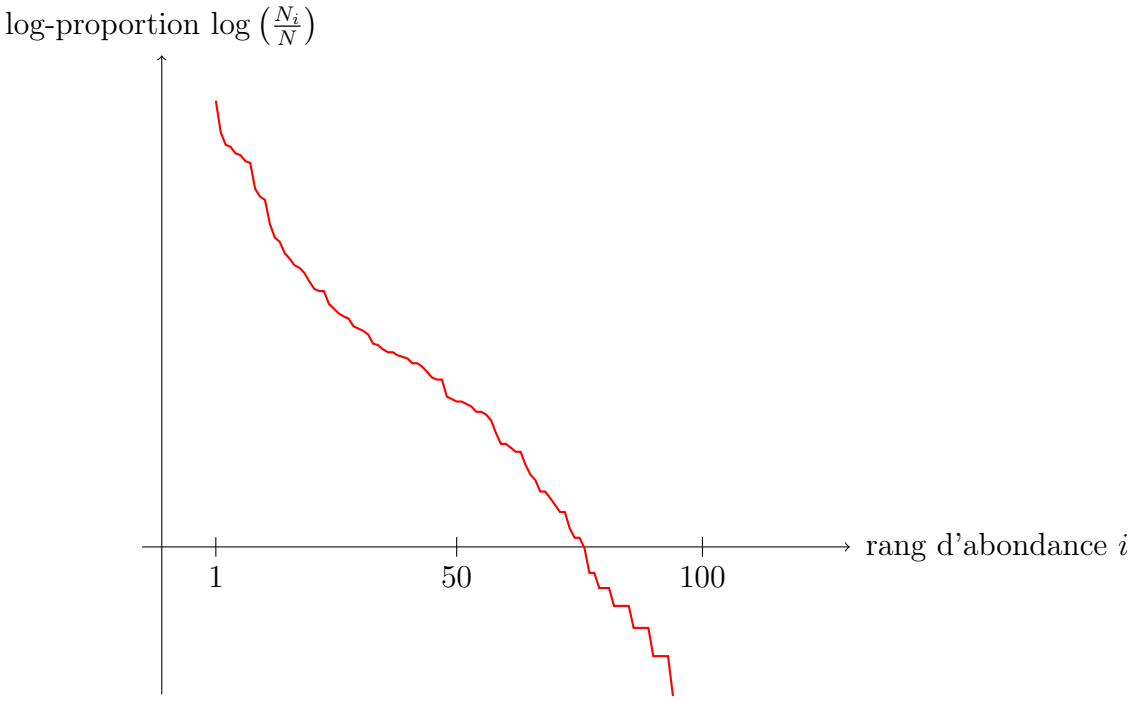
THÉORIE NEUTRE UNIFIÉE DE LA BIODIVERSITÉ

mots-clés : chaînes de Markov, surmartingales, lois invariantes, mesures d'Ewens.

L'objectif de ce texte est de proposer un modèle mathématique qui explique la répartition des individus d'un écosystème parmi différentes espèces. Considérons un écosystème complexe, par exemple une forêt tropicale ou un amas océanique de plancton. On suppose fixé un nombre total N d'individus, qui se répartissent en un nombre S d'espèces : $N = \sum_{i=1}^{S} N_i$, N_i désignant le nombre d'individus dans l'espèce i. Supposons un instant qu'on ait ordonné les espèces de sorte que $N_1 \geq N_2 \geq \cdots \geq N_i \geq N_{i+1} \geq \cdots \geq N_S$ (attention, cette hypothèse ne sera plus valable à partir de la section 1). Le graphe suivant représente alors la fonction logarithme de la proportion $\frac{N_i}{N}$, dans le cas où l'on observe 10000 individus répartis en une centaine d'espèces :



Les spécialistes de la biodiversité ont observé que les écosystèmes complexes donnaient presque tous la même forme de courbe "en S" pour la répartition des espèces en fonction du rang d'abondance (ces observations ont été faites pour la première fois dans les années 1950 par Frank Preston). Ainsi, les répartitions par espèces des arbres dans une forêt, des poissons dans une mer ou des oiseaux sur une île donnent à peu près toutes la même courbe (à une renormalisation près suivant les deux axes). Un modèle expliquant cette répartition universelle a été proposé par Stephen Hubbell en 2001, dans le livre *The Unified Neutral Theory of Biodiversity and Biogeography* ; l'objectif de ce texte est de comprendre certains aspects mathématiques de cette théorie.

