Chapitre 1

Dynamiques aléatoires : chaines de Markov

Pour modéliser l'évolution au cours du temps (la dynamique) de systèmes biologiques, par exemple celle d'un organisme, d'une substance, d'un ecosystème, on choisit souvent des modèles aléatoires. Les plus simples de ces modèles aléatoires sont les chaînes de Markov qui, dans le cas particulier étudié ici sont faciles à utiliser car il n'y a pratiquement aucun prérequis mathématique. Elles nous donneront l'occasion d'une première familiarisation avec le calcul matriciel que nous approfondirons lors des leçons suivantes.

1.1 La plus simple des dynamiques aléatoires

On modélise avec des chaînes de Markov l'évolution au cours du temps de quantités X qui peuvent prendre un nombre fini d'états $X = x_1$, $X = x_2$, ..., $X = x_n$ et qui passent de l'état x_i à l'instant t à l'état x_j à l'instant suivant $t+1$ avec une probabilité p_{ij} donnée. Les nombres $p_{ij} = P(X_{t+1} = x_j / X_t = x_i)$ vérifient donc $0 \leq p_{ij} \leq 1$ et $\sum_{j=0}^{n} p_{ij} = 1$ (puisque si la chaîne est dans l'état x_i à un instant, elle sera nécessairement dans l'un des états possibles $x_1, \ldots, x_n$ l'instant suivant et donc $p_{i1} + p_{i2} + \ldots + p_{in} = 1$). ; L'expression $P(X_{t+1} = x_j / X_t = x_i)$ s'appelle une *probabilité conditionnelle* et représente la "probabilité que la quantité X vaille x_j à l'instant $t+1$ *sachant qu*'elle valait x_i à l'instant t".

Pour définir une chaîne de Markov il faut donc deux ingrédients de base :

1. L'*espace des états* $S := \{x_1, \ldots, x_n\}$ connu que l'on supposera fini
2. La *matrice de transition* (ou de passage)

$$\mathbb{P} = (p_{ij})_{1 \leq i \leq n, 1 \leq j \leq n} = \begin{array}{cc} \begin{array}{cccc} x_1 & x_2 & \ldots & x_n \end{array} & \\ \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ \vdots & & & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{pmatrix} & \begin{array}{c} x_1 \\ \vdots \\ x_n \end{array} \end{array}$$

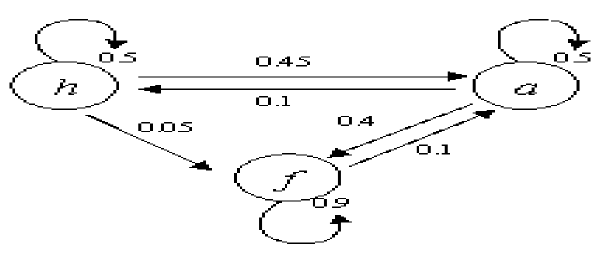


Fig. 1.1 – Diagramme en points et flèches correspondant à l'exemple de la dynamique de la forêt naturelle à trois états, *herbe*, *arbustes* et *forêt*, étudiée ci dessous.

A noter que cette matrice est appelée *matrice stochastique* parce que ses coefficients sont tous compris entre 0 et 1 et la somme des coefficients de chaque ligne vaut 1 (ce qui n'est pas vrai en général pour les colonnes).

On peut aussi représenter une chaîne de Markov $(S, \mathbb{P})$ par un *diagramme en points et flèches* comme indiqué par la figure (1.1) correspondant à l'exemple ci-dessous. Dans ces diagrammes, chaque état est représenté par un point et chaque coefficient p_{ij} non nul de la matrice de transition par une flèche allant de l'état i à l'état j.

