Modèles statistiques pour l'écologie

Examen de 3 heures

28 mars 2023

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

1 Algorithme EM pour la loi Gamma-Poisson

On considère un couple de variables aléatoire (Z, Y) dont la loi est définie de la façon suivante :

$$Z \sim \mathcal{G}\text{am}(a, a), \qquad (Y \mid Z) \sim \mathcal{P}(\lambda Z_i). \tag{1}$$

On rappelle que la densité de la loi Gamma $\mathcal{G}\text{am}(a, b)$ est, pour $u \in \mathbb{R}^+$,

$$\mathcal{G}\text{am}(u; a, b) = \frac{b^a}{\Gamma(a)}\, u^{a-1}\, e^{-bu}$$

et on rappelle certains de ces moments : si $U \sim \mathcal{G}\text{am}(a, b)$, alors

$$\mathbb{E}(U) = a/b, \qquad \mathbb{V}(U) = a/b^2, \qquad \mathbb{E}[\log(U)] = \psi(a) - \log(b)$$

où $\psi(x)$ est la fonction *digamma*, définie comme la dérivée de la fonction $\log(\Gamma(x))$. On donne, sans démonstration, la propriété suivante.

Proposition 1. *Pour tout $c < 0$, l'équation $\psi(x) - \log(x) = c$ admet une unique solution qu'on sait déterminer (au moins numériquement).*

Propriétés de la loi jointe. On considère un couple (Z, Y) de loi donnée en (1).

1. Montrer que les deux premiers moments marginaux de Y sont

$$\mathbb{E}(Y) = \lambda, \qquad \mathbb{V}(Y) = \lambda\left(1 + \frac{\lambda}{a}\right).$$

En quoi peut-on dire que la loi marginale de Y est surdispersée ?

Solution. Puisque $\mathbb{E}(Z) = 1$ et $\mathbb{V}(Z) = 1/a$, on a

$$\mathbb{E}(Y) = \mathbb{E}_Z[\mathbb{E}(Y \mid Z)] = \mathbb{E}[\lambda Z] = \lambda\frac{a}{a},$$
$$\mathbb{V}(Y) = \mathbb{V}_Z[\mathbb{E}(Y \mid Z)] + \mathbb{E}_Z[\mathbb{V}(Y \mid Z)] = \lambda^2\mathbb{V}[Z] + \lambda\mathbb{E}[Z] = \frac{\lambda^2}{a} + \lambda > \lambda.$$

La variance marginale de Y est donc strictement supérieure à celle d'une loi de Poisson de même espérance.

2. Déterminer la loi de Z sachant Y.

Solution. La densité de Z conditionnellement à $Y = y$ vaut

$$\frac{\mathcal{G}\mathrm{am}(z; a, a)\mathcal{P}(y; \lambda z)}{\Pr\{Y = y\}} = \frac{1}{\Pr\{Y = y\}} \frac{a^a}{\Gamma(a)} \frac{1}{y!} z^{a-1} e^{-bz} e^{-\lambda z} (\lambda z)^y \propto z^{a+y-1} e^{-(a+\lambda)z}$$

où on reconnaît la densité de la loi $\mathcal{G}\mathrm{am}(a + y, a + \lambda)$.

3. En déduire que

$$\mathbb{E}(Z \mid Y = y) = (a + y)/(a + \lambda), \qquad \mathbb{E}[\log(Z) \mid Y = y] = \Psi(a + y) - \log(a + \lambda).$$

Solution. C'est une application directe à la loi $\mathcal{G}\mathrm{am}(a + y, a + \lambda)$ des formules des moments données dans l'énoncé.

Estimation par l'algorithme EM. La loi Gamma-Poisson peut être utilisée pour modéliser les abondances (i.e. le nombre d'individus observés) d'espèces différentes présentes dans un même site. On suppose alors que n couples (Y_i, Z_i) de loi (1) sont associés à chacune des espèces et qu'on observe seulement les abondances $Y = (Y_i)_{1 \le i \le n}$. On note de plus $Z = (Z_i)_{1 \le i \le n}$ les n variables latentes qui leur sont respectivement associées. On cherche à estimer le paramètre $\theta = (a, \lambda)$ au moyen de l'algorithme EM.