1. Hypothèses du modèle neutre

Dans le modèle neutre de Hubbell, un écosystème ou communauté d'individus correspond à un espace géographique délimité, et contient un nombre d'individus N qui est

fixe au cours du temps. Il a été observé qu'indépendamment de la répartition en diverses espèces, tous les espaces géographiques d'un même type (par exemple toutes les forêts tropicales) avaient cette propriété, avec N proportionnel à l'aire A de l'espace géographique considéré. Si l'on veut que cette règle soit respectée à tout instant t, alors il faut que la répartition des individus entre espèces soit un "jeu à somme nulle", c'est-à-dire que toute mort d'un individu d'une espèce i soit compensée au même instant par la naissance d'un autre individu, qui peut être dans l'espèce i ou dans une autre espèce j. Le modèle neutre consiste à supposer que les chances de survie ou de mort d'un individu ne sont pas influencées par l'espèce à laquelle il appartient. Par ailleurs, on autorise l'apparition de nouvelles espèces, à un taux $\nu \in (0,1)$ très faible. En particulier, le nombre $S = S(t)$ d'espèces dans l'écosystème pourra varier au cours du temps (phénomènes de spéciation et d'extinction d'une espèce).

Ces hypothèses sont modélisées par une chaîne de Markov dont l'espace des états est l'ensemble $\mathfrak{C}_N$

$$\mathfrak{C}_N = \left\{ (N_1, \ldots, N_R) \mid R \geq 1,\ N_i \geq 0 \text{ pour tout } i \in [\![1, R]\!],\ \sum_{i=1}^{R} N_i = N \right\}$$

des *compositions* de taille N. Par exemple, si $N = 10$, alors $(3,3,4)$ et $(1,5,1,0,2,1)$ sont deux éléments de $\mathfrak{C}_N$. On discrétise le temps t, et si $t \in \mathbb{N} = \{0,1,2,\ldots\}$, on note $c(t) = (N_1(t), N_2(t), \ldots, N_{R(t)}(t))$ la composition qui représente la répartition des espèces dans la population au temps t. À l'instant $t+1$, la nouvelle répartition $c(t+1)$ est choisie comme suit :

(1) On choisit au hasard un individu de la population qui meurt. La probabilité que cet individu soit issu de l'espèce i avec $i \in [\![1, R(t)]\!]$ est égale à $\frac{N_i(t)}{N}$.

(2) Avec probabilité ν, l'individu de type i qui est mort à l'étape 1 est remplacé par un nouvel individu d'une nouvelle espèce (phénomène de spéciation). Dans ce cas, la nouvelle composition $c(t+1)$ s'écrit :

$$c(t+1) = (N_1(t), \ldots, N_{i-1}(t), N_i(t) - 1, N_{i+1}(t), \ldots, N_{R(t)}(t), 1).$$

(3) Avec probabilité $1-\nu$, l'individu de type i qui est mort à l'étape 1 est remplacé par un individu d'une espèce j avec $j \in [\![1, R(t)]\!]$. La probabilité de choix de j est $\frac{N_j(t)}{N}$. Les indices i et j étant choisis, la nouvelle composition $c(t+1)$ s'écrit :

$$c(t+1) = \begin{cases} (N_1(t), \ldots, N_i(t) - 1, \ldots, N_j(t) + 1, \ldots, N_{R(t)}(t)) & \text{si } j \neq i, \\ c(t) & \text{si } j = i. \end{cases}$$

