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 ?

> *Solution.* Z est l'espérance du comptage observé Y et est donc proportionnelle à l'abondance de l'espèce. La variabilité de Z représente les différences d'abondance entre les espèces présentes.

2. Montrer que la loi marginale de y est donnée, pour $y \in \mathbb{N}$, par

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

> *Solution.* On écrit la loi jointe
>
> $$\begin{aligned} p(y, z) &= p_Z(z)\, p_{Y|Z}(y \mid z) = \mathcal{G}\text{am}(z; a, b)\, \mathcal{P}(y; z) \\ &= \frac{b^a}{\Gamma(a)}\, z^{a-1}\, e^{-bz}\, e^{-z}\, \frac{z^y}{y!} = \frac{b^a}{\Gamma(a) y!}\, z^{a+y-1}\, e^{-(b+1)z}. \end{aligned}$$
>
> On obtient alors $p(y)$ de y en intégrant la loi jointe par rapport à z, soit
>
> $$p(Y) = \int p(y, z)\, \mathrm{d}z = \frac{b^a}{\Gamma(a) y!} \int z^{a+y-1} e^{-(b+1)z}\, \mathrm{d}z$$
>
> or, la constante de normalisation de la loi Gamma nous rappelle que $\int z^{\alpha-1} e^{-\beta z}\, \mathrm{d}z = \Gamma(\alpha)/\beta^\alpha$, ce qui donne le résultat en prenant $\alpha = a + y$ et $\beta = b + 1$.

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

> *Solution.* Il suffit de calculer $p(0)$ avec la formule de la question précédente.

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 $p^+(y)$.

Solution. p^+ est la loi de Y conditionnellement au fait que Y est strictement positif, soit

$$p^+(y) = \Pr\{Y = y | Y > 0\} = \Pr\{Y = y\}/\left(1 - \Pr\{Y = 0\}\right)$$
$$= p(y)/(1 - p(0)) = p(y) \Big/ \left(1 - \left(\frac{b}{b+1}\right)^a\right).$$

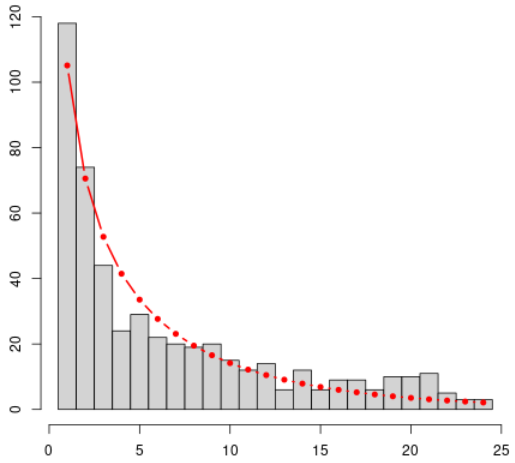
5. Proposer *succinctement* une méthode d'estimation des paramètres a et b à partir des effectifs $Y_1^+, \ldots Y_m^+$ observés.

Solution. On peut, par exemple, estimer a et b par maximum de vraisemblance en maximisant

$$\sum_{j=1}^{m} \log p^+(Y_j^+)$$

en fonction a et b.

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.

Solution. Application numérique : $\widehat{p}(0) = 0.32$.

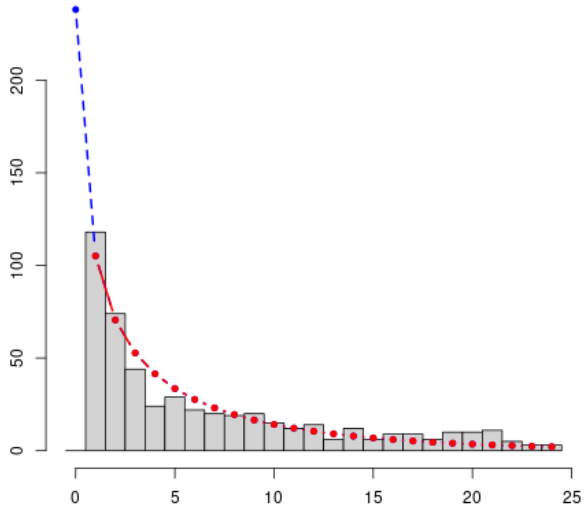
7. En déduire un estimateur des moments et l'estimation correspondante $\widehat{n}_0$ du nombre d'espèces présentes mais non-observées et une estimation $\widehat{n}$ du nombre total d'espèces présentes.

