

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 :

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

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 :

$$\Sigma = AA^\top + \sigma^2 I_p, \quad \Gamma = A^\top A + \sigma^2 I_q.$$

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

Questions préliminaires.

1. Montrer que $A^\top \Sigma^{-1} = \Gamma^{-1} A^\top$.

Solution. En notant $B = \sigma^{-1}A$, on a

$$\begin{aligned} \sigma A^\top \Sigma^{-1} &= B^\top (I_p + BB^\top)^{-1} = B^\top \left(\sum_{k \geq 0} (BB^\top)^k / k! \right) \\ &= B^\top \left(I_p + B \left(\sum_{k \geq 1} (B^\top B)^{k-1} / k! \right) B^\top \right) = \left(I_p + B^\top B \left(\sum_{k \geq 1} (B^\top B)^{k-1} / k! \right) \right) B^\top \\ &= \left(\sum_{k \geq 0} (B^\top B)^k / k! \right) B^\top = (I_p + B^\top B)^{-1} B^\top = \sigma \Gamma^{-1} A^\top. \end{aligned}$$

Alternativement, on peut remarquer que Σ et Γ sont inversibles et que, donc,

$$A^\top \Sigma^{-1} = \Gamma^{-1} A^\top \quad \Leftrightarrow \quad A^\top = \Gamma^{-1} A^\top \Sigma \quad \Leftrightarrow \quad \Gamma A^\top = A^\top \Sigma$$

qui se vérifie facilement.

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

Solution. On vérifie que

$$\begin{aligned} \Gamma (I_q - A^\top \Sigma^{-1} A) &= (A^\top A + \sigma^2 I_q) (I_q - A^\top \Sigma^{-1} A) \\ &= A^\top A + \sigma^2 I_q - A^\top A A^\top \Sigma^{-1} A - \sigma^2 A^\top \Sigma^{-1} A \\ &= A^\top A + \sigma^2 I_q - A^\top (A A^\top + \sigma^2 I_p) \Sigma^{-1} A = \sigma^2 I_q \end{aligned}$$

et que $(I_q - A^\top \Sigma^{-1} A) \Gamma = \sigma^2 I_q$.

Estimation par EM.

3. Écrire la log-vraisemblance complète $\log p_\theta(Y, Z)$ du modèle (1).

Solution. En omettant les termes ne dépendant pas de θ , on a

$$\log p_\theta(Y, Z) = \log p_\theta(Z) + \log p_\theta(Y | Z) = -\frac{1}{2} \sum_i \|Z_i\|^2 - \frac{np}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_i \|Y_i - AZ_i\|^2.$$

4. Déterminer la loi jointe d'un couple (Y_i, Z_i) pour $1 \leq i \leq n$ quelconque.

Solution. Les couples $\{(Y_i, Z_i)\}_{1 \leq i \leq n}$ sont iid de loi normale

$$\begin{bmatrix} Y_i \\ Z_i \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} 0_p \\ 0_q \end{bmatrix}, \begin{bmatrix} \Sigma & A \\ A^\top & I_q \end{bmatrix} \right).$$

5. En déduire que

$$M_i := \mathbb{E}(Z_i | Y) = \Gamma^{-1} A^\top Y_i, \quad Q_i := \mathbb{E}(Z_i Z_i^\top | Y) = \sigma^2 \Gamma^{-1} + M_i M_i^\top. \quad (2)$$

Solution.

a) L'indépendance des couples (Y_i, Z_i) nous assure que $M_i = \mathbb{E}(Z_i | Y_i)$ et $\mathbb{V}(Z_i | Y) = \mathbb{V}(Z_i | Y_i)$.

b) La loi jointe de (Y_i, Z_i) établie à la question précédente implique que $M_i = A^\top \Sigma^{-1} Y_i$ et $\mathbb{V}(Z_i | Y_i) = I_q - A^\top \Sigma^{-1} A$.

c) Les questions préliminaires impliquent que $M_i = \Gamma^{-1} A^\top Y_i$ et $\mathbb{V}(Z_i | Y) = \sigma^2 \Gamma^{-1}$.

d) On obtient finalement Q_i en utilisant l'identité $\mathbb{E}(Z_i Z_i^\top | Y) = \mathbb{V}(Z_i | Y) + M_i M_i^\top$.