4. Écrire la log-vraisemblance complète $\log p_\theta(Y, Z)$

Solution. On écrit directement

$$\begin{aligned}\log p_\theta(Y, Z) &= \sum_i \log \mathcal{G}\mathrm{am}(Z_i; a, a) + \log \mathcal{P}(Y_i; \lambda Z_i) \\ &= na \log a - n \log \Gamma(a) + \sum_i (a + Y_i - 1) \log Z_i \\ &\quad + \log \lambda \sum_i Y_i - \sum_i (a + \lambda) Z_i - \sum_i \log(Y_i!)\end{aligned}$$

5. En déduire son espérance conditionnelle à Y : $\mathbb{E}_\theta[\log p_\theta(Y, Z) \mid Y]$.

Solution. On a

$$\begin{aligned}\mathbb{E}_\theta[\log p_\theta(Y, Z) \mid Y] &= na \log a - n \log \Gamma(a) + \sum_i (a + Y_i - 1)\nu_i \\ &\quad + \log \lambda \sum_i Y_i - \sum_i (a + \lambda)\mu_i + \mathrm{cst},\end{aligned}$$

en notant $\mu_i = \mathbb{E}_\theta(Z_i \mid Y)$ et $\nu_i = \mathbb{E}_\theta(\log Z_i \mid Y)$.

6. Étape E : calculer, pour une valeur courante θ^h du paramètre, les moments conditionnels des variables Z_i nécessaires au calcul de $\mathbb{E}_{\theta^h}[\log p_\theta(Y, Z) \mid Y]$.

Solution. Puisque les couples (Y_i, Z_i) sont iid, on a

$$\mu_i^h := \mathbb{E}_{\theta^h}(Z_i \mid Y_i) = (a^h + Y_i)/(a + \lambda^h),$$
$$\nu_i^h := \mathbb{E}_{\theta^h}(\log Z_i \mid Y_i) = \psi(a^h + Y_i) - \log(a^h + Y_i)$$

d'après les formules des moments conditionnels établis à la question 3.

7. Étape M : donner une formule de mise à jour explicite pour λ^{h+1} et une équation vérifiée par a^{h+1} et qu'on sait résoudre.

Solution. En notant $\overline{Y}$ la moyenne des Y_i (et respectivement $\overline{\mu}^h$ et $\overline{\nu}^h$), on a

$$\mathbb{E}_{\theta^h}[\log p_\theta(Y, Z) \mid Y] = n\left(a \log a - \log \Gamma(a) + \overline{\nu}^h a + \overline{Y} \log \lambda - \overline{\mu}^h(a + \lambda)\right) + \text{cst}$$

dont la dérivé par rapport à λ vaut $n(\overline{Y}/\lambda - \overline{\mu}^h)$, qui s'annule pour

$$\lambda^{h+1} = \overline{Y}/\overline{\mu}^h,$$

et dont la dérivée par rapport à a :

$$n\left(1 + \log a - \psi(a) + \overline{\nu}^h - \overline{\mu}^h\right)$$

s'annule pour la solution a^h de l'équation

$$\psi(a) - \log(a) = \overline{\nu}^h - \overline{\mu}^h + 1.$$

D'après la proposition 1, cette équation admet une unique solution si $\overline{\nu}^h - \overline{\mu}^h + 1 < 0$, ce qui est vrai en remarquant qu'on a, terme à terme,

$$\nu_i + 1 - \mu_i = \mathbb{E}(\log Z_i \mid Y) + 1 - \mu_i \leq \log \mu_i + 1 - \mu_i \leq 0$$

(puisque $\log(x) \leq x - 1$ avec égalité seulement pour $x = 1$) et qu'on a donc l'inégalité stricte $\overline{\nu}^h < \overline{\mu}^h - 1$ dès que les Y_i ne sont pas tous égaux entre eux.

2 Communautés de poissons de la rivière Fatala

Cet exercice porte principalement sur l'interprétation de résultats d'analyses. Les réponses doivent être argumentées tout en restant synthétiques.

On s'intéresse à la distribution de $p = 13$ espèces de poisson le long du cours de la rivière Fatala (Guinée-Conakry). On considère quatre stations situées respectivement à 3, 17, 33 et 46 km de l'embouchure de la rivière. Dans chaque station, trois ou quatre pêches ont été effectuées indépendamment, selon le même protocole en six dates : avril, juin, août, octobre, décembre 1993 et février 1994. Baran [1995] a ainsi collecté $n = 95$ vecteurs de comptages $Y_i = (Y_{ij})_{1 \leq j \leq p}$ (pour $1 \leq i \leq n$) où Y_{ij} désigne le nombre d'individus de l'espèce j observés lors de la pêche i.

