Chaînes de Markov et génome

Etienne Pardoux

Leçon 2

5 Chaînes de Markov homogènes irrédutibles

On veut maintenant énoncer l'équivalent de la loi des grands nombres, pour les chaînes de Markov. Il nous faut d'abord ajouter une condition.

Définition 5.1 *On dit que la chaîne de Markov $(X_1, \ldots, X_n)$ de matrice de transition P est irréductible si $\forall x, y \in E$, $\exists k \geq 1$ tel que*

$$(P^k)_{xy} > 0,$$

i.e. si $\exists k$ tel que la chaîne passe avec probabilité non nulle de x à y en k itérations.

Théorème 5.2 *Soit $(X_k,\ k \geq 1)$ une chaîne de Markov à valeurs dans un ensemble fini E, de matrice de transition P irréductible. Alors il existe une unique probabilité π invariante par P, i.e. telle que $\pi = \pi P$, qui vérifie $\pi_x > 0$, $\forall x \in E$. Si $(X_k,\ k \geq 1)$ est une chaîne de Markov de loi initiale π et de matrice de transition P, alors π est la loi de X_k pour tout $k \geq 1$. Enfin on a le théorème ergodique (généralisation de la loi forte des grands nombres) : pour tout $f : E \to \mathbb{R}$,*

$$\frac{1}{n}\sum_{k=1}^{n} f(X_k) \to \sum_{x \in E} f(x)\pi_x \quad p.s.,$$

quelle que soit la loi de X_1.

Si on note à nouveau $\mu(k)$ la loi de X_k, il résulte du théorème ergodique, en prenant l'espérance, que pour toute probabilité $\mu(1)$ sur E,

$$\frac{1}{n}\sum_{k=1}^{n} \mu(k) \to \pi,$$

quand $n \to \infty$. Une question naturelle à se poser est de savoir si l'on a ou non $\mu(n) \to \pi$, quand $n \to \infty$. Ce n'est pas toujours vrai sous les hypothèses ci–dessus. Il faut en outre supposer que

Définition 5.3 *On dit qu'une chaîne de Markov de matrice de transition P irréductible est apériodique ssi pour un couple (x, y) dans $E \times E$ (et alors pour tous les couples, par l'irréductibilité), il existe N tel que $(P^n)_{xy} > 0$, pour tout $n \geq N$.*

Remarque 5.4 *On peut admettre que la chaîne des nucléotides successifs est irréductible et apériodique. Mais les chaînes périodiques ne sont pas que des curiosités mathématiques que l'on ne rencontrerait pas dans des modèles simples. Considérons la chaîne cachée (i.e. dont les valeurs ne sont pas données par la lecture des nucléotides), qui indique dans quelle plage (non codante, codante à l'endroit, codante à l'envers) se trouve le nucléotide que l'on lit. Codons par 0 l'état "non codant", 1 l'état "codant à l'endroit", 2 l'état "codant à l'envers". Quand on est à l'état 0, soit on y reste, soit on passe dans l'état 1, soit on passe dans l'état 2. Il est assez raisonnable de supposer que l'on ne passe pas directement de l'état 1 dans l'état 2 (et vice versa), sans repasser dans l'état 0. Considérons maintenant la chaîne qui décrit la suite des plages visitées, en "gommant" les longueurs de ces plages (donc en particulier la matrice de transition correspondante P a tous ses termes diagonaux nuls). Alors*

$$(P^n)_{00} = \begin{cases} 0, & \text{si } n \text{ est impair} \\ 1, & \text{si } n \text{ est pair,} \end{cases}$$

et la chaîne que nous venons de décrire est périodique.

Théorème 5.5 *Soit $(X_k, k \geq 1)$ une chaîne de Markov de matrice de transition P irréductible et apériodique. On désigne pour tout $n \geq 1$ par $\mu(n)$ la loi de probabilité de X_n. Alors $\mu(n) \to \pi$ quand $n \to \infty$.*

Donc dans le cas irréductible et apériodique on peut admettre que, en tout cas à partir d'un certain rang, la suite des X_k est identiquement distribuée, de loi commune l'unique probabilité invariante π de la chaîne.