Notons qu'il y a un choix dans l'écriture de $c(0)$: ainsi, $(2,3,5)$ et $(5,3,0,2)$ représentent les mêmes répartitions d'espèces pour une population de $N = 10$ individus. On convient de fixer une écriture pour $c(0)$ sans espèce i de taille 0 : ainsi, $(2,3,5)$ et $(5,2,3)$ sont deux compositions autorisées pour décrire la même population initiale, mais $(5,2,0,3)$ n'est pas autorisée pour $c(0)$. Si $c(0)$ est choisie, alors les règles écrites ci-dessus déterminent ensuite de façon unique $c(1), c(2), \ldots$, et elles permettent de suivre l'évolution d'une espèce i au sein de la population. Par ailleurs, $R(t)$ représente le nombre total d'espèces observées avec au moins un individu à un temps $s \in [\![0, t]\!]$, et $S(t) = \sum_{i=1}^{R(t)} 1_{N_i(t) \geq 1}$ est le nombre total d'espèces au temps t.

2. Évolution d'une espèce fixée

Dans cette section, on s'intéresse à l'évolution d'une espèce i fixée, par exemple celle qui a été numérotée 1 au temps $t = 0$.

Théorème 1. *On suppose $\nu > 0$. Le processus à temps discret $(N_1(t))_{t\geq 0}$ est une chaîne de Markov d'espace d'états $[\![0, N]\!]$, avec pour seul état absorbant 0. C'est aussi une sur-martingale positive. Par conséquent,*

$$\mathbb{P}\left[\lim_{t\to\infty} N_1(t) = 0\right] = \mathbb{P}[N_1(t) = 0 \text{ pour } t \text{ assez grand}] = 1.$$

On introduit le temps aléatoire T défini par :

$$T = \inf\{t \in \mathbb{N} \mid N_1(t) = 0\}.$$

D'après le théorème 1, T est fini presque sûrement, et on peut montrer que

$$\mathbb{P}[T > t] \leq \mathbb{E}[N_1(t)] = N_1(0)\,\left(1 - \frac{\nu}{N}\right)^t,$$

et donc que $\mathbb{E}[T] \leq \frac{N\,N_1(0)}{\nu}$.

En transposant ce résultat à une nouvelle espèce i qui apparaît à un certain temps aléatoire $\widetilde{T} \geq 1$ et qui meurt à un temps $\widetilde{T} + T$, on en déduit :

Proposition 2. *La durée de vie moyenne d'une nouvelle espèce i dans un écosystème de taille N est plus petite que $\frac{N}{\nu}$.*

En pratique, le coefficient ν est très petit, de l'ordre de $O(\frac{1}{N})$. On introduira plus loin la constante $\theta = \frac{N\nu}{1-\nu}$, et on observe pour θ des valeurs entre 1 et quelques dizaines (par exemple, $\theta \approx 50$ pour une forêt tropicale). La constante θ étant fixée, la durée de vie d'une espèce est donc de l'ordre de $\frac{N^2}{\theta}$.

3. Évolution de la répartition des espèces

Si $c = (N_1, N_2, \ldots, N_r)$ est une composition, la *partition* de taille N qui lui est associée est la suite p obtenue à partir de c en retirant les parts de taille 0, et en ordonnant les parts de façon décroissante. Par exemple, si $c = (1, 5, 1, 0, 2, 1)$, alors $p = (5, 2, 1, 1, 1)$. On note $p(t)$ la partition associée à la composition $c(t)$ définie précédemment, et $\mathfrak{P}_N$ l'ensemble des partitions de taille N (compositions sans part égale à 0, et avec des parts classées par ordre décroissant) :

$$\mathfrak{P}_N = \left\{ (N_1 \geq N_2 \geq \cdots \geq N_R) \mid R \geq 1,\ N_i \geq 1 \text{ pour tout } i \in [\![1, R]\!],\ \sum_{i=1}^{R} N_i = N \right\}.$$