L'objectif est d'analyser les effets environnementaux (résumés par la distance à l'embouchure et la date) sur les populations des différentes espèces, ainsi que les interactions qu'elles entretiennent entre elles. Les analyses sont menées dans le cadre du modèle Poisson log-normal.

Classification des pêches. On souhaite tout d'abord effectuer une classification non-supervisée des pêches prenant éventuellement en compte un vecteur de descripteurs $x_i \in \mathbb{R}^d$ pour chaque pêche. On pose pour cela le modèle de mélange suivant

$$\begin{aligned} (C_i)_{1\le i\le n} \text{ iid :} &\qquad C_i \sim \mathcal{M}_K(0,\pi), \\ (Z_i)_{1\le i\le n} \text{ iid :} &\qquad Z_i \sim \mathcal{N}_p(0,\Sigma), \\ (Y_{ij})_{1\le i\le n, 1\le j\le p} \text{ indépendants} \mid (Z_i),(C_i) : &\qquad (Y_{ij} \mid Z_{ij}, C_i = k) \sim \mathcal{P}(\exp(\mu_{kj} + x_i^\top \beta + Z_{ij})) \end{aligned} \tag{2}$$

où

— C_i désigne le groupe (non-observé) auquel appartient la pêche i,
— Z_i le vecteur gaussien latent associé à la pêche i,
— x_i le vecteur des descripteurs pour la pêche i et
— Y_{ij} l'effectif de l'espèce j observé lors de cette même pêche.

Les paramètres de ce modèle sont : le vecteur de proportions π, la matrice de covariance Σ, l'ensemble des coefficients $(\mu_{kj})_{1\le k\le K, 1\le j\le p}$ réunis dans la matrice μ de dimension $K \times p$ et le vecteur des coefficients de régression $\beta \in \mathbb{R}^d$, soit $\theta = (\pi, \Sigma, \mu, \beta)$.

On a considéré cinq modèles incluant respectivement comme descripteurs : aucun descripteur (noté 'cst', $d = 1$), la distance ('distance', $d = 4$), la date ('date', $d = 6$), la distance et la date ('distance+date', $d = 9$) et enfin la distance, la date et leur effet croisé ('distance*date', $d = 24$).

On choisit de déterminer le nombre de groupes K au moyen du critère BIC.

1. *Déterminer le nombre de paramètres D_K intervenant dans la pénalité BIC pour un modèle à K groupes incluant un vecteur de descripteurs x_i de dimension d.*

> *Solution.* $D_K = (K-1) + p(p+1)/2 + Kp + d$.

Pour chacun des cinq modèles considérés on donne le nombre de groupes $\widehat{K}$ choisi par BIC et le critère $BIC_{\widehat{K}}$ correspondant :

	cst	distance	date	distance+date	distance*date
$\widehat{K}$	6	4	3	2	2
$BIC_{\widehat{K}}$	-2419	-2340	-2424	-2338	-2506

Les tableaux suivants donnent de plus la répartition des pêches dans les six groupes obtenus avec le modèle 'cst' avec leur distance d'une part et leur date d'autre part (les tirets correspondent à des effectifs nuls) :

	distance			
k	km03	km17	km33	km46
1	-	1	6	20
2	-	4	11	-
3	3	-	-	-
4	17	1	-	-
5	-	12	7	4
6	3	6	-	-

	date					
k	apr93	jun93	aug93	oct93	dec93	feb94
1	-	5	6	8	4	4
2	1	4	6	-	4	-
3	-	3	-	-	-	-
4	4	1	4	4	1	4
5	10	2	-	3	-	8
6	1	-	-	1	7	-

Les tableaux suivants donnent la même répartition pour les deux groupes obtenus avec le modèle 'distance*date' :

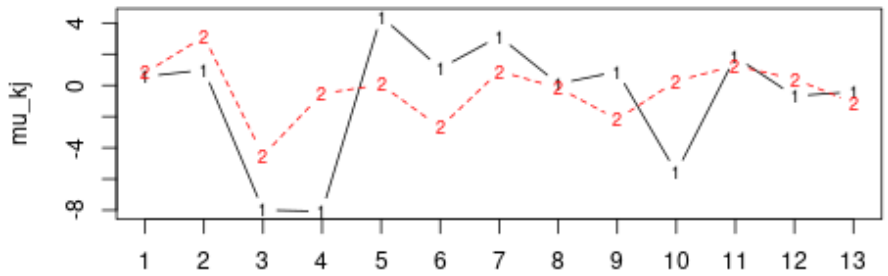
	distance			
k	km03	km17	km33	km46
1	8	13	9	11
2	15	11	15	13

	date					
k	apr93	jun93	aug93	oct93	dec93	feb94
1	6	7	8	6	8	6
2	10	8	8	10	8	10

2. *Comment expliquez-vous la variation du nombre estimé de groupes entre les différents modèles ? Comparer la composition des groupes issus des modèles 'cst' et 'distance*date'*