6 Chaîne de Markov réversible

Soit $(X_1, \ldots, X_n)$ une chaîne de Markov de matrice de transition P. Posons $\hat{X}_k = X_{n+1-k}$. C'est un exercice facile sur les probabilités conditionnelles de montrer que la *suite retournée* $(\hat{X}_1, \ldots, \hat{X}_n) = (X_n, \ldots, X_1)$ est une chaîne de Markov. En général cette nouvelle chaîne de Markov n'est pas homogène. $(\hat{X}_1, \ldots, \hat{X}_n)$ est une chaîne homogène si $(X_1, \ldots, X_n)$ est initialisée avec sa probabilité invariante π. Dans ce cas, d'après la formule de Bayes, la matrice $\hat{P}$ de la chaîne retournée est donnée par

$$\hat{P}_{xy} = \frac{\pi_y P_{yx}}{\pi_x}.$$

On dit que la chaîne $(X_1, \ldots, X_n)$ est *réversible* si $\hat{P} = P$, ce qui est équivalent à ce que la *relation d'équilibre ponctuel* (en Anglais *detailed balance equation*) suivante soit satisfaite

$$\pi_x P_{xy} = \pi_y P_{yx}, \quad \forall x \neq y,$$

autrement dit si la quantité $\pi_x P_{xy}$ est symmétrique en x, y.

Remarque 6.1 *Etant donnée une chaîne irréductible de matrice de transition P, cette chaîne possède une unique probabilité invariante, qui forcément satisfait la relation $\pi P = \pi$. Le couple (π, P) satisfait ou non la relation d'équilibre ponctuel (i.e. toutes les chaînes irréductibles ne sont pas réversibles). Pour construire une chaîne irréductible et non réversible, il suffit de choisir une matrice P irréductible, telle pour un certain couple $x \neq y$, $P_{xy} = 0 < P_{yx}$.*

D'un autre coté, si P est une matrice de transition et π une probabilité sur E, telles que la relation d'équilibre ponctuel soit satisfaite, alors π est une probabilité invariante, comme on le vérifie en sommant la relation d'équilibre ponctuel par rapport à x (ou à y).

7 Digression : chaînes en temps continu et chaînes sur les arbres

On va faire une digression sur les chaînes de Markov sur les arbres, qui sont couramment utilisées comme modèle en phylogénie. On verra au passage l'utilité de la notion de chaîne réversible. Dans ce cadre, la notion naturelle est celle de chaîne de Markov en temps continu, que nous allons introduire.

7.1 Chaîne de Markov en temps continu

Lorsque l'on modélise une séquence de nucléotides, on fait bien sûr appel à une collection indexées par les entiers de variables aléatoires . Lorsque l'on étudie l'évolution au cours du temps de cette même séquence, ou pour simplifier d'un site de cette séquence, il est naturel d'indexer les variables aléatoires par un paramètre t qui varie dans $[0, T]$.

On appelle *chaîne de Markov en temps continu* ou *processus markovien de sauts* une collection $(X_t,\ 0 \leq t \leq T)$ de v.a. à valeurs dans un ensemble fini E (E pourraît être dénombrable, mais le cas fini nous suffira), telle que pour tout n, $0 \leq t_1 < t_2 < \cdots < t_n \leq T$, $(X_{t_1}, \ldots, X_{t_n})$ est une chaîne de Markov. On ne considèrera que le cas homogène, où la transition de X_s à X_t $(s < t)$ ne dépend que de $t - s$.

Pour motiver la description que nous allons donner des chaînes de Markov en temps continu, revenons au cas d'une chaîne en temps discret de matrice de transition P. Considérons un état $x \in E$, tel que $p = P_{xx} > 0$. Alors la loi du temps de séjour de la chaîne à l'état x est la loi géométrique de paramètre p. En effet, si $X_1 = x$, et si $S = \inf\{k,\ X_{k+1} \neq x\}$,

$$\begin{aligned} \mathbb{P}(S = \ell | X_1 = x) &= \mathbb{P}(X_2 = x, \ldots, X_\ell = x, X_{\ell+1} \neq x | X_1 = x) \\ &= p^{\ell-1}(1-p). \end{aligned}$$

