République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université Mohammed Seddik Ben Yahia - Jijel

Faculté des Sciences Exacte et Informatique
Département de Mathématique

Mémoire de fin de cycle

Présenté pour l'obtention du diplôme de

Master

Spécialité : Mathématiques.

Option : Probabilité et Statistique.

Thème

Modèles de Markov Cachés : Application en Biologie

Présenté par :

Kimouche Achouak

Devant le jury composé de :

Sellami Nawel	M.A.A Université Mohammed Seddik Ben Yahia - Jijel	Président
Cheraitia Hassen	M.C.B Université Mohammed Seddik Ben Yahia - Jijel	Encadreur
Yakoubi Fatma	M.A.A Université Mohammed Seddik Ben Yahia - Jijel	Examinateur

Promotion **2020/2021**

♡ R*emerciements* ♡

Quelles mots en préambule de cette étude, qui met un point d'orgue à une année riche et intense

Tout d'abord, nous remercions **Allah** le tout puissant pour son aide et pour nous avoir guidé pour mener à bien ce travail

La première personne que nous tenons à remercier est notre encadreur **M**r

∗**Cheraitia Hassen**∗

pour l'orientation, la confiance, la patience qui ont constitué un apport considérable sans lequel ce travail n'aurrait pas pu étre mené au bon port

Un grand merci également aux membres du jury, **M**me la présidente **Sellami Nawel** ainsi que l'examinatrice **M**me **Yakoubi Fatma** pour l'honneur qu'elles nous ont fait en acceptant de juger notre mémoire

Nos vifs remerciements vont à tous enseignants qui nous ont suivi tout au long de nos 5 ans d'études à l'université

Enfin, nous remercions toutes les personnes qui auraient contribué d'une manière ou d'une autre à la réalisation de ce travail.

Dédicace

Je dédie ce travail :

A ma chère mère ***Saliha*** *, pour son amour, ses encouragements et sacrifiees .*

A mon cher père ***Ali*** *, pour son soutien , son affection et la confiance q'uil m'accordé .*

A mon âme soeur ***Nada*** *.*

A mes frères : ***Aniss , Hamza , Fadi*** *.*

A mon grand-père ***"Cherif "*** *qui je souhaite une bonne santé .*

A tous les nombres de ma famille et surtout mon oncle ***"Abdelali"*** *, ma tante* ***"Nassima"****.*

A tous mes amis et mes camarades .

A promo prob-stat 2020/2021 .

Et tous ceux qui m'aiment .

K.Achouak

Table des matières

Liste des tableaux

Table des figures

Notations

- CM : chaînes de Markov .
- CMTD ; chaînes de Markov à temps discret .
- CMTC :chaînes de Markov à temps continu .
- ssi : si et seulement si .
- c-à-d : - c'est à dire .
- HMM : hidden Markov chain.
- tq : tels que
- E : espérance .
- $E(x/y)$: espérance conditionnelle (l'espérance de x sachant y) .

Résumé

Les modèles de Markov cachés (CMC) du nom du mathématicien russe Andrey Andreyevich Markov, qui a développé une grande partie de leur théorie statistique .Les chaînes de Markov cachées sont des modèles statistiques permettant de capturer des informations cachées à partir des symboles séquentiels observés.

Depuis les années 80,les modèles de Markov cachés ont été utilisé pour résoudre divers problèmes d'analyse de séquence biologiques notamment l'alignement de séquences , la prédiction des gènes , la prédiction de la structure des protéines et bien d'autre.

Dans ce mémoire , une séquence de nucléotides d'ADN a été modélisée par un modèle de Markov caché , cette séquence a été considérée comme une suite (émissions) générée par l'un de deux états cachés : AT-rich et GC-rich . L'ensemble des algorithmes de base de CMC ont été appliquées pour l'estimation des probabilités.

Abstract

Hidden Markov models (HMM) , named after the Russian mathematician Andrey Andreyevich Markov , who developed much of relevant statistical theory .HMM are statistical models to capture hidden information from observable sequential symbols.

Since 1980 , HMM have been used to resolve various problems of biological sequence analysis : sequence alignment , gene prediction ,protien structure prediction and many others.

In this thesis , an AND sequence of nucleotides has been modeled by a HMM , this sequence was considered as a series of outputs (or emission) generated by one of two intemal(hidden) states AT-rich and GC-rich . A set of HMM inferance algorithms were applied for estimation the probabilities .

Intoduction Général

Dans la théorie de probabilité ,les processus stochastique sont des outils très importants qui permettent de modéliser des différents phénomènes dans de nombreux domaines.

Après les travaux de Markov , les processus devient notés Markovien. Ces dernier sont des suites de variables aléatoires qui ne sont pas indépendantes et identiquement distribuées . Ils sont basés sur un principe fondamental qui est " le futur dépend seulement de présent , pas de passé".Les modèles de Markov s'intéresse qu'à l'étude des états observables.Ces processus ne sont pas suffisants dans des situations où ces états ne sont pas directement observables. Pour cela, on s'intéresse aux chaînes de Markov cachés(CMC).

Les modèles de Markov cachés ont été introduits et étudiés ou début des années 1970, ils sont utilisés pour la première fois dans la reconnaissance vocale et ont été appliqués avec succès à l'analyse des séquences biologiques depuis la fin des années 80.

Un modèle CMC utilise un processus de Markov qui contient des paramètres cachés et observés. Dans ce modèle, les paramètres observés sont utilisés pour identifier les paramètres cachés, il s'agit d'un processus doublement stochastique. Dans ce type de modèle, les variables cachées contrôlent la mécanisme de génération des données , ainsi les attributs sont directement affectés par les variables cachées. Le modèle CMC utilise des algorithmes d'inférences pour estimer la probabilité d chaque état à chaque position le long de données.

Les chaîns de Markov cachés ont été utilisées pour résoudre divers problèmes d'analyse de séquences biologiques (Won et al 2007, Durbin et al 1998, Pachter et al.2002) : l'alignement des séquences par paires et multiples, la prédiction de la structure des protéines, l'identification des ARNnc.....

Nous visons au cours de ce mémoire à mettre en place d'une modélisation d'une séquence d'ADN comportant 04 nucléotides A (Adénine),T (Thymine),G (Guanine),C (Cytosine) par les modèles de Markov cachés.

Afin d'atteindre cet objectif, nous allons répartir notre mémoire en trois chapitres :

Dans le premier chapitre intitulé " Les chaînes de Markov ", nous commencerons par des diverses définitions tels que les processus stochastique et ces différentes types. Puis, nous présenterons la théorie des processus markoviens à temps discrets et à temps continus avec quelques exemples.

Dans le deuxième chapitre qui à pour titre ; " Les chaînes de Markov cachées à temps discret " présente de manière formelle les CMC et les trois problèmes fondamentaux : " Évaluation, Décodage et Apprentissage" qu'ils permettent de résoudre avec les algorithmes de "Forward-Backward, de Viterbi et de Baum-Welch".

Finalement , **dans le troisième chapitre** intitulé "Simulations et Applications" , est la mise en application de tout ce qui est établi dans les chapitres précédents . On va appliquer les CMC et leurs propriétés sur une séquence d'ADN .

Chapitre 1

Les chaînes de Markov

1.1 Processus stochastique

Très souvent , lorsque nous étudions un phénomène qui dépend du hasard ; il y a lieu de prendre en compte l'évolution de ce phénomène au cours du temps . Nous avons vu que chaque observation d'un phénomène réel est modélisée par une variable aléatoire réelle ; l'étude de cette dernière évoluant dans le temps va donc être modélisée par une famille de variables aléatoires appelée processus stochastique .

Définition 1.1.1.

Un processus stochastique est une famille de variable aléatoires $\{X(t), t \in T\}$ *; définie sur un espace de probabilité commun (* $\Omega, \mathcal{B}, \mathbb{P}$*) où l'ensemble* T *est un sous-ensemble du réel à valeur dans E(espace d'état).*

- L'ensemble T *représente le temps où bien le domaine d'évolution .*

*** $T = \mathbb{N}$ *; on parlera alors de processus stochastique de temps discret .*

*** $T = \mathbb{R}$ *; on dira alors que* $X(t), t \geq 0$ *est un processus stochastique à temps continu .*

Remarque 1.1.1.

• *Les processus aléatoires peuvent être classés selon la dénombrabilité ou non de l'espace d'état et du domaine d'évolution , le tableau ci-dessous représente les quatre cas :*

T / E	Discret	Continu
Discret	processus à temps discret à espace d'état discret .	processus à temps continu à espace d'état discret .
Continu	processus à temps discret à espace d'état continu .	processus à temps continu à d'état continu .

TABLE 1.1 – classification des processus aléatoires

Définition 1.1.2.

Un processus aléatoire peut définie comme une application de (Ω, T) *dans E lorsque :*

$$X : (\Omega, T) \longrightarrow E$$
$$(\omega, t) \longrightarrow X(\omega, t)$$

Remarque 1.1.2.

- *Pour* $t \in T$ *fixé ,* $X(., t)$ *est une variable aléatoire réelle , ie :*

$$X : \Omega \longrightarrow E$$
$$\Omega \longrightarrow X(\omega)$$

- *Pour* $\omega \in \Omega$ *fixé ; donc l'application* $X(\omega, .)$ *définie comme suite :*

$$X : T \longrightarrow E$$
$$t \longrightarrow X_t(\omega)$$

est une trajectoire de ω *.*

1.1.1 Les différents processus aléatoire :

Il ya trois éléments qui fait une différence des processus aléatoire qui sont :

1- Le domaine d'évolution T.

2- Espace d'état E.

3- Les relations de dépendances entre les variables aléatoire .

quelque exemples des processus aléatoire :

1/- Processus de comptage :

Un processus de comptage $(N_t)_{t\in T\subset\mathbb{R}}$ est un processus croissant

(si $s < t$ alors $N_s < N_t$) à valeur dans E= $\mathbb{N}$.

2/- Processus accroissement indépendants :

Un processus croissant $(X_t)_{t\geq 0}$ est dit à accroissement indépendant si pour tout n$\in \mathbb{N}^*$, et pour tous $t_1, t_2, ..., t_n$ tel que $t_1 < t_2 < ... < t_n$; les accroissements $X_{t_1} - X_{t_0}, X_{t_2} - X_{t_1}, ...,$ $X_{t_n} - X_{t_{n-1}}$ sont des variables aléatoire indépendantes .

3/- Processus homogène dans le temps :

Le processus $(X_t)_{t\geq 0}$ est dit homogène , si pour tout $t \in T$ et tout s , la loi de $X_{t+s} - X_s$ ne dépend pas de s .

4/- Processus strictement stationnaire :

On dit que $(X_t, t \in T)$ est strictement stationnaire si les fonctions de répartition des familles de variables aléatoires $(X_{t_1}, X_{t_2}, ..., X_{t_n})'$ et $(X_{t_1+h}, X_{t_2+h}, ..., X_{t_n+h})'$ sont les mêmes $\forall h \in T$.

5/- Processus faiblement stationnaire (stationnaire en seconde-ordre) :

Il est est dit faiblement stationnaire s'il vérifié les propriétés suivants :

• $\forall t \in T; E(X_t) = m < \infty$.

• $\forall t \in T$; la variance existe ie : $E(X_t^2) =< \infty$

• $cov(X_t, X_{t+h}) = E(X_t X_{t+h}) - E(X_t)E(X_{t+h}) \quad , \forall h \in T$

Remarque 1.1.3.

• *Pour un processus du 2^{nd} ordre .la stationnarité strict implique la stationnarité faible (la réciproque est fausse).*

6/- Processus de Poisson :

Un processus aléatoire N_t ; $t \geqslant 0$ à valeur entières est un processus de Poisson de paramétré $\lambda > 0$ si :

◦ N_t est un processus de comptage à accroissements indépendants et stationnaires .

◦ la variable N_t suit la loi de Poisson de paramètre λt

$$\forall n \geqslant 0, \mathbb{P}(N_t = n) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}$$

1.2 Chaîne de Markov à temps discret (CMTD)

Définition 1.2.1.

Une chaîne de Markov à temps discret est un processus stochastique à temps discret définie sur un espace dénombrable et vérifiant la propriété de Markov :

$\forall j, i, i_{n-1}, ..., i_0 \in E$

$$\mathbb{P}(X_{n+1} = j / X_0 = i_0, ..., X_n = i) = \mathbb{P}(X_{n+1} = j / X_n = i)$$

Définition 1.2.2. ***(Chaîne de Markov homogène) :***

Chaîne de Markov à temps discret est dit homogène si pour tout $(i, j) \in E^2$ et tout instant n.

$$\mathbb{P}(X_{n+1} = j / X_n = i) = \mathbb{P}(X_{n+1+k} = j / X_{n+k} = i), \forall k \geq 0$$

donc l'homogénéité d'une chaîne de Markov précise donc la probabilité de l'état i à l'état j reste la même à travers le t temps .

Définition 1.2.3. ***(Probabilité et matrice de transition) :***

- ***En 1 étape :***

On définie la probabilité de transition de l'état i à l'état j la probabilite :
$\forall n \geq 0, \forall (i, j) \in E^2$

$$\mathrm{p}_{ij} = \mathbb{P}(X_{n+1} = j / X_n = i) = \mathbb{P}(X_1 = j / X_0 = i_0)$$

La matrice $\mathrm{P} = \begin{pmatrix} \mathrm{p}_{00} & \mathrm{p}_{01} & .. & .. \\ \mathrm{p}_{10} & \mathrm{p}_{11} & .. & .. \\ .. & .. & .. & .. \\ & & & \end{pmatrix}$

dont les coefficients sont les probabilités de transition p_{ij} est appelée matrice de transition (passage) de la chaîne .C'est une matrice finie ou dénombrable , suivant que l'ensemble des états fini ou dénombrable .

- ***En K étapes :***

La probabilité conditionnelle d'aller de i à j en k étapes exactement est :

$$\mathrm{p}_{ij}^{(k)} = \mathbb{P}(X_{n+k} = j / X_0 = i_0) = \mathbb{P}(X_{n+k} = j / X_n = i) \; ; \forall n \geq 0$$

Cette probabilité est indépendant de n car le processus est homogène est appelée la probabilité de transition en k étapes de i à j.

La matrice $\mathrm{P}^{(k)}$ *dont l'élément* (i,j) *est égale* $\mathrm{p}_{ij}^{(k)}$ *est appelée la matrice de transition en k étapes.*

Remarque 1.2.1.

- *On a évidement :* $\mathrm{p}_{ij}^{(1)} = \mathrm{p}_{ij}$.
- *Pour k=0 ; on a* $\mathrm{P}^{(0)} = \mathbf{I}$.

Définition 1.2.4. ***(Matrice stochastique) :***

Une matrice carre est dit stochastique si :

1- tout les coefficient sont positifs ou nuls ie : $\forall (i,j) \in E^2$; $\mathrm{p}_{ij}^{(k)} \geq 0$.

2- les coefficients du chacune des lignes somment à 1 ie :$\forall i \in E; \sum_{j\in E} \mathrm{p}_{ij}^{(k)} = 1$.

proprieté 1.2.1.

a)- la matrice $\mathrm{P}^{(k)}$*est une matrice stochastique .*

b)- P *admet 1 comme valeur propre .*

c)- On peut associer à cette valeur propre (1) un vecteur noté V *tout les coordonnées sont égale à 1.*

Démonstration.

a)- 1- les nombre $\mathrm{p}_{ij}^{(k)}$ sont des probabilité donc positifs .

2- On a :

$$\begin{aligned}
\sum_{j\in E} \mathrm{P}_{ij}^{(k)} &= \sum_{j\in E} \mathbb{P}(X_{n+k} = j / X_n = i) \\
&= \sum_{j\in E} \frac{\mathbb{P}(X_{n+k} = j \ , \ X_n = i)}{\mathbb{P}(X_n = i)} \\
&= \sum_{j\in E} \frac{\mathbb{P}(X_n = i)}{\mathbb{P}(X_n = i)} \\
&= 1.
\end{aligned}$$

donc la matrice e transition est stochastique.

b)- Soit $U = (1,1,...,1)^t \in \mathbb{R}^n$ Alors, la deuxième condition qui définit la matrice P équivaut à $AU = A$.

Donc 1 est bien valeur propre de A, et U est un vecteur propre associé.

c)- On considérant comme un vecteur colonne on a :

P.V = V ssi pou tout i∈ E , la relation satisfaite ,donc il suffit de prendre $v_i = 1 \forall i \in E$.

Théorème 1.2.1.

Pour tout $k \geq 0$ *;* $\mathrm{P}^{(k)} = \mathrm{P}^k$

Démonstration.

On a :

$$\begin{aligned}
&\mathrm{p}_{ij}^{(k)} = \mathbb{P}[X_k = j / X_0 = i] \\
&= \sum_{m \in E} \mathbb{P}[X_k = j / X_{k-1} = m] \mathbb{P}[X_{k-1} = m / X_0 = i] \\
&= \sum_{m \in E} \mathrm{p}_{mj}^{(1)} \mathrm{p}_{im}^{(k-1)} \\
&= \sum_{m \in E} (\mathrm{p}^{\mathrm{k}-1})_{im} (\mathrm{p})_{mj} \\
&= (\mathrm{p}^{\mathrm{k}})_{ij}
\end{aligned}$$

Corollaire 1.2.2. ***(Équation de Chapman -Kolmogorov) :***

pour tout $(i, j) \in E^2$*, et tout couple* (n, m) *d'entier positifs .*

On a l'identité : $\mathbb{P}[X_{n+m} = j / X_0 = i] = \sum_{k \in E} \mathbb{P}[X_n = k / X_0 = i] \mathbb{P}[X_m = j / X_0 = k]$.

où encore :

$$\mathrm{p}_{ij}^{(n+m)} = \sum_{k \in E} \mathrm{p}_{ik}^{(n)} \mathrm{p}_{kj}^{(m)}$$

et aussi en notation matricielle :

$$\mathrm{P}^{(n+m)} = \mathrm{P}^{(n)} \mathrm{P}^{(m)}$$

Démonstration. Soient k l'état de la chaîne au temps m.

$$\begin{aligned}
\mathrm{p}_{ij}^{(n+m)} &= \mathbb{P}[X_{n+m} = j / X_0 = i] \\
&= \sum_{k \in E} \mathbb{P}[X_{n+m} = j, X_n = k / X_0 = i] \\
&= \sum_{k \in E} \frac{\mathbb{P}[X_{n+m} = j, X_n = k, X_0 = i]}{\mathbb{P}[X_0 = i]} \\
&= \sum_{k \in E} \frac{\mathbb{P}[X_{n+m} = j / X_n = k, X_0 = i]\mathbb{P}[X_n = k / X_0 = i]\mathbb{P}[X_0 = i]}{\mathbb{P}[X_0 = i]} \\
&= \sum_{k \in E} \mathbb{P}[X_{n+m} = j / X_n = k, X_0 = i]\mathbb{P}[X_n = k / X_0 = i] \\
&= \sum_{k \in E} \mathbb{P}[X_n = k / X_0 = i]\mathbb{P}[X_{n+m} = j / X_n = k] \\
&= \sum_{k \in E} \mathrm{p}_{ik}^{(n)} \mathrm{p}_{kj}^{(m)}
\end{aligned}$$

Définition 1.2.5. ***(Graphe de transition) :***

La matrice de transition P *d'une chaîne de Markov peut être représentée par un graphe orienté* **G** *; ses sommets sont les états de la chaîne et il y a une flêche ,étiquetée* p_{ij} *, être les sommes i et j ssi* $\mathrm{p}_{ij} \geq 0$.

Si E est fini cette représentation est particulièrement utile et parlant.

1.2.1 Classification des états :

Les états d'une chaîne de Markov se répartissent en classe que l'on définit à partir de la matrice de transition .

Relations de communication entre états :

Définition 1.2.6. ***(Accessibilité) :***

On dit que l'état j est accessible à partir de l'état i , s'il existe un entier $n \geq 0$ *tq* $\mathrm{p}_{ij}^{(n)}$ *, on note :* $i \rightsquigarrow j$ *ou bien l'état j est accessible depuis l'état i ,s'il existe dans* **G** *, au moins un chemin du i à j .*

Remarque 1.2.2.