> *Solution.* Le mélange modélise une hétérogénétité non expliquée par la partie régression du modèle. De façon attendue, l'introduction des covariables réduit le nombre de groupes nécessaires. On peut remarque qu'il reste une part d'hétérogénéité ($K = 2$) après introduction de l'ensemble des covariables disponibles ('distance+date' ou 'distance*date').
> Dans le modèle 'cst', les pêches effectuées à une même distance sont majoritairement classées dans un même groupe : les groupes reconstituent en grande partie la structure géographique. L'effet existe aussi pour les dates (voir janvier 1993), mais est moins marqué. La classification du modèle 'cst' reconstitue pour partie la structure spatio-temporelle.
> Celle du modèle le plus riche ('distance*date') est indépendante de la structure statio-temporelle et révèle une hétérogénéité due à un effet non décrit par les covariables disponibles.

La figure suivante représente les coordonnées des deux coefficients $\widehat{\mu}_{kj}$ obtenus avec le modèle 'distance*date' (noir : $k = 1$, rouge : $k = 2$) associées à chacune des $p = 13$ espèces (en abscisse).



3. *Comment s'interprète le coefficient μ_{kj} ?*
Que peut-on dire de l'abondance de l'espèce 10 dans les pêches du groupe 1 par rapport à celles du groupe 2.

> *Solution.* La coordonnée μ_{kj} contrôle l'abondance moyenne de l'espèce j dans une pêche du groupe k. Une valeur fortement négative de μ_{kj} indique une espèce j quasiment absente dans les pêches du groupe k (exemple : espèces 3 et 4 dans le groupe 1).
> L'espèce 10 est plus abondante dans les pêches du groupe 2 que du groupe 1 et cet écart n'est pas dû aux effets spatio-temporels pris en compte par le vecteurs x_i.

4. *Quel modèle choisiriez-vous finalement ?*

> *Solution.* Le modèle préférable au vu du critère BIC est le modèle 'distance-date' avec $K = 2$ groupes.

Interactions entre espèces. On cherche maintenant à déterminer les interactions directes entre espèces (i.e. ne résultant ni des effets environnementaux, ni des interactions avec d'autres espèces). On considère pour cela le modèle Poisson log-normal

$$\begin{aligned} (Z_i)_{1\leq i\leq n} \text{ iid :} \qquad & Z_i \sim \mathcal{N}_p(0, \Sigma), \\ (Y_{ij})_{1\leq i\leq n, 1\leq j\leq p} \text{ indépendants } | \ (Z_i) : \qquad & (Y_{ij} \mid Z_{ij}) \sim \mathcal{P}(\exp(x_i^\top \beta + Z_{ij})), \end{aligned}$$

dans lequel on suppose que la matrice de précision $\Omega = \Sigma^{-1}$ est creuse. Le réseau d'interactions entre les espèces est alors défini, selon les propriétés des modèles graphiques gaussiens, comme le support de la matrice $\Omega = [\omega_{jk}]_{1\leq j,k\leq p}$:

$$\{j \text{ et } k \text{ en interaction directe}\} \qquad \Leftrightarrow \qquad \omega_{jk} \neq 0.$$

Pour estimer le paramètre $\theta = (\beta, \Sigma)$, en forçant Ω à être creuse, on maximise en θ et q, le critère régularisé

