Modèles à bariables latentes pour l'écologie

Examen du 24 mars 2025

Durée : 2 heures

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

1 Estimation d'un nombre d'espèces

On souhaite estimer le nombre n total d'espèces de papillons présentes dans un site particulier. On organise pour cela une campagne de capture au cours de laquelle on observe m espèces. On note Y_j^+ le nombre d'individus capturés appartenant à l'espèce $j \in \{1, \dots m\}$.

Pour estimer n, il faut de plus estimer le nombre n_0 d'espèces présentes dans le site, mais dont aucun individu n'a été capturé. On aura alors $n = m + n_0$. On se propose pour cela de définir un modèle pour le nombre d'individus capturés d'une espèce, en autorisant celui-ci à être nul.

Loi Poisson-Gamma. La loi Poisson-Gamma, notée $Y \sim \mathcal{PG}\text{am}(a, b)$ est construite de la façon suivante : on tire d'abord une variable Z selon une loi Gamma de paramètres a et b, puis on tire la variable Y selon une loi de Poisson de paramètre Z :

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

On rappelle que la densité de la loi Gamma est définie pour $z \in \mathbb{R}^+$ et vaut

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

On rappelle également que la fonction de probabilité de la loi de Poisson est $\mathcal{P}(y; \lambda) = e^{-\lambda}\lambda^y / y!$ pour $y \in \mathbb{N}$.

Probabilité de ne pas observer une espèce.

1. Comment interpréter la variable latente Z ?
2. Montrer que la loi marginale de y est donnée, pour $y \in \mathbb{N}$, par

$$p(y) = \mathbb{P}\{Y = y\} = \frac{\Gamma(a + y)}{\Gamma(a) y!}\, \frac{b^a}{(b+1)^{a+y}}.$$

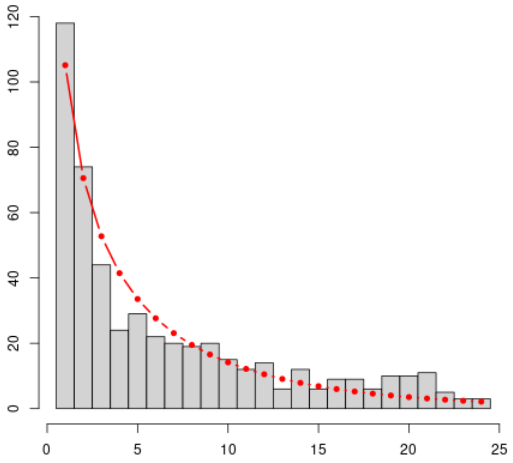
3. En déduire que, selon la loi Poisson-Gamma (1), la probabilité de ne pas observer une espèce vaut $[b/(b+1)]^a$.

Modèle pour l'ensemble des espèces. On note maintenant Y_i $(1 \leq i \leq n)$ le nombre (éventuellement nul) d'invididus capturés pour chacune des n espèces présentes. On suppose que les Y_i sont iid et de loi $\mathcal{PG}\text{am}(a, b)$. On note p^+ la loi du nombre Y_j^+ d'individus capturés pour une espèce j $(1 \leq j \leq m)$ dont on a observé au moins un individu.

4. Donner $p^+(y)$.
5. Proposer *succinctement* une méthode d'estimation des paramètres a et b à partir des effectifs $Y_1^+, \dots Y_m^+$ observés.

Estimation du nombre d'espèces. La figure suivante donne la distribution des effectifs Y_j^+ de $m = 501$ espèces des papillons capturés dans une forêt de Malaisie [Fisher et al., 1943]. Pour 118 espèces, on a observé un seul individu, pour 74

espèces on en a observé deux et ainsi de suite.



On a estimé à partir de ces données les valeurs des paramètres a et b et on a obtenu

$$\widehat{a} = 0.49, \qquad \widehat{b} = 0.11.$$

La courbe donne la distribution p^+ pour les paramètres $(\widehat{a}, \widehat{b})$.

6. Donner une estimation de la probabilité $p(0)$ qu'une espèce présente ne soit pas observée.
7. En déduire un estimateur des moments et l'estimation correspondante $\widehat{n}_0$ du nombre d'espèces présentes mais non-observées et une estimation $\widehat{n}$ du nombre total d'espèces présentes.

2 Approximation variationnelle pour la loi Poisson log-normale unidimensionnelle

Modèle. On considère le modèle Poisson log-normal univarié défini de la façon suivante :

$$Z \sim \mathcal{N}(0, \sigma^2), \qquad Y \mid Z \sim \mathcal{P}(e^{\mu+Z})$$

dont le paramètre est $\theta = (\mu, \sigma^2)$.
On cherche une approximation de la loi conditionnelle $p_\theta(Z \mid Y)$ au sein de la classe des lois normales :

$$\mathcal{Q} = \{q = \mathcal{N}(m, s^2), m \in \mathbb{R}, s^2 \in \mathbb{R}^{*+}\}.$$

On rappelle que la divergence de Küllback-Leibler entre deux distributions f et g vaut

$$KL[f\|g] = \mathbb{E}_{X\sim f}\left(\log \frac{f(X)}{g(X)}\right).$$

Approximation variationnelle. On considère tout d'abord la divergence

$$D_1(q;p) = KL[q\|p(\cdot \mid Y)].$$

1. Montrer que, si U suit une loi normale $\mathcal{N}(\nu, \gamma^2)$, alors $\mathbb{E}(e^U) = \exp(\nu + \gamma^2/2)$.
2. Écrire, à une constante additive près, $D_1(q;p)$ en fonction des paramètres (m, s^2) de la loi q.
3. En déduire que les paramètres m_1 et s_1^2 de la loi q_1 qui minimisent $D_1(q;p)$ au sein de $\mathcal{Q}$ satisfont

$$\left\{ \begin{array}{rcl} m_1 & = & \sigma^2\left(Y - e^{\mu+m_1+s_1^2/2}\right), \\ 1/s_1^2 & = & e^{\mu+m_1+s_1^2/2} + 1/\sigma^2. \end{array} \right. \tag{2}$$

4. Montrer que le système (2) admet un unique couple de solutions (m_1, s_1^2).

Approximation par propagation de l'espérance. On considère maintenant la divergence

$$D_2(q;p) = KL[p(\cdot \mid Y)\|q].$$

5. Écrire, à une constante additive près, $D_2(q;p)$ en fonction des paramètres (m, s^2) de la loi q.
   (*On pourra exprimer les résultats sous la forme de moments conditionnels sous la loi p : $\mathbb{E}_p[f(Z) \mid Y]$.*)
6. En déduire les paramètres optimaux m_2 et s_2^2 de la loi q_2 qui minimise $D_2(q;p)$ au sein de $\mathcal{Q}$.
   (*On pourra encore exprimer les résultats sous la forme $\mathbb{E}_p[f(Z) \mid Y]$.*)

Références

R. A Fisher, A S. Corbet, and C. B Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, pages 42–58, 1943.