Cette propriété est intimement liée au fait que la propriété de Markov impose que la loi de S soit "sans mémoire", au sens où

$$\mathbb{P}(S > k + \ell | S > k) = \mathbb{P}(S > \ell),$$

propriété qui est caractéristique de la loi géométrique. Les lois sur $\mathbb{R}_+$ qui vérifient une propriété analogue sont les lois exponentielles, et il n'est pas difficile de se convaincre que les temps de séjour dans les différents états d'une chaîne de Markov en temps continu suivent des lois exponentielles. On dit que la v. a. S suit la loi exponentielle de paramètre $\lambda > 0$ si

$$\mathbb{P}(S > t) = \exp(-\lambda t), \quad \forall t > 0.$$

Une chaîne de Markov en temps continu est entièrement décrite par son *générateur infinitésimal* Q, qui est une matrice $d \times d$ (avec $d = |E|$), dont tous les termes hors diagonaux sont positifs ou nuls, et les termes diagonaux sont strictement négatifs (ou nul si l'état correspondant est absorbant). Supposons que la chaîne parte de x (i.e. $X_0 = x$). La chaîne reste à l'état x pendant un temps aléatoire T_1, de loi exponentielle de paramètre $q_x = -Q_{xx}$. L'état dans lequel la chaîne aboutit à l'issue du saut qui se produit à l'instant T_1 est choisi indépendemment de la valeur de T_1, suivant la loi de probabilité $(q_x^{-1} Q_{xy},\ y \in E \backslash \{x\})$, et ainsi de suite... On a alors

$$\mathbb{P}(X_t = y | X_0 = x) = P(t)_{xy},$$

où la matrice $P(t)$ est donnée par

$$P(t) = \exp(tQ) = \sum_{k=1}^{\infty} \frac{t^k}{k!} Q^k.$$

La suite des valeurs à l'issue des sauts est une chaîne de Markov appelée la *chaîne incluse*, de matrice de transition P dont les termes diagonaux sont nuls et les termes hors diagonaux sont donnés par $P_{xy} = q_x^{-1} Q_{xy}$. La chaîne en temps continu est dite irréductible ssi sa chaîne incluse l'est, et alors la chaîne en temps continu possède une unique probabilité invariante π, qui est la solution de l'équation $\pi Q = 0$. On a encore le théorème ergodique :

$$\frac{1}{t} \int_0^t f(X_s) ds \to \sum_x f(x) \pi_x \text{ p. s., quand } t \to \infty.$$

Enfin une chaîne en temps continu irréductible est *réversible* ssi la relation d'équilibre ponctuel

$$\pi_x Q_{xy} = \pi_y Q_{yx}, \quad \forall x \neq y$$

est satisfaite. Dire qu'une chaîne en temps continu est réversible, c'est dire que $(X_t,\ 0 \leq t \leq T)$ et $(\hat{X}_t = X_{T-t},\ 0 \leq t \leq T)$ ont même loi, donc évoluent suivant le même mécanisme.

7.2 Chaîne de Markov sur un arbre avec racine

La chaîne part de la racine (qui joue le rôle de l'instant initial 0) dans un certain état, disons x. Elle évolue jusqu'au premier noeud qui se trouve à la distance r de la racine, comme une chaîne en temps continu pendant l'intervalle de temps r. Notons y l'état de la chaîne en ce noeud. Sur chaque branche qui part de ce noeud court une chaîne en temps continu, partant de y, de telle sorte que les différentes chaînes sur les différentes branches sont mutuellement indépendantes, jusqu'au prochain noeud, et ainsi de suite...

7.3 Chaîne de Markov sur un arbre sans racine

Dans le cas d'un arbre sans racine, la même construction peut être faite en partant de n'impote quel point de l'arbre (soit d'un noeud, soit d'un point arbitraire d'une branche arbitraire), à condition d'utiliser une chaîne réversible (i.e. une chaîne irréductible dont le couple générateur infinitésimal – probabilité invariante satisfait la relation d'équilibre ponctuel).