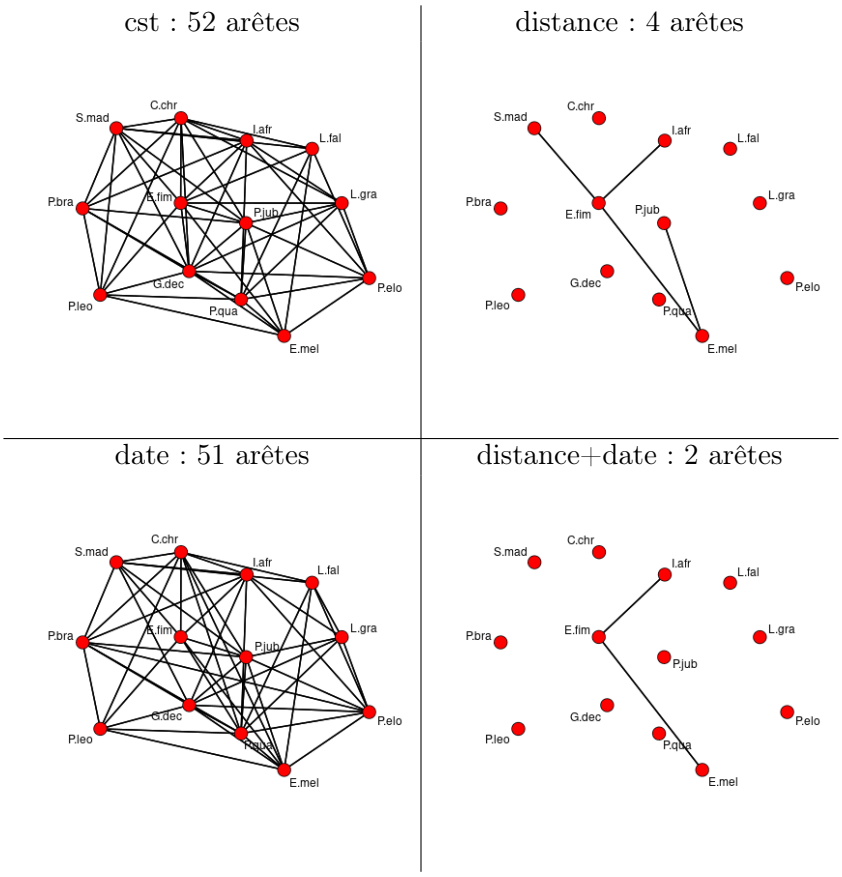
$$\mathcal{J}_Y(\theta, q) - \lambda \sum_{1 \leq j < k \leq p} |\omega_{jk}| \tag{3}$$

où q est une distribution pour Z et $\mathcal{J}_Y(\theta, q)$ est la borne inférieure de la log-vraisemblance $\log p_\theta(Y)$ utilisée dans l'algorithme EM variationnel. Le paramètre de régularisation λ est choisi au moyen d'un critère de type BIC.

5. *Rappeler la borne inférieure $\mathcal{J}_Y(\theta, q)$ pour le modèle Poisson log-normal.*
   Dans quelle classe $\mathcal{Q}$ la distribution q est elle choisie ?

> *Solution.* La vraisemblance $\log p_\theta(Y)$ n'étant pas calculable, on maximise la borne inférieure
>
> $$\mathcal{J}_Y(\theta, q) = \log p_\theta(Y) - KL[q(Z) \| p_\theta(Z \mid Y)]$$
>
> où $q(Z)$ est de la forme $q(Z) = \prod_{1 \leq i \leq n} q_i(Z_i)$ et chaque q_i est une loi normale de dimension p : $q_i = \mathcal{N}_p(m_i, S_i)$.

On a utilisé le critère pénalisé (3) pour le quatre premiers modèles vus à la questions 2 : 'cst', 'distance', 'date' et 'distance+date'. La figure suivante donne les réseaux estimés avec chacun d'eux, ainsi que le nombre d'arêtes (interactions) qu'ils contiennent. Les noeuds sont annotés par les noms abrégés des espèces.



6. *Comparer les réseaux obtenus pour les différents modèle et commenter l'effet des covariables sur la structure du réseau estimé.*
   Quel est l'effet environnemental principal ?

Solution. Comment attendu, l'introduction de covariables rend systématiquement plus creux le réseaux estimé. En notant A le nombre d'arêtes, on a

$$A(\text{cst}) > A(\text{date}) > A(\text{distance}) > A(\text{distance+date}).$$

La non-prise en compte de l'environnement fait apparaître comme interaction le résultat des effets abiotiques sur les différentes espèces.
L'introduction de l'effet temporel (date) n'a qu'un effet modéré sur la structure du réseau alors que l'introduction de l'effet spatial (distance) réduit drastiquement le nombre d'arêtes : l'effet principal est clairement l'effet distance (ce qui est cohérent avec les valeurs des critères BIC donnés à la question 2).
Le réseau obtenu avec le modèle de plus complet ne tient pourtant pas compte de l'hérérogénéité résiduelle détectée par le modèle de mélange : les interactions inférées peuvent en fait n'être pas directes, mais résulter de l'effet d'un facteur environnemental non pris en compte.

Références

E. Baran. *Dynamique spatio-temporelle des peuplements de poissons estuariens en Guinée (Afrique de l'Ouest)*. PhD thesis, Thèse de Doctorat, Université de Bretagne Occidentale, 1995.