Si l'on connait la distribution initiale des différents états (c'est-à-dire la proportion d'individus de la population étudiée se trouvant dans chacun des états x_i, que l'on appelle la *loi de probabilité initiale* π_0), l'étude de la chaîne de Markov va permettre de calculer, à partir de cette répartition

S	x_1	x_2	....	x_n
π_0	$\pi_0(x_1)$	$\pi_0(x_2)$	....	$\pi_0(x_n)$

$\pi_0(x_i)$ à l'instant $t = 0$, c'est-à-dire à partir des nombres $\pi_0(x_i) := P(X_0 = x_i)$, quels états la population va atteindre à l'instant $t = 1$ et avec quelles probabilités π_1, puis à l'instant $t = 2$ et ainsi de suite. En d'autres termes, on va ainsi calculer la loi π_t pour tous les $t > 0$ et ainsi modéliser la dynamique de cette population.

1.2 Un exemple en écologie

On s'interesse au développement d'une forêt naturelle en région tempérée sur une parcelle en friche (par exemple par abandon d'une zône cultivée ou suite à un incendie). Notre modèle simplifié comporte 3 états. L'état 1 est celui d'une végétation constituée d'herbes ou d'autres espèces pionnières ; l'état 2 correspond à la présence d'arbustes dont le développement rapide nécessite un ensoleillement maximal et l'état 3 celui d'arbres plus gros qui peuvent se développer dans un environnement semi ensoleillé. Si l'on note h, a, f ces trois états (pour *herbe, arbustes, forêt*), on a donc ici $S = \{h, a, f\}$. Sur la parcelle on repère au sol un grand nombre de points (un millier) répartis sur un maillage régulier et on enregistre à intervalle de temps fixé (tous les 3 ans) l'état de la végétation en chacun de ces points.

L'observation de l'ensemble de ces points à l'instant initial t_0 permet de déterminer les proportions initiales de chacun des 3 états $\pi_0 = (\pi_0(h), \pi_0(a), \pi_0(f))$. Pour cela, on relève pour chacun d'eux l'état dans lequel il se trouve et on calcule la proportion de points dans chacun des états possible. On peut voir ces proportions comme les probabilités pour un point quelconque de la parcelle d'être dans l'un de ces états à l'instant initial.

Dans ce modèle, on suppose connues les 9 probabilités

$$p_{ij} = P(X_1 = j / X_0 = i)$$

pour chaque valeur $i \in \{h, a, f\}$ et $j \in \{h, a, f\}$, probabilités pour un point quelconque de passer de l'état i à l'état j. On a (par exemple) :

$$\mathbb{P} = \begin{array}{c} \\ \\ \end{array}\begin{array}{cccc} \quad h & a & f & \\ \left(\begin{array}{ccc} 0,5 & 0,45 & 0,05 \\ 0,1 & 0,5 & 0,4 \\ 0 & 0,1 & 0,9 \end{array}\right) & & & \begin{array}{c} h \\ a \\ f \end{array} \end{array}$$

d'où le diagramme en points et flèches de la figure (1.1).

On peut ainsi calculer la probabilité de n'importe quelle succession d'états, appelée *trajectoire* de la chaîne de Markov. Par exemple la probabilité qu'en un point de la parcelle on observe la succession d'états (h, h, a, f, f) est égale à

$$\begin{aligned} P & \quad (X_0 = h, X_1 = h, X_2 = a, X_3 = f, X_4 = f) \\ = & \quad \pi_0(h) P(X_1 = h / X_0 = h) P(X_2 = a / X_1 = h) P(X_3 = f / X_2 = a) P(X_4 = f / X_3 = f) \\ = & \quad \pi_0(h) p_{hh} p_{ha} p_{af} p_{ff} = \pi_0(h)(0,5)(0,45)(0,4)(0,9) = 0,081 \pi_0(h). \end{aligned}$$