- *Tout état j est accessible depuis lui même .*

proprieté 1.2.2.

La relation d'accessibilité entre états est réflexive et transitive .

Démonstration.

- **Réflexive** : comme $\mathrm{p}_{ii}^{(0)} = \mathbb{P}(X_0 = i/X_0 = i) = 1$, pour tout état i, on a $i \longrightarrow i$.

- **Transitive** : on suppose que : $i \rightsquigarrow k$ et $k \rightsquigarrow j$ alors :

$\exists m, n \geqslant 0$ tels que : $\mathrm{p}_{ik}^{(n)} > 0$ et $\mathrm{p}_{kj}^{(m)} > 0$.

d'après l'équation de Chapman-Kolmogorov :

$$\mathrm{p}_{ij}^{(n+m)} = \sum_{k \in E} \mathrm{p}_{ik}^{(n)} \mathrm{p}_{kj}^{(m)} > 0$$

donc ; $i \rightsquigarrow j$.

Définition 1.2.7. ***(États communiquants) :***

On dit que i et j communiquant et l'on écrit : $i \leftrightsquigarrow j$, si on a à la fois $i \rightsquigarrow j$ et $j \rightsquigarrow i$.

proprieté 1.2.3.

la relation de communication entre états est une relation d'équivalence .

Remarque 1.2.3.

- *Tout état d'une chaîne de Markov communique avec lui même puisque $\forall i \in E$, $\mathrm{p}_{ii}^{(0)} = 1$.*
- *Un état est appelé état de retour ,s'il existe $n \neq 1$ tel que $\mathrm{p}_{ii}^{(n)} \neq 0$.*
- *Il existe des états i tel que $\mathrm{p}_{ii}^{(n)} = 0$, ils sont appelés états de retour .*
- *L'ensemble E d'état se partitionne en classe d'équivalence , disjointes et non vide , dit classe indécomposables . Si C_1 et C_2 sont deux classe distinctes de .En revanche tous les états d'une même classe communiquent .*
- *Certaines classes peuvent ne compter qu'un seul élément ; ces sont les singletons ; on mentionne par exemple :*

- *Un état de retour* $\mathrm{p}_{ii}^{(0)} = 1$; $\mathrm{p}_{ii}^{(n)} = 0 \ \forall n \geq 1$.
- *Un état absorbant* $\mathrm{p}_{ii}^{(0)} = 1$; $\mathrm{p}_{ii}^{(n)} = 1 \ \forall n \geq 1$.

Définition 1.2.8. ***(Irréductible) :***

Une chaîne de Markov est dit irréductible si elle ne contient qu'une seul classe d'équivalence. Autrement dit si tous les états communiquent ente eux .

Définition 1.2.9. ***(Absorbant) :***

Un état i d'une CM est dit absorbant , si la chaîne ne peut plus quitter cet état une fois qu'elle y est entrée ,en autre terme si $\mathrm{P}_{ii} = 1$.

Remarque 1.2.4.

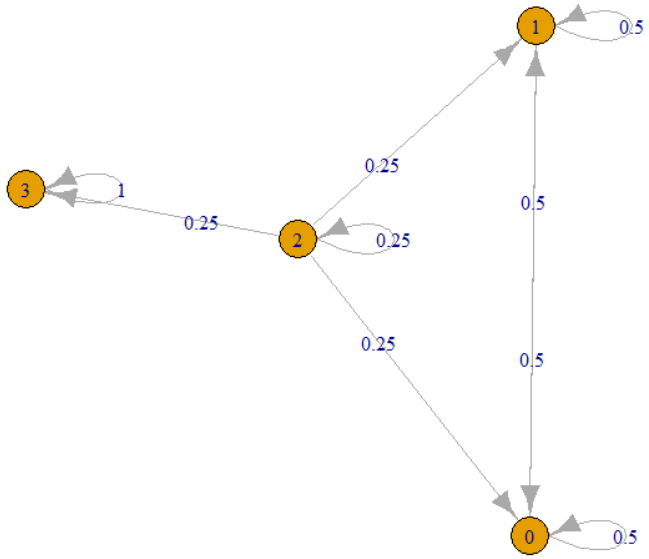
• *On dit que CM est absorbant si elle comprend au moins un état absorbant et que on peut passe de n'importe état à un état absorbant .*

Example 1.2.1.

Considerons la CM avec E={ 0,1,2,3} et :

$$\mathrm{P} = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

le graphe est :



- *La chaîne comporte 3 classe : {0,1},{2} ,{3} .*
- *L'état 3 est absorbant ($\mathrm{P}_{33} = 1$) .*

État récurrents et transients :

Définition 1.2.10. ***(Temps d'atteinte) :***

Soit i un état quelconque dans l'ensemble E de la chaîne de Markov $\{X_n, n \in \mathbb{N}\}$; on appelle temps d'atteinte (ou temps de premier passage à l'état i) , la variable aléatoire T_i définie par :

$$T_i = inf\{n \geq 1; X_n = i\}.$$

Considérons deux états i et j dans l'ensemble E. Notons f_{ij} la probabilité que ,parant de i; la chaîne passe au moins une fois par l'état j; $f_{ij} = \mathbb{P}(T_j < \infty / X_0 = i)$.

Soit $f_{ij}^{(n)}$ la probabilité que parant de l'état i la chaîne aille , pour la première fois , à l'état j l'instant n; ie : $f_{ij}^{(n)} = \mathbb{P}(T_j = n / X_0 = i) = \mathbb{P}(X_n = j, X_k \neq j \forall k = 1, ..., n-1 / X_0 = i)$.

On pose avec convention $f_{ij}^{(0)} = 0$.

Théorème 1.2.3.

Pour tout i et j et tout $n \geq 1$ on a :

$$\mathrm{p}_{ij}^{(n)} = \sum_{k=0}^{n} f_{ij}^{(n)} \mathrm{p}_{jj}^{(n-k)} \ .$$

Démonstration.

$\forall (i,j) \in E^2, ona :$

$$\begin{aligned}
\mathrm{p}_{ij}^{(n)} &= \mathbb{P}(X_n = j / X_0 = i) \\
&= \mathbb{P}(X_n = j; \bigcup_{k=1}^{n} (T_j = k) / X_0 = i) \\
&= \sum_{k=1}^{n} \mathbb{P}((X_n = j); (T_j = k) / X_0 = i) \\
&= \sum_{k=1}^{n-1} \mathbb{P}((X_n = j); (T_j = k) / X_0 = i) + f_{ij}^{(n)} \\
&= \sum_{k=1}^{n-1} \mathbb{P}((X_n = j) / (T_j = k); X_0 = i) \mathbb{P}((T_j = k) / X_0 = i) + f_{ij}^{(n)} \\
&= \sum_{k=1}^{n-1} \mathbb{P}((X_n = j) / (X_k = j)) \mathbb{P}((T_j = k) / X_0 = i) + f_{ij}^{(n)} \\
&= \sum_{k=1}^{n-1} \mathrm{p}_{jj}^{(n-k)} f_{ij}^{(n)} + f_{ij}^{(n)}.
\end{aligned}$$

Comme on a : $\mathrm{p}_{jj}^{(0)} = 1$ et $f_{ij}^{(0)}$.

On peut écrire :

$$\mathrm{p}_{ij}^{(n)} = \sum_{k=0}^{n} f_{ij}^{(k)} \mathrm{p}_{jj}^{(n-k)}$$

.

Remarque 1.2.5.

• *D'après le théorème permet de déterminer les* $f_{ij}^{(n)}$ *par récurrence à partir de probabilités* $\mathrm{p}_{ij}^{(n)}$ *:*

$$f_{ij}^{(1)} = \mathrm{p}_{ij}^{(1)}.$$
$$f_{ij}^{(n)} = \mathrm{p}_{ij}^{(n)} - \sum_{k=1}^{n-1} f_{ij}^{(k)} \mathrm{p}_{jj}^{(n-k)}; n \geq 2$$

• *posons :* $f_{ij} = \mathbb{P}(T_j < \infty / X_0 = i) = \sum_{n\geq 1} f_{ij}^{(n)}$.

Définition 1.2.11.

- *Un état i est dit transient ou transitoire si $f_{ii} < 1$.*
- *Un état i est dit récurrent si $f_{ii} = 1$.*

Remarque :

• Si $\mu_i = E[T_i/X_0 = i] < \infty (\mu_i = \sum_{n=1}^{\infty} f_{ii}^n)$ moyenne de retour ,alors état i est récurent positif .

• Si $\mu_i = inf$ alors état i est récurrent nul.

Définition 1.2.12.

La variable aléatoire $N_i = \sum_i^{\infty} 1_i(X_n)$ dénombre les passages de chaîne de Markov par l'état i .

Définition 1.2.13.

Le nombre moyen de retours à l'état i ; définie par l'espérance conditionnelle $E(N_i/X_0 = i)$ est égale : $E(N_i/X_0 = i) = \sum_{n\geq 1} \mathrm{p}_{ii}^{(n)}$

Démonstration.

$E(1_i(X_n)/X_0) = \mathbb{P}(X_n = i/X_0 = i) = \mathrm{p}_{ii}^{(n)}$

Corollaire 1.2.4.

- *i transitoire ssi $\sum_{n\geq 1} \mathrm{p}_{ii}^{(n)}$ converge .*
- *i récurrent ssi $\sum_{n\geq 1} \mathrm{p}_{ii}^{(n)}$ diverge .*

Proposition 1.2.5.

- *Tout état de non-retour est un état transient .*
- *Tout état absorbant est un état récurrent .*
- *Une chaîne à un nombre fini d'états elle à au moins un état récurrent .*

3/- La périodicité :

Définition 1.2.14.

Soit i un état de retour ; on appelle période de i le **PGCD** *de tous les entier $n \geq 1$; pour lesquels $\mathrm{p}_{ii}^{(n)} > 0$ ie : $d(i) = \mathbf{PGCD}\{n \geq 1 \ tq \ \mathrm{P}_{ii}^{(n)} > 0\}$.*

- Si : $d(i) = d \geq 2$, on dit que i est d-périodique .

- Si :$d(i) = 1$, on dit que i est apériodique.

- Si : i est un état non retour , on pose $d(i) = +\infty$.

Théorème 1.2.6.

Si i est périodique et $i \leftrightsquigarrow j, i \neq j$ alors $d(i) = d(j)$.

Définition 1.2.15. ***(Ergodicité) :***

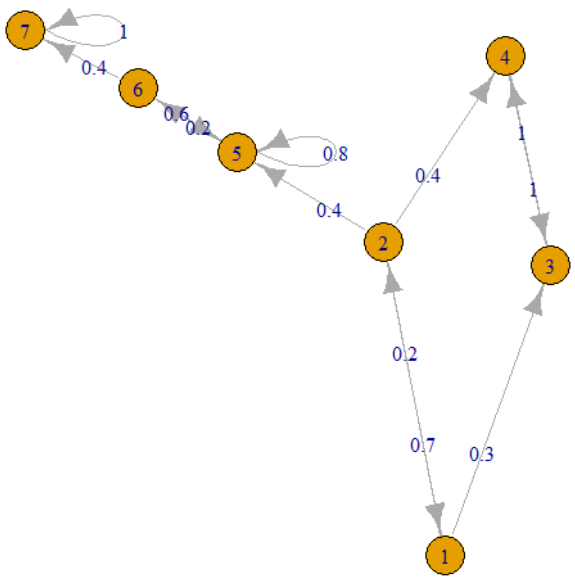
Un état récurrent positif apériodique est dit ergodique.C'est le cas en particulier d'un état i tq $\mathrm{p}_{ii} = 1$ qui est absorbant .

Example 1.2.2.

Pour la matrice de transition

$$\mathrm{P} = \begin{pmatrix} 0 & 0.7 & 0.3 & 0 & 0 & 0 & 0 \\ 0.2 & 0 & 0 & 0.4 & 0.4 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.8 & 0.2 & 0 \\ 0 & 0 & 0 & 0 & 0.6 & 0 & 0.4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

- *Le graphe représentatif de la chaîne est :*



- *La chaîne possède quatre classes :*

$$C_1 = \{1, 2\}\ ; C_2 = \{3, 4\}\ ; C_3 = \{5, 6\}\ ; C_4 = \{7\}$$

- *la classification des classes et ses états :*

Classe	États de la classe	Classification de la classe et de ses états	Période
C_1	1,2	transitoire	-
C_2	3,4	persistant (absorbant)	2
C_3	5,6	transitoire	-
C_4	7	persistant	1

TABLE 1.2 – Exemple 1.2.2.

- La chaîne est réductible, mais ni absorbante, ni ergodique .

1.2.2 Distribution des états d'une chaîne de Markov

Définition 1.2.16.

La distribution des états d'une CM après n transition est note $\pi^{(n)}$ *.Cette distribution est un vecteur de probabilité contenant la loi de variable aléatoire* X_n *, ie :*

$$\pi^{(n)} = \mathbb{P}(X_n = i), \forall i \in E$$

avec : $\pi^{(n)} = (\pi_1^{(n)}, \pi_2^{(n)}, ...)$ *dont la somme des termes vaut 1.*

Pour calculer le vecteur , il faut connaitre soit la valeur par X_0 *, c-à-d :l'état initial du processus qui a une distribution initial* $\pi^{(0)}$ *.*

Remarque 1.2.6.

• *Si l'état initial est connu avec certitude et est égale à i , on a simplement* $\pi_i^{(0)} = 1$ *et* $\pi_j^{(0)} = 0$*, pour tout* $i \neq j$*.*

1.2.3 Comportement transitoire

Théorème 1.2.7.

Soit la matrice de transition P *d'une CM et la distribution de son état initial .Pour tout* $n \geq 1$ *on a :*

$$\pi^{(n)} = \pi^{(n-1)}\mathrm{P} \quad et \quad \pi^{(n)} = \pi^{(0)}\mathrm{P}^{(n)}$$

Démonstration.

On a : $\forall j \in E$

$$\begin{aligned}\pi_j^{(1)} &= \mathbb{P}(X_1 = j) \\ &= \sum_{i \in E} \mathbb{P}(X_1 = j / X_0 = i)\mathbb{P}(X_0 = i) \\ &= \sum_{i \in E} \mathrm{p}_{ij}\pi^{(0)} = \sum_{i \in E} \pi^{(o)}\mathrm{p}_{ij}.\end{aligned}$$

et $\pi^{(1)} = \pi^{(0)}\mathrm{P}^{(n)}$.

La chaîne étant homogène, on obtient immédiatement le premier résultat qui est $\pi^{(n)} = \pi^{(n-1)}\mathrm{P}, \forall n > 0$.

Pour démontrer le seconde il suffit dr résoudre l'équation de récurrence précédente par substitution .

1.2.4 Distribution invariante

Définition 1.2.17.

Une distribution est invariante ou stationnaire si : $\pi = \pi\mathrm{P}$

proprieté 1.2.4.

Si $\lim_{n\longrightarrow\infty}\pi^{(n)}$ *existe ; alors la limite est une distribution invariante.*

Théorème 1.2.8.

Une CM possède toujours au moins une distribution invariante .

Théorème 1.2.9.

Une CM possède autant de distribution invariantes linéairement indépendants que la multiplicité de la valeur propre 1 de sa matrice de transition .

Théorème 1.2.10.

La distribution $\pi^{(n)}$ *des états d'une CM converge vers une distribution (invariante)* π^* *indépendante de la distribution initial ssi la suite des puissance de la matrices de transition de la chaîne converge vers une matrice (stochastique)* $\mathtt{P}^*$ *dont toutes les lignes sont égale entre elle . De plus si tel est le cas , chaque ligne* $\mathtt{P}^*$ *de est égale à* π^* *.*

Démonstration.

$\exists \quad \pi^* \perp\!\!\!\perp \pi^0 \Longleftrightarrow \lim_{n\longrightarrow\infty} \mathrm{P}^{(n)} = \mathrm{P}^*$

- La condition est nécessaire car si indépendant de $\pi^{(0)}$ et $\lim_{n\longrightarrow\infty}\pi^{(n)} = \pi^{(*)}$, il suffit de considérer successivement la distributions intailles : $\pi_1^{(0)} = (1,0,0,...,0), \pi_2^{(0)} = (0,1,0,...,0), ...$
$\pi_s^{(0)} = (0,0,0,...,1)$, pour obtenir :

$$\pi^{(*)} = \lim_{n\longrightarrow\infty}\pi^{(n)} = \lim_{n\longrightarrow\infty}\pi_i\mathtt{P}^n = \lim_{n\longrightarrow\infty}(\mathtt{P}^n)_i = (\mathtt{P}^*)_i$$

ainsi $\mathtt{P}^*$ existe et tout ses lignes sont égales à π^* .

- Les conditions suffisante si P^* existe et $\mathtt{p}_{ij}^* = \mathtt{p}_j^*, \forall i \in E$ on a :

$$\lim_{n\longrightarrow\infty}\pi^{(n)} = \pi^{(0)}\mathrm{P}^n = \pi^{(0)}\lim_{n\longrightarrow\infty}\mathrm{P}^n = \pi^{(0)}\mathrm{P}^*$$

et la limite existe π^* de plus ; $\pi_j^* = \sum_{i\in E}\pi_i^{(0)}\mathtt{p}_{ij}^* = \sum_{i\in E}\pi_i^{(0)}\mathtt{p}_j^* = \mathtt{p}_j^*\sum_{i\in E}\pi_i^{(0)} = \mathtt{p}_j^*$
et π^* est indépendant de $\pi^{(0)}$ et identique à m'improte quelle ligne de P^* .

Remarque 1.2.7.

• *Si* $\pi^{(*)} = \lim_{n\longrightarrow\infty}\pi^n$ *, on parlera de distribution asymptotique stationnaire ou invariante.*

1.2.5 Comportant asymptotique des chaînes irréductible et apériodique

Théorème 1.2.11.

soit P *la matrice d'une chaîne irréductible et apériodique les propriété suivantes sont vérifiées :*

- *La matrice* P^n *tend vers une matrice stochastique* P *ie :* $\mathrm{P}^n \underset{n\longrightarrow\infty}{\longrightarrow} \mathrm{P}^*$.
- *Les lignes de* P^* *sont toutes égales entre elles.*
- $\mathrm{p}_{ij} > 0$ *pour tout* $i, j \in E$.
- *Pour toute distribution initial ;* $\pi^{(0)}$

$$\lim_{n\longrightarrow\infty} \pi^{(n)} = \lim_{n\longrightarrow\infty} \pi^{(0)}\mathrm{P}^n = \pi^{(*)}.$$

- π^* *la solution de système :*

$$\begin{cases} \pi\mathrm{P} = \pi \\ \pi 1 = 1 \end{cases}$$

- π^* *est égal n'importe quelle ligne de la matrice .*
- *Pour tout* $i \in E$: $\pi_i^* = \frac{1}{\mu_i}$ *, ou est l'ésprerance du nombre de transitions entre deux visite successives de l'état i .*

Remarque 1.2.8.

- *Pour n suffisamment grand ; on a* $\pi^{(n)} \simeq \pi^*$ *et* π^* *est la probabilité que la chaîne se trouve dans l'état i à un instant quelconque .*

  Cette valeur représente aussi la proportion du temps passé dans l'état i.

1.2.6 Comportant asymptotique des chaîne réductible

1.2.6.1 Chaîne réductible

Théorème 1.2.12.

Pour tout état initial i , la probabilité de se retrouver dans un état persistant à l'étape n tend vers 1 lorsque n tend vers l'infini .

Démonstration.

