Modèles statistiques à variables latentes pour l'écologie

Examen de 2 heures

29 mars 2022

Les notes de cours et une calculatrice sont autorisées, à l'exclusion de tout autre appareil électronique (téléphone compris).

1 Algorithme EM pour l'ACP probabiliste

Modèle et notations. On considère le modèle d'analyse en composantes principales (ACP) probabiliste suivant :

$$\begin{aligned} &\{Z_i\}_{1\leq i\leq n} \text{ iid} : && Z_i \sim \mathcal{N}(0_q, I_q) \\ &\{Y_i\}_{1\leq i\leq n} \text{ indépendants} \mid \{Z_i\}_{1\leq i\leq n} : && (Y_i \mid Z_i) \sim \mathcal{N}(AZ_i, \sigma^2 I_p) \end{aligned} \tag{1}$$

où $q < p$, A est de dimension $p \times q$, les variables Z_i sont latentes alors que les Y_i sont observées. On note

- $Z = [Z_{ik}]_{1\leq i\leq n, 1\leq k\leq q}$ la matrice $n \times q$ contenant les variables latentes,
- $Y = [Y_{ij}]_{1\leq i\leq n, 1\leq j\leq p}$ la matrice $n \times p$ contenant les variables observées,
- $\theta = (A, \sigma^2)$ l'ensemble des paramètres de ce modèle,
- Σ et Γ les matrices :

$$\Sigma = AA^\intercal + \sigma^2 I_p, \qquad \Gamma = A^\intercal A + \sigma^2 I_q.$$

On se propose d'établir un algorithme EM pour l'estimation de θ.

Questions préliminaires.

1. Montrer que $A^\intercal \Sigma^{-1} = \Gamma^{-1} A^\intercal$.
2. Montrer que $I_q - A^\intercal \Sigma^{-1} A = \sigma^2 \Gamma^{-1}$.

Estimation par EM.

3. Écrire la log-vraisemblance complète $\log p_\theta(Y, Z)$ du modèle (1).
4. Déterminer la loi jointe d'un couple (Y_i, Z_i) pour $1 \leq i \leq n$ quelconque.
5. En déduire que

$$M_i := \mathbb{E}(Z_i \mid Y) = \Gamma^{-1} A^\intercal Y_i, \qquad Q_i := \mathbb{E}(Z_i Z_i^\intercal \mid Y) = \sigma^2 \Gamma^{-1} + M_i M_i^\intercal. \tag{2}$$

6. Écrire l'espérance conditionnelle de la log-vraisemblance complète $\mathbb{E}_\theta(\log p_\theta(Y, Z) \mid Y)$ en fonction des M_i et Q_i.
7. En déduire les formules de mise à jour à l'étape h de A^h et $(\sigma^2)^h$ en fonction des moments conditionnels M_i^{h-1} et Q_i^{h-1} calculés à l'étape précédente.

Estimation alternative.

8. En combinant les formules obtenues à la question précédente avec l'équation (2), montrer que les estimateurs du maximum de vraisemblance $\widehat{A}$ et $\widehat{\sigma}^2$ satisfont les équations de point fixe suivantes :

$$\widehat{A} = S\widehat{A}\left(\widehat{\sigma}^2 I_q + \widehat{\Gamma}^{-1} \widehat{A}^\intercal S \widehat{A}\right)^{-1}, \qquad \widehat{\sigma}^2 = \operatorname{tr}\left(S - S\widehat{A}\widehat{\Gamma}^{-1}\widehat{A}\right)/p.$$

où $\widehat{\Gamma} = \widehat{A}^\intercal \widehat{A} + \widehat{\sigma}^2 I_q$ et S est la matrice de covariance empirique : $S = \left(\sum_i Y_i Y_i^\intercal\right)/n$.

9. En déduire un algorithme alternatif à EM pour l'estimation de $\theta = (A, \sigma^2)$ par maximum de vraisemblance.

2 Distribution jointe d'absence et d'abondance d'espèces

Modèle et notations. On s'intéresse à la présence et à l'abondance de p espèces animales dans n sites. On observe pour cela

— Y_{ij} = le nombre (éventuellement nul) d'individus de l'espèce j observés dans le site i ($Y_{ij} \in \mathbb{N}$) et

— x_i = vecteur de covariables environnementales (incluant une constante) décrivant le site i ($x_i \in \mathbb{R}^d$).