Par exemple,

$$\mathfrak{P}_4 = \{(4), (3, 1), (2, 2), (2, 1, 1), (1, 1, 1, 1)\}.$$

Théorème 3. *On suppose toujours $\nu > 0$. Le processus à temps discret $(p(t))_{t\in\mathbb{N}}$ est une chaîne de Markov irréductible et apériodique sur l'ensemble fini $\mathfrak{P}_N$. Il existe donc une mesure de probabilité μ sur $\mathfrak{P}_N$ telle que*

$$\lim_{t\to\infty} \mathbb{P}[p(t) = p] = \mu(p)$$

pour toute partition $p \in \mathfrak{P}_N$.

Si $p \in \mathfrak{P}_N$, la suite des multiplicités de p est la suite $(m_1(p), m_2(p), \ldots, m_N(p))$ telle que p ait $m_i(p)$ parts de taille i. Par exemple, la suite des multiplicités de $p = (5, 2, 1, 1, 1)$ est

$$(3, 1, 0, 0, 1, 0, 0, 0, 0, 0).$$

Théorème 4. *La mesure invariante μ sur $\mathfrak{P}_N$ est la mesure d'Ewens de paramètre θ, qui s'écrit*

$$\mu(p) = \mu_{N,\theta}(p) = \frac{N!\,\theta^{m_1+m_2+\cdots+m_N}}{1^{m_1}2^{m_2}\cdots N^{m_N}\,(m_1)!\,(m_2)!\cdots(m_N)!\,\theta(\theta+1)\cdots(\theta+N-1)},$$

où $m = m(p)$ est la suite des multiplicités de p, et où $\theta = \frac{N\nu}{1-\nu}$ est la constante de biodiversité de l'écosystème.

Dans ce qui suit, on admet ce résultat et on étudie un autre modèle aléatoire qui fait apparaître cette mesure de probabilité. Ce modèle permettra en particulier d'estimer $S(t)$ pour t grand, et de simuler la mesure d'Ewens $\mu_{N,\theta}$.

4. Mesures d'Ewens et modèles de permutations aléatoires

Un entier N étant fixé, on note $\mathfrak{S}_N$ l'ensemble des permutations de taille N. On rappelle que toute permutation $\sigma \in \mathfrak{S}_N$ s'écrit comme produit de cycles à supports disjoints : par exemple, la permutation 914362857 qui envoie 1 sur 9, 2 sur 1, 3 sur 4, *etc.* se décompose en le produit de cycles $(1, 9, 7, 8, 5, 6, 2) \circ (3, 4)$. Le type cyclique d'une permutation $\sigma \in \mathfrak{S}_N$ est la partition $p(\sigma) \in \mathfrak{P}_N$ dont les parts sont les tailles des cycles de σ. Par exemple, la permutation 914362857, qui est de taille 9, a pour type cyclique $(7, 2)$.

Proposition 5. *Étant donnée une partition $p \in \mathfrak{P}_N$, le nombre de permutations $\sigma \in \mathfrak{S}_N$ qui ont ce type cyclique est égal à*

$$\frac{N!}{1^{m_1}2^{m_2}\cdots N^{m_N}\,(m_1)!\,(m_2)!\cdots(m_N)!},$$

où $m = m(p)$ est la suite des multiplicités de p.

Un paramètre θ étant fixé, on note $\rho_{N,\theta}$ la mesure de probabilité sur $\mathfrak{S}_N$ qui donne à une permutation σ une probabilité proportionnelle à $\theta^{\ell(\sigma)}$, $\ell(\sigma)$ étant le nombre de cycles de σ (les points fixes étant considérés comme des cycles de taille 1). En termes de type cyclique, notons que $\ell(\sigma) = m_1 + m_2 + \cdots + m_N$ si $m = m(p(\sigma))$. L'algorithme suivant permet de créer une permutation aléatoire $\sigma \in \mathfrak{S}_N$ suivant la loi $\rho_{N,\theta}$.