Soit $\mathtt{p}_i(m) = \mathbb{P}[X_m \quad transitoire/X_0 = i]$ si $s = |E|$, on a $\mathtt{p}_i(s) < 1 \ \forall i \in E^{'}$ car il est possible d'atteindre un état persistant en au plus s transition depuis n'importe quel état initial . Ainsi

$$\mathtt{p} = \underset{i \in E^{'}}{max}(\mathtt{p}_i(s)) < 1$$

et $\forall i \in E^{'}$

$$\mathtt{p}_i(2s) = \sum_{j\ tansitoire} \mathtt{p}_{ij}^{(s)} \mathtt{p}_j(s) \leq \sum_{j\ tansitoire} \mathtt{p}_{ij}^{(s)} \mathtt{p} = \mathtt{p}_i(s)\mathtt{p} \leq \mathtt{p}^2$$

Ainsi $\mathtt{p}_i(ks) \leq \mathtt{p}_k$ et $\lim\limits_{k \longrightarrow \infty} \mathtt{p}_i(ks) = 0$.

D'autre part , pour tout $n \geqslant 0$;

$$\begin{aligned} \mathtt{p}_i(n+1) &= \sum_{j\ tansitoire} \mathtt{p}_{ij}^{(n+1)} = \sum_{j\ tansitoire} \mathtt{p}_{ik}^{(n)} \mathtt{p}_{kj} \\ &= \sum_{j\ tansitoire} \mathtt{p}_{ik}^{(n)} (\sum_{j\ tansitoire} \mathtt{p}_{kj}) \leq \mathtt{p}_i(n) \end{aligned}$$

ce qui permet de conclure que $\lim\limits_{n \longrightarrow \infty} \mathtt{p}_i(n) = 0$.

Corollaire 1.2.13.

$$\lim_{n \longrightarrow \infty} \mathrm{p}_{ij}^{(n)} = 0 \quad \forall j\ transitoire\ , \forall i.$$

1.2.6.2 Forme canonique

La matrice de transition d'une chaîne de Markov réductible est sous forme canonique si :

1- Les sommets d'une classe (persistante) sont numérotés consécutivement .

2- Les sommets persistants sont numérotés en premier.

Si une chaîne a k classes persistantes, sa matrice sous forme canonique ressemble à

$$\mathrm{P} = \left(\begin{array}{c|c} \mathtt{P}_1 \ ...0 & 0 \\ 0 \ ... \ \mathtt{P}^k & 0 \\ \hline R_1 \ ... \ R_k & \mathrm{Q} \end{array} \right)$$

Chaque sous-matrice $\mathtt{P}$ définit une chaîne de Markov irréductible sur l'ensemble des états de la classe persistante C_i.

Remarque 1.2.9.

• *Pour obtenir une matrice de transition sous forme canonique, il est important de numéroter consécutivement les états de chaque classe persistante.*

• *La matrice de transition, sous forme canonique, d'une chaîne absorbante est*

$$\mathrm{P} = \left(\begin{array}{c|c} I & 0 \\ \hline R & Q \end{array}\right)$$

Corollaire 1.2.14. ***(Matrice fondamentale)***

Soit Q telle que $\lim\limits_{n\longrightarrow\infty} Q^n = 0$ alors :

$$(I-Q)^{-1} = I + Q + Q^2 + ... = \sum_{n=0}^{\infty} Q^n$$

.

Corollaire 1.2.15.

Soit P *la matrice de transition, sous forme canonique, d'une chaîne de Markov absorbante. Alors :*

$$\lim_{n\longrightarrow\infty} \mathrm{P}^n = \lim_{n\longrightarrow\infty} \left(\begin{array}{c|c} I & 0 \\ \hline \sum_{k=0}^{n-1} Q^k R & Q^n \end{array}\right) = \left(\begin{array}{c|c} I & 0 \\ \hline (I-Q)^{-1}R & 0 \end{array}\right)$$

- La matrice $\mathsf{N} = (I-Q)^{-1}$ st appelée la matrice fondamentale de la chaîne de Markov.

1.2.6.3 Étude d'une chaîne réductible

L'étude d'une chaîne de Markov réductible se décompose en deux étapes.

1) Étude des classes persistantes :

- On applique les résultats obtenus pour les chaînes irréductibles afin de déterminer la période et la distribution stationnaire de chacune de sous-chaînes associées aux classes persistantes.

2) Étude des classes transitoires :

- On rend la chaîne absorbante soit en contractant les classes persistantes en un seul état, soit en rendant absorbants tous les états persistants .

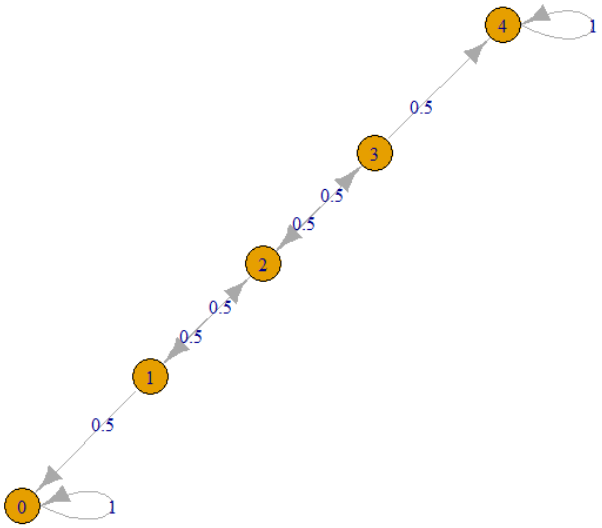
- On calcule ensuite les temps moyens avant absorption (données par la matrice fondamentale N) et les probabilités d'absorption (données par la matrice $\mathsf{B} = \mathsf{NR}$) .

Example 1.2.3.

Soit E={0,1,2,3,4} et la matrice de transition :

$$\mathrm{P} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

- Le graphe :



- On permute les lignes/colonnes pour avoir d'abord les états transitoires puis ensuite les états absorbants :

$$\mathrm{P} = \left(\begin{array}{ccc|cc} 0 & 1/2 & 0 & 1/2 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 1/2 \\ \hline 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array}\right) = \left(\begin{array}{c|c} Q & R \\ \hline 0 & I \end{array}\right)$$

Alors :

$$I - Q = \begin{pmatrix} 1 & -1/2 & 0 \\ -1/2 & 1 & -1/2 \\ 0 & -1/2 & 1 \end{pmatrix}$$

Puis :

$$N = (I - Q)^{-1} = \begin{pmatrix} 3/2 & 1 & 1/2 \\ 1 & 2 & 1 \\ 1/2 & 1 & 3/2 \end{pmatrix}$$

Et :

$$B = NR = (I - Q)^{-1}R = \begin{pmatrix} 3/2 & 1 & 1/2 \\ 1 & 2 & 1 \\ 1/2 & 1 & 3/2 \end{pmatrix} \begin{pmatrix} 1/2 & 0 \\ 0 & 0 \\ 0 & 1/2 \end{pmatrix}$$
$$= \begin{pmatrix} 3/4 & 1/4 \\ 1/2 & 1/2 \\ 1/4 & 3/4 \end{pmatrix}$$

1.3 Chaîne de Markov à temps continue (CMTC)

Définition 1.3.1.

*Une chaîne de Markov à temps continue est un processus stochastique à temps continu , défini sur un espace d'état E' fini ou dénombrable et vérifiant la propriété de Markov (**sans mémoire**) :*

$$\mathbb{P}[X_{t+u} = j/X_s, 0 \leqslant s \leqslant t] = \mathbb{P}[X_{t+u} = j/X_t], \forall j \in E'; \forall t, u \geqslant 0$$

Définition 1.3.2.

Une CMTC est homogène si les probabilité précédentes sont indépendant de t c-à-d :

$$\mathbb{P}[X_{t+u} = j/X_t = i] = \mathbb{P}[X_u = j/X_0 = i], \forall j \in E'; \forall t, u > 0$$

Définition 1.3.3.

Pour une chaîne de Markov homogène , nous noterons :

- *Les probabilités de transition au temps t par :*

$$\mathtt{p}_{ij}(t) = \mathbb{P}[X_t = j / X_0 = i]$$

- *la matrice de passage (transition) au temps t par :*

$$\mathrm{P}(t) = \mathtt{p}_{ij}(t)$$

Remarque 1.3.1.

- $\mathrm{P}(0) = \mathbf{I}$.
- *La matrice de transition est stochastique .*

Proposition 1.3.1. ***(Équation de Chapman - Kolmogorov)***

Soit $(X_t)_t$ un processus de Markov à temps continu de matrice de transition $\mathrm{P}(t)$, soit $u, t \geqslant 0$ alors :

$$\mathrm{P}(t+u) = \mathrm{P}(t)\mathrm{P}(u)$$

Autrement dit ;

$$\mathtt{p}_{ij}(t+u) = \textstyle\sum_{k\in E} \mathtt{p}_{ik}(t)\mathtt{p}_{kj}(u)$$

Démonstration.

$$\begin{aligned}\mathtt{p}_{ij}(t+u) &= \mathbb{P}[X_{t+u} = j / X_0 = i] \\ &= \sum_{k\in E} \mathbb{P}[X_{t+u} = j / X_t = k, X_0 = i]\end{aligned}$$

d'aprés proprieté de Markov

$$= \sum_{k\in E} \mathbb{P}[X_{t+u} = j / X_t = k]\mathbb{P}[X_t = k / X_0 = i]$$

par homogénitié

$$\begin{aligned}&= \sum_{k\in E} \mathbb{P}[X_u = j / X_0 = k]\mathbb{P}[X_t = k / X_0 = i] \\ &= \sum_{k\in E} \mathtt{p}_{ik}(t)\mathtt{p}_{kj}(u)\end{aligned}$$

1.3.1 Classification des CMTC

Définition 1.3.4.

- L'état j est accessible depuis l'état i s'il existe $t > 0$ tel que : $\mathtt{p}_{ij}(t) > 0$.

- Les états i et j communiquent s'ils sont accessibles l'un depuis l'autre, c-à-d s'il existe $t_1 \geqslant 0$ et $t_2 \geqslant 0$ tel que : $\mathtt{p}_{ij}(t_1)$ et $\mathtt{p}_{ij}(t_2)$.

- Une chaîne de Markov à temps continu est irréductible si tous ses états communiquent deux à deux.

- Un état i est absorbant si $\mathtt{p}_{ii}(t) = 1$, pour tout $t \geqslant 0$.

Remarque 1.3.2.

• *Une CMTC n'est jamais périodique .*

1.3.2 Structure des chaînes de Markov à temps continu

Définition 1.3.5. ***(temps de séjour) :***

Supposons que l'état d'une chaîne de Markov à l'instant t soit égal à i $(X_t = i)$ .Elle va rester dans cet état pendant une durée aléatoire $\mathcal{T}_i$

• *Dont la loi ne dépend pas la valeur de t car le processus est homogène.*

• *Dont la loi est sans mémoire car, le processus étant markovien, son évolution au-delà de t est indépendante de son passé une fois l'état X_t connu.*

- Le temps de séjour $\mathcal{T}_i$ dans l'état i est une variable aléatoire exponentielle de paramètre α_i ne dépendant que de i.

Définition 1.3.6. ***(Probabilité de passage) :***

Lorsque la chaîne de Markov quitte l'état i, elle se déplace dans l'état j avec probabilité $\mathtt{q}_{ij}$. Cette probabilité est :

• *Indépendante de la valeur de t (le processus homogène) .*

• *indépendante de la valeur de $\mathcal{T}_i$ (le processus markovien) .*

- La suite des états visités par une chaîne de Markov à temps continu forme une chaîne de Markov à temps discret. Cette dernière est appelée chaîne de Markov sous-jacente ou induite et sa matrice de transition sera notée $\mathcal{Q} = \mathtt{q}_{ij}$.

Définition 1.3.7. ***(La loi exponentielle :)***

Une variable aléatoire X est une variable aléatoire exponentielle de paramètre $\alpha(\alpha > 0)$ si :

$$\mathbb{P}[X \leqslant x] = \mathbb{F}(x) = \begin{cases} 1 - e^{-\alpha x} & ; x \geqslant 0 \\ 0 & ; x < 0 \end{cases}$$

de densité $\mathbb{F}(x)' = f(x) = \begin{cases} \alpha e^{-\alpha x} & ; x \geqslant 0 \\ 0 & ; x < 0 \end{cases}$

avec : $\mathtt{E}(x) = \frac{1}{\alpha}$ *;* $var(x) = \frac{1}{\alpha^2}$

Remarque 1.3.3.

• *La loi exponentielle est la seule loi continue sans mémoire .*

• *Si* $\alpha = 0$ *la variable exponentielle est prendre une seul valeur l'infini. Elle vérifie :*
$\mathbb{P}[X > t] = 1, \forall t > 0; \forall u > 0$

• *Si* $\alpha = \infty$ *la variable ne prennent qu'une seul valeur Zéro.*

• *L'état i est absorbant ssi le temps de séjour en i est une variable aléatoire exponentielle de paramètre* α_i *égale* 0 *.*

Définition 1.3.8. ***(Les chaînes régulières)***

Une chaîne de Markov à temps continu est régulière si, avec probabilité 1,le nombre de transitions qu'elle effectue dans un intervalle de temps fini est fini .

Autrement dit ; s'il existe une constante $c < \infty$ *tq :* $0 \leqslant \alpha_i < c, \forall i \in E$ *suffit à assurer la régularité d'une CM .En particulier , tout chaîne possédant un nombre fini d'état est régulière.*

1.3.3 Intensités de transition de passage

Théorème 1.3.2.

Les probabilités de transition d'une chaîne de Markov à temps continu, homogène et régulière admettent une dérivée à droite en $t = 0$ *égale à :*

$$a_{ij} = \frac{d\mathtt{P}_{ij}}{dt} = \lim_{t \longrightarrow 0^+} \frac{\mathtt{P}_{ij}\mathtt{P}_{ij}^{(0)}}{t} = \begin{cases} -\alpha_i & si \quad i = j \\ \alpha_i \mathrm{q}_{ij} & si \quad i \neq j \end{cases}$$

• *Pour t petit, on a* $\mathbb{P}[X_t = j / X_0 = i] = \mathtt{p}_{ij}(t) = a_{ij}t + o(t)$ $i \neq j$ *et* $a_{ij} = \alpha_i \mathrm{q}_{ij}$ *est appelée intensité de transition de i à j .*

• *On a* $\mathbb{P}[X_t \neq j / X_0 = i] = 1 - \mathtt{p}_{ii}(t) = -a_{ii}t + o(t)$ *et* $a_{ij} = -\alpha_i$ *est appelée intensité de passage hors de i .*

Démonstration.

• **Nous ne traiterons que le cas i=j (cas $i \neq j$ de façon similaire).**

On suppose qu'à l'instant initial t = 0 et au temps t, la chaîne se trouve dans l'état i . Alors soit le processus n'a pas quitté l'état i, ce qui signifie que le premier temps de saut $\mathtt{T}_1$ survient après t ; soit le processus a quitté l'état i et j est revenu, ce qui implique au moins deux sauts avant l'instant t. On a donc :

$$\mathbb{P}(X_t = i/X_0 = i) = \mathbb{P}(\mathtt{T}_1 > t/X_0 = i) + \mathbb{P}(\mathtt{T}_1 < t; \mathtt{T}_2 < t; X_t = i/X_0 = i).$$

Sachant que $X_0 = i$, l'instant du premier saut $\mathtt{T}_1$ suit la loi exponentielle de paramètre α_i. On a alors :

$$\mathbb{P}(\mathtt{T}_1 > t/X_0 = i) = e^{-\alpha_i t}$$

Lorsque t tend vers 0, on a donc :

$$\mathbb{P}(\mathtt{T}_1 > t/X_0 = i) = 1 - \alpha_i t + 0(t)$$

On a aussi :

$$\begin{aligned}\mathbb{P}(\mathtt{T}_1 < t; \mathtt{T}_2 < t; X_t = i/X_0 = i) &\leqslant \mathbb{P}(\mathtt{T}_1 < t; \mathtt{T}_2 < t/X_0 = i) \\ &\leqslant \mathbb{P}(\mathtt{T}_1 < t; \mathtt{T}_2 - \mathtt{T}_1 < t/X_0 = i) \\ &\leqslant \sum_{Z \in E'} \mathbb{P}(\mathtt{T}_1 < t; \mathtt{T}_2 - \mathtt{T}_1 < t, X_{\mathtt{T}_1}/X_0 = i)\end{aligned}$$

Soit $z \in E$. Par définition d'une chaîne de Markov en temps continu, on a :

$$\begin{aligned}\mathbb{P}(\mathtt{T}_1 < t; \mathtt{T}_2 - \mathtt{T}_1 < t, X_{\mathtt{T}_1} = k/X_0 = i) &= \mathbb{P}(\mathtt{X}_{\mathtt{T}_1} = k, \mathtt{T}_2 - \mathtt{T}_1 < t, X_{\mathtt{T}_1}/\mathtt{T}_1 < t; X_0 = i)\mathbb{P}(\mathtt{T}_1 < t/X_0 = i) \\ &= \mathbb{P}(\mathtt{X}_{\mathtt{T}_1} = k, \mathtt{T}_2 - \mathtt{T}_1 < t/\mathtt{T}_1 \leqslant t, X_0 = i) \\ &= (1 - e^{-\alpha_k t})\mathtt{q}_{ij}(1 - e^{-\alpha_i t}))\end{aligned}$$

En utilisant l'inégalité de convexité $e^{-u} \geqslant 1-u$ et le fait que la suite $(\alpha_i)_{i\in E}$ soit bornée par une constante α_0, on a, quand t tend vers 0 :

$$\begin{aligned}
\mathbb{P}(\mathtt{T}_1 < t; \mathtt{T}_2 < t/X_t = i/X_0 = i) &\leqslant \mathbb{P}(\mathtt{T}_1 < t; \mathtt{T}_2 - \mathtt{T}_1 < t, X_{\mathtt{T}_1} = k/X_0 = i) \\
&\leqslant \sum_{k\in E'} (1 - e^{-\alpha_k t})\mathtt{q}_{ik}(1 - e^{-\alpha_i t})) \\
&\leqslant \sum_{k\in E'} \alpha_k \mathtt{q}_{ik} \alpha_i t^2 \\
&\leqslant \alpha_0 \alpha_i t^2 \sum_{k\in E'} \mathtt{q}_{ik} \\
&\leqslant \alpha_0 \alpha_i t^2 = 0(t)
\end{aligned}$$

Ainsi , on bntient la limite récherchée

$$\frac{\mathbb{P}(\mathtt{X}_T = i/X_0 + i) - 1}{t} = -\alpha_i + 0(1)$$

Définition 1.3.9. ***(Générateur de Markov)***

La matrice $\mathbf{A} = (a_{ij})_{ij}$ *est appelée la matrice génératrice de la chaîne qui satisfait les conditions suivant :*

- $0 \leqslant a_{ij} < \infty \quad , \forall i \in E^{'}$.
- $a_{ij} \geqslant 0 \quad , \forall i \neq j$
- $\sum_{j\in E'} a_{ij} = 0 \quad , \forall i \in E^{'}$

- On associe à la matrice génératrice $\mathbf{A}$ *un graphe représentatif* $\mathbf{G} = (E^{'}, V)$
où $V = \{(i,j)/a_{ij} > 0\}$.

Théorème 1.3.3. ***(Équation de Kolmogorov)***

On considère un processus markovien de sauts de probabilités de transition P *et de générateur* $\mathbf{A}$. *On a la relation :*

$$\mathrm{P}^{'}(t) = \mathbf{A}\mathrm{P}(t) = \mathrm{P}(t)\mathbf{A}$$

où $\mathrm{P}^{'}(t)$ *désigne la dérivée de* $\mathrm{P}(t)$ *au point* t *si* $t > 0$ *et la dérivée à droite si* $t = 0$ *:*

$$\mathrm{P}^{'}(t) = \lim_{h\longrightarrow\infty} \frac{\mathrm{P}(t+h) - \mathrm{P}(t)}{h}$$

Démonstration.

• **Nous ne traiterons le démonstration que dans le cas où $E^{'}$ est fini**.