(Noter que $Z_i Z_i^\top \neq \|Z_i\|^2$ et que $\mathbb{E}(\|Z_i\|^2 | Y) = \text{tr}(\mathbb{V}(Z_i | Y)) + M_i^\top M_i$)

6. Écrire l'espérance conditionnelle de la log-vraisemblance complète $\mathbb{E}_\theta(\log p_\theta(Y, Z) | Y)$ en fonction des M_i et Q_i .

Solution. On a

$$\begin{aligned}\mathbb{E}_\theta(\log p_\theta(Y, Z) | Y) &= -\frac{np}{2} \log \sigma^2 - \frac{1}{2} \sum_i \text{tr}(Q_i) \\ &\quad - \frac{1}{2\sigma^2} \sum_i \|Y_i\|^2 + \frac{1}{\sigma^2} \sum_i M_i^\top A^\top Y_i - \frac{1}{2\sigma^2} \sum_i \text{tr}(A^\top A Q_i) \\ &= -\frac{np}{2} \log \sigma^2 - \frac{1}{2} \sum_i \text{tr}(Q_i) \\ &\quad - \frac{1}{2\sigma^2} \sum_i (\|Y_i - AM_i\|^2 + \text{tr}(A^\top A \mathbb{V}(Z_i | Y)))\end{aligned}$$

7. En d eduire les formules de mise  a jour  a l' etape h de A^h et $(\sigma^2)^h$ en fonction des moments conditionnels M_i^{h-1} et Q_i^{h-1} calcul es  a l' etape pr ecedente.

Solution. On calcule les d eriv ees de $f^{h-1}(\theta) := \mathbb{E}_{\theta^{h-1}}[\log p_\theta(Y, Z) | Y]$ par rapport  a σ^2 et A :

$$\begin{aligned}\partial_{\sigma^2} f^{h-1} &= -\frac{np}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_i (\|Y_i - AM_i\|^2 + \text{tr}(A^\top A \mathbb{V}(Z_i | Y))), \\ \partial_A f^{h-1} &= \frac{1}{\sigma^2} \left(\sum_i Y_i M_i^\top - A \sum_i Q_i \right)\end{aligned}$$

qui s'annulent respectivement pour

$$\begin{aligned}(\sigma^2)^h &= \frac{1}{np} \sum_i (\|Y_i - AM_i\|^2 + \text{tr}(A^\top A \mathbb{V}(Z_i | Y))), \\ A^h &= \left(\sum_i Y_i M_i^\top \right) \left(\sum_i Q_i \right)^{-1}.\end{aligned}$$

Estimation alternative.

8. En combinant les formules obtenues  a la question pr ecedente avec l' equation (2), montrer que les estimateurs du maximum de vraisemblance \hat{A} et $\hat{\sigma}^2$ satisfont les  equations de point fixe suivantes :

$$\hat{A} = S \hat{A} \left(\hat{\sigma}^2 I_q + \hat{\Gamma}^{-1} \hat{A}^\top S \hat{A} \right)^{-1}, \quad \hat{\sigma}^2 = \text{tr} \left(S - S \hat{A} \hat{\Gamma}^{-1} \hat{A} \right) / p.$$

o u $\hat{\Gamma} = \hat{A}^\top \hat{A} + \hat{\sigma}^2 I_q$ et S est la matrice de covariance empirique : $S = (\sum_i Y_i Y_i^\top) / n$.