(1) On part de la permutation identité $\sigma_1 = \mathrm{id}_{[1,N]}$, et pour chaque $i \in [\![1, N]\!]$, on tire au hasard une variable de Bernoulli B_i, avec $\mathbb{P}[B_i = 1] = 1 - \mathbb{P}[B_i = 0] = \frac{\theta}{\theta+i-1}$. Les variables $B_1, B_2, B_3, \ldots, B_N$ sont supposées indépendantes.

(2) Pour passer de la permutation σ_{i-1} à la permutation σ_i, on procède comme suit :
   — Si $B_i = 1$, on pose $\sigma_i = \sigma_{i-1}$.
   — Si $B_i = 0$, on tire au hasard un entier $m_i \in [\![1, i-1]\!]$, et on pose $\sigma_i = \sigma_{i-1} \circ (i, m_i)$.

Théorème 6. *Dans l'algorithme ci-dessus, à chaque étape i, la permutation σ_i appartient à $\mathfrak{S}_i$ et suit la loi $\rho_{i,\theta}$. En particulier, $\sigma = \sigma_N$ suit la loi $\rho_{N,\theta}$, et cette loi s'écrit :*

$$\rho_{N,\theta}(\sigma) = \frac{\theta^{\ell(\sigma)}}{\theta(\theta+1)\cdots(\theta+N-1)}.$$

Le nombre de cycles de σ_N est $B_1 + B_2 + \cdots + B_N$.

Corollaire 7. *La loi image de la mesure $\rho_{N,\theta}$ par l'application $\sigma \in \mathfrak{S}_N \mapsto p(\sigma) \in \mathfrak{P}_N$ est la mesure d'Ewens $\mu_{N,\theta}$. Le nombre de parts d'une partition aléatoire p choisie suivant la mesure $\mu_{N,\theta}$ suit la loi d'une somme $B_1 + B_2 + \cdots + B_N$ de variables de Bernoulli indépendantes de paramètres $1, \frac{\theta}{\theta+1}, \frac{\theta}{\theta+2}, \ldots, \frac{\theta}{\theta+N-1}$.*

En combinant le corollaire 7 et le théorème 4, on voit que pour t grand, le nombre $S(t)$ d'espèces dans l'écosystème a une loi proche de celle d'une somme $B_1 + B_2 + \cdots + B_N$, et donc que

$$\mathbb{E}[S(t)] \simeq_{\substack{t \text{ grand,} \\ N \text{ grand}}} \theta \log N.$$

Par ailleurs, l'algorithme permet de simuler la mesure d'Ewens, et donc de retrouver le graphe présenté au début du texte. Le paramètre θ de biodiversité est alors essentiellement proportionnel au nombre d'espèces que l'on observe, et il correspond donc à l'échelle horizontale pour la courbe (rang d'abondance, log-proportion).

Remarque 8. *Une preuve du théorème 4 peut être donnée en utilisant un modèle markovien de permutations aléatoires $(\sigma(t))_{t\in\mathbb{N}}$ qui se projette sur le processus $(p(t))_{t\in\mathbb{N}}$ par l'application type cyclique, et qui a pour mesure invariante $\rho_{N,\theta}$. Cette construction donne une explication simple au fait que les mesures d'Ewens sont les lois limites des répartitions d'espèces dans un écosystème.*

Questions

Pour la rédaction des programmes, on pourra utiliser n'importe quel langage de programmation, ou éventuellement donner une description détaillée de l'algorithme (pseudo-code). Les questions sont organisées par parties du texte, mais des parties ultérieures peuvent aider à les résoudre.

I.1 Commenter les hypothèses du modèle neutre, et discuter de l'intérêt de la constante de biodiversité θ.

I.2 On note $\mathfrak{C}_N^*$ l'ensemble des compositions de taille N sans parts de taille 0. Par exemple, $\mathfrak{C}_4^* = \{(4), (3,1), (2,2), (1,3), (2,1,1), (1,2,1), (1,1,2), (1,1,1,1)\}$. Montrer que pour tout $N \geq 1$, $\mathfrak{C}_N^*$ a pour cardinal 2^{N-1}.