La relation de Chapman-Kolmogorov s'écrit :

$$\forall(i,j) \in E_2^{'} \quad ; \quad \mathrm{p}_{ij}(t+s) = \sum_{k\in E^{'}} \mathrm{p}_{ik}(t)\mathrm{p}_{kj}(s) \ .$$

On dérive cette formule par rapport à s :

$$\forall(i,j) \in E_2^{'} \quad ; \quad \lim_{h\longrightarrow\infty} \frac{\mathrm{p}_{ij}(t+s+h)}{h} = \sum_{k\in E^{'}} \mathrm{p}_{ik}(t)(\lim_{h\longrightarrow\infty} \frac{\mathrm{p}_{kj}(s+h)-\mathrm{p}_{kj}(s)}{h}) \ .$$

On calcule cette expression en s = 0 ;

pour tout $i; j \in E^{'}$:

$$\lim_{h\longrightarrow\infty} \frac{\mathrm{p}_{ij}(t+h)-\mathrm{p}_{ij}(t)}{h} = \sum_{k\in E^{'}} \mathrm{p}_{ik}(t)(\lim_{h\longrightarrow\infty} \frac{\mathrm{p}_{kj}(h)-\mathrm{p}_{kj}(0)}{h}).$$

$$\mathrm{p}^{'}(t) = \sum_{k\in E^{'}} \mathrm{p}_{kj}(t)\mathtt{A}_{ij}.$$

On en déduit alors que :

$$\mathrm{P}^{'}(t) = \mathrm{P}(t)\mathbf{A}$$

La seconde égalité s'obtient en dérivant par rapport à t et en faisant t = 0 .

Remarque 1.3.4.

• $\mathrm{P}^{'}(t) = \mathrm{P}(t)\mathbf{A}$ *connues sous le nom d'équitation de future .*

• $\mathrm{P}^{'}(t) = \mathbf{A}\mathrm{P}(t)$ *connues sous le nom d'équitation de passé.*

• *L'unique solution des équations de Kolmogorov est, pour la condition initiale* $\mathrm{P}(0) = \mathbf{I}$ *est :*

$$\mathrm{P}(t) = e^{\mathbf{A}t} = \sum_{k=0}^{\infty} \frac{(\mathbf{A}t)^k}{k!} \quad t > 0 \quad ; \quad (\mathbf{A}^0 = \mathbf{I})$$

1.3.4 Distribution initiale et comportement transitoire

Comme dans le cas à temps discret, l'état initial (au temps t = 0) de la chaîne est choisi selon une distribution initiale donnée par un vecteur de probabilités $\pi_i(0)$ vérifiant :

$$\pi_i(0) = \mathbb{P}[X_0 = i]$$

La probabilité d'observer le processus dans l'état i au temps t est alors :

$$\pi_i(t) = \mathbb{P}[X_t = i] = \sum_{j\in E^{'}} \mathbb{P}[X_0 = j]\mathbb{P}[X_t = i/X_0 = j] = \sum_{j\in E^{'}} \pi_j(0)\mathrm{p}_{ij}(t)$$

et $\pi(t) = \pi(0)\mathrm{P}(t)$.

1.3.5 Comportement asymptotique des chaînes homogènes , régulières et irréductibles

- Si une chaîne possède une distribution asymptotique unique, elle doit être irréductible (ou du moins ne posséder qu'une classe persistante).

- Pour toute chaîne irréductible, $\lim_{t\longrightarrow\infty}\pi_j(t)$ existe et est indépendante de $\pi(0)$

$$\lim_{t\longrightarrow\infty}\pi_j(t)=\lim_{t\longrightarrow\infty}\mathrm{p}_{ij}(t)=\pi_j(*)\quad\forall j\in E^{'}\quad\text{tq j dépendant de i .}$$

De plus, si $\lim_{t\longrightarrow\infty}\pi_i(t)=\lim_{t\longrightarrow\infty}\mathrm{p}_{ij}(t)$ existe alors $\lim_{t\longrightarrow\infty}\pi_i^{'}(t)=\lim_{t\longrightarrow\infty}\mathrm{p}_{ij}(t)=0$

Partant des équations du passé $\mathrm{p}_{ij}^{'}(t)=\sum_{k\in E'}\mathrm{p}_{ij}(t)a_{ij}\qquad\forall i,j\in E^{'};\forall t>0$

lorsque $t\longrightarrow\infty$, les probabilités stationnaire doivent vérifier

$$\sum_{k\in E'}\pi_k^*a_{kj}=0\ \forall j\in E^{'}$$

- Sous forme matricielle, la distribution π est stationnaire si elle est solution du système :

$$\begin{cases}\pi\mathbf{A}=0\\ \pi 1=1\end{cases}$$

Théorème 1.3.4.

Soit $X_t, t\geqslant 0$ une chaîne de Markov à temps continu , homogène, régulière et irréductible. Les propriétés suivantes sont vérifiées :

- $\mathrm{P}(t)\longrightarrow\mathrm{P}^*$.
- *Les lignes de* P^* *sont toutes égales à un même vecteur* π^*.
- *Soit* $\pi_j^*=0$ *pour tout* $j\in E^{'}$ *et la chaîne est transitoire ou récurrente nulle .*
- *Soit* $\pi_j^*>0$ *pour tout* $j\in E^{'}$ *et la chaîne est transitoire ou récurrente non nulle .*
- *Si la chaîne est récurrente non nulle, elle est ergodique.*

Dans ce cas, le vecteur π^* est une distribution de probabilités et est la solution unique du système

$$\begin{cases}\pi\mathbf{A}=0\\ \pi 1=1\end{cases}$$

avec $\pi\mathbf{A}$ est l'équation de Bilan et s'écrivant aussi :

$$-\pi_i a_{ii}=\sum_{j\neq i}\pi_j a_{ji}\ \forall i\in E^{'}.$$

Chapitre 2

Les chaînes de Markov cachées à temps discret

Les modèles de Markov cachés sont des outils statistiques permettant de modéliser des phénomènes stochastiques .Ces modèles sont utilisés dans nombreux domaine tels que la reconnaissance et la synthèse de la parole , la biologie , la prédiction de séries temporelles,... .
Pour pouvoir utiliser ces modèles efficacement , il est nécessaire d'en connaitre les principes .

Ce chapitre a pour objectif d'établir les principes ,les notations utilisés et les principaux algorithmes qui constituent la théorie des CMC.

2.1 Historique

Les modèles de Markov cachés ont une longue histoire derrière eux. En 1913, les premiers applications ont été développées par Markov pour l'analyse du langage permettent à A.A Markov de concevoir la théorie des chaînes de Markov .De 1948 à 1951 , Shannon conçoit la théorie de l'information en utilisant les chaînes de Markov.

Ces travaux ont été utilisés régulièrement mais les premières applications exploitables furent réalisées dans le années 60, telles que les modèles probabilistes d'urnes par Neuwirtch, le calcul direct du maximum de vraisemblance ou l'observation de la suite d'états dans une chaîne de Markov .Ceci a permet à la communauté scientifique d'exploiter pleinement le potentiel de ces modèles.

C'est dans les années 70 que des chercheurs ont apporté des algorithmes puissants permettant de résoudre les problèmes fondamentaux en CMC.

2.2 Notions de base

A fin d'utiliser une chaîne de Markov cachée, il est indispensable de spécifier les caractéristiques ci-après. On parle d'éléments de cette chaîne :

• N est le nombre d'états cachés dans le modèle. On note l'ensemble d'états cachés par $S = \{S_1, S_2, ..., S_N\}$ et l'état au temps t par y_t .

• M est le nombre de symboles distincts observables par états. On note ces symboles par o_k où $k = 1, 2, ..., M$, et l'observation au temps t par O_t .

• $A = \{a_{ij}\}$ est la matrice de transition des états cachés où :

$$a_{ij} = \mathbb{P}[y_{t+1} = S_j / y_t = S_i] \ ; \qquad \forall 1 \leqslant i, j \leqslant N$$

• $B = \{b_{S_i}(o_k)\}$ est la matrice de probabilité des observations k dans l'état S_i où :

$$b_{S_i}(o_k) = \mathbb{P}[O_t = o_k / y_t = S_i] \ ; \qquad \forall 1 \leqslant i \leqslant N \, ; \qquad \forall 1 \leqslant j \leqslant M$$

La matrice B contient les probabilités d'observer au temps t le symbole k sachant qu'au même instant le modèle est dans l'état caché S_i .

• $\pi = \{\pi_i\}$ la distribution de l'état initial du modèle où :

$$\pi_i = \mathbb{P}[y_1 = S_i] \ ; \quad \forall 1 \leqslant i \leqslant N$$

Ce vecteur contient la probabilité qu'au moment initial ($t = 1$), le modèle se trouve dans l'état caché S_i .

• T longueur de la séquence d'observations O .

Donc, en utilisant une notation compacte, une CMC est noté par $\lambda = (\pi, A, B)$ désignant les paramètres complets d'un modèle de Markov caché .

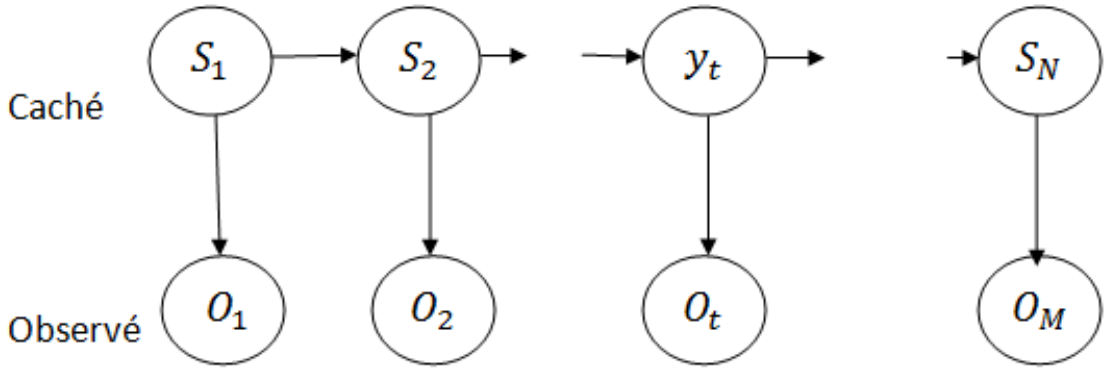
2.3 Définitions

Définition 2.3.1.

Un modèle de Markov Caché ou HMM (Hidden Model Markov) est un processus doublement stochastique dont une composante est une chaîne de Markov non observable. Ce processus peut être observé à travers un autre ensemble de processus qui produit une suite d'observations. Plus simplement, c'est un modèle qui décrit les états d'un processus markovien à l'aide des probabilités de transition et des probabilités d'observation par états.

Définition 2.3.2. ***(Graphe d'indépendance d'une chaîne de Markov cachée)***

Graphiquement, on représente une chaîne de Markov cachée par le graphe suivant :



2.4 Génération d'une séquence par CMC

Pour des valeurs données de N, M, A, B et π la CMC peut être utilisé pour générer une séquence d'observation $O(t) = o_1 o_2 ... o_T$ de la manière suivant :

Algorithme de génération d'une séquence d'état :

1- Pour $i = 1 : N$ choisir aléatoirement un état initial $y_1 = S_i$ si selon la loi π .

2-Définit $t = 1$;

3- Pour k = 1 : M , choisir $O_t = O_K$. selon la distribution des observations dans l'état S_i , c 'est à dire selon $b_{S_i}(o_k)$;

4- Pour $t = 1 : T$ et $j = 1 : N$, choisir l 'état $Y_{t+l} = S_i$ selon les probabilités de transitions de l'état S_i, c'est à dire selon a_{ij} ;

5- Définir $t = t + 1$; si $t < T$ alors à retour à l'étape (3) Sinon fin de la procédure.

2.5 Les trois problèmes fondamentaux des CMC

Les CMC pour être utiles , nécessitent la résolution de plusieurs problèmes principaux :

2.5.1 Problème 1 : Évaluation

Étant donnés une CMC ; $\lambda = (\pi, A, B)$ et une séquence observée $O = \{o_1, o_2, ..., o_M\}$ quelle est la vraisemblance $\mathbb{P}(O/\lambda)$ que le modèle λ génère O ?

Il existe plusieurs techniques pour évaluer cette probabilité d'observation d'une séquence de longueur T : procédure « Forward-Backward ».

2.6.1.1 Algorithme Forward

Pour une représentation à cet algorithme , il est nécessaire de définir les variables **Forward**. pour tout $i = \overline{1, N}$; $t = \overline{2, T}$:
$\alpha_t(i)$ définie par :

$$\alpha_t(i) = \mathbb{P}(O(t), y_t = S_i/\lambda) \ , t = 1, 2, ..., T \text{ et } i = 1, 2, ..., N$$

Pour t=1 (l' instant initial), on a :

$$\begin{aligned}\alpha_1(i) &= \mathbb{P}(O_1, y_1 = S_i/\lambda) \\ &= \mathbb{P}(y_1 = S_i/\lambda)\mathbb{P}(O_1/y_1 = S_i, \lambda) \\ &= \pi_i b_{S_i}(o_1)\end{aligned}$$

on a par récurrent la relation est vérifiée pour tout $t = \overline{1, T-1}$, $j = \overline{1, N}$

$$\begin{aligned}\alpha_{t+1}(j) &= \mathbb{P}(O(t+1), y_{t+1} = S_j/\lambda) \\ &= \sum_{i=1}^{N} \mathbb{P}(O(t) \wedge O_{t+1}, y_t = S_i, y_{t+1} = S_j/\lambda) \\ &= \sum_{i=1}^{N} \mathbb{P}(O(t), y_t = S_i/\lambda)\mathbb{P}(O_{t+1}, y_{t+1} = S_j/O(t), y_t = S_i, \lambda) \\ &= \sum_{i=1}^{N} \alpha_t(i)\mathbb{P}(O_{t+1}, y_{t+1} = S_j/y_t = S_i, \lambda) \\ &= \sum_{i=1}^{N} \alpha_t(i)\mathbb{P}(O_{t+1}/y_{t+1} = S_j, \lambda)\mathbb{P}(y_{t+1} = S_j/y_t = S_i, \lambda) \\ &= \sum_{i=1}^{N} \alpha_t(i) a_{ij} b_{S_j}(o_{t+1})\end{aligned}$$

Par cette méthode progressive, en sommant la quantité a $\alpha_t(i)$ à chaque état S_i, on obtient la probabilité de n 'observer que la séquence $O(t)$ compte tenu du CMC de paramètre λ.

$$\sum_{i=1}^{N} \alpha_t(i) = \mathbb{P}(O(t)/\lambda)$$

Donc **l'algorithme de Forward** donnée par :

POUR $i = 1 : N$ FAIRE
  $\alpha_1(i) = \pi_i b_{S_i}(o_1)$
  POUR $t = 1 : T - 1$ FAIRE
    POUR $j = 1 : N$ FAIRE
      $\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i) a_{ij}] b_{S_j}(O_{t+1})$
    FIN POUR
  FIN POUR
FIN POUR
FINALISATION $\mathbb{P}(O(T)/\lambda) = \sum_{i=1}^{N} \alpha_t(i)$

2.6.1.2 Algorithme Backward

Bien que le problème du calcul de la vraisemblance soit résolu , nous allons aussi présenter l'algorithme **Backward** .

Les variables **Backward** sont :

- $\beta_t(i)$ définie par :

$$\beta_t(i) = \mathbb{P}(O_{t+1:T}/y_t = S_i, \lambda) \ , t = T-1, T-2, ...1$$

avec : $\beta_t(i)$ est la probabilité d 'observer la séquence partielle ultérieure $O_{t+1:T}$, sachant que la CMC de paramètre λ était dans l'état S_i à l'instant t .L'algorithme **Backward** ne produit qu'une seule information, $B_t(i)$.

On choisit :

$$\beta_T(i) = 1$$

On a par récurrent :

$$\begin{aligned}
\beta_t(i) &= \mathbb{P}(O_{t+1:T}/y_t = S_i, \lambda) \\
&= \sum_{j=1}^{N} \mathbb{P}(O_{t+1:T}, y_{t+1} = S_j/y_t = S_i, \lambda) \\
&= \sum_{j=1}^{N} \mathbb{P}(O_{t+1} \wedge O_{t+2:T}, y_{t+1} = S_j/y_t = S_i, \lambda) \\
&= \sum_{j=1}^{N} \mathbb{P}(y_{t+1} = S_j/y_t = S_i, \lambda)\mathbb{P}(O_{t+2:T}/y_{t+1} = S_j, y_t = S_i, \lambda)\mathbb{P}(O_{t+1}/y_{t+1} = S_j, O_{t+2:T}, y_t = S_i, \lambda) \\
&= \sum_{j=1}^{N} \mathbb{P}(y_{t+1} = S_j/y_t = S_i, \lambda)\mathbb{P}(O_{t+2:T}/y_{t+1} = S_j, \lambda)\mathbb{P}(O_{t+1}/y_{t+1} = S_j, \lambda) \\
&= \sum_{j=1}^{N} a_{ij}\beta_{t+1}(j)b_{S_j}(O_{t+1})
\end{aligned}$$

L'algorithme de Backward donnée par :

POUR $i = 1 : N$ FAIRE
  $\beta_T(i) = 1$
  POUR $t = T - 1 : (-1) : 1$ FAIRE
    POUR $j = 1 : N$ FAIRE
      $\beta_t(i) = \sum_{j=1}^{N} a_{ij}\beta_{t+1}(j)b_{S_j}(O_{t+1})$
    FIN POUR
  FIN POUR
FIN POUR

2.5.2 Problème 2 :Décodage (le calcul de chemin optimal)

Il s'agit de déterminer le meilleur chemin correspondant à l'observation c-à-d de trouver le modèle λ ayant la meilleur suite d'état S , alors on réponds à la question suivante :
Comment choisir une séquence d'état $y = (y_1, y_2, ..., y_T)$ qui est optimal ?

Pour trouver une séquence d'observation $O(o_1, o_2, ..., o_M)$, définit la variable intermédiaire $\delta_t(i)$ comme la probabilité du meilleur chemin amenant à l'état S_t à l'instant t , en étant guidé par t premier observation :

Soit les états S_i et S_j aux instants respectifs t-1 et t , on a par récursive :

$$\begin{aligned}
\delta_t(j) &= \underset{y(t-1)}{max}\mathbb{P}(y(t-1), y_t = S_j, O_t/\lambda) \\
&= \underset{y(t-1)}{max}\mathbb{P}(y_t = S_j/y_{t-1} = S_i, \lambda)\ \mathbb{P}(O_t/y_t = S_j, \lambda)\ \mathbb{P}(y(t-2), y_{t-1} = S_i, O_{t-1}/\lambda) \\
&= \underset{y(t-1)}{max}\mathbb{P}(y_t = S_j/y_{t-1} = S_i, \lambda)\ \mathbb{P}(O_t/y_t = S_j, \lambda)\ \underset{y(t-2)}{max}\ \mathbb{P}(y(t-2), y_{t-1} = S_i, O_{t-1}/\lambda) \\
&= \underset{S_i}{max}[a_{ij} b_{S_j}(O_t)\delta_{t-1}(i)] \\
&= \underset{S_i}{max}[a_{ij}\delta_{t-1}(i)] b_{S_j}(O_t)
\end{aligned}$$

Pour t=1 , on a :

$$\begin{aligned}
\delta_1(i) &= \underset{S_i}{max}\mathbb{P}(y_1 = S_i, O_1/\lambda) \\
&= \underset{S_i}{max}[\mathbb{P}(y_1 = S_i/\lambda)\mathbb{P}(O_1/y_1 = S_i, \lambda)] \\
&= \underset{S_i}{max}[\pi_i b_{S_j}(O_1)] \\
&= \pi_i b_{S_j}(O_1)
\end{aligned}$$

L'algorithme de Vetirbi donnée par :

POUR $1 : N$ FAIRE

$\delta_1(i) = \pi_i b_{S_j}(O_1)$

$\psi_1(i) = 0$

POUR $t = 2 : T$ FAIRE

POUr $j = 1 : N$ FAIRE

$\delta_t(i) = \max_{S_i}[a_{ij}\delta_{t-1}(i)]b_{S_j}(O_t)$

$\psi_t(S_j) = \underset{S_i}{argmax}[a_{ij}\delta_{t-1}(i)]$

FIN POUR

FIN POUR

$\mathbb{P}^* = \max_{S_i} \delta_T(i)$

$S^*_{i,T} = \underset{S_i}{argmax}\ \delta_T(i)$

FIN POUR

% Construction de la séquence d 'états.