Solution. Par la méthode des moments : le nombre m d'espèces observées est la réalisation d'un variable binomiale $M \sim \mathcal{B}(n, 1 - p(0))$, d'espérance $\mathbb{E}M = n(1 - p(0))$, ce qui suggère

$$\widehat{n} = m/(1 - \widehat{p}(0)) = 739.$$

Le nombre d'espèce non-observée vaut $n_0 = n - m$, soit un estimation $\widehat{n}_0 = 238$.
Comme le montre la figure suivante, cette méthode revient à prolonger la distribution estimée $\widehat{p}^+$ en 0 pour obtenir

une estimation de la distribution p.



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 $\mathbb{E}(e^U) = \exp(\nu + \gamma^2/2)$.

Solution. Le calcul est direct :

$$\begin{aligned}
\mathbb{E}(e^U) &= \frac{1}{\gamma\sqrt{2\pi}} \int_{\mathbb{R}} \exp\left(u - \frac{(u-\nu)^2}{2\gamma^2}\right) \mathrm{d}u \\
&= \frac{1}{\gamma\sqrt{2\pi}} \int_{\mathbb{R}} \exp -\left(\frac{u^2 - 2u(\nu+2s^2) + \nu^2}{2\gamma^2}\right) \mathrm{d}u \\
&= \frac{1}{\gamma\sqrt{2\pi}} \int_{\mathbb{R}} \exp -\left(\frac{[u-(\nu+2s^2)]^2 + \nu^2 - (\nu+\gamma^2)^2}{2\gamma^2}\right) \mathrm{d}u \\
&= \exp\left(\frac{(\nu+\gamma^2)^2 - \nu^2}{2\gamma^2}\right) \frac{1}{\gamma\sqrt{2\pi}} \underbrace{\int_{\mathbb{R}} \exp -\left(\frac{(u-(\nu+2s^2))^2}{2\gamma^2}\right) \mathrm{d}u}_{=\gamma\sqrt{2\pi}} \\
&= \exp(\nu + \gamma^2/2).
\end{aligned}$$

2. Écrire, à une constante additive près, $D_1(q;p)$ en fonction des paramètres (m, s^2) de la loi q.

Solution. Puisque la loi marginale $p(Y)$ ne dépend pas de q, on a

$$\begin{aligned} D_1(q;p) &= \mathbb{E}_q\left[\log q(Z) - \log p(Y,Z)\right] + \text{cst} \\ &= \mathbb{E}_q\left[\log q(Z) - \log p(Z) - \log p(Y \mid Z)\right] + \text{cst} \\ &= \mathbb{E}_q\left[-\frac{1}{2}\log s^2 - \frac{(Z-m)^2}{2s^2} + \frac{Z^2}{2\sigma^2} + e^{\mu+Z} - Y(\mu+Z)\right] + \text{cst} \\ &= -\frac{1}{2}\log s^2 + \frac{m^2+s^2}{2\sigma^2} + e^{\mu+m+s^2/2} - Y(\mu+m) + \text{cst}. \end{aligned}$$

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

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

Solution. Les dérivées de $D_1(q;p)$ par rapport à m et s^2 (plutôt que s) valent

$$\partial_m D_1(q;p) = \frac{m}{\sigma^2} + e^{\mu+m+s^2/2} - Y, \qquad \partial_{s^2} D_1(q;p) = -\frac{1}{2s^2} + \frac{1}{2\sigma^2} + \frac{1}{2}e^{\mu+m+s^2/2}.$$

qui s'annulent sous les conditions indiquées.

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

Solution. On fait l'étude des fonctions $m \to \partial_m D_1(q;p)$ et $s^2 \to \partial_{s^2} D_1(q;p)$.

