

1 Modèles de Poisson avec excès de zéros

Données

On s'intéresse à l'abondance d'espèces de poissons dans $n = 89$ sites répartis dans la mer de Barents. Les données sont disponibles dans le fichier `BarentsFish.csv` dont les 4 premières colonnes correspondent à quatre covariables environnementales (latitude, longitude, profondeur, température) et les 30 colonnes suivantes aux abondances de 30 espèces.

Dans la suite, on ne considérera qu'une espèce de poisson, par exemple la 20ème ('Se_ma' = *Sebastes marinus* = *Golden redfish* = Sébaste doré) et on notera

$$Y_i = \text{abondance de sébastes dorés dans la station } i \quad (1 \leq i \leq n).$$

1.1 Modèle sans covariable

On considère d'abord un modèle sans covariable prévoyant l'absence ou la présence de l'espèce au site i et, conditionnellement à sa présence, une abondance poissonnienne. On pose :

$$\begin{aligned} \{Z_i\}_{1 \leq i \leq n} &\text{ iid,} & Z_i &\sim \mathcal{B}(\pi), \\ \{Y_i\}_{1 \leq i \leq n} &\text{ indépendants } | \{Z_i\}, & Y_i | Z_i &\sim Z_i \delta_0 + (1 - Z_i) \mathcal{P}(\lambda). \end{aligned} \quad (1)$$

La variable latente Z_i est donc l'indicatrice d'absence de l'espèce dans le site i . L'objectif est d'implémenter un algorithme EM afin d'obtenir l'estimateur du maximum de vraisemblance (EMV) de $\theta = (\pi, \lambda)$.

1. Écrire la vraisemblance complète $\log p_\theta(Y, Z)$ du modèle (1) en fonction de θ .
2. Écrire l'étape E.
3. Écrire l'étape M.
4. Proposer une valeur initiale pour le paramètre θ .
5. Coder l'algorithme EM.
6. Comparer ce modèle au modèle de Poisson simple

$$\{Y_i\}_{1 \leq i \leq n} \text{ iid,} \quad Y_i \sim \mathcal{P}(\lambda). \quad (2)$$

1.2 Modèle avec covariables

On considère maintenant un modèle analogue au modèle (1) mais prenant en compte les covariables environnementales. On note x_i le vecteur comprenant ces covariables pour le site i , ainsi qu'une terme constant :

$$x_i = [1 \text{ latitude}_i \text{ longitude}_i \text{ profondeur}_i \text{ température}_i]^\top.$$

On pose :

$$\begin{aligned} \{Z_i\}_{1 \leq i \leq n} &\text{ indépendants,} & Z_i &\sim \mathcal{B}(\pi_i), & \log \left(\frac{\pi_i}{1 - \pi_i} \right) &= x_i^\top \alpha, \\ \{Y_i\}_{1 \leq i \leq n} &\text{ indépendants } | \{Z_i\}, & Y_i | Z_i &\sim Z_i \delta_0 + (1 - Z_i) \mathcal{P}(\lambda_i), & \log \lambda_i &= x_i^\top \beta. \end{aligned} \quad (3)$$

Les vecteurs α et β contiennent les coefficients de régression permettant de prédire respectivement l'absence et l'abondance conditionnelle à la présence de l'espèce en chaque site.

1. Écrire la vraisemblance complète $\log p_\theta(Y, Z)$ du modèle (1) en fonction du paramètre $\theta = (\alpha, \beta)$.
2. Écrire l'étape E.
3. Écrire l'étape M.
4. Proposer une valeur initiale pour le paramètre θ .
5. Coder l'algorithme EM.
6. Comparer les trois modèles (1), (2) et (3) ainsi que le modèle de régression poissonnienne

$$\{Y_i\}_{1 \leq i \leq n} \text{ indépendants,} \quad Y_i \sim \mathcal{P}(\lambda_i), \quad \log \lambda_i = x_i^\top \beta. \quad (4)$$