Modèles à variables latentes pour l'écologie et la biologie

Examen 2025

Certaines études de biodiversité consistent à relever la présence d'espèces d'intérêt sur un grand nombre de site géographiques. Plus précisément, pour n sites géographiques, on relève la présence ou l'absence de p espèces (déterminées à l'avance). Les données se présentent sous la forme d'une table $\boldsymbol{y}$ de taille $n \times p$ telle que $y_{ij} = 1$ si l'espèce j a été observée sur le site i, $y_{ij} = 0$ sinon.

	Espèce 1	$\cdots$	Espèce j	$\cdots$	Espèce p
Site 1	1	$\cdots$	0	$\cdots$	1
$\vdots$					
Site i	1	$\cdots$	1	$\cdots$	0
$\vdots$					
Site n	0	$\cdots$	1	$\cdots$	1

Table 1: Table de présence-abscence

On suppose que $\boldsymbol{y}$ est la réalisation d'un modèle probabiliste. Vous allez devoir considérer plusieurs modèles pour analyser ces données et écrire leurs méthodes d'inférence. On suppose que les n sites sont indépendants.

Partie 1. Modèle de mélange

La première possibilité est de considérer les espèces indépendantes. Ainsi $\forall j = 1, \dots, p$ et $\forall i = 1, \dots, n$:

$$Y_{ij} \sim \mathcal{B}\text{ern}(\alpha_j)$$

Cependant, l'hypothèse que les espèces ont la même fréquence sur tous les sites est peu réaliste. On propose donc un modèle de mélange sur les sites. Supposons que les sites soient répartis dans K classes. Le modèle de mélange s'écrit de la façon suivante:

$$\begin{aligned} Z_i &\sim_{i.i.d.} \mathcal{C}\text{at}(\pi_1, \dots, \pi_K) \\ Y_{ij} \mid Z_i = k &\sim_{i.i.d.} \mathcal{B}\text{ern}(\alpha_{kj}) \end{aligned} \tag{1}$$

où $\mathcal{C}$at est la loi catégorielle: $\mathbb{P}(Z_i = k) = \pi_k \in [0, 1]$, $\sum_{k=1}^{K} \pi_k = 1$. $\mathcal{B}$ern est la loi de Bernoulli.

1. Ecrire le DAG du modèle.
2. Ecrire la densité marginale d'un vecteur ligne $(y_1, \dots, y_p)$. En déduire la log-vraisemblance des observations $\log p_\theta(\boldsymbol{y})$ pour θ le vecteur de paramètres à spécifier.

$$f_\theta(y_1,\dots,y_p) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{p} \alpha_{kj}^{y_j}(1-\alpha_{kj})^{1-y_j}$$

Donc:

$$\log p_\theta(\boldsymbol{y}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \prod_{j=1}^{p} \alpha_{kj}^{y_{ij}}(1-\alpha_{kj})^{1-y_{ij}}$$

3. Après avoir spécifié le vecteur des variables latentes $\boldsymbol{Z}$, écrire la log-vraisemblance complète $\log p_\theta(\boldsymbol{y}, \boldsymbol{Z})$

$$\begin{aligned}
\log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) &= \log p_\theta(\boldsymbol{y} \mid \boldsymbol{Z}) + \log p_\theta(\boldsymbol{Z}) \\
&= \sum_{i=1}^{n} \log \prod_{j=1}^{p} \alpha_{Z_i j}^{y_{ij}}(1-\alpha_{Z_i j})^{1-y_{ij}} + \sum_{i=1}^{n} \log \pi_{Z_i} \\
&= \sum_{i,j=1}^{n,p} \sum_{k=1}^{K} Z_{ik} \left[y_{ij} \log \alpha_{kj} + (1-y_{ij}) \log(1-\alpha_{kj}) \right] + \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \log \pi_k
\end{aligned}$$

avec $Z_{ik} = \mathbb{1}_{\{Z_i = k\}}$.

