3 Structure d'un réseau trophique

Problème

On veut analyser la structure d'un réseau d'interactions *asymétriques* entre espèces, plus précisément de relations trophiques. On considère pour cela n espèces et, pour chaque couple $1 \leq i, j \leq n$, on note

$$Y_{ij} = \begin{cases} 1 & \text{si l'espèce } i \text{ est prédatrice de l'espèce } j, \\ 0 & \text{sinon.} \end{cases}$$

(On supposera ici qu'une espèce n'est pas prédatrice d'elle même, ce qui n'est souvent pas vrai.)

Kéfi et al. [2016] ont recueilli un tel réseau dans la zone intertidale (alternativement couverte et découverte par la marée) de la côte rocheuse du Chili. Le réseau porte sur $n = 106$ espèces et est disponible dans le fichier `chilean_TI.csv` (accessible sur le moodle du cours)

Objectif. On cherche à distinguer des catégories d'espèces jouant des rôles différents dans le réseau, correspondants à différent niveaux trophiques.

3.1 Modèle à blocs stochastiques asymétrique

On se propose d'utiliser le modèle à blocs stochastiques à K groupes suivant

$$\begin{aligned} Z = \{Z_t\}_{1\leq t\leq n} &\sim \mathcal{M}_K(\pi), \\ \{Y_{ij}\}_{1\leq i,j\leq n} \text{ indépendants } \mid Z : \qquad (Y_{ij} \mid Z_i = k, Z_j = \ell) &\sim \mathcal{B}(\gamma_{k\ell}) \end{aligned} \tag{6}$$

où π désigne le vecteur des probabilités d'appartenance à chaque groupe et $\gamma_{k\ell}$ la probabilité qu'une espèce du groupe k interagisse avec une espèce du groupe ℓ. Les paramètres du modèle à K états sont donc

$$\theta = (\pi = (\pi_k)_{1\leq k\leq K}, \gamma = (\gamma_{k\ell})_{1\leq k,\ell\leq K}).$$

Dans le cas d'un réseau trophique, "interagir" signifie "être un prédateur de". La matrice $Y = [Y_{ij}]_{1\leq i,j\leq n}$ (le "réseau") est donc carrée, mais pas symétrique.

Estimation des paramètres.

1. Écrire la vraisemblance complète de ce modèle.
2. En déduire son espérance conditionnelle aux données observées pour une valeur courante du paramètre notée $\theta^{(h)}$. On notera $\tau_{ik}^{(h)} = \mathbb{P}_{\theta^{(h)}}\{Z_i = k \mid Y\}$ et $\eta_{ijk\ell}^{(h)} = \mathbb{P}_{\theta^{(h)}}\{Z_i = k, Z_j = \ell \mid Y\}$.
3. En supposant les quantités $\tau_{ik}^{(h)}$ et $\eta_{ijk\ell}^{(h)}$ connues, en déduire la valeur $\theta^{(h+1)}$ qui maximise $\mathbb{E}_{\theta^{(h)}}(\log p_\theta(Z, Y) \mid Y)$ en θ.

Approximation variationnelle.

La loi conditionnelle $p_\theta(Z \mid Y)$ n'étant pas calculable, on choisit de l'approcher par un loi factorisable, c'est à dire par la loi $\widetilde{q}(Z)$ appartenant à la classe

$$\mathcal{Q} = \left\{ q : q(Z) = \prod_{i=1}^{n} q_i(Z_i) \right\}$$

et qui minimise la distance de Küllback-Leibler $KL(q(Z)\|p_\theta(Z \mid Y))$. Dans la suite on marque par un "tilde" $(\widetilde{\cdot})$ les probabilités calculées sous cette loi, notamment :

$$\widetilde{\tau}_{ik} = \mathbb{P}_{\widetilde{q}}\{Z_i = k\}, \qquad \widetilde{\eta}_{ijk\ell} = \mathbb{P}_{\widetilde{q}}\{Z_i = k, Z_j = \ell\} = \widetilde{\tau}_{ik}\widetilde{\tau}_{j\ell}.$$

Il s'agit dès lors d'estimer le paramètre θ en maximisant la borne inférieure de la log-vraisemblance (ELBO)

$$ELBO(Y, \theta, q) = \log p_\theta(Y) - KL(q(Z)\|p_\theta(Z \mid Y)) = \mathbb{E}_q[\log p_\theta(Y, Z)] + \mathcal{H}(q(Z))$$

où $\mathcal{H}$ désigne l'entropie : $\mathcal{H}(q(Z)) = -\mathbb{E}_q[\log q(Z)]$.