Mais on ne cherche pas seulement à calculer la probabilité particulière de chaque trajectoire de notre chaîne de Markov, on voudrait plus généralement déterminer l'évolution des proportions des trois états entre le premier et le deuxième instant, entre le deuxième et le troisième, et plus généralement savoir comment vont évoluer ces proportions à l'avenir. Voici comment on procède. Pour calculer les probabilités π_1 des trois états à l'instant $t = 1$, c'est-à-dire pour calculer

$$\pi_1 := (P(X_1 = h), P(X_1 = a), P(X_1 = f)) = (\pi_1(h), \pi_1(a), \pi_1(f)),$$

on remarque que $\pi_1(h)$ est égal à

$$P(X_1 = h/X_0 = h)P(X_0 = h) + P(X_1 = h/X_0 = a)P(X_0 = a) + P(X_1 = h/X_0 = f)P(X_0 = f)$$

ce qui peut s'écrire ici $\pi_1(h) = 0,5 \cdot \pi_0(h) + 0,1 \cdot \pi_0(a) + 0 \cdot \pi_0(f)$ compte tenu des valeurs des probabilités de transition données par la matrice $\mathbb{P}$. On remarque que $\pi_1(h)$ est le *produit scalaire* du vecteur π_0 avec la première colonne de la matrice $\mathbb{P}$. De même, on vérifie que $\pi_1(a)$ est le produit scalaire du vecteur π_0 avec la deuxième colonne de la matrice $\mathbb{P}$ et que $\pi_1(f)$ est le produit scalaire du vecteur π_0 avec la troixième colonne de la matrice $\mathbb{P}$. On résume cela en disant que le vecteur π_1 est le produit du vecteur π_0 par la matrice $\mathbb{P}$, ce qui s'écrit simplement $\pi_1 = \pi_0 \cdot \mathbb{P}$ ou simplement $\pi_1 = \pi_0\mathbb{P}$, comme un produit de deux nombres (mais ici il s'agit d'un vecteur et d'une matrice). Cette formule très courte signifie que le vecteur $\pi_1 = (\pi_1(h), \pi_1(a), \pi_1(f))$ est le produit de la matrice $\mathbb{P}$ par le vecteur $\pi_0 = (\pi_0(h), \pi_0(a), \pi_0(f))$, ce qui s'écrit encore de façon matricielle :

$$(\pi_1(h), \pi_1(a), \pi_1(f)) = (\pi_0(h), \pi_0(a), \pi_0(f)) \begin{pmatrix} 0,5 & 0,45 & 0,05 \\ 0,1 & 0,5 & 0,4 \\ 0 & 0,1 & 0,9 \end{pmatrix}$$

1.3 Puissances de $\mathbb{P}$ et loi stationnaire

Pour une chaîne de Markov d'espace d'états S et de matrice de transition $\mathbb{P}$, l'évolution au cours du temps de la loi de probabilité initiale π_0 est donnée par $\pi_1 = \pi_0\mathbb{P}$, $\pi_2 = \pi_1\mathbb{P} = (\pi_0\mathbb{P})\mathbb{P} = \pi_0\mathbb{P}^2$,... et plus généralement $\pi_t = \pi_{t-1}\mathbb{P} = \ldots = \pi_0\mathbb{P}^n$. En particulier, la matrice de transition pour passer de l'état t à l'état $t+2$ est égale à $\mathbb{P} \times \mathbb{P} = \mathbb{P}^2$, et plus généralement, celle pour passer de l'état t à l'état $t+k$ est égale à $\mathbb{P}^k$. C'est une caractéristique importante des chaînes de Markov que la matrice de transition $\mathbb{P}$ élevée à la puissance k contient les probabilités de transitions de chacun des états vers les autres en exactement k intervalles de temps.

Une loi de probabilité π sur l'espace des états S est appelée *stationnaire* pour la chaîne de Markov $(S, \mathbb{P})$ si la chaîne laisse la loi inchangée, c'est-à-dire si l'on a $\pi\mathbb{P} = \pi$. Trouver les lois stationnaires d'une chaîne de Markov, s'il en existe, permet souvent de décrire plus facilement sa dynamique, comme nous le verrons à la leçon suivante.