4. Déterminer la loi conditionnelle des variables latentes $\mathbb{P}(\boldsymbol{Z} \mid \boldsymbol{y})$. On veillera à bien justifier les indépendances entre variables aléatoires.

$$\begin{aligned}
\mathbb{P}(\boldsymbol{Z} = \boldsymbol{z} \mid \boldsymbol{y}) &= \prod_{i=1}^{n} \mathbb{P}(Z_i = z_i \mid y_{i1},\dots,y_{ip}) \\
\tau_{ik} &= \mathbb{P}(Z_i = k \mid y_{i1},\dots,y_{ip}) \\
&= \frac{\mathbb{P}(y_{i1},\dots,y_{ip} \mid Z_i = k)\mathbb{P}(Z_i = k)}{f_\theta(y_{i1},\dots,y_{ip})} \\
&= \frac{\pi_k \prod_{j=1}^{p} \alpha_{kj}^{y_{ij}}(1-\alpha_{kj})^{1-y_{ij}}}{\sum_{\ell=1}^{K} \pi_\ell \prod_{j=1}^{p} \alpha_{\ell j}^{y_{ij}}(1-\alpha_{\ell j})^{1-y_{ij}}}
\end{aligned}$$

5. Ecrire l'étape Expectation de l'algorithme EM pour un paramètre courant $\theta^{(h)}$

$$\begin{aligned}
Q(\theta \mid \theta^{(h)}) &= \mathbb{E}_{\theta^{(h)}} \left[\log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) \mid \boldsymbol{y} \right] \\
&= \sum_{i,j=1}^{n,p} \sum_{k=1}^{K} \tau_{ik}^{(h)} \left[y_{ij} \log \alpha_{kj} + (1-y_{ij}) \log(1-\alpha_{kj}) \right] \\
&\quad + \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(h)} \log \pi_k
\end{aligned}$$

6. Ecrire l'étape Maximisation de l'algorihme EM permettant d'obtenir $\theta^{(h+1)}$

$$\begin{aligned}\frac{\partial}{\partial \alpha_{kj}} Q(\theta \mid \theta^{(h)}) &= \sum_{i=1}^{n} \tau_{ik}^{(h)} \left[y_{ij} \frac{1}{\alpha_{kj}} - (1 - y_{ij}) \frac{1}{1 - \alpha_{kj}} \right] = 0 \\ &= \frac{S_{kj}}{\alpha_{kj}} - \frac{N_k - S_{kj}}{1 - \alpha_{kj}} = 0\end{aligned}$$

avec $S_{kj} = \sum_{i=1}^{n} \tau_{ik}^{(h)} y_{ij}$ et $N_k = \sum_{i=1}^{n} \tau_{ik}^{(h)}$.

$$\alpha_{kj}^{(h+1)} = \frac{S_{kj}}{N_k}.$$

Pour π_k sous la contrainte $\sum_{k=1}^{K} \pi_k = 1$, cf cours:

$$\pi_k^{(h+1)} = \frac{N_k}{n}$$

7. Proposer une façon d'initialiser l'algorithme EM.

CAH sur les données brutes

8. Proposer des critères pour choisir le nombre de clusters K.

Nb parameters $= p \times K + (K - 1)$

9. Proposer une méthode de clustering des sites.

$$\widehat{Z}_i = \arg\max_k \tau_{ik}$$

10. Proposer un modèle permettant d'obtenir un bi-clustering des sites ET des espèces. Comment peut-on choisir entre ce dernier modèle et le modèle (1).

LBM : comparaison des ICL

Partie 2. Modèle Bernoulli Log Normal

Le modèle de mélange permet de tenir compte d'une hétérogénéité entre sites. Supposons qu'on dispose de covariables environnementales permettant de décrire ces sites: $\forall i = 1, \ldots, n, x_i = (x_i^1, \ldots, x_i^d) \in \mathbb{R}^d$ avec $x_i^1 = 1$ pour tout i. On propose d'utiliser ces covariables pour modéliser une hétérogénité de présence de chaque espèce entre espèces. Le modèle de régression logistique classique spécifierait que:

$$\mathbb{P}(Y_{ij} = 1) = \frac{1}{1 + e^{-x_i^\top \alpha_j}} \quad \text{où} \quad x_i^\top \alpha_j = \sum_{l=1}^{d} x_i^l \alpha_j^l. \tag{2}$$

où la fonction de lien $x \mapsto \frac{1}{1+e^{-x}}$ est une bijection croissante de $\mathbb{R}$ dans $[0, 1]$. Ainsi la probabilité de présence d'une espèce j est fonction d'une combinaison linéaire des covariables environnementales dont les paramètres $\alpha_j = (\alpha_j^1, \ldots, \alpha_j^d)$ sont propres à l'espèce j. Selon le même principe que le

modèle Poisson Log Normal vu en cours, on propose d'introduire une dépendance entre espèces grâce à un effet aléatoire Z_{ij}. On note $Z_i = (Z_{i1}, \dots, Z_{ip})$

11. Ecrire ce modèle. Lister les paramètres à estimer ainsi que leur dimension.

$$\begin{aligned} Z_i &\sim_{i.i.d} \mathcal{N}_p(0, \Sigma) \\ Y_{ij}|Z_{ij} &\sim \mathcal{B}\text{ern}\left(\frac{1}{1+e^{-x_i^\top \alpha_j - Z_{ij}}}\right) \end{aligned} \tag{3}$$

12. Quel paramètre encode la dépendance entre espèces? Pour quelle valeur de ce paramètre les espèces sont-elles supposées indépendantes? Est-ce qu'on retombe sur le modèle classique de régression logistique (2).

13. Donner une expression de la densité d'une ligne du tableau $f_\theta(y_i)$ sous forme intégrale. A-t-elle une forme explicite?

$$\begin{aligned} p_\theta(y_i, Z_i) &= \prod_{j=1}^{p} (1 + e^{-x_i^\top \alpha_j - Z_{ij}})^{-y_{ij}} (1 + e^{+x_i^\top \alpha_j + Z_{ij}})^{-y_{ij}+1} \\ p_\theta(Z_i) &= \frac{1}{|\Sigma|^{1/2}\sqrt{2\pi}^d} e^{-\frac{1}{2} Z_i^\top \Sigma^{-1} Z_i} \\ p_\theta(y_i) &= \int_{\mathbb{R}^d} \prod_{j=1}^{p} (1 + e^{-x_i^\top \alpha_j - Z_{ij}})^{-y_{ij}} (1 + e^{+x_i^\top \alpha_j + Z_{ij}})^{-y_{ij}+1} \frac{1}{|\Sigma|^{1/2}\sqrt{2\pi}^d} e^{-\frac{1}{2} Z_i^\top \Sigma^{-1} Z_i} dZ_i \end{aligned}$$

Pas d'expressison explicite

14. Ecrire la log-vraisemblance complète $p_\theta(\boldsymbol{y}, \boldsymbol{Z})$.

$$\begin{aligned} \log p_\theta(\boldsymbol{y}, \boldsymbol{Z}) &= \sum_{i=1}^{n} \sum_{j=1}^{p} -y_{ij} \log(1 + e^{-x_i^\top \alpha_j - Z_{ij}}) + (1 - y_{ij}) \log(1 + e^{+x_i^\top \alpha_j + Z_{ij}}) \\ &\quad - \frac{np}{2} \log(2\pi) - \frac{n}{2} \log|\Sigma| - \sum_{i=1}^{n} \frac{1}{2} Z_i^\top \Sigma^{-1} Z_i \end{aligned}$$

15. L'EM n'est pas envisageable. Pourquoi?

16. Vers quel algorithme vous-tournez vous? Que proposez-vous comme approximation de $\boldsymbol{Z} \mid \boldsymbol{y}$?

17. Sous cette approximation, quelle est la fonction de coût que vous allez optimiser? Pouvez vous en obtenir une expression explicite sans intégrale?

18. En utilisant le cours sur les autoencodeurs, proposez une solution pratique et fournissez une version grossière de votre algorithme d'estimation.