8 Statistique des chaînes de Markov homogènes $M1$

8.1 Estimation de la mesure invariante

On sait que $\frac{1}{n}\sum_1^n \mathbf{1}_{\{X_k=x\}} \to \pi_x$ p.s. quand $n \to \infty$. Donc

$$\frac{1}{n}\sum_1^n \mathbf{1}_{\{X_k=x\}}$$

est un estimateur de π_x, $x \in E$.

8.2 Estimation de la matrice de transition

On va montrer que

$$\frac{\sum_1^n \mathbf{1}_{\{X_k=x,X_{k+1}=y\}}}{\sum_1^n \mathbf{1}_{\{X_k=x\}}} \to P_{x,y}, n \to \infty.$$

Notons tout d'abord que la fraction ci–dessus vaut

$$\frac{\frac{1}{n}\sum_1^n \mathbf{1}_{\{X_k=x,X_{k+1}=y\}}}{\frac{1}{n}\sum_1^n \mathbf{1}_{\{X_k=x\}}}.$$

Il suffit de considérer le numérateur. Plus précisément, il vaut

$$\frac{1}{n}\sum_{k \text{ pair}} \mathbf{1}_{\{X_k=x,X_{k+1}=y\}} + \frac{1}{n}\sum_{k \text{ impair}} \mathbf{1}_{\{X_k=x,X_{k+1}=y\}}$$

Considérons par exemple la quantité

$$\frac{2}{n} \sum_{k \text{ pair}, \, 1 \leq k \leq n} \mathbf{1}_{\{X_k = x, X_{k+1} = y\}}$$

On remarque que la chaîne $\{(X_1, X_2), (X_3, X_4), (X_5, X_6), \ldots\}$ est une chaîne de Markov à valeurs dans $E \times E$, de matrice de transition de (x, y) à (x', y') donnée par $P_{y,x'} P_{x'y}$, et de probabilité invariante $\tilde{\pi}_{xy} = \pi_x P_{xy}$.

Donc

$$\frac{2}{n} \sum_{k \text{ pair}, \, 1 \leq k \leq n} \mathbf{1}_{\{X_k = x, X_{k+1} = y\}} \to \pi_x P_{xy}.$$

Finalement

$$\frac{\sum_1^n \mathbf{1}_{\{X_k = x, X_{h+1} = y\}}}{\sum_1^n \mathbf{1}_{\{X_k = x\}}} \to P_{x,y}$$

p.s., quand $n \to \infty$.

9 Statistique des chaînes de Markov Mk

Pour simplifier, on va se contenter de décrire le modèle $M2$. Dans ce cas, ce qui remplace la matrice P est une matrice de transition de $E \times E$ dans E, qui donne la loi de probabilité de X_{k+1}, sachant le couple (X_{k-1}, X_k). Dans le cas $E = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$, on a donc une matrice transition à 16 lignes (indexées par les dinucléotides $\{\mathtt{aa}, \mathtt{ac}, \ldots, \mathtt{gt}, \mathtt{tt}\}$) et 4 colonnes (indexées par $\{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$).

Remarque 9.1 *On peut aussi se ramener à un modèle $M1$ sur l'espace d'état $E \times E$, puisque si $(X_1, X_2, \ldots, X_n)$ est une chaîne de Markov d'ordre 2 à valeurs dans E, $((X_1, X_2), (X_2, X_3), \ldots, (X_{n-1}, X_n))$ est une chaîne de Markov d'ordre 1 à valeurs dans $E \times E$. On se ramène à une matrice de transition carrée, et on peut introduire la notion de probabilité invariante...*

On estime la probabilité de transition $P_{xy,z}$ à l'aide de la quantité

$$\frac{\sum_{k=1}^n \mathbf{1}_{\{X_k = x, X_{k+1} = y, X_{k+2} = z\}}}{\sum_{k=1}^n \mathbf{1}_{\{X_k = x, X_{k+1} = y\}}},$$

qui converge p.s. vers $P_{xy,z}$ quand $n \to \infty$. Remarquons que cette statistique inclut le décompte des trinucléotides, donc en particulier des codons, ce qui fait que les chaînes d'ordre 2 sont très utilisées pour modéliser les régions codantes de l'ADN.