Solution.  a convergence, on doit avoir $\hat{A} = A^h = A^{h-1}$ et $\hat{\sigma}^2 = (\sigma^2)^h = (\sigma^2)^{h-1}$, soit

$$\begin{aligned}\hat{A} &= \left(\sum_i Y_i M_i^\top \right) \left(\sum_i Q_i \right)^{-1} = \left(\sum_i Y_i Y_i^\top \right) \hat{A} \hat{\Gamma}^{-1} \left(n \hat{\Gamma}^{-1} + \hat{\Gamma}^{-1} \hat{A} \sum_i Y_i Y_i^\top \hat{A}^\top \hat{\Gamma}^{-1} \right)^{-1} \\ &= S \hat{A}^\top \left(\hat{\sigma}^2 I_q + \hat{\Gamma}^{-1} \hat{A}^\top S \hat{A} \right)^{-1}\end{aligned}$$

et idem pour $\hat{\sigma}^2$.

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

Solution. L'algorithme du point fixe consistant à itérer les équations précédentes jusqu'à convergence.

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

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

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
- $\tilde{Y} = [\tilde{Y}_{ij}]_{1 \leq i \leq n, 1 \leq j \leq p}$ 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.

Solution.

$$\begin{aligned} (Z_i)_i \text{ iid} : & & Z_i &\sim \mathcal{N}(0, \Sigma), \\ (Y_{ij})_{ij} \text{ indep.} \mid (Z_i) : & & (Y_{ij} \mid Z_{ij}) &\sim \mathcal{P}(\exp(x_i^\top \beta_j + Z_{ij})). \end{aligned}$$

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.

Solution.

$$\begin{aligned} (\tilde{Z}_i)_i \text{ iid} : & & \tilde{Z}_i &\sim \mathcal{N}(0, \tilde{\Sigma}), \\ (\tilde{Y}_{ij})_{ij} \text{ indep.} \mid (\tilde{Z}_i) : & & (\tilde{Y}_{ij} \mid \tilde{Z}_{ij}) &\sim \mathcal{B}((1 + \exp(-x_i^\top \tilde{\beta}_j - \tilde{Z}_{ij}))^{-1}). \end{aligned}$$

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.

Solution. En supposant que les processus de colonisation (présence/absence) et d'abondance

sont gouvernés par des paramètres distincts :

$$\begin{aligned}
 (\tilde{Z}_i)_i \text{ iid} : & \quad \tilde{Z}_i \sim \mathcal{N}(0, \tilde{\Sigma}), \\
 (\tilde{Y}_{ij})_{ij} \text{ indep.} \mid (\tilde{Z}_i) : & \quad (\tilde{Y}_{ij} \mid \tilde{Z}_{ij}) \sim \mathcal{B}((1 + \exp(-x_i^\top \tilde{\beta}_j - \tilde{Z}_{ij}))^{-1}), \\
 (Z_i)_i \text{ iid} : & \quad Z_i \sim \mathcal{N}(0, \Sigma), \\
 (Y_{ij})_{ij} \text{ indep.} \mid (Z_i), (\tilde{Y}_{ij}) : & \quad (Y_{ij} \mid Z_{ij}, \tilde{Y}_{ij}) \sim \tilde{Y}_{ij} \delta_0 + (1 - \tilde{Y}_{ij}) \mathcal{P}(\exp(x_i^\top \beta_j + Z_{ij})).
 \end{aligned}$$

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

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

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.

Solution.

- π_k = proportions d'individu de la population issue du groupe k .
- γ_{kja} = probabilité qu'un individu issu de la population k porte l'allèle a au marqueur j .

2. Discuter les hypothèses d'indépendances.

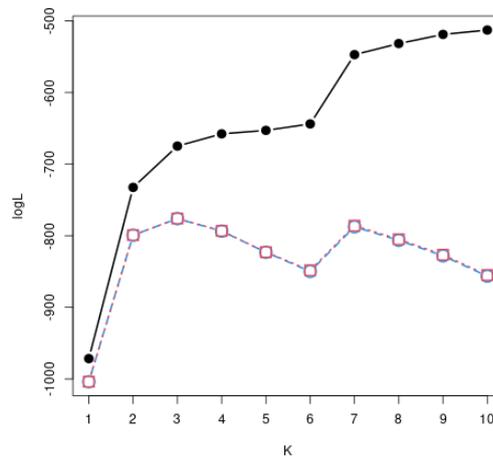
Solution.