1.4 Exercices

Exercice 1 : Dans l'exemple décrit ci-dessus,

1. Calculer la probabilité des trajectoires suivantes (h, a, f, h), (h, a, f, a), (a, a, a).
2. Calculer la distribution des états π_1 à l'instant $t = 1$ si l'on suppose $\pi_0 = (1, 0, 0)$. Interpréter.
3. Montrer qu'une distribution uniforme $\pi_0 = (1/3, 1/3, 1/3)$ n'est pas une distribution stationnaire pour cette chaîne de Markov. Interprétez ce résultat.
4. Y-a-t-il une distribution stationnaire pour cette chaîne de Markov ?

Réponses 1. $P(h\ \ a\ \ f\ \ h) = \pi_0(h) \cdot 0,45 \cdot 0,4 \cdot 0 = 0$, $P(h\ \ a\ \ f\ \ a) = \pi_0(h) \cdot 0,45 \cdot 0,4 \cdot 0,1 = 0,018 \cdot \pi_0(h)$ et $P(a\ \ a\ \ a) = \pi(a) \cdot 0,5 \cdot 0,5 = 0,25 \cdot \pi_0(a)$

2. $\pi_1 = \pi_0.\mathbb{P} = (1,0,0).\begin{pmatrix} 0,5 & 0,45 & 0,05 \\ 0,1 & 0,5 & 0,4 \\ 0 & 0,1 & 0,9 \end{pmatrix} = (0,5 \quad 0,45 \quad 0,05)$ Il y a donc, après trois ans, la moitié de la parcelle recouverte de d'herbe, 45% d'arbustes et 5% de foret, si l'on suppose qu'au debut il n'y avait que de l'herbe.

3. $\pi_0.\mathbb{P} = (\frac{1}{3} \; \frac{1}{3} \; \frac{1}{3}).\begin{pmatrix} 0,5 & 0,45 & 0,05 \\ 0,1 & 0,5 & 0,4 \\ 0 & 0,1 & 0,9 \end{pmatrix} = (0,2 \;\; 0,35 \;\; 0,45) \neq \pi_0$ Donc la distribution uniforme (même proportion de chacun des trois type) n'est pas stationnaire, ce qui signifie que s'il y a au depart le même pourcentage de chacun des trois types, la distribution, est modifiée après trois ans.

4. On doit résoudre l'équation $\pi * .\mathbb{P} = \pi*$, c'est-à-dire, si on note $\pi* = (p \;\; q \;\; r)$, on doit trouver trois nombres compris entre 0 et 1 et de somme égale à 1, tels que :

$$\left\{ \begin{array}{rcl} 0,5 \;\; p + 0,1 \;\; q & = & p \\ 0,45 \;\; p + 0,5 \;\; q + 0,1 \;\; r & = & q \\ 0,05 \;\; p + 0,4 \;\; q + 0,9 \;\; r & = & r \end{array} \right. \tag{1.1}$$

   On trouve la solution $\pi* = (\frac{2}{53} \; \frac{10}{53} \; \frac{41}{53})$.

Exercice 2 : Une souris se déplace dans un labyrinthe représenté ci-dessous qui comporte 5 compartiments numérotés de 1 à 5. On suppose qu'elle change de compartiment à chaque instant $t = 0, 1, 2, 3, \ldots$ et que, lorsqu'elle se trouve dans un compartiment ayant k portes ($k = 1$, 2 ou 3), elle choisit l'une des portes avec la probabilité $\frac{1}{k}$, ses choix étant indépendants à chaque instant de ceux qu'elle a fait auparavant.

Le cheminement de la souris peut être décrit par une chaîne de Markov $(X_t)_{t=0,1,2,..}$ dont les états sont les 5 compartiments et la matrice de transition $\mathbb{P}$ la matrice des probabilités de passage d'un compartiment à un autre. Par exemple $p_{12} = \frac{1}{2}$ car le compartiment 1 contient 2 portes dont l'une vers le compartiment 2.

