

## 4 Estimation de la richesse spécifique par modèle de mélange

### Problème

Une mesure simple de la diversité d'un milieu est sa richesse spécifique, c'est-à-dire le nombre  $n$  d'espèces différentes vivant dans ce milieu. Une difficulté de la mesure de cette richesse vient du fait que, lors de l'échantillonnage, il est possible qu'aucun individu d'une espèce pourtant présente ne soit effectivement observé.

On considère une expérience d'écologie microbienne lors de laquelle on a dénombré les individus appartenant à  $m = 514$  espèces bactériennes ( $1 \leq i \leq m$ ). On note  $y_i$  le nombre d'individus de l'espèce  $i$  qui ont été observés. Le tableau `microbial.csv` contient, pour chaque effectif  $j$  compris entre 1 et 50, le nombre d'espèces  $n_j$  pour lesquelles  $j$  individus ont été observés :

$$n_j = \sum_{i=1}^m \mathbb{I}\{y_i = j\}.$$

**Objectif.** Il s'agit ici d'estimer la richesse spécifique de ce milieu en estimant le nombre  $n_0$  d'espèces bactériennes présentes mais n'ayant pas été observées ( $y_i = 0$ ). On se propose pour cela de supposer

— que les abondances  $X_i$  des espèces  $n$  suivent le modèle de mélange de lois géométriques *non tronquées* :

$$(X_i)_{1 \leq i \leq n} \text{ iid} : X_i \sim \sum_{k=1}^K \eta_k \mathcal{G}(\gamma_k).$$

— et qu'on observe l'abondance  $Y_i = X_i$  si  $X_i > 0$  et rien sinon.

On note  $m = \sum_{i=1}^n \mathbb{I}\{X_i > 0\}$  le nombre d'espèces effectivement observées. Dans la suite de l'exercice, on travaillera conditionnellement à ce nombre.

### 4.1 Mélange de lois géométriques tronquées

**Notation.** Dans toute la suite, on définit la loi géométrique  $\mathcal{G}(\gamma)$  de la façon suivante :

$$Y \sim \mathcal{G}(\gamma) \quad \Leftrightarrow \quad p(y; \gamma) := \mathbb{P}_\gamma\{Y = y\} = \gamma(1 - \gamma)^y \quad \text{pour } y \geq 0.$$

On note de plus  $\mathcal{G}^+(\gamma)$  la distribution géométrique de paramètre  $\gamma$  tronquée en 0 et  $p^+$  sa distribution :

$$p^+(y; \gamma) := \mathbb{P}_\gamma\{Y = y \mid Y > 0\} \quad \text{pour } Y \sim \mathcal{G}(\gamma) \text{ et } y > 0.$$

**Modèle de mélange.** On considère ici le modèle de mélange à  $K$  classes suivant

$$\begin{aligned} (Z_i)_{1 \leq i \leq m} \text{ iid} : & Z_i \sim \mathcal{M}(1, \pi), \\ (Y_i)_{1 \leq i \leq m} \text{ indep.} \mid (Z_i) : & (Y_i \mid Z_i = k) \sim \mathcal{G}^+(1, \gamma_k) \end{aligned}$$

où  $\pi \in [0, 1]^K$ ,  $\sum_{k=1}^K \pi_k = 1$  et  $\gamma \in [0, 1]^K$ . Le paramètre de ce modèle est donc  $\theta = (\pi, \gamma)$ . Dans la suite, on notera de façon compacte

$$(Y_i)_{1 \leq i \leq m} \text{ iid} : Y_i \sim \sum_{k=1}^K \pi_k \mathcal{G}^+(\gamma_k). \quad (9)$$

### Rappels.

- Dans un modèle de mélange, l'entropie conditionnelle vaut  $\mathcal{H}_\theta(Z \mid Y) = -\sum_{i=1}^m \sum_{k=1}^K \tau_{ik} \log \tau_{ik}$  où  $\tau_{ik} = \mathbb{P}_\theta\{Z_i = k \mid Y_i\}$ .
- On a de plus :  $\log p_\theta(Y) = \mathbb{E}_\theta[\log p_\theta(Y, Z) \mid Y] + \mathcal{H}_\theta(Z \mid Y)$ .

### Questions.

1. Calculer la probabilité  $p(0; \gamma)$  et en déduire que, pour  $y > 0$ ,  $p^+(y; \gamma) = \gamma(1 - \gamma)^{y-1}$ .
2. Écrire la vraisemblance complète  $\log p_\theta(Y, Z)$  du modèle (9) en fonction de  $\theta$ .  
En déduire son espérance conditionnelle à  $Y$ .
3. Écrire l'étape E de l'algorithme EM visant à obtenir l'estimateur du maximum de vraisemblance de  $\theta$ .
4. Écrire l'étape M de ce même algorithme.

5. Implémenter cet algorithme EM en R.

A titre de vérification, on pourra utiliser les données contenues dans le fichier `simul-n1000-K3-seed1.csv` qui ont été simulées avec les paramètres suivants :

$$K = 3, \quad m = 1000, \quad \pi = (1/6, 1/3, 1/2), \quad \gamma = (1/2, 1/9, 1/28)$$

et pour lesquelles les estimateurs du maximum de vraisemblance sont

$$\hat{\pi} = (0.22, 0.38, 0.40), \quad \hat{\gamma} = (0.44, 0.08, 0.03).$$

6. Dénombrer les paramètres indépendants  $d_K$  pour un modèle (9) à  $K$  classes.

7. En déduire le critère  $BIC$

$$BIC(K) = \log p_{\hat{\theta}}(Y) - d_K \log(m)/2.$$

## 4.2 Estimation du nombre d'espèces

### Questions.

1. Montrer que, conditionnellement à  $m$ , les abondances observées  $Y_i$  suivent le modèle de mélange (9) avec

$$\pi_k = \frac{\eta_k(1 - \gamma_k)}{\sum_{\ell=1}^K \eta_{\ell}(1 - \gamma_{\ell})}.$$

2. Importer le fichier `microbial.csv` et construire le vecteur des abondances observées  $y = (y_i)_{1 \leq i \leq m}$  (de longueur  $m$ ) à partir de son contenu.

A titre de vérification, le nombre moyen d'individus parmi les espèces observées vaut  $\bar{y} = 1.91$ .

3. Estimer les paramètres du modèle (9) pour le vecteur  $y$  au moyen de l'algorithme EM précédemment implémenté pour  $K$  allant de 1 à 5.

4. Choisir le nombre de classes  $K$  au moyen du critère  $BIC$ .

5. Donner une estimation de  $p(0; \gamma_k)$  pour  $1 \leq k \leq K$ .

6. Donner une estimation des proportions  $\eta_k$  pour  $1 \leq k \leq K$ .

7. En déduire une estimation de la probabilité de ne pas être observée pour une espèce quelconque.

8. Proposer un estimateur du nombre  $n_0$  d'espèces présentes mais non observées et en déduire une estimation de la richesse spécifique  $n$ .

**Questions subsidiaires.** L'exercice précédent fournit de fait un estimateur  $\hat{n}_K$  pour chaque valeur de  $K$ . On rappelle que le critère BIC approche, à une constante indépendante de  $K$  près  $\log \mathbb{P}\{K \mid Y\}$ .

16. Proposer un estimateur  $\hat{n}$  de  $n$  combinant les différents estimateurs  $\hat{n}_K$ .

17. En déduire une nouvelle estimation de la richesse spécifique  $n$ .