POUR $t = T - 1 : (-1) : 1$ FAIRE

$S^*_{i,T} = \psi_{T+1}(S^*_{i,t+1})$

FIN POUR

2.5.3 Problème 3 : Apprentissage

Ce problème s'occupe d'optimiser les paramètre λ dans l'objectif de maximiser la probabilité d'observation $\mathbb{P}(O(T)/\lambda)$; donc trouver $\widehat{\lambda}$? $\quad (tq : \quad \widehat{\lambda} = \underset{\lambda}{argmax}\mathbb{P}(O/\lambda))$

Le fait de la longueur de la suite d'observation soit fini , il n'existe pas de solutions analytique direct pour construire le modèle .

Donc , on choisit $\lambda = (\pi, A, B)$.

En utilisant le procédure itérative telle que **Baum-Welch**.

L'algorithme de Baum-Welch donnée par :

Appliquer les algorithmes Forward-Backward sur la CMC initial de paramètres arbitraires que l'on note $\lambda(0), z = 0$ (z itération) .

(*) Faire $z = z + 1$

Pour $t : 1 : T$ faire

  Pour $i : N$ faire

    Pour $j = 1 : N$ faire

      $\theta_t(i, j)$

    Fin Pour

    $\gamma_t(i)$

  Fin Pour

Fin Pour

% Calculer les fréquences espérées

$\sum_{t=1}^{T-1} \theta_t(i, j)$;

$\sum_{t=1}^{T-1} \gamma_t(i)$

% Ré-estimer les paramètre du modèle

$\widehat{\pi} = \pi(i)$

$\widehat{A} = \widehat{a}_{ij}$

$\widehat{B} = \widehat{b}_{S_i}(K)$

% On pose : $\lambda(z + 1) = (\widehat{\pi}, \widehat{A}, \widehat{B})$

% Retour à l'étape (*) , tant qu 'il y a augmentation de la probabilité $\mathbb{P}(O(T)/\lambda(z))$ ou tant qu 'il y a encore des itérations à faire.

Avec :

$$\begin{aligned}
\theta_t(i, j) &= \mathbb{P}(y_t = S_i, y_{t+1} = S_j / O(T), \lambda) \\
&= \frac{\mathbb{P}(y_t = S_i, y_{t+1} = S_j, O(T)/\lambda)}{\mathbb{P}(O(T)/\lambda)} \\
&= \frac{\mathbb{P}(y_t = S_i, y_{t+1} = S_j, O(t), O_{t+1}, O_{t+2:T}/\lambda)}{\mathbb{P}(O(T)/\lambda)} \\
&= \frac{b_{S_j}(O_{t+1})\beta_{t+1}(j)a_{ij}\alpha_t(i)}{\sum_{m=1}^{N}\sum_{n=1}^{N} b_{S_n}(O_{t+1})\beta_{t+1}(n)a_{mn}\alpha_t(m)}
\end{aligned}$$

$$\begin{aligned}\gamma_t(i) &= \frac{\mathbb{P}(y_t = S_i, O(T)/\lambda)}{\mathbb{P}(O(T)/\lambda)} \\ &= \frac{\alpha_t(i)\beta_t(i)}{\mathbb{P}(O(T)/\lambda)} \\ &= \frac{\beta_t(i)\alpha_t(i)}{\sum_{m=1}^{N}\sum_{n=1}^{N} b_{S_n}(O_{t+1})\beta_{t+1}(n)a_{mn}\alpha_t(m)}\end{aligned}$$

Dans les deux égalité précédent on utilisé les formule suivantes :

- $\mathbb{P}(y_t = S_i, y_{t+1} = S_j, O(t), O_{t+1}, O_{t+2:T}/\lambda) = \mathbb{P}(O_{t+1}/, y_{t+1} = S_j, \lambda)$
$$\begin{aligned}&\mathbb{P}(y_t = S_i, y_{t+1} = S_j, O(t), O_{t+2:T}/\lambda) \\ &= b_{S_j}(O_{t+1})\mathbb{P}(O_{t+2:T}/y_{t+1} = S_j, \lambda) \\ &\mathbb{P}(y_t = S_i, y_{t+1} = S_j, O(t)/\lambda) \\ &= b_{S_j}(O_{t+1})\beta_{t+1}(j)\mathbb{P}(y_{t+1} = S_j/y_t = S_i,\ \lambda) \\ &\mathbb{P}(y_t = S_i, O(t)/\lambda) \\ &= b_{S_j}(O_{t+1})\beta_{t+1}(j)a_{ij}\alpha_t(i)\end{aligned}$$

- $$\begin{aligned}\mathbb{P}(O(T)/\lambda) &= \sum_{m=1}^{N}\sum_{n=1}^{N}\mathbb{P}(y_t = S_m, y_{t+1} = S_n, O(T)/\lambda) \\ &= \sum_{m=1}^{N}\sum_{n=1}^{N} b_{S_n}(O_{t+1})\beta_{t+1}(n)a_{mn}\ \alpha_t(m)\end{aligned}$$

alors les estimations sont :

$$\widehat{\pi} = \gamma_1(i) \quad ;\ \widehat{a}_{ij} = \frac{\sum_{t=1}^{T-1}\theta_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)} \quad ;\ \widehat{b}_{S_j}(K) = \frac{\sum_{t=1\cap O_t=k}^{T-1}\gamma_t(i)}{\sum_{t=1}^{T-1}\gamma_t(i)}$$

2.6 Les avantage et les inconvénients des modèles de Markov cachés

2.6.1 Les avantages

- Séparation franche entre données et algorithmes.
- Variabilité de la forme.
- Base mathématique solide pour comprendre son fonctionnement.
- Reconnaissance réalisée par un simple calcul de probabilité cumulée.

2.6.2 Les inconvénients

- Dégradation des performances si l'apprentissage n'est pas suffisant .
- Le choix à priori de la typologie des modèles (nombre d'état, transitions autorisées et règles de transitions).

Chapitre 3

Simulations et Application

Le but de ce chapitre est de renforcer les résultats théoriques étudiés précédemment d'une façon numérique, en utilisant le langage R version 3.6.3.

3.1 Simulations

3.1.1 Simulation des chaînes de Markov

On considére dans cette partie une chaîne de Markov dont la matrice de transition est la suivante :

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0.1 & 0 & 0.5 & 0.2 & 0.2 \\ 0 & 0.5 & 0 & 0.5 & 0 \\ 0.5 & 0 & 0.5 & 0 & 0 \end{pmatrix}$$

######## Définir la chaîne de Markov
> statesNames <− c("1","2","3","4","5")
> statesNames
[1]"1""2""3""4""5"
> cm1 <− new("markovchain", transitionMatrix=matrix(c(1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0.1, 0, 0.5, 0.2, 0.2, 0, 0.5, 0, 0.5, 0, 0.5, 0, 0.5, 0, 0),byrow=TRUE, nrow=5, dimnames=list(statesNames,statesNames)))

> cm1

Unnamed Markov chain

A 5 - dimensional discrete Markov Chain defined by the following states :

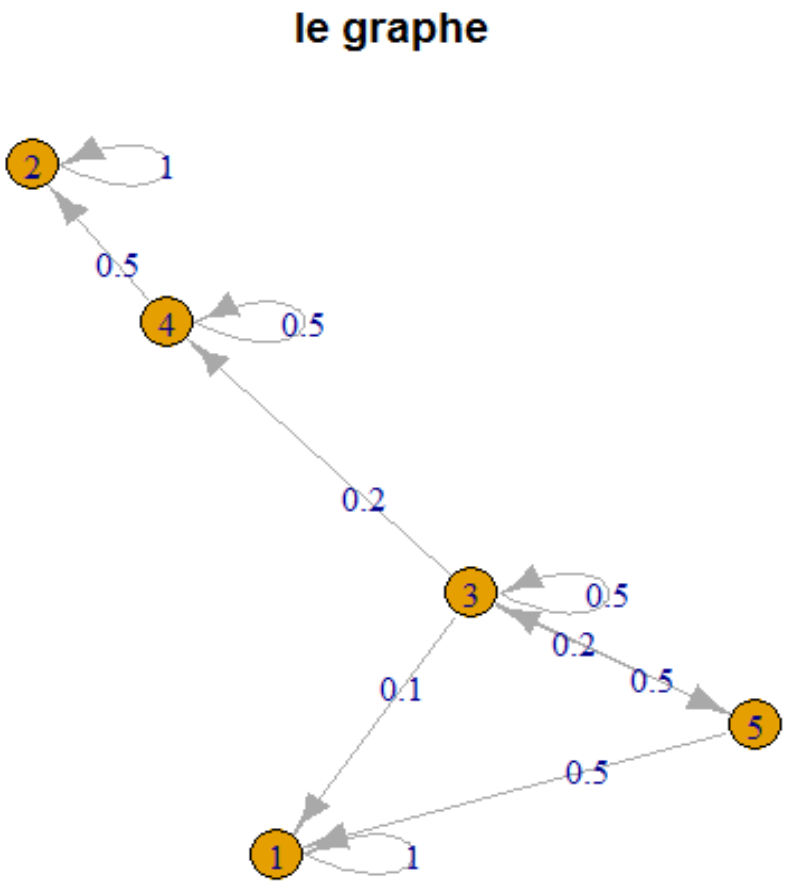
1, 2, 3, 4, 5

The transition matrix (by rows) is defined as follows :

	1	2	3	4	5
1	1.0	0.0	0.0	0.0	0.0
2	0.0	1.0	0.0	0.0	0.0
3	0.1	0.0	0.5	0.2	0.2
4	0.0	0.5	0.0	0.5	0.0
5	0.5	0.0	0.5	0.0	0.0

######## Le graphe représentatif de la chaîne

> plot(cm1,main =" le graphe ")



######## Les passages

> firstPassage(cm1, "3", 10) ### premier passage

	1	2	3	4	5
1	1	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0
5	0	0	0	0	0
6	0	0	0	0	0
7	0	0	0	0	0
8	0	0	0	0	0
9	0	0	0	0	0
10	0	0	0	0	0

> firstPassage(cm1, "1", 10)

	1	2	3	4	5
1	0.100000000	0.00000000	0.5	0.200000000	0.200000000
2	0.150000000	0.10000000	0.1	0.100000000	0.100000000
3	0.085000000	0.10000000	0.0	0.070000000	0.050000000
4	0.057500000	0.08500000	0.0	0.045000000	0.025000000
5	0.037250000	0.06500000	0.0	0.029500000	0.012500000
6	0.024375000	0.04725000	0.0	0.019250000	0.006250000
7	0.015912500	0.03325000	0.0	0.012575000	0.003125000
8	0.010393750	0.02291250	0.0	0.008212500	0.001562500
9	0.006788125	0.01556250	0.0	0.005363750	0.000781250
10	0.004433438	0.01046313	0.0	0.003503125	0.000390625

> meanFirstPassageTime(cm1)

Markov chain needs to be ergodic (= irreducile) for this method to work ### réductible

######## Classification des états

> is.regular(cm1)

[1] FALSE

> is.irreducible(cm1) ### irréductibilité.

[1] FALSE

> is.accessible(cm1) ### accessibilité.

	1	2	3	4	5
1	*TRUE*	*FALSE*	*FALSE*	*FALSE*	*FALSE*
2	*FALSE*	*TRUE*	*FALSE*	*FALSE*	*FALSE*
3	*TRUE*	*TRUE*	*TRUE*	*TRUE*	*TRUE*
4	*FALSE*	*TRUE*	*FALSE*	*TRUE*	*FALSE*
5	*TRUE*	*TRUE*	*TRUE*	*TRUE*	*TRUE*

> period(cm1) ### la période

[1] 0

> steadyStates(cm1) ### états stables.

	1	2	3	4	5
[1,]	0	1	0	0	0
[2,]	1	0	0	0	0

> absorbingStates(cm1) ### les états absorbants.

[1] "1""2"

> transientStates(cm1) ### les états transitions.

[1] "3""4""5"

> recurrentStates(cm1) ### les états récurrents.

[1] "1""2"

> meanRecurrenceTime(cm1)

> absorptionProbabilities(cm1)

	1	2
3	0.50	0.50
4	0.00	1.00
5	0.75	0.25

> meanAbsorptionTime(cm1)

3	4	5
4	2	3

######## La distribution conditionnelle de l'état suivant, étant donné l'état actuel

> conditionalDistribution(cm1, "1")

1	2	3	4	5
1	0	0	0	0

> conditionalDistribution(cm1, "2")

1	2	3	4	5
0	1	0	0	0

```
> conditionalDistribution(cm1, "3")
  1   2   3   4   5
0.1 0.0 0.5 0.2 0.2
> conditionalDistribution(cm1, "4")
  1   2   3   4   5
0.0 0.5 0.0 0.5 0.0
> conditionalDistribution(cm1, "5")
  1   2   3   4   5
0.5 0.0 0.5 0.0 0.0
```

On va simuler un exemple qui déjà existe dans la base de R qui est le "**blanden**" .En ligne, on trouve les revenus trimestriels du père lorsque son fils avait 16 ans, et en colonne les revenus trimestriels du fils lorsque atteint 30 ans.

On va utiliser le package "markovchain" , qu'il faut l'installer , les concepts utilisés sont : rmarkovchain ,markovchainFit , createSequenceMatrix .

Pour différent tailles N =100 ,500 ,1000 , on a les tableaux suivants où :

- P : la probabilité de passage de l'exemple "**blanden**" .
- $\hat{P}$: la probabilité de passage d'une série simulé de N taille .
- Écat-type : erreur de l'estimation .
- $Z_c = \frac{|\hat{P}-P|}{\acute{E}cat-type}$

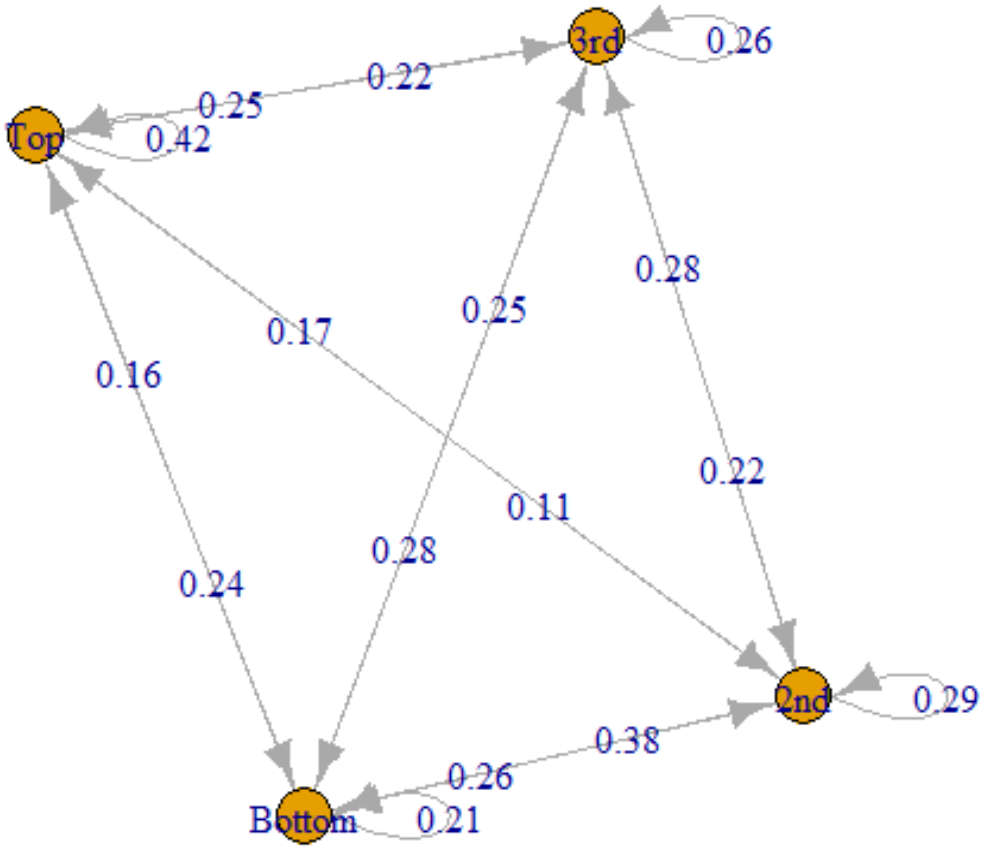


FIGURE 3.1 – Le diagramme de la chaîne de Markov pour données **blanden**

- **Pour N=100**

la série est :

[1] "3rd" "2nd" "Top" "2nd" "2nd" "2nd" "Top" "Top" "2nd" "3rd" "Bottom" "Top"
[13] "Bottom" "3rd" "Bottom" "3rd" "Top" "Top" "2nd" "2nd" "Bottom" "Bottom" "3rd"
[24] "2nd" "2nd" "3rd" "2nd" "3rd" "Top" "Top" "2nd" "Bottom" "2nd" "Bottom" "3rd"
[37] "3rd" "Top" "2nd" "3rd" "Bottom" "2nd" "2nd" "Top" "2nd" "Bottom" "3rd" "Bottom"
[49] "Bottom" "3rd" "Top" "2nd" "2nd" "3rd" "Bottom" "Top" "Top" "Top" "Top" "2nd"
[61] "Top" "Top" "Bottom" "3rd" "Top" "Top" "Top" "Bottom" "3rd" "2nd" "Top" "2nd"
[73] "Bottom" "Top" "Bottom" "2nd" "Bottom" "3rd" "2nd" "Top" "Top" "Top" "2nd" "3rd"

[84] "Bottom" "Top" "Top" "2nd" "2nd" "2nd" "Top" "2nd" "3rd" "3rd" "3rd" "Bottom" [96] "2nd" "3rd" "2nd" "Bottom" "Bottom" .

	N=100	2nd	3rd	Bottom	Top
2nd	P	0.2772277	0.2574257	0.2475248	0.2178218
	$\hat{P}$	0.3437500	0.3125000	0.1562500	0.1875000
	É-T	0.10364452	0.09882118	0.06987712	0.07654655
	Z_c	0.6418313	0.5573127	1.3062187	0.3961224
3rd	P	0.2626263	0.2828283	0.2121212	0.2424242
	$\hat{P}$	0.3500000	0.2000000	0.1000000	0.3500000
	É-T	0.13228757	0.10000000	0.07071068	0.13228757
	Z_c	0.6604831	0.8282830	1.5856332	0.8131966
Bottom	P	0.2900000	0.2200000	0.3800000	0.1100000
	$\hat{P}$	0.4117647	0.1764706	0.2941176	0.1176471
	É-T	0.15563243	0.10188534	0.13153341	0.08318903
	Z_c	0.78238642	0.42723909	0.65293221	0.09192438
Top	P	0.1700000	0.2500000	0.1600000	0.4200000
	$\hat{P}$	0.2333333	0.1000000	0.1333333	0.5333333
	É-T	0.08819171	0.05773503	0.06666667	0.13333333
	Z_c	0.7181321	2.5980761	0.4000005	0.8499998
log vraisemblance		125.4696			

TABLE 3.1 – Probabilités et écart-type pour la chaîne de Markov simulée avec N=100 .

- **Pour N=500**

la série est :

[1] "2nd" "Bottom" "2nd" "2nd" "Top" "Top" "Top" "Bottom" "Bottom" "2nd" "Bottom"
[12] "Top" "2nd" "2nd" "3rd" "3rd" "2nd" "Bottom" "3rd" "2nd" "3rd" "3rd" "Bottom" "3rd"
[25] "Top" "2nd" "3rd" "Bottom" "3rd" "Top" "Top" "Bottom" "Top" "Top" "2nd" "Top"
[37]"Top" "Top" "Bottom" "2nd" "2nd" "2nd" "Bottom" "Bottom" "3rd" "3rd" "Top" "Top"
[49] "Top" "Bottom" "2nd" "Bottom" "2nd" "Bottom" "2nd" "2nd" "2nd" "2nd" "3rd"
[...].....
[...].....
[...].....
[480] "Top" "2nd" "Top" "Bottom" "Bottom" "2nd" "2nd" "2nd" "Bottom" "2nd" "Bottom"
[491] "3rd" "Top" "Top" "Bottom" "Bottom" "Bottom" "3rd" "Bottom" "Bottom" "3rd" .