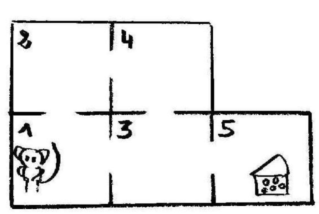
1. Ecrire la matrice de transition.

2. Calculer les probabilités des cheminements suivants : $(X_0 = 1, X_1 = 2, X_2 = 1, X_3 = 3, X_4 = 5)$ et $(X_0 = 1, X_1 = 2, X_2 = 3, X_3 = 4, X_4 = 5)$

3. Calculer la probabilité que la souris, partant à l'instant initial du compartiment 1, atteigne le compartiment 5, en 2 étapes, en 3 étapes, en 4 étapes.

4. On ne considère plus désormais que la souris part nécessairement du compartiment 1. On étudie une autre distribution initiale sur l'ensemble des états de cette chaîne de Markov définie de la façon suivante : la probabilité de chaque compartiment est proportionnelle au nombre de portes du compartiment. Préciser qu'elle est cette distribution initiale.

5. Cette loi de probabilité est-elle une loi stationnaire pour la chaîne de Markov considérée ? Pourquoi ?



Réponses 1. La matrice de transition est la suivante :

$$\mathbb{P} = \begin{pmatrix} 0 & 0,5 & 0,5 & 0 & 0 \\ 0,5 & 0 & 0 & 0,5 & 0 \\ \frac{1}{3} & 0 & 0 & \frac{1}{3} & \frac{1}{3} \\ 0 & 0,5 & 0,5 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

2. On trouve $P(1 \;\; 2 \;\; 1 \;\; 3 \;\; 5) = \pi_0(1) \;\; \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{24} \cdot \pi_0(1)$ et $P(1 \;\; 2 \;\; 3 \;\; 4 \;\; 5) = \pi_0(1) \cdot \frac{1}{2} \cdot 0 \cdot \frac{1}{3} \cdot 0 = 0$
3. C'est un calcul de probabilité conditionnelle. On a $P(X_2 = 5/X_0 = 1) = P(1 \;\; 3 \;\; 5) = \pi_0(1) \cdot \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6} \cdot \pi_0(1)$, $P(X_3 = 5/X_0 = 1) = P(1 \;\; 3 \;\; 5) = 0$ car il n'existe pas de chemin allant de la case 1 à la case 5 en deux étapes et enfin $P(X_4 = 5/X_0 = 1) = P(1 \;\; 2 \;\; 4 \;\; 3 \;\; 5) + P(1 \;\; 2 \;\; 1 \;\; 3 \;\; 5) + P(1 \;\; 3 \;\; 4 \;\; 3 \;\; 5) + P(1 \;\; 3 \;\; 1 \;\; 3 \;\; 5) + P(1 \;\; 3 \;\; 5 \;\; 3 \;\; 5) = \frac{1}{24} \cdot \pi_0(1) + \frac{1}{24} \cdot \pi_0(1) + \frac{1}{36} \cdot \pi_0(1) + \frac{1}{36} \cdot \pi_0(1) + \frac{1}{18} \cdot \pi_0(1) = \frac{7}{36} \cdot \pi_0(1)$.
4. En considérant que la probabilité d'être dans une case est proportionnelle au nombre de portes de cette case, la distribution initiale est $\pi_0 = (\frac{1}{5} \;\; \frac{1}{5} \;\; \frac{3}{10} \;\; \frac{1}{5} \;\; \frac{1}{10}$. En effet $\pi_0(1)$ est proportionnelle à 2, donc de la forme $\pi_0(1) = 2\lambda$, de même on aura $\pi_0(2) = 2\lambda$, $\pi_0(3) = 3\lambda$, $\pi_0(4) = 2\lambda$ et $\pi_0(5) = \lambda$. Comme on doit aussi avoir $\pi_0(1) + \pi_0(2) + \pi_0(3) + \pi_0(4) + \pi_0(5) = 1$, il faut prendre $\lambda = \frac{1}{10}$.
5. C'est bien une distribution stationnaire car on peut s'assurer qu'elle verifie $\pi_0 \cdot \mathbb{P} = \pi_0$.