Question.

4. Déterminer l'équation de point fixe satisfaite par les $(\widetilde{\tau}_{ik}^{(h)})_{1\leq i\leq n, 1\leq k\leq K}$ pour une valeur $\theta^{(h)}$ du paramètre et en déduire la loi approchée $\widetilde{q}_i^{(h)}$.

3.2 Implémentation de l'algorithme VEM

1. Écrire une fonction `Mstep` prenant en arguments les données Y et la valeur courante des probabilités conditionnelles approchées $\widetilde{\tau}^{(h)} = (\widetilde{\tau}_{ik}^{(h)})_{1 \leq i \leq n, 1 \leq k \leq K}$ et qui retourne les estimations obtenues à la question 3.
2. Écrire une fonction `VEstep` prenant en arguments les données Y et la valeur courante du paramètre $\theta^{(h)}$ et qui retourne, au moyen de l'équation établie à la question 4,
   — la matrice des probabilités conditionnelles conditionnelles approchées $\widetilde{\tau}^{(h)}$ et
   — les log-densités estimées $\log \phi_{ijk\ell}^{(h)} = \left(Y_{ij} \log \gamma_{k\ell}^{(h)} + (1 - Y_{ij}) \log(1 - \gamma_{k\ell}^{(h)}) \right)$.
3. Écrire une fonction `ELBO` prenant en arguments les données Y, la valeur courante du paramètre $\theta^{(h)}$, les probabilités conditionnelles approchées $\widetilde{\tau}_{ik}^{(h)}$ et les log-densités estimées $\log \phi_{ijk\ell}^{(h)}$ et qui retourne la borne inférieur $ELBO(Y, \theta^{(h)}, \widetilde{q}^{(h)})$.
4. A partir de méthodes que vous connaissez, proposer une initialisation des probabilités conditionnelles $\tau_{ik}^{(0)}$. Écrire une fonction `InitSBM` prenant en arguments les données Y et le nombre d'états K et qui retourne ces valeurs.
5. Écrire une fonction `SBM` prenant en arguments les données Y et le nombre d'états K et utilisant l'algorithme VEM et qui retourne
   — l'estimation par maximum de vraisemblance approché $\widehat{\theta}$ de θ,
   — les probabilités conditionnelles approchées $\widehat{\tau}_{ik}$ et
   — la borne inférieur de la log-vraisemblance $ELBO(Y, \widehat{\theta}, \widetilde{q})$.

3.3 Application

1. Appliquer la fonction `SBM` aux données de Kéfi et al. [2016] pour différentes valeurs de K et déterminer un critère ICL pour choisir le K optimal $\widehat{K}$.
2. Interpréter les résultats pour $\widehat{K}$. Le fichier `chilean_metadata.csv` (accessible sur le moodle du cours) contient différentes caractéristiques des espèces, notamment le phylum auquel elles appartiennent. Peut-on mettre ces caractéristiques en lien avec les groupes déterminés par le modèle à blocs stochastiques ?

Rendu attendu

Vous enverrez à l'adresse `stephane.robin@sorbonne-universite.fr` un fichier '.R' contenant l'implémentation en R de l'algorithme EM. Ce programme devra :

— prendre en entrée un fichier de la même forme que `chilean_TI.csv`,
— tracer l'évolution de l'ELBO au cours des itérations de l'algorithme EM,
— afficher la valeur du critère ICL variationnel pour $K = 1 \dots 10$,
— afficher les estimations des paramètres du modèles pour le K optimal.