{\sum_{i=1}^{n} g(j|X=X_i;\theta_k)}.
\end{aligned}$$

3.2 Pourquoi ça marche ?

Il est plus difficile d'apporter une réponse à cette question. Le résultat suivant ouvre la voie en montrant que la (log-)vraisemblance est croissante le long de l'algorithme.

Theorème 3.4. *La suite $(\theta_k)_k$ construite par l'algorithme EM vérifie la propriété de stabilité numérique suivante :*

$$L(\overline{X};\theta_{k+1}) \geqslant L(\overline{X};\theta_k), \tag{3}$$

où $L(\overline{X};\theta)$ est définie par (2).

Démonstration. Pour alléger les notations, posons $Q(\theta,\theta_k) = L_c(\overline{X};\theta,\theta_k)$. Puisque

$$h(x,z;\theta) = f(x;\theta)g(z|X=x;\theta),$$

la fonction Q s'écrit encore

$$\begin{aligned}
Q(\theta,\theta_k) &= \mathbb{E}(L(\overline{X},\overline{Z};\theta)|\overline{X};\theta_k) \\
&= \mathbb{E}(L(\overline{X};\theta)|\overline{X};\theta_k) + \mathbb{E}(L(\overline{Z}|\overline{X};\theta_k)|\overline{X};\theta_k) \\
&= L(\overline{X};\theta) + H(\theta,\theta_k),
\end{aligned}$$

où $L(\overline{Z}|\overline{X};\theta) = \ln g(\overline{Z}|\overline{X};\theta)$ et $H(\theta,\theta_k) = \mathbb{E}(L(\overline{Z}|\overline{X};\theta_k)|\overline{X};\theta_k)$.

Lors de l'étape M, la fonction $Q(\theta,\theta_k)$ est maximisée par rapport à θ donc, en particulier,

$$Q(\theta_{k+1},\theta_k) \geqslant Q(\theta_k,\theta_k).$$

On en déduit que

$$L(\overline{X};\theta_{k+1}) + H(\theta_{k+1},\theta_k) \geqslant L(\overline{X};\theta_k) + H(\theta_k,\theta_k). \tag{4}$$

Il reste à obtenir la majoration $H(\theta_{k+1},\theta_k) \leqslant H(\theta_k,\theta_k)$. Celle ci est une conséquence de l'inégalité de Jensen :

$$\begin{aligned} H(\theta_{k+1};\theta_k) - H(\theta_k;\theta_k) &= \mathbb{E}(L(\overline{Z}|\overline{X};\theta_{k+1})|\overline{X};\theta_k) - \mathbb{E}(L(\overline{Z}|\overline{X};\theta_k)|\overline{X};\theta_k) \\ &= \mathbb{E}(\ln(g(\overline{Z}|\overline{X};\theta_{k+1})/g(\overline{Z}|\overline{X};\theta_k))|\overline{X};\theta_k) \\ &\leqslant \ln \mathbb{E}(g(\overline{Z}|\overline{X};\theta_{k+1})/g(\overline{Z}|\overline{X};\theta_k)|\overline{X};\theta_k). \end{aligned}$$

Enfin, on a

$$\mathbb{E}(g(\overline{Z}|\overline{X};\theta_{k+1})/g(\overline{Z}|\overline{X};\theta_k)|\overline{X};\theta_k) = \int \frac{g(\overline{Z}|\overline{X};\theta_{k+1})}{g(\overline{Z}|\overline{X};\theta_k)} g(\overline{Z}|\overline{X};\theta_k) d\overline{Z} = 1.$$

Puisque $H(\theta_{k+1},\theta_k) \leqslant H(\theta_k,\theta_k)$, la relation (4) implique (3). □

Ce résultat permet de montrer que la suite $(\theta_k)_k$ converge vers un maximum local ou un point selle de la vraisemblance.

4 L'algorithme EM

Étant données les observations $x_1,\dots,x_n$ et des valeurs initiales des paramètres, l'algorithme consiste à répéter les calculs suivants.

— Étant donnés à l'étape k, les trois vecteurs

$$\alpha_k = (\alpha_k(1),\dots,\alpha_k(J)), \quad m_k = (m_k(1),\dots,m_k(J)) \quad \text{et} \quad v_k = (v_k(1),\dots,v_k(J)),$$

(v pour variance), on définit la matrice $H^{(k)}$ de taille $n \times J$ par

$$H_{ij}^{(k)} = \frac{\alpha_k(j)\gamma_{m_k(j),v_k(j)}(X_i)}{\sum_{l=1}^{J} \alpha_k(l)\gamma_{m_k(l),v_k(l)}(X_i)}.$$

— Étant donnée $H^{(k)}$, on obtient α_{k+1}, m_{k+1} et v_{k+1} par les relations suivantes :

$$\begin{aligned} \alpha_{k+1}(j) &= \frac{1}{n}\sum_{i=1}^{n} H_{ij}^{(k)} \\ m_{k+1}(j) &= \frac{\sum_{i=1}^{n} X_i H_{ij}^{(k)}}{\sum_{i=1}^{n} H_{ij}^{(k)}} \\ v_{k+1}(j) &= \frac{\sum_{i=1}^{n} (X_i - m_j)^2 H_{ij}^{(k)}}{\sum_{i=1}^{n} H_{ij}^{(k)}}. \end{aligned}$$

Remarque 4.1. Contrairement au cas où l'on observe en même temps X et Z, il peut exister des maxima locaux qui vont piéger l'algorithme. Cette procédure peut s'avérer très sensible au point de départ choisi.

Remarque 4.2. On ne sait pas combien d'itérations sont nécessaires pour approcher de manière satisfaisante un maximum local. Il n'y a pas en général de procédure pour prévoir le nombre d'itérations à opérer.

5 Suggestions

1. On pourra commenter le choix du modèle paramétrique utilisé.
2. On pourra démontrer la proposition 2.1.
3. On pourra démontrer la proposition 3.2.
4. On pourra démontrer la proposition 3.3.
5. On pourra démontrer le théorème 3.4.
6. On pourra illustrer par la simulation la convergence de l'algorithme EM.
7. On pourra tenter d'illustrer par la simulation la remarque 4.1.
8. On pourra tenter d'illustrer par la simulation la remarque 4.2.