Exercice 3 : Modélisation d'un brin d'ADN La modélisation la plus simple d'un brin d'ADN, enchainement "désordonné" de nucleotides de l'un des 4 types adenine (a), cytosine (c), guanine (g) et thymine (t), est de le considérer comme une trajectoire d'une chaîne de Markov X_n à quatre états $S = \{a, c, g, t\}$ dont la matrice de transition $\mathbb{P}$ fournit les probabilités que l'un de ces états succède à un autre. Ainsi le brin *aagc* est la trajectoire $X_0 = a, X_1 = a, X_2 = g, X_3 = c$.

1. En supposant que la matrice $\mathbb{P}$ est donnée par $\mathbb{P} = \begin{pmatrix} 0 & 0,3 & 0 & 0,7 \\ 0,6 & 0 & 0,4 & 0 \\ 0 & 0,8 & 0 & 0,2 \\ 0,3 & 0 & 0,7 & 0 \end{pmatrix}$ tracer le graphe en points et flèches associé à cette chaîne de Markov.
2. Calculer les probabilités des trajectoires suivantes en fonction de la probabilité initiale de l'état c : *cgcata* et *cgct*.
3. La distribution initiale $\pi_0 = (\frac{1}{8} \; ; 0 \; ; \frac{7}{8} \; ; 0)$ est-elle une distribution stationnaire pour cette chaîne de Markov ?
4. Reprendre les deux questions précédentes en prenant cette fois pour $\mathbb{P}$ la matrice

$$\mathbb{P} = \begin{pmatrix} 0,3 & 0 & 0,7 & 0 \\ 0 & 0,8 & 0 & 0,2 \\ 0,1 & 0 & 0,9 & 0 \\ 0 & 0,4 & 0 & 0,6 \end{pmatrix}$$

5. Supposons que vous puissiez observer un très long brin d'ADN. Indiquer quelle méthode vous choisiriez pour estimer les probabilités de transitions figurant dans la matrice $\mathbb{P}$.

Exercice 4 : Une foret à deux espèces On suppose[1] que l'on s'intéresse à une forêt composée de deux espèces d'arbres, E1 et E2. Lorsqu'un arbre meurt, un nouveau grandit à

[1]Exemple extrait du livre "Mathematical Models in Biology", E.S. Allman et J.A. Rhodes, Cambridge University Press, 2004

sa place mais il peut être de l'une ou l'autre des deux espèces. Ceux de la première espèce ayant une longue durée de vie, on suppose que 1% d'entre eux meurt chaque année alors que ce taux est de 5% pour la deuxième espèce. Mais ces derniers grandissant plus rapidement réussiront plus souvent à occuper une place laissée vacante : on suppose que 75% des places vacantes sont prises par un arbre de la deuxième espèce contre seulement 25% pour un arbre de la première espèce.

1. Expliquer comment l'on peut modéliser la dynamique ce cette foret par une chaîne de Markov $(X_t)_{t\geq 0}$ à deux états $E1$ et $E2$ et justifier la formule suiante :

$$P(X_{t+1} = E_1/X_t = E_1) = 0,99 + 0,01 \cdot 0,25 = 0,9925.$$

2. En déduire la matrice de transition $\mathbb{P}$ de la chaîne de Markov .
3. Tracer un diagramme en points et flèches.
4. Si l'on commence avec une population de 10 arbres de l'espèce $E1$ et 990 de l'espèce $E2$, combien aura-t-on d'arbres de l'espèce $E1$ après une étape, deux étapes ?
5. $\pi_0 = (0,01 \;\; 0,99)$ est-elle une distribution stationnaire pour cette chaîne de Markov ?
6. Reprendre les deux questions précédentes si l'on suppose qu'il y a au depart une proportion de cinq arbres de la première espèce contre trois de la seconde.