	N=500	2nd	3rd	Bottom	Top
2nd	P	0.2772277	0.2574257	0.2475248	0.2178218
	$\hat{P}$	0.3023256	0.2403101	0.2170543	0.2403101
	É-T	0.04841084	0.04316096	0.04101940	0.04316096
	Z_c	0.5184355	0.3965528	0.7428314	0.5210334
3rd	P	0.2626263	0.2828283	0.2121212	0.2424242
	$\hat{P}$	0.2521739	0.2260870	0.2000000	0.3217391
	É-T	0.04682752	0.04433930	0.04170288	0.05289359
	Z_c	0.2232106	1.2797067	0.2906562	1.4995182
Bottom	P	0.2900000	0.2200000	0.3800000	0.1100000
	$\hat{P}$	0.2095238	0.3619048	0.3142857	0.1142857
	É-T	0.04467063	0.05870870	0.05471012	0.03299144
	Z_c	1.8015461	2.4171000	1.2011361	0.1299034
Top	P	0.1700000	0.2500000	0.1600000	0.4200000
	$\hat{P}$	0.2600000	0.1333333	0.1333333	0.4733333
	É-T	0.04163332	0.02981424	0.02981424	0.05617433
	Z_c	2.1617301	3.9131200	0.8944283	0.9494248
log vraisemblance		658.8806			

TABLE 3.2 – Probabilités et écart-type pour la chaîne de Markov simulée avec N=500 .

- **Pour N=1000**

la série est :

[1] "Top" "3rd" "2nd" "Top" "3rd" "Bottom" "Bottom" "3rd" "3rd" "3rd" "Bottom" "3rd"

[13] "3rd" "3rd" "Top" "3rd" "Top" "2nd" "3rd" "2nd" "2nd" "2nd" "3rd" "Top"

[25] "Bottom" "Top" "Bottom" "Bottom" "Bottom" "3rd" "Bottom" "3rd" "3rd" "3rd"

[...].....

[...].....

[600] "Top" "2nd" "Top" "2nd" "Top" "2nd" "Bottom" "Bottom" "3rd" "2nd" "Bottom"

[612]"2nd" "Top" "Top" "2nd" "Top" "2nd" "Top" "2nd" "Top" "Top" "Top" "Bottom" "3rd"

[...].....

[970] "Top" "Top" "3rd" "3rd" "Top" "2nd" "2nd" "Bottom" "3rd" "Bottom" "3rd" "Bottom"

[982] "3rd" "Bottom" "2nd" "Bottom" "2nd" "3rd" "3rd" "3rd" "2nd" "Bottom" "3rd"

[993] "2nd" "3rd" "Bottom" "3rd" "3rd" "Bottom" "2nd" "3rd" .

	N=1000	2nd	3rd	Bottom	Top
2nd	P	0.2772277	0.2574257	0.2475248	0.2178218
	$\hat{P}$	0.2623574	0.2471483	0.2205323	0.2699620
	É-T	0.03158412	0.03065497	0.02895731	0.03203859
	Z_c	0.4708157	0.3352605	0.9321480	1.6274187
3rd	P	0.2626263	0.2828283	0.2121212	0.2424242
	$\hat{P}$	0.3026316	0.2105263	0.2236842	0.2631579
	É-T	0.03643256	0.03038686	0.03132205	0.03397354
	Z_c	1.0980645	2.3793837	0.3691649	0.6102897
Bottom	P	0.2900000	0.2200000	0.3800000	0.1100000
	$\hat{P}$	0.2500000	0.3125000	0.3125000	0.1250000
	É-T	0.03340766	0.03735089	0.03735089	0.02362278
	Z_c	1.1973302	2.4765139	1.8071859	0.6349803
Top	P	0.1700000	0.2500000	0.1600000	0.4200000
	$\hat{P}$	0.2429577	0.1619718	0.1549296	0.4401408
	É-T	0.02924868	0.02388144	0.02335651	0.03936739
	Z_c	2.4943929	3.6860508	0.2170872	0.5116113
log vraisemblance		1342.26			

TABLE 3.3 – Probabilités et écart-type pour la chaîne de Markov simulée avec N=1000 .

3.1.2 Simulations des chaînes de Markov cachées

On va simuler une CMC , donc on utilise le package "HMM" , précisément les concepts : initHMM, simHMM.

Soit la matrice de transition A , matrice d'émission B et probabilité initiale π :

$$\mathrm{A} = \begin{array}{c} \\ r \\ c \\ s \end{array} \begin{array}{c} \begin{array}{ccc} r & c & s \end{array} \\ \begin{pmatrix} 0.5 & 0.3 & 0.2 \\ 0.4 & 0.2 & 0.4 \\ 0 & 0.3 & 0.7 \end{pmatrix} \end{array} ; \mathrm{B} = \begin{array}{c} \\ r \\ c \\ s \end{array} \begin{array}{c} \begin{array}{cc} h & a \end{array} \\ \begin{pmatrix} 0.9 & 0.1 \\ 0.6 & 0.4 \\ 0.2 & 0.8 \end{pmatrix} \end{array}$$

Tout d'abord, on simule par la chaîne de Markov caché ci-dessus des séquences de tailles N= 50,100, et 300 avec $\pi = (0.2, 0.2, 0.6)$:

États	observations
"s" "s" "s" "s" "s" "s" "s" "s" "c" "r"	"a" "a" "a" "a" "a" "a" "a" "a" "a" "h"
"c" "r" "s" "s" "s" "s" "s" "s" "s" "c"	"a" "h" "a" "a" "a" "a" "a" "a" "a" "a"
"s" "s" "s" "c" "s" "c" "c" "s" "s" "s"	"a" "a" "a" "h" "a" "h" "a" "a" "a" "a"
"s" "c" "s" "c" "c" "s" "s" "s" "s" "s"	"h" "a" "a" "h" "a" "a" "a" "a" "a" "a"
"s" "c" "r" "r" "r" "r" "c" "s" "s" "s"	"a" "a" "h" "h" "h" "a" "a" "h" "a" "a"

TABLE 3.4 – séquence générée pour N=50 .

États	observations
"s" "s" "s" "s" "s" "c" "s" "s" "c" "r"	"h" "a" "a" "a" "h" "a" "a" "a" "h" "h"
"c" "r" "r" "c" "r" "c" "s" "s" "s" "c"	"h" "h" "h" "a" "h" "h" "a" "a" "h" "a"
"r" "s" "s" "s" "c" "c" "s" "s" "s" "c"	"h" "a" "a" "a" "a" "h" "a" "h" "a" "h"
"s" "s" "s" "c" "r" "r" "r" "c" "c" "s"	"a" "a" "a" "a" "a" "a" "h" "h" "h" "a"
"c" "c" "c" "c" "r" "c" "r" "r" "c" "r"	"a" "a" "h" "h" "h" "a" "h" "h" "a" "h"
"s" "s" "c" "s" "s" "c" "s" "c" "s" "s"	"a" "a" "a" "a" "a" "a" "a" "h" "a" "a"
"s" "c" "r" "c" "c" "r" "r" "s" "s" "s"	"a" "h" "h" "h" "a" "h" "h" "a" "a" "a"
"s" "s" "c" "s" "s" "s" "c" "s" "c" "s"	"a" "a" "a" "a" "a" "a" "h" "a" "a" "a"
"c" "c" "r" "s" "c" "r" "r" "c" "r" "r"	"h" "a" "a" "a" "h" "h" "h" "h" "h" "h"
"r" "r" "r" "c" "r" "c" "r" "c" "s" "c"	"h" "h" "h" "h" "a" "a" "h" "h" "a" "h"

TABLE 3.5 – séquence générée pour N=100 .

États	observations
”c” ”r” ”r” ”s” ”c” ”s” ”s” ”s” ”s” ”c”	”a” ”h” ”h” ”a” ”h” ”h” ”a” ”a” ”a” ”a”
”c” ”r” ”s” ”s” ”s” ”s” ”s” ”s” ”c” ”r”	”h” ”a” ”a” ”a” ”a” ”a” ”a” ”a” ”h” ”h”
”s” ”c” ”r” ”c” ”c” ”s” ”s” ”s” ”s” ”s”	”a” ”a” ”h” ”h” ”a” ”a” ”a” ”a” ”h” ”h”
”c” ”s” ”c” ”r” ”r” ”r” ”c” ”r” ”r” ”s”	”a” ”a” ”a” ”h” ”h” ”h” ”a” ”h” ”a” ”a”
”s” ”s” ”s” ”s” ”s” ”s” ”s” ”s” ”s” ”c”	”a” ”a” ”a” ”h” ”a” ”a” ”a” ”a” ”a” ”a”
”s” ”c” ”c” ”s” ”s” ”s” ”c” ”s” ”s” ”s”	”a” ”a” ”h” ”a” ”a” ”a” ”a” ”a” ”a” ”a”
”s” ”s” ”c” ”s” ”s” ”s” ”s” ”s” ”s” ”s”	”h” ”a” ”a” ”a” ”a” ”a” ”a” ”a” ”a” ”a”
”s” ”s” ”s” ”s” ”s” ”c” ”c” ”c” ”r” ”r”	”a” ”a” ”a” ”a” ”a” ”a” ”a” ”h” ”h” ”h”
”c” ”r” ”r” ”r” ”r” ”c” ”s” ”c” ”c” ”c”	”a” ”h” ”h” ”h” ”h” ”a” ”a” ”a” ”h” ”a”
”c” ”s” ”c” ”s” ”s” ”s” ”s” ”c” ”c” ”r”	”a” ”a” ”h” ”a” ”a” ”a” ”a” ”a” ”a” ”h”
”r” ”r” ”r” ”r” ”c” ”r” ”r” ”c” ”c” ”r”	”a” ”h” ”a” ”h” ”h” ”h” ”h” ”a” ”a” ”h”
”r” ”c” ”s” ”s” ”s” ”c” ”s” ”s” ”c” ”s”	”h” ”a” ”a” ”a” ”a” ”a” ”a” ”a” ”a” ”a”
”c” ”r” ”r” ”c” ”c” ”s” ”c” ”r” ”c” ”r”	”h” ”h” ”h” ”h” ”h” ”a” ”a” ”h” ”a” ”a”
”c” ”s” ”s” ”s” ”s” ”c” ”r” ”s” ”s” ”s”	”h” ”a” ”a” ”h” ”a” ”h” ”h” ”a” ”a” ”a”
”s” ”s” ”c” ”s” ”c” ”s” ”c” ”r” ”c” ”r”	”a” ”a” ”a” ”a” ”a” ”a” ”a” ”h” ”h” ”h”
”r” ”r” ”c” ”r” ”r” ”r” ”r” ”s” ”s” ”s”	”a” ”h” ”a” ”a” ”h” ”h” ”h” ”a” ”a” ”a”
”s” ”s” ”s” ”c” ”r” ”c” ”r” ”r” ”c” ”r”	”a” ”h” ”a” ”h” ”h” ”h” ”h” ”h” ”h” ”h”
”s” ”s” ”s” ”s” ”s” ”c” ”r” ”s” ”c” ”r”	”a” ”a” ”a” ”a” ”h” ”a” ”h” ”a” ”h” ”h”
”c” ”r” ”r” ”r” ”c” ”s” ”s” ”s” ”s” ”c”	”a” ”h” ”h” ”h” ”h” ”a” ”a” ”a” ”a” ”h”
”r” ”c” ”r” ”r” ”c” ”r” ”c” ”s” ”c” ”s”	”h” ”a” ”a” ”h” ”a” ”h” ”h” ”a” ”h” ”a”
”c” ”c” ”r” ”c” ”r” ”r” ”r” ”c” ”c” ”s”	”a” ”h” ”a” ”a” ”h” ”h” ”h” ”h” ”a” ”a”
”c” ”s” ”c” ”s” ”s” ”s” ”s” ”s” ”c” ”c”	”h” ”a” ”h” ”a” ”a” ”a” ”a” ”a” ”a” ”a”
”r” ”r” ”r” ”c” ”r” ”c” ”c” ”r” ”s” ”s”	”h” ”h” ”a” ”h” ”h” ”h” ”a” ”h” ”h” ”a”
”s” ”s” ”s” ”s” ”s” ”s” ”s” ”c” ”r” ”r”	”a” ”a” ”h” ”a” ”a” ”a” ”a” ”a” ”h” ”h”
”r” ”s” ”c” ”r” ”s” ”c” ”r” ”c” ”r” ”r”	”h” ”a” ”a” ”a” ”a” ”h” ”h” ”a” ”h” ”h”
”c” ”r” ”r” ”c” ”r” ”r” ”r” ”r” ”r” ”c”	”a” ”a” ”h” ”a” ”h” ”h” ”h” ”h” ”h” ”a”
”s” ”s” ”s” ”c” ”s” ”c” ”c” ”r” ”s” ”c”	”a” ”a” ”h” ”a” ”a” ”h” ”h” ”h” ”a” ”h”
”c” ”s” ”s” ”c” ”s” ”c” ”r” ”r” ”c” ”r”	”h” ”h” ”a” ”h” ”a” ”a” ”h” ”h” ”a” ”h”
”s” ”s” ”s” ”s” ”s” ”s” ”s” ”s” ”s” ”s	a” ”a” ”a” ”h” ”a” ”a” ”a” ”h” ”a” ”a”
” ”s” ”c” ”s” ”s” ”c” ”s” ”s” ”s” ”s” ”s”	”a” ”a” ”a” ”a” ”a” ”a” ”a” ”a” ”a” ”a”

TABLE 3.6 – séquence générée pour N=300 .

En suite , on simule d'autre séquence avec cette fois $\pi = (0.5, 0.5, 0)$ pour N= 50,100,et 300 :

États	observations
”c” ”r” ”r” ”r” ”s” ”s” ”c” ”r” ”s” ”s”	”h” ”h” ”h” ”h” ”a” ”a” ”h” ”h” ”a” ”a”
”s” ”s” ”c” ”s” ”c” ”s” ”s” ”c” ”r” ”r”	”a” ”a” ”a” ”a” ”a” ”a” ”h” ”h” ”h” ”h”
”c” ”r” ”r” ”c” ”r” ”r” ”r” ”c” ”s” ”s”	”h” ”h” ”a” ”a” ”h” ”h” ”h” ”h” ”a” ”a”
”c” ”s” ”c” ”r” ”s” ”c” ”s” ”s” ”c” ”s”	”a” ”a” ”a” ”h” ”a” ”h” ”a” ”a” ”h” ”a”
”c” ”c” ”r” ”s” ”s” ”s” ”s” ”s” ”s” ”s”	”a” ”h” ”h” ”a” ”a” ”a” ”a” ”a” ”h” ”a”

TABLE 3.7 – séquence générée pour N=50 .

États	observations
”c” ”s” ”c” ”r” ”c” ”r” ”c” ”s” ”s” ”c”	”a” ”a” ”a” ”h” ”a” ”h” ”h” ”a” ”a” ”h”
”r” ”r” ”s” ”s” ”c” ”s” ”s” ”s” ”c” ”r”	”h” ”h” ”a” ”a” ”a” ”a” ”a” ”a” ”a” ”h”
”r” ”r” ”r” ”r” ”c” ”r” ”s” ”s” ”s” ”s”	”a” ”h” ”h” ”h” ”h” ”h” ”a” ”a” ”a” ”a”
”c” ”s” ”s” ”s” ”c” ”s” ”s” ”s” ”s” ”s”	”a” ”a” ”a” ”a” ”h” ”a” ”a” ”h” ”a” ”a
”s” ”s” ”s” ”s” ”c” ”s” ”s” ”s” ”s” ”c”	”a” ”a” ”a” ”a” ”h” ”a” ”h” ”a” ”a” ”a”
”c” ”r” ”c” ”s” ”c” ”c” ”r” ”r” ”c” ”s”	a” ”h” ”h” ”a” ”h” ”h” ”h” ”h” ”a” ”a”
”s” ”c” ”r” ”s” ”c” ”r” ”s” ”s” ”s” ”c”	”a” ”h” ”h” ”a” ”h” ”h” ”a” ”h” ”a” ”h”
”c” ”s” ”s” ”c” ”c” ”c” ”c” ”r” ”r” ”c”	”a” ”a” ”a” ”a” ”h” ”a” ”h” ”h” ”h” ”h”
”r” ”r” ”r” ”r” ”c” ”s” ”s” ”c” ”c” ”r”	”a” ”h” ”a” ”h” ”a” ”a” ”a” ”h” ”h” ”a”
”r” ”r” ”c” ”s” ”s” ”s” ”s” ”s” ”s” ”s”	”h” ”h” ”a” ”a” ”a” ”a” ”a” ”a” ”a” ”a”

TABLE 3.8 – séquence générée pour N=100 .

États	observations
”c” ”s” ”s” ”s” ”c” ”s” ”s” ”s” ”c” ”r”	”a” ”h” ”a” ”a” ”a” ”h” ”a” ”a” ”h” ”h”
”s” ”s” ”s” ”s” ”c” ”s” ”s” ”s” ”s” ”s”	”a” ”a” ”a” ”a” ”a” ”a” ”h” ”a” ”a” ”a”
”c” ”r” ”r” ”c” ”r” ”s” ”s” ”s” ”c” ”s”	”a” ”h” ”h” ”h” ”h” ”a” ”a” ”a” ”a” ”a”
”s” ”s” ”c” ”s” ”s” ”c” ”c” ”r” ”r” ”r”	a” ”a” ”h” ”a” ”h” ”a” ”a” ”h” ”h” ”a”
”s” ”s” ”c” ”s” ”c” ”s” ”s” ”s” ”s” ”s”	”a” ”a” ”h” ”a” ”a” ”h” ”a” ”a” ”a” ”a”
”s” ”s” ”s” ”s” ”s” ”c” ”s” ”s” ”c” ”s”	”a” ”a” ”a” ”a” ”a” ”a” ”a” ”a” ”a” ”a”
”c” ”s” ”c” ”s” ”c” ”s” ”s” ”c” ”s” ”s”	”h” ”a” ”a” ”a” ”a” ”a” ”a” ”a” ”a” ”a”
”c” ”c” ”r” ”c” ”r” ”r” ”r” ”s” ”c” ”c”	”h” ”h” ”h” ”h” ”h” ”a” ”a” ”a” ”a” ”a”
”r” ”c” ”s” ”s” ”c” ”s” ”c” ”s” ”s” ”s”	”h” ”a” ”a” ”a” ”h” ”a” ”h” ”a” ”h” ”a”
”s” ”c” ”r” ”s” ”c” ”s” ”s” ”c” ”r” ”r”	”a” ”a” ”h” ”a” ”a” ”h” ”a” ”a” ”h” ”a”
”s” ”s” ”s” ”s” ”c” ”c” ”s” ”s” ”s” ”s”	”a” ”a” ”a” ”a” ”a” ”a” ”a” ”a” ”a” ”a”
”s” ”s” ”c” ”c” ”r” ”c” ”r” ”r” ”r” ”r”	”a” ”a” ”h” ”a” ”a” ”h” ”a” ”h” ”h” ”h”
”s” ”s” ”s” ”c” ”r” ”s” ”s” ”s” ”s” ”s”	”a” ”a” ”a” ”h” ”h” ”a” ”a” ”a” ”h” ”a”
”s” ”c” ”r” ”c” ”r” ”c” ”r” ”c” ”c” ”c”	”a” ”a” ”a” ”a” ”h” ”a” ”h” ”h” ”a” ”a”
”r” ”r” ”r” ”c” ”s” ”c” ”c” ”s” ”s” ”c”	”h” ”a” ”h” ”a” ”a” ”h” ”a” ”a” ”h” ”a”
”s” ”s” ”s” ”c” ”r” ”r” ”c” ”r” ”c” ”r”	”h” ”a” ”a” ”a” ”a” ”h” ”a” ”h” ”a” ”a”
”c” ”r” ”s” ”s” ”s” ”c” ”r” ”c” ”c” ”s”	”h” ”h” ”a” ”a” ”a” ”h” ”h” ”a” ”a” ”a”
”s” ”s” ”s” ”s” ”c” ”r” ”r” ”c” ”s” ”c”	”a” ”a” ”a” ”a” ”h” ”a” ”h” ”a” ”a” ”h”
”c” ”s” ”s” ”s” ”s” ”s” ”s” ”s” ”c” ”s”	”a” ”h” ”a” ”a” ”a” ”a” ”a” ”a” ”h” ”a”
”s” ”s” ”s” ”s” ”c” ”s” ”c” ”s” ”s” ”c”	”a” ”a” ”a” ”a” ”h” ”a” ”h” ”a” ”h” ”a”
”r” ”c” ”r” ”r” ”r” ”r” ”r” ”c” ”c” ”s”	”h” ”a” ”a” ”h” ”h” ”h” ”h” ”a” ”a” ”a”
”s” ”s” ”s” ”s” ”c” ”r” ”c” ”s” ”s” ”c”	”a” ”a” ”a” ”a” ”h” ”a” ”h” ”a” ”a” ”h”
”s” ”s” ”s” ”s” ”s” ”c” ”r” ”r” ”c” ”s”	”a” ”a” ”a” ”a” ”a” ”a” ”a” ”h” ”h” ”a”
”s” ”c” ”r” ”c” ”r” ”c” ”s” ”s” ”s” ”s”	”a” ”h” ”h” ”a” ”a” ”a” ”a” ”a” ”a” ”a”
”s” ”s” ”c” ”s” ”s” ”s” ”s” ”s” ”c” ”c”	”a” ”a” ”a” ”a” ”a” ”a” ”a” ”a” ”a” ”h”
”s” ”s” ”s” ”s” ”s” ”s” ”s” ”s” ”s” ”s”	”a” ”a” ”a” ”a” ”h” ”a” ”a” ”a” ”a” ”a”
”s” ”s” ”s” ”s” ”s” ”s” ”s” ”s” ”c” ”r”	”a” ”a” ”a” ”a” ”a” ”a” ”a” ”h” ”h” ”a”
”s” ”s” ”s” ”s” ”s” ”s” ”c” ”s” ”c” ”c”	”a” ”a” ”a” ”a” ”a” ”a” ”h” ”a” ”a” ”h”
”s” ”s” ”s” ”s” ”s” ”s” ”s” ”s” ”s” ”s”	”a” ”a” ”a” ”h” ”a” ”a” ”a” ”h” ”a” ”a”
”c” ”s” ”s” ”s” ”c” ”c” ”s” ”s” ”s” ”s”	”a” ”a” ”a” ”a” ”a” ”a” ”a” ”a” ”a” ”a”

TABLE 3.9 – séquence générée pour N=300 .