- L'indépendance des Z_i suppose que les individus ne sont pas apparentés.
- L'indépendance conditionnelles des Y_{ij} suppose que les marqueurs sont indépendants.

Questions.

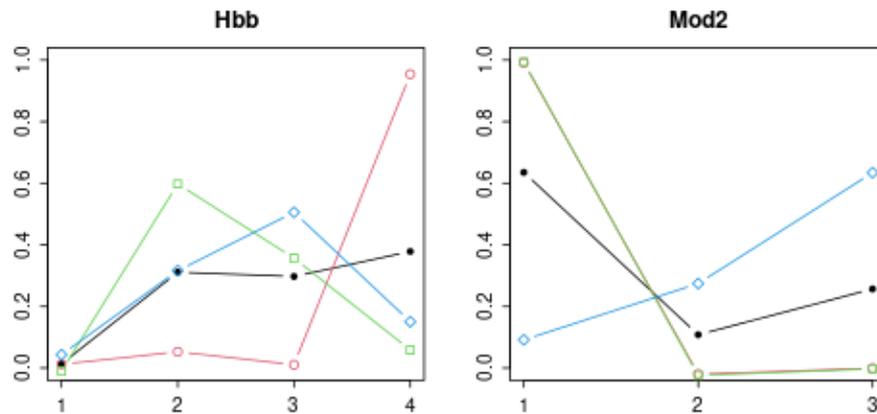
3. La figure suivante donne les valeurs de la log-vraisemblance (\bullet), du critère BIC (\square) et du critère

ICL (\circ) du modèle (3) pour K allant de 1 à 10 groupes.



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

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**. Abscisse = allèle du marqueur, ordonnée = fréquence. Légende : \bullet = fréquence de chaque allèle dans l'échantillon total, \circ = fréquence estimée dans le groupe 1, \square = dans le groupe 2, \diamond = 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

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

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 $M_{pi} : \Pr\{Y_{i,Gpd1} = Y_{i,Mpi} = 2 \mid Z_i = k\}$ pour chaque $k = 1, 2, 3$.

En déduire une estimation de la probabilité marginale $\Pr\{Y_{i,Gpd1} = Y_{i,Mpi} = 2\}$.

Solution. Du fait de l'hypothèse d'indépendance conditionnelle des marqueurs, on a

$$\Pr\{Y_{i,Gpd1} = Y_{i,Mpi} = 2 \mid Z_i = k\} = \Pr\{Y_{i,Gpd1} = 2 \mid Z_i = k\} \times \Pr\{Y_{i,Mpi} = 2 \mid Z_i = k\}$$

qui donne $\widehat{\Pr}\{Y_{i,Gpd1} = Y_{i,Mpi} = 2 \mid Z_i = k\} = 0$ ($k = 1$), 0 ($k = 2$), 0.167 ($k = 3$).

En intégrant sur le groupe caché

$$\Pr\{Y_{i,Gpd1} = Y_{i,Mpi} = 2\} = \sum_k \pi_k \Pr\{Y_{i,Gpd1} = Y_{i,Mpi} = 2 \mid Z_i = k\},$$

on obtient $\widehat{\Pr}\{Y_{i,Gpd1} = Y_{i,Mpi} = 2\} = 0.068$.

Solution.

— L'indépendance des Z_i suppose que les individus ne sont pas apparentés.

— L'indépendance conditionnelles des Y_{ij} suppose que les marqueurs sont indépendants.

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

Solution. L'hypothèse d'indépendance marginale des marqueurs dans la population amènerait à l'estimation

$$\widetilde{\Pr}\{Y_{i,Gpd1} = Y_{i,Mpi} = 2\} = 0.324 \times 0.73 = 0.237$$

qui diffère notablement de $\widehat{\Pr}\{Y_{i,Gpd1} = Y_{i,Mpi} = 2\} = 0.068$.

Le modèle de mélange ne suppose qu'une indépendance conditionnelle des marqueurs, mais la structure en groupes induit une dépendance marginale.

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	<i>castaneus</i>	<i>domesticus</i>	<i>musculus</i>	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.