I.3 On note comme dans le texte $R(t)$ le nombre de parts (éventuellement nulles) de la composition $c(t)$. Décrire le processus aléatoire $(R(t))_{t\in\mathbb{N}}$, et déterminer le comportement asymptotique de $\frac{R(t)}{t}$.

I.4 Écrire un programme qui calcule le processus $(c(t))_{t\in\mathbb{N}}$, par exemple avec $N = 10$ et $\nu = 0.1$. On pourra représenter l'évolution de la composition c par une courbe pour chaque espèce i, et une couleur différente par espèce.

II.1 Montrer que $(N_1(t))_{t\in\mathbb{N}}$ est une surmartingale positive, et calculer $\mathbb{E}[N_1(t)]$ en fonction de $N_1(0)$. Pourquoi ceci implique-t-il que $(N_1(t))_{t\in\mathbb{N}}$ converge presque sûrement vers 0 lorsque t tend vers l'infini ?

II.2 Démontrer à l'aide du texte l'inégalité $\mathbb{E}[T] \leq \frac{N\,N_1(0)}{\nu}$. Écrire un programme qui simule la variable aléatoire T, et utiliser ce programme pour vérifier cette inégalité.

III.1 Démontrer entièrement le théorème 3.

III.2 Pour $N = 4$, écrire la matrice de transition de $(p(t))_{t\in\mathbb{N}}$ et vérifier que la mesure d'Ewens $\mu_{4,\theta}$ est bien invariante pour la chaîne de Markov $(p(t))_{t\in\mathbb{N}}$. Utiliser également le programme de la question I.4 pour vérifier que cette mesure est la loi limite du processus $(p(t))_{t\in\mathbb{N}}$ (avec par exemple $N = 4$ et $\nu = 0.5$).

IV.1 Démontrer la proposition 5. Étant donnée une partition $p = (p_1 \geq p_2 \geq \cdots \geq p_l)$ de taille N, on pourra considérer l'application

$\mathfrak{S}_N \to \mathfrak{S}_N$

$\sigma \mapsto (\sigma(1), \sigma(2), \ldots, \sigma(p_1)) \circ (\sigma(p_1 + 1), \ldots, \sigma(p_1 + p_2)) \circ \cdots \circ (\sigma(N - p_l + 1), \ldots, \sigma(N))$

qui est une surjection sur l'ensemble des permutations de type cyclique p.

IV.2 Montrer par récurrence sur N que toute permutation $\sigma \in \mathfrak{S}_N$ s'écrit de manière unique sous la forme

$$\sigma = (1, n_1) \circ (2, n_2) \circ \cdots \circ (N, n_N),$$

où chaque n_i appartient à $[\![1, i]\!]$ (dans cette écriture, si $n_i = i$, alors la transposition (i, i) est l'identité). Montrer que de plus, $\ell(\sigma) = \text{card}\,\{i \in [\![1, N]\!] \mid i = n_i\}$. Dans l'algorithme décrit dans la section 4 pour la construction de σ_N, si l'on a $\sigma_N = (1, n_1) \circ \cdots \circ (N, n_N)$, calculer les probabilités $\mathbb{P}[n_i = i]$ et $\mathbb{P}[n_i = j]$ avec $j < i$. En déduire une preuve du théorème 6.

IV.3 Utiliser le corollaire 7 et l'algorithme qui engendre des permutations aléatoires de loi $\rho_{N,\theta}$ pour simuler une partition $p = (p_1 \geq p_2 \geq \cdots \geq p_l)$ qui suit la mesure d'Ewens $\mu_{N,\theta}$. On prendra $N = 10000$ et $\theta \in \{1, 2, 5, 10\}$. Dessiner pour chaque valeur du paramètre θ la fonction $\log(\frac{p_i}{N})$, et commenter.