Discussion des résultats des simulations

Les estimations des probabilités de transition et leurs écarts- types selon les tailles d'échantillon de la série simulé sont montrées aux tableaux ci dessus.

On remarque que les estimations des probabilités de transition convergent vers les vraies valeurs lorsque la taille de l'échantillon de la série simulée a un grand nombre de valeurs. Soit lorsque N=1000, les probabilités estimées sont toutes très proches des vraies probabilités de la chaîne. Dans un moins de mesure, c'est également le cas lorsque N=500. Cependant, lorsque N=100, les estimations des probabilités de transition sont parfois éloignées des vraies valeurs.

Les chaînes comportent des estimateurs des probabilités éloignées des vraies valeurs pour certains taille possèdent des écarts-types énormes due aux difficultés des estimations, de plus lorsque N augmente les écarts-types diminuent.

3.2 Application en Biologie : évolution d'une séquence d'ADN

3.2.1 Par un modèle multinomial :

Le modèle multinomial est le plus simple modèle qui prévoit l'évolution d'une séquence d'ADN , ce modèle assume :

- La séquence se produit par un processus aléatoire .
- La probabilité de choisir l'un des quatre nucléotides dépend d'une distribution de probabilité prédéterminée.
- La somme des probabilités des quatre types différents de nucléotides doit être égale à 1.

En utilisant le R pour générer une séquence d'ADN par le modèle multinomial en utilisant la fonction sample() .On suppose que les probabilités des quatre nucléotides $\mathbb{P}_A$,$\mathbb{P}_C$, $\mathbb{P}_G$ et $\mathbb{P}_T$ sont : 0.4,0.1,0.1,0.4 respectivement on a :

$>$ nucleotides $< -c("A","C","G","T")$ ## Définir les noms nucléotide .

$>$ probabilities1 $< -c(0.4, 0.1, 0.1, 0.4)$ ## donner les valeur de probabilité de chaque nucléotide.

$>$ seqlength < -100 ## donner la longueur de séquence.

$>$ sample(nucleotides, seqlength, rep=TRUE, prob=probabilities1) ## généré la séquence .

[1] "A" "C" "G" "T" "T" "C" "T" "A" "T" "C" "T" "T" "C" "T" "A" "A" "T" "T"
[19] "T" "G" "T" "T" "T" "T" "T" "G" "C" "A" "A" "A" "C" "A" "T" "T" "A" "C"
[37]"T" "G" "T" "T" "A" "A" "T" "G" "C" "A" "T" "G" "C" "G" "A" "A" "A" "G"
[55] "T" "T" "A" "G" "A" "T" "T" "A" "A" "A" "G" "A" "T" "A" "T" "A" "A" "C"
[73] "T" "A" "C" "C" "A" "T" "T" "C" "T" "C" "A" "A" "A" "G" "T" "A" "G" "G"
[91] "T" "T" "T" "T" "T" "A" "T" "A" "A" "A".

La séquence générée par ce modèle multinomail est dominée par les deux nucléotides A et T par rapport à G et C, ce résultat peut être justifié en terme des grandes probabilités de A et T.

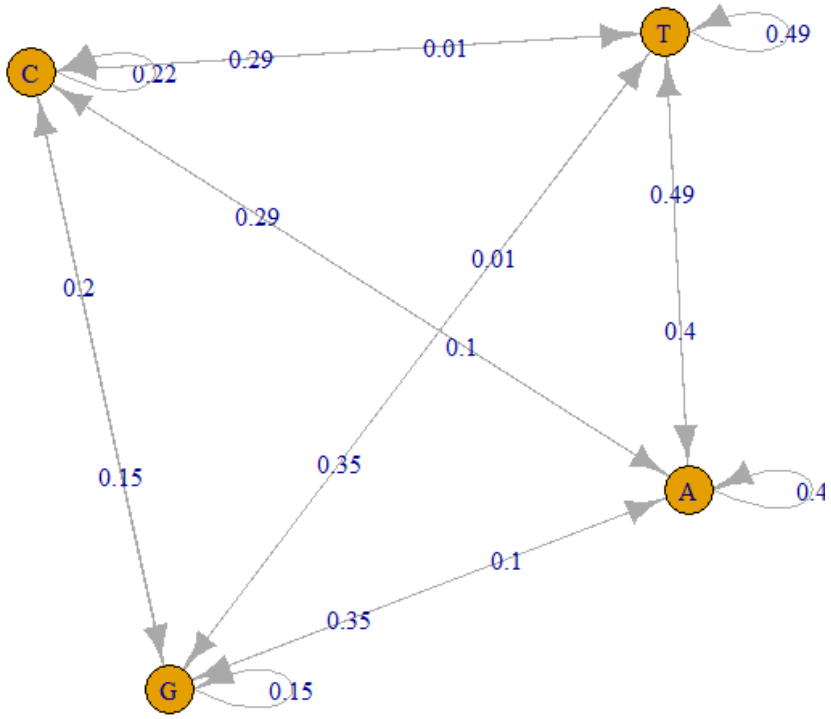
3.2.2 Par un modèle de Markov :

Le modèle multinomial est un bon modèle pour l'évolution de nombreuses séquences d'ADN. Cependant, pour quelques séquences le modèle multinomial n'est pas le modèle adéquat à cause de formation des parties répétitives dans la séquence générée. Le modèle de chaîne de Markov simple peut palier cette insuffisance.

Soit la matrice de transition :

$$\mathrm{P} = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.4 & 0.1 & 0.1 & 0.4 \\ 029 & 0.22 & 0.2 & 0.29 \\ 0.35 & 0.15 & 0.015 & 0.35 \\ 0.49 & 0.01 & 0.01 & 0.49 \end{pmatrix} \end{array}$$

Le diagramme de la chaîne est :



- Pour la probabilité initiale : $\pi = (0.25, 0.25, 0.25, 0.25)$:

[1] "G" "A" "A" "T" "C" "T" "T" "A" "A" "T" "T" "A" "T" "A" "A" "C" "A" "A"
[19] "T" "T" "T" "T" "A" "G" "T" "T" "A" "T" "A" "T" "A" "T" "A" "A" "T" "T"
[37] "T" "A" "T" "A" "G" "T" "T" "A" "T" "A" "T" "A" "A" "T" "T" "T" "T" "A"
[55] "C" "T" "T" "T" "A" "T" "T" "T" "T" "T" "A" "A" "T" "A" "T" "A" "G" "T"
[73]"A" "T" "A" "G" "G" "G" "A" "A" "T" "T" "A" "A" "T" "T" "T" "A" "A" "A"
[91] "C" "T" "T" "T" "A" "T" "C" "T" "T" "A".

- Pour la probabilité initiale : $\pi = (0.3, 0.1, 0.1, 0.3)$ **:**

[1] "T" "T" "T" "T" "T" "A" "A" "T" "A" "T" "A" "A" "A" "A" "A" "A" "A" "T"
[19] "A" "A" "T" "A" "A" "G" "A" "T" "T" "T" "T" "A" "T" "A" "A" "G" "A" "T"
[37] "A" "T" "T" "T" "T" "A" "A" "G" "T" "T" "A" "A" "A" "T" "T" "A" "G" "T"
[55] "T" "T" "T" "T" "A" "A" "T" "T" "T" "T" "T" "A" "C" "A" "A" "A" "G" "A"
[73] "A" "G" "A" "A" "T" "A" "T" "A" "C" "T" "T" "T" "T" "C" "A" "A" "A" "T"
[91] "C" "C" "G" "A" "A" "A" "G" "A" "A" "C".

Lorsque on génére une séquence d'ADN par le modèle de Markov , le nucléotide choisit dans la position actuelle dépend du nucléotide choisit de la position précédente. Il faut notée que nucléotide A est fréquemment situé après le nucléotide T ce qui est normale parce que $\mathbb{P}_{TA} = 0.49$.

3.2.3 Par un modèle de Markov caché :

Le modèle de Markov caché , le nucléotide généré à la position actuelle dépend de son état caché précédent . Par exemple , un modèle de Markov caché pour modéliser la position par le bais d'une séquence appartenant à l'un des deux états suivants : AT-rich , GC-rich

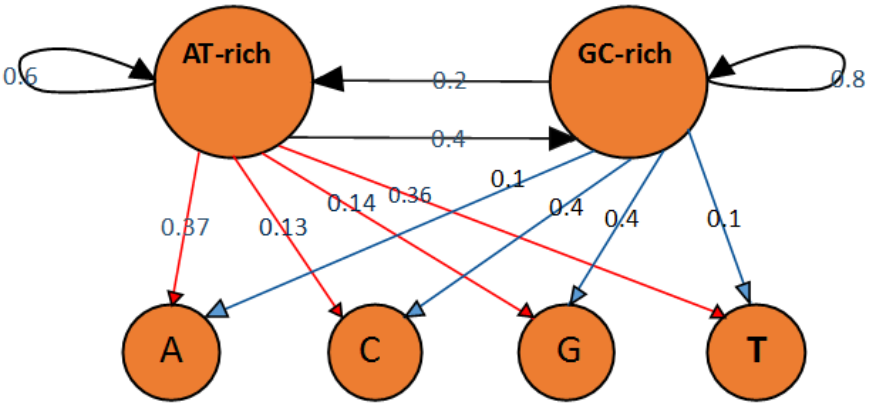
Soit la matrice de transition :

$$\mathrm{A} = \begin{array}{c} \\ AT-rich \\ GC-rich \end{array} \begin{array}{c} \begin{array}{cc} AT-rich & GC-rich \end{array} \\ \begin{pmatrix} 0.6 & 0.4 \\ 0.2 & 0.8 \end{pmatrix} \end{array}$$

Soit la matrice d'émission :

$$\mathrm{B} = \begin{array}{c} \\ AT-rich \\ GC-rich \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.37 & 0.13 & 0.14 & 0.36 \\ 0.10 & 0.40 & 0.40 & 0.10 \end{pmatrix} \end{array}$$

Le diagramme représentative est :



- Pour une probabilité initiale :$\pi = c(0.5, 0.5)$:

[1] "Position 1 , State GC-rich , Nucleotide = T"

[1] "Position 2 , State AT-rich , Nucleotide = A"

[1] "Position 3 , State GC-rich , Nucleotide = C"

[1] "Position 4 , State GC-rich , Nucleotide = C"

[1] "Position 5 , State GC-rich , Nucleotide = C"

[1] "Position 6 , State GC-rich, Nucleotide = C"

[1] "Position 7 , State GC-rich , Nucleotide = C"

[1] "Position 8 , State GC-rich , Nucleotide = G"

[1] "Position 9 , State AT-rich , Nucleotide = T"

[1] "Position 10 , State GC-rich , Nucleotide = C"

[1] "Position 11 , State AT-rich , Nucleotide = G"

[1] "Position 12 , State GC-rich , Nucleotide = A"

[1] "Position 13 , State GC-rich , Nucleotide = G"

[1] "Position 14 , State GC-rich , Nucleotide = G"

[1] "Position 15 , State AT-rich , Nucleotide = A"

[1] "Position 16 , State AT-rich , Nucleotide = T"

[1] "Position 17 , State AT-rich , Nucleotide = A"

[1] "Position 18 , State AT-rich , Nucleotide = A"

[1] "Position 19 , State GC-rich , Nucleotide = G"

[1] "Position 20 , State GC-rich , Nucleotide = G"

[1] "Position 21 , State GC-rich , Nucleotide = C"

[1] "Position 22 , State GC-rich , Nucleotide = G"

[1] "Position 23 , State AT-rich , Nucleotide = A"

[1] "Position 24 , State AT-rich , Nucleotide = C"

[1] "Position 25 , State GC-rich , Nucleotide = C"

[1] "Position 26 , State AT-rich , Nucleotide = A"

[1] "Position 27 , State GC-rich , Nucleotide = A"

[1] "Position 28 , State AT-rich , Nucleotide = A"

[1]"Position 29 , State AT-rich , Nucleotide = G"

[1] "Position 30 , State GC-rich , Nucleotide = C"

[1] "Position 31 , State GC-rich , Nucleotide = T"

[1] "Position 32 , State GC-rich , Nucleotide = G"

[1] "Position 33 , State AT-rich , Nucleotide = T"
[1] "Position 34 , State AT-rich , Nucleotide = A"
[1] "Position 35 , State GC-rich , Nucleotide = G"
[1] "Position 36 , State AT-rich , Nucleotide = A"
[1] "Position 37 , State GC-rich , Nucleotide = C"
[1] "Position 38 , State GC-rich , Nucleotide = C"
[1] "Position 39 , State GC-rich , Nucleotide = G"
[1] "Position 40 , State GC-rich , Nucleotide = C"
[1] "Position 41 , State AT-rich , Nucleotide = A"
[1] "Position 42 , State AT-rich , Nucleotide = A"
[1] "Position 43 , State AT-rich , Nucleotide = A"
[1] "Position 44 , State GC-rich , Nucleotide = C"
[1] "Position 45 , State AT-rich , Nucleotide = T"
[1] "Position 46 , State GC-rich , Nucleotide = G"
[1] "Position 47 , State GC-rich , Nucleotide = G"
[1] "Position 48 , State GC-rich , Nucleotide = C"
[1] "Position 49 , State AT-rich , Nucleotide = T"
[1] "Position 50 , State AT-rich , Nucleotide = T"
[1] "Position 51 , State AT-rich , Nucleotide = C"
[1] "Position 52 , State GC-rich , Nucleotide = G"
[1] "Position 53 , State GC-rich , Nucleotide = A"
[1] "Position 54 , State GC-rich , Nucleotide = C"
[1] "Position 55 , State GC-rich , Nucleotide = A"
[1] "Position 56 , State GC-rich , Nucleotide = G"
[1] "Position 57 , State AT-rich , Nucleotide = T"
[1]"Position 58 , State AT-rich , Nucleotide = A"
[1] "Position 59 , State GC-rich , Nucleotide = G"
[1] "Position 60 , State GC-rich , Nucleotide = T"
[1] "Position 61 , State GC-rich , Nucleotide = G"
[1] "Position 62 , State GC-rich , Nucleotide = G"
[1] "Position 63 , State GC-rich , Nucleotide = T"
[1] "Position 64 , State AT-rich, Nucleotide = A"
[1] "Position 65 , State AT-rich , Nucleotide = T"

[1] "Position 66 , State GC-rich , Nucleotide = C"
[1] "Position 67 , State GC-rich , Nucleotide = G"
[1] "Position 68 , State GC-rich , Nucleotide = A"
[1] "Position 69 , State GC-rich , Nucleotide = T"
[1] "Position 70 , State GC-rich , Nucleotide = G"
[1] "Position 71 , State GC-rich , Nucleotide = C"
[1] "Position 72 , State GC-rich , Nucleotide = T"
[1] "Position 73 , State AT-rich , Nucleotide = A"
[1] "Position 74 , State GC-rich , Nucleotide = T"
[1] "Position 75 , State AT-rich , Nucleotide = G"
[1] "Position 76 , State AT-rich , Nucleotide = A"
[1] "Position 77 , State AT-rich , Nucleotide = C"
[1] "Position 78 , State GC-rich , Nucleotide = G"
[1] "Position 79 , State GC-rich , Nucleotide = T"
[1] "Position 80 , State GC-rich , Nucleotide = G"
[1]"Position 81 , State GC-rich , Nucleotide = C"
[1] "Position 82 , State GC-rich , Nucleotide = C"
[1] "Position 83 , State AT-rich , Nucleotide = A"
[1] "Position 84 , State AT-rich , Nucleotide = A"
[1] "Position 85 , State GC-rich , Nucleotide = C"
[1] "Position 86 , State GC-rich , Nucleotide = A"
[1] "Position 87 , State GC-rich , Nucleotide = C"
[1] "Position 88 , State AT-rich , Nucleotide = C"
[1] "Position 89 , State AT-rich , Nucleotide = T"
[1] "Position 90 , State GC-rich , Nucleotide = A"
[1] "Position 91 , State GC-rich , Nucleotide = A"
[1] "Position 92 , State GC-rich , Nucleotide = C"
[1] "Position 93 , State GC-rich , Nucleotide = T"
[1] "Position 94 , State GC-rich , Nucleotide = C"
[1] "Position 95 , State AT-rich , Nucleotide = G"
[1] "Position 96 , State GC-rich , Nucleotide = G"
[1] "Position 97 , State GC-rich , Nucleotide = G"
[1] "Position 98 , State GC-rich , Nucleotide = C"

[1] "Position 99 , State GC-rich , Nucleotide = C"

[1] "Position 100 , State GC-rich , Nucleotide = C"

- Pour une probabilité initiale : $\pi = c(0.75, 0.25)$ **:**

[1] "Position 1 , State AT-rich , Nucleotide = A"

[1] "Position 2 , State AT-rich , Nucleotide = A"

[1]"Position 3 , State AT-rich , Nucleotide = C"

[1] "Position 4 , State GC-rich , Nucleotide = C"

[1] "Position 5 , State GC-rich , Nucleotide = G"

[1] "Position 6 , State GC-rich , Nucleotide = A"

[1] "Position 7 , State AT-rich , Nucleotide = A"

[1] "Position 8 , State GC-rich , Nucleotide = G"

[1] "Position 9 , State GC-rich , Nucleotide = G"

[1] "Position 10 , State GC-rich , Nucleotide = C"

[1] "Position 11 , State GC-rich , Nucleotide = C"