On définit $\widetilde{Y}_{ij}$ la variable indicatrice d'absence de l'espèce j dans le site i :

$$\widetilde{Y}_{ij} = \mathbb{I}\{Y_{ij} = 0\}$$

et on note

— $Y = [Y_{ij}]_{1 \leq i \leq n, 1 \leq j \leq p}$ la matrice $n \times p$ des abondances,

— $X = [x_{ik}]_{1 \leq i \leq n, 1 \leq k \leq d}$ la matrice $n \times d$ des covariables

— $\widetilde{Y} = [\widetilde{Y}_{ij}]_{1 \leq i \leq n, 1 \leq j \leq p}$ la matrice $n \times p$ des absences.

Questions.

1. Rappeler le modèle Poisson log-normal permettant de décrire les abondances en fonction des covariables environnementales et des interactions entre espèces.
2. Proposer un modèle analogue au modèle Poisson log-normal permettant de décrire les absences en fonction des covariables environnementales et des interactions entre espèces.
3. Proposer un modèle décrivant conjointement les absences et les abondances en fonction des covariables environnementales et des interactions entre espèces.
   Tracer le modèle graphique orienté associé à ce modèle et interpréter chacun de ses paramètres.

3 Classification non supervisée de génotypes

Modèle et notations. On considère un échantillon de $n = 74$ souris (*mus musculus*) dont on a relevé le génotype pour $p = 15$ marqueurs génétiques (nommés `Aat`, `Amy`, `Es1`, `Es2`, `Es10`, `Hbb`, `Gpd1`, `Idh1`, `Mod1`, `Mod2`, `Mpi`, `Np`, `Pgm1`, `Pgm2` et `Sod`, qui peuvent être vus comme des variables catégorielles). On cherche à identifier des individus issus de groupes génétiquement distincts. On se propose d'utiliser à cette fin un modèle de mélange de lois multinomiales.

On note

— Y_{ij} le génotype de l'individu i au marqueur j pour $1 \leq i \leq n$ et $1 \leq j \leq p$,

— m_j le nombre d'allèles du j-ème marqueurs ($1 \leq j \leq p$).

On suppose le modèle de mélange à K groupes suivant

$$\begin{aligned} (Z_i)_{1 \leq i \leq n} \text{ iid} : & \qquad Z_i \sim \mathcal{M}(1, \pi), \\ (Y_i)_{1 \leq i \leq n, 1 \leq j \leq p} \text{ independants} \mid (Z_i) : & \qquad (Y_{ij} \mid Z_i = k) \sim \mathcal{M}(1, \gamma_{kj}) \end{aligned} \tag{3}$$

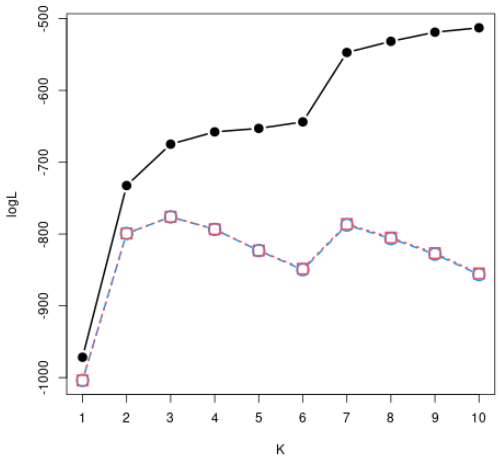
où $\pi \in [0,1]^K$, $\sum_{k=1}^K \pi_k = 1$, et pour chaque $1 \leq j \leq p$, $\gamma_{kj} \in [0,1]^{m_j}$, $\sum_{a=1}^{m_j} \gamma_{kja} = 1$.

Questions.