— La dérivée seconde de $D_1(q;p)$ par rapport à m vaut $\partial^2_{m^2} D_1(q;p) = \sigma^{-2} + \exp(\mu+m+s^2/2)$ et est donc strictement positive. De plus, les limites quand $m \to -\infty$ et $m \to +\infty$ de $\partial_m D_1(q;p)$ sont respectivement $-\infty$ et $+\infty$. Quelque soit s^2, $\partial_m D_1(q;p)$ s'annule en une unique valeur $m_1(s^2)$.

— De même, la dérivée seconde de $D_1(q;p)$ par rapport à s^2 vaut $\partial^2_{(s^2)^2} D_1(q;p) = s^{-4}/2 + e^{\mu+m+s^2/2}/4$ et est donc également strictement positive. De même, les limites quand $s^2 \to -0$ et $s^2 \to +\infty$ de $\partial_{s^2} D_1(q;p)$ sont respectivement $-\infty$ et $+\infty$. Quelque soit m, $\partial_m D_1(q;p)$ s'annule en une unique valeur $s_1^2(m)$.

Alternative : On peut écrire la jacobienne de la fonction D_1 qui vaut

$$J_{m,s^2}(D_1) = \left[\begin{array}{cc} a+e & e/2 \\ e/2 & b+e/4 \end{array} \right]$$

où les trois coefficients $a = \sigma^{-2}$, $b = s^{-4}/2$ et $e = \exp(\mu + m + s^2/2)$ sont strictement positifs. On observe que

$$\delta := |J_{m,s^2}(D_1)| = ab + \left(\frac{a}{4} + b\right)e, \qquad \tau := \operatorname{tr}(J_{m,s^2}(D_1)) = a + b + \frac{5e}{4}$$

sont tous les deux strictement positifs et que le discriminant de son polynôme caratéristique $P(\lambda) = \lambda^2 - \tau\lambda + \delta$ vaut (après calcul)

$$\Delta = \tau^2 - 4\delta = \left(a - b + \frac{3e}{4}\right)^2 + e^2$$

et est donc également strictement positif. $J_{m,s^2}(D_1)$ est donc strictement définie positive en tout point (m, s^2) : D_1 est une fonction strictement convexe de (m, s^2) et admet donc un unique minimum.

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 : $\mathbb{E}_p[f(Z) \mid Y]$.*)

Solution. La divergence de Küllback-Leibler est ici une espérance sous la loi p, conditionnellement à Y. Puisque la

loi marginale $p(Y)$ ne dépend pas de q, on a de nouveau

$$\begin{aligned} D_2(q;p) &= \mathbb{E}_p\left[\log p(Y,Z) - q(Z) \mid Y\right] + \text{cst} \\ &= \mathbb{E}_p\left[\log p(Z) + \log p(Y \mid Z) - q(Z) \mid Y\right] + \text{cst} \\ &= \frac{1}{2}\log s^2 + \frac{1}{2s^2}\mathbb{E}_p\left[(Z-m)^2 \mid Y\right] + \text{cst}. \end{aligned}$$

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* $\mathbb{E}_p[f(Z) \mid Y]$.)

Solution. On sait que l'espérance minimise la perte quadratique $\mathbb{E}_p[(Z-m)^2 \mid Y]$, la valeur optimale de m est donc $m_2 = \mathbb{E}_p(Z \mid Y)$.
De plus, la dérivée de $D_2(q;p)$ par rapport à s^2

$$\partial_{s^2} D_2(q;p) = \frac{1}{2s^2} - \frac{\mathbb{E}_p[(Z-m)^2 \mid Y]}{2s^4}$$

s'annule, puisque $m_2 = \mathbb{E}_p[Z \mid Y]$, pour $s_2^2 = \mathbb{E}_p[(Z - \mathbb{E}_p[Z \mid Y])^2 \mid Y] = \mathbb{V}_p(Z \mid Y)$.
On obtient donc

$$m_2 = \mathbb{E}_p(Z \mid Y), \qquad s_2^2 = \mathbb{V}_p(Z \mid Y).$$

Alternative : On peut invoquer les propriétés générales de l'approximation par propagation de l'espérance pour les classes d'approximation prises dans la famille exponentielle, en se souvenant que les paramètre canoniques de la loi normale $\mathcal{N}(m, s^2)$ sont m/s^2 et $1/s^2$.

Remarque : Ces deux moments n'admettent pas de forme explicite et ne sont donc calculables que par intégration numérique.

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.