[1] "Position 12 , State GC-rich , Nucleotide = T"

[1] "Position 13 , State GC-rich , Nucleotide = C"

[1] "Position 14 , State GC-rich , Nucleotide = C"

[1] "Position 15 , State GC-rich , Nucleotide = C"

[1] "Position 16 , State GC-rich , Nucleotide = C"

[1] "Position 17 , State GC-rich , Nucleotide = C"

[1] "Position 18 , State GC-rich , Nucleotide = C"

[1] "Position 19 , State GC-rich , Nucleotide = G"

[1] "Position 20 , State GC-rich , Nucleotide = T"

[1] "Position 21 , State GC-rich , Nucleotide = G"

[1] "Position 22 , State GC-rich , Nucleotide = G"

[1] "Position 23 , State GC-rich , Nucleotide = C"

[1] "Position 24 , State GC-rich , Nucleotide = G"

[1] "Position 25 , State GC-rich , Nucleotide = G"

[1] "Position 26 , State GC-rich , Nucleotide = G"

[1] "Position 27 , State GC-rich , Nucleotide = C"

[1] "Position 28 , State GC-rich , Nucleotide = G"

[1] "Position 29 , State AT-rich , Nucleotide = A"

[1] "Position 30 , State AT-rich , Nucleotide = T"
[1] "Position 31 , State AT-rich , Nucleotide = C"
[1] "Position 32 , State AT-rich , Nucleotide = A"
[1] "Position 33 , State AT-rich , Nucleotide = C"
[1] "Position 34 , State AT-rich , Nucleotide = A"
[1] "Position 35 , State GC-rich , Nucleotide = G"
[1] "Position 36 , State GC-rich , Nucleotide = G"
[1] "Position 37 , State GC-rich , Nucleotide = G"
[1] "Position 38 , State GC-rich , Nucleotide = T"
[1] "Position 39 , State GC-rich , Nucleotide = G"
[1] "Position 40 , State AT-rich , Nucleotide = G"
[1] "Position 41 , State GC-rich , Nucleotide = C"
[1] "Position 42 , State GC-rich , Nucleotide = C"
[1] "Position 43 , State GC-rich , Nucleotide = G"
[1] "Position 44 , State AT-rich , Nucleotide = C"
[1] "Position 45 , State GC-rich , Nucleotide = C"
[1] "Position 46 , State AT-rich , Nucleotide = G"
[1] "Position 47 , State GC-rich , Nucleotide = G"
[1] "Position 48 , State GC-rich , Nucleotide = T"
[1] "Position 49 , State GC-rich , Nucleotide = C"
[1] "Position 50 , State GC-rich , Nucleotide = C"
[1] "Position 51 , State AT-rich , Nucleotide = T"
[1] "Position 52 , State AT-rich , Nucleotide = G"
[1] "Position 53 , State AT-rich , Nucleotide = C"
[1] "Position 54 , State AT-rich , Nucleotide = A"
[1] "Position 55 , State AT-rich , Nucleotide = A"
[1] "Position 56 , State AT-rich , Nucleotide = C"
[1] "Position 57 , State AT-rich , Nucleotide = T"
[1] "Position 58 , State AT-rich , Nucleotide = T"
[1] "Position 59 , State GC-rich , Nucleotide = T"
[1] "Position 60 , State AT-rich , Nucleotide = A"
[1] "Position 61 , State AT-rich , Nucleotide = T"
[1] "Position 62 , State GC-rich , Nucleotide = C"

[1] "Position 63 , State GC-rich , Nucleotide = C"
[1] "Position 64 , State GC-rich , Nucleotide = T"
[1] "Position 65 , State GC-rich , Nucleotide = G"
[1] "Position 66 , State GC-rich , Nucleotide = G"
[1] "Position 67 , State GC-rich , Nucleotide = G"
[1] "Position 68 , State GC-rich , Nucleotide = C"
[1] "Position 69 , State AT-rich , Nucleotide = T"
[1] "Position 70 , State GC-rich , Nucleotide = G"
[1] "Position 71 , State GC-rich , Nucleotide = T"
[1] "Position 72 , State AT-rich , Nucleotide = G"
[1] "Position 73 , State AT-rich , Nucleotide = A"
[1] "Position 74 , State AT-rich , Nucleotide = A"
[1] "Position 75 , State GC-rich , Nucleotide = G"
[1] "Position 76 , State AT-rich , Nucleotide = T"
[1] "Position 77 , State AT-rich , Nucleotide = A"
[1] "Position 78 , State GC-rich , Nucleotide = C"
[1] "Position 79 , State GC-rich , Nucleotide = T"
[1] "Position 80 , State AT-rich , Nucleotide = A"
[1] "Position 81 , State GC-rich , Nucleotide = G"
[1]"Position 82 , State AT-rich , Nucleotide = A"
[1] "Position 83 , State AT-rich , Nucleotide = A"
[1] "Position 84 , State AT-rich , Nucleotide = A"
[1] "Position 85 , State GC-rich , Nucleotide = C"
[1] "Position 86 , State GC-rich , Nucleotide = G"
[1] "Position 87 , State GC-rich , Nucleotide = G"
[1] "Position 88 , State GC-rich , Nucleotide = C"
[1] "Position 89 , State GC-rich , Nucleotide = G"
[1] "Position 90 , State GC-rich , Nucleotide = G"
[1] "Position 91 , State GC-rich , Nucleotide = C"
[1] "Position 92 , State GC-rich , Nucleotide = C"
[1] "Position 93 , State GC-rich , Nucleotide = G"
[1] "Position 94 , State AT-rich , Nucleotide = C"
[1] "Position 95 , State GC-rich , Nucleotide = C"

[1] "Position 96 , State GC-rich , Nucleotide = C"

[1] "Position 97 , State GC-rich , Nucleotide = C"

[1] "Position 98 , State GC-rich , Nucleotide = C"

[1] "Position 99 , State GC-rich , Nucleotide = G"

[1] "Position 100 , State GC-rich , Nucleotide = C"

Algorithmes

Pour la séquence d'ADN suivante :
"G" "T" "G" "C" "G" "T" "G" "G" "A" "G" "C" "A" "G" "C" "G" "A" "C" "G" "A" "C" "A" "T" "G" "C" "C" "G" "T" "G" "T" "A".

Soit probabilité initiale $\pi = (0.5, 0.5)$.

On applique les trois algorithme fondamentaux :

1- Algorithme de Forward -Backward

1-1 Algorithme de Forward

state	1	2	3	4	5	6	7	8
$AT-rich$	0.07	0.02952	0.00300608	0.0005136934	0.0001528407	0.0001291909	$1.40154e-05$	$2.768346e-06$
$GC-rich$	0.20	0.01880	0.01073920	0.0039175168	0.0013357963	0.0001129773	$5.68233e-05$	$2.042592e-05$

state	9	10	11	12	13	14
$AT-rich$	$2.126091e-06$	$2.274462e-07$	$4.110214e-08$	$3.309442e-08$	$3.551537e-09$	$6.439697e-10$
$GC-rich$	$1.744807e-06$	$8.985129e-07$	$3.239155e-07$	$2.755733e-08$	$1.411345e-08$	$5.084551e-09$

state	15	16	17	18	19	20
$AT-rich$	$1.964609e-10$	$1.716411e-10$	$1.719091e-11$	$3.523530e-12$	$2.744421e-12$	$2.728830e-13$
$GC-rich$	$1.730091e-09$	$1.462657e-10$	$7.426761e-11$	$2.651618e-11$	$2.262236e-12$	$1.163023e-12$

state	21	22	23	24	25	26
$AT-rich$	$1.466437e-13$	$3.915996e-14$	$3.686541e-15$	$5.684525e-16$	$1.495641e-16$	$5.137212e-17$
$GC-rich$	$1.039571e-13$	$1.418232e-14$	$1.080394e-14$	$4.047106e-15$	$1.386026e-15$	$4.674587e-16$

state	27	28	29	30
$AT-rich$	$4.475340e-17$	$4.863930e-18$	$2.475132e-18$	$6.810041e-19$
$GC-rich$	$3.945158e-17$	$1.978505e-17$	$1.777361e-18$	$2.411942e-19$

TABLE 3.10 – Probabilités d'émissions estimées par l'algorithme Forward.

1-2 Algorithme de Backward

state	1	2	3	4	5	6
$AT-rich$	$4.350105e-18$	$1.558821e-17$	$4.714373e-17$	$1.662431e-16$	$8.491536e-16$	$3.069704e-15$
$GC-rich$	$3.088455e-18$	$2.457630e-17$	$7.267585e-17$	$2.136048e-16$	$5.932140e-16$	$4.652441e-15$
state	7	8	9	10	11	12
$AT-rich$	$1.062118e-14$	$5.452031e-14$	$1.926524e-13$	$6.536973e-13$	$3.434680e-12$	$1.204908e-11$
$GC-rich$	$1.360953e-14$	$3.775924e-14$	$2.937871e-13$	$8.608861e-13$	$2.411201e-12$	$1.899462e-11$
state	13	14	15	16	17	18
$AT-rich$	$3.645387e-11$	$1.287637e-10$	$6.604497e-10$	$2.332060e-09$	$8.429347e-09$	$4.566160e-08$
$GC-rich$	$5.616847e-11$	$1.650644e-10$	$4.580369e-10$	$3.568306e-09$	$1.046607e-08$	$2.871108e-08$
state	19	20	21	22	23	24
$AT-rich$	$1.692219e-07$	$1.046812e-06$	$4.179212e-06$	$1.495207e-05$	$4.403579e-05$	0.0001364610
$GC-rich$	$2.023583e-07$	$5.473160e-07$	$2.975679e-06$	$2.373913e-05$	$7.033165e-05$	0.0002086989

state	25	26	27	28	29	30
$AT-rich$	0.0004940234	0.002685729	0.01026061	0.062752	0.262	1
$GC-rich$	0.0006120448	0.001677639	0.01173594	0.031184	0.154	1

TABLE 3.11 – Probabilités d'émissions estimées par l'algorithme Backward.

2- Algorithme de Viterbi

On garde le même modèle de Markov caché cité précédemment, le programme de l'algorithme Viterbi (voir annexe) nous a fournit le résultat suivant :

[1] "Positions 1 - 2 Most probable state = AT-rich"

[1] "Positions 3 - 21 Most probable state = GC-rich"

[1] "Positions 22 - 22 Most probable state = AT-rich"

[1] "Positions 23 - 29 Most probable state = GC-rich"

[1] "Positions 30 - 30 Most probable state = AT-rich"

3- Algorithme de Baum-Welch

L'application de l'algorithme de Baum-Welch sur le modèle de Markov caché traité nous a conduit aux résultats suivants :

La matrice de transition estimée :

$$\mathrm{A} = \begin{array}{c} \\ AT-rich \\ GC-rich \end{array} \begin{array}{c} \begin{array}{cc} AT-rich & GC-rich \end{array} \\ \begin{pmatrix} 0.4345201 & 0.5654799 \\ 0.2280259 & 0.7719741 \end{pmatrix} \end{array}$$

La matrice d'émission estimée :

$$\mathrm{B} = \begin{array}{c} \\ AT-rich \\ GC-rich \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.1730538 & 0.09023724 & 0.4621759 & 0.2745330 \\ 0.2115375 & 0.29460242 & 0.3733783 & 0.1204818 \end{pmatrix} \end{array}$$

Conclusion

Le modèle de Markov Caché est un outil probabiliste dans lequel les données observés (comme dans notre cas la séquence ADN) sont modélisées comme une série de sortie (ou émission) générées par l'un de plusieurs états internes (cachés).

Ce travail est juste une petite introduction aux modèles CMC et comment les utilise afin de modéliser une séquence ADN, on a donné quelques outils de base sur les processus stochastiques, les chaînes de Markov et les chaînes de Markov cachés.

Dans la partie pratique, on a commencé tout d'abord à mettre en place de quelques simulations des chaîne de Markov simple et cachés afin de comprendre le comportement stochastique de ces modèles. Par la suite, une application empirique a été entamé dont l'objet est de modéliser l'évolution d'une séquence ADN par trois approches : modèle multinomial, modèle de Markov, et modèle de Markov cachés. Cette démarche était pour but d'évoquer l'avantage de modèle de Markov caché par rapport aux modèles classiques en terme de la prise en compte d'une quantité importante d'informations, et ca capacité à bien comprendre le comportement dynamique de la série étudiés. Cette application a été clôturé par l'application de multiples algorithmes d'inférence des modèle de Markov cachés à savoir : Forward, Backward, Viterbi , et Baum-Welch.

Nous espérons avoir répondu au problème posé et les résultats trouvés seront d'une utilité pertinente.

Bibliographie

[1] **ALAN RUGG** : Processus Stochastiques : avec applications aux phénomènes d'attente et fiabilité . Lausanne,Suisse (1989).

[2] **Amélie Guilbaut** et **Axel Belotti** :Chaînes de Markov cachées , Sciences et Technologies Master 1 ISN , Université de Lille , 11 mai 2018.

[3] **BELBEDJ FARID** : Les modèles de Markov cachés et leur application dans un processus industriel , Mémoire de magistère , Département de génie industriel , Université Hadj Lakhdar de Btana , Année Universitaire 2014 - 2015 .

[4] **BERBECHE Kamal** : Modèles de Markov Cachés : Application à La Reconnaissance Automatique de la Parole ,Mémoire de magistère , Département d'Électronique, Université Mouloud Mammeri de Tizi Ouzou.

[5] CHAÎNES DE MARKOV , Cours de" INGENIEUR, 1ère année" , Université Paris-Est-Créteil .

[6] Cours sur internet : Les chaînes de Markov cachées , "chapitre 2.pdf " .

[7] Cours sur internet : MODÈLE DE MARKOV CACHES , "chapitre2.pdf ".

[8] **Dominique Foata** et **Aimé Fuchs** : Processus Stochastiques .

[9] **GRELA Fabrice** : Introduction aux processus d'exclusions ,Rapport de stage , Institut Camille Jordan Université de Lyon 1, 2016 .

[10] **TOUCHE Nassim** : Chaîne de Markov à temps discrète , Cours de Troisième année Mathématiques Appliquées ,Université Abderrahmane Mira de Béjaia , Année Universitaire 2017 - 2018 .

[11] **Jean-François Hêche** : Chaîne de Markov finies et homogènes à temps discret , Recherche Opérationnelle , École Polytechnique Fédérale de Lausanne .

[12] **Jean-François Hêche** : Chaines de Markov finies et homogènes à temps continu , Recherche Opérationnelle, École Polytechnique Fédérale de Lausanne .

[13] **JEAN-BAPTISTE VOUMA LEKOUNDJI** : MODÈLES DE MARKOV CACHÉS , Université du QUÉBEC À MONTRÉAL , Septembre 2014 .

[14] **Sébastien Aupetit** Contributions aux Modèles de Markov Cachés :métaheuristiques d'apprentissage,nouveaux modèles et visualisation de dissimilarité , Université François Rabelais-Tours,2005.Français, tel-00168392 .

Annexe

Les fonctions dans R

######### **Par multinational**

> nucleotides < − c(”A”, ”C”, ”G”, ”T”) # Définie les noms nucléotide

> probabilities1 < − c(0.4, 0.1, 0.1, 0.4) # donner lev valeur de probabilité de chaque nucléotide

> seqlength < − 100 # donner la longueur de séquence

> sample(nucleotides, seqlength, rep=TRUE, prob=probabilities1) # généré la séquence .

######### **Par chaîne de Markov**

>library(markovchain)

> E=statesNames = c (”A”,”C”,”G”,”T”)

> E

> mc < −new(”markovchain”, states=statesNames, transitionMatrix=matrix($c(0.4, 0.1, 0.1, 0.4, 0.29, 0.22, 0.20, 0.29, 0.35, 0.15, 0.15, 0.35, 0.49, 0.01, 0.01, 0.49)$, nrow =4,byrow=TRUE, dimnames=list(statesNames, statesNames)))

> mc

> plot(mc)

> ###

> nucleotides < − c(”A”, ”C”, ”G”, ”T”) # Définir l’alphabet des nucléotides

> afterAprobs < − c(0.4,0.1,0.1,0.4) # donner les valeurs des probabilités, où le nucléotide précédent était ”A”

> afterCprobs < − c(0.29,0.22,0.20,0.29) # Sdonner les valeurs des probabilités, où le nucléotide précédent était ”C”

> afterGprobs < − c(0.35,0.15,0.15,0.35) # donner les valeurs des probabilités, où le nucléotide précédent était ”G”

> afterTprobs < − c(0.49,0.01,0.01,0.49) # donner les valeurs des probabilités, où le nucléotide précédent était "T"
> mytransitionmatrix < − matrix(c(afterAprobs, afterCprobs, afterGprobs, afterTprobs), 4, 4, byrow = TRUE) #créer une matrice 4 x 4
> rownames(mytransitionmatrix) < − nucleotides
> colnames(mytransitionmatrix) < − nucleotides
> mytransitionmatrix ### afficher la matrice de transition
> ##
> generatemarkovseq < − function(transitionmatrix, initialprobs, seqlength)
+ {
+ nucleotides < − c("A", "C", "G", "T")
+ mysequence < − character() # Créer un vecteur pour stocker la nouvelle séquence
+ # Choisissez le nucléotide pour la première position dans la séquence :
+ firstnucleotide < − sample(nucleotides, 1, rep=TRUE, prob=initialprobs)
+ mysequence[1] < − firstnucleotide # Stocker le nucléotide pour la première position de la séquence
+ for (i in 2 : seqlength)
+ {
+ prevnucleotide< − mysequence[i-1] # Obtenir le nucléotide précédent dans la nouvelle séquence
+ # Obtenez les probabilités du nucléotide actuel, étant donné le nucléotide précédent "prevnucleotide" :
+ probabilities < − transitionmatrix[prevnucleotide,]
+ # Choisissez le nucléotide à la position actuelle de la séquence :
+ nucleotide < − sample(nucleotides, 1, rep=TRUE, prob=probabilities)
+ mysequence[i] < − nucleotide # Stocker le nucléotide pour la position actuelle de la séquence
+ }
+ return(mysequence)
+ }
> #### pour probabilité intaille (0.25, 0.25, 0.25, 0.25)
> myinitialprobs < − c(0.25, 0.25, 0.25, 0.25)
> generatemarkovseq (mytransitionmatrix, myinitialprobs, 100)
> ##### pour probabilité intaille (0.3,0.1,0.1,0.3)

> myinitialprobs <− c(0.3,0.1,0.1,0.3)

> generatemarkovseq(mytransitionmatrix, myinitialprobs, 100)

> ##### **Par HMM**.

> ##### la matrice de transition .

> states <− c("AT-rich", "GC-rich") # Définir les noms des états

> ATrichprobs <− c(0.4,0.6) # Définir les probabilités d'états de commutation, où l'état précédent était "AT-rich".

> GCrichprobs <− c(0.8,0.2) # Définir les probabilités d'états de commutation, où l'état précédent était "GC-rich".

> thetransitionmatrix <− matrix(c(ATrichprobs, GCrichprobs), 2, 2, byrow = TRUE) # Créer une matrice 2 x 2 .

> rownames(thetransitionmatrix) <− states

> colnames(thetransitionmatrix) <− states

> thetransitionmatrix # afficher la matrice de transition.

> ##### la matrice d'émission.

> nucleotides <− c("A", "C", "G", "T") # Définir l'alphabet des nucléotides

> ATrichstateprobs <− c(0.13,0.37,0.36,0.14) # Définir les valeurs des probabilités, pour l'état AT-rich .

> GCrichstateprobs <− c(0.4,0.1,0.1,0.4) # Définir les valeurs des probabilités, pour l'état GC-rich .

> theemissionmatrix <− matrix(c(ATrichstateprobs, GCrichstateprobs), 2, 4, byrow = TRUE) # Créer une matrice 2 x 4.

> rownames(theemissionmatrix) <− states

> colnames(theemissionmatrix) <− nucleotides

> theemissionmatrix # afficher la matrice d'émission

> ##### Fonction pour générer une séquence d'ADN, étant donné un HMM et la