1. Interpréter chacun des paramètres π_k et γ_{kja} de ce modèle.
2. Discuter les hypothèses d'indépendances.

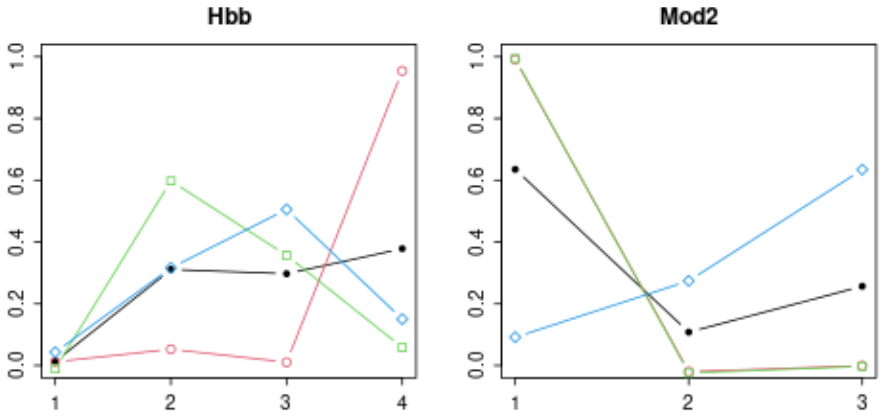
Questions.

3. La figure suivante donne les valeurs de la log-vraisemblance (•), du critère BIC (□) et du critère ICL (◯) du modèle (3) pour K allant de 1 à 10 groupes.



Justifier le choix de $\widehat{K} = 3$.

4. La figure suivante donne les estimations des fréquences alléliques γ_{ja} pour le modèle à $K = 3$ classes pour les marqueurs `Hbb` et `Mod2`. Abcsisse = allèle du marqueur, ordonnée = fréquence. Légende : • = fréquence de chaque allèle dans l'échantillon total, ○ = fréquence estimée dans le groupe 1, □ = dans le groupe 2, ◇ = dans le groupe 3.



Quels groupes du mélange chacun de ces marqueurs permet-il le mieux de distinguer ?

5. Les proportions estimées pour le modèle à 3 groupes valent

$$\widehat{\pi} = [0.324, 0.270, 0.406].$$

Pour les marqueurs `Gpd1` et `Mpi`, on obtient les estimations suivantes pour les fréquences alléliques γ_{kja} dans chaque groupe :

Marqueur Gpd1	$a=1$	$a=2$	$a=3$	$a=4$	$a=5$	$a=6$
$k=1$	0	0	0	1	0	0
$k=2$	0.05	0.95	0	0	0	0
$k=3$	0.033	0.167	0.2	0.301	0.167	0.133
Population	0.027	0.324	0.081	0.446	0.068	0.054

Marqueur Mpi	$a=1$	$a=2$	$a=3$	$a=4$
$k=1$	0	1	0	0
$k=2$	0.05	0	0.4	0.55
$k=3$	0	1	0	0
Population	0.014	0.73	0.108	0.149

A partir de ces valeurs, donner une estimation, selon de modèle (3), de la probabilité conditionnelle qu'un individu du groupe k porte simultanément les deuxièmes allèles des marqueurs `Gpd1` et `Mpi` : $\Pr\{Y_{i,Gpd1} = Y_{i,Mpi} = 2 \mid Z_i = k\}$ pour chaque $k = 1, 2, 3$.
En déduire une estimation de la probabilité marginale $\Pr\{Y_{i,Gpd1} = Y_{i,Mpi} = 2\}$.

6. Comparer ce résultat avec les fréquences alléliques moyennes dans la population et commenter.

Comparaison avec des sous-espèces connues. On sait par ailleurs que les 74 individus de l'échantillon appartiennent en fait à trois sous-espèces connues (*castaneus*, *domesticus* et *musculus*) et à une population vivant près du lac Casitas (Californie). Le tableau suivant croise l'appartenance à ces populations avec les groupes obtenus par le modèle de mélange :

	Casitas	*castaneus*	*domesticus*	*musculus*	Total
$k = 1$	1	0	23	0	24
$k = 2$	0	11	0	9	20
$k = 3$	29	0	1	0	30
Total	30	11	24	9	74

Questions.

7. Les marqueurs permettent-ils de distinguer les sous-espèces connues entre elles ?
Quelles sont les sous-espèces génétiquement les plus proches ?
8. La population vivant près du lac Casitas peut-elle être rattachée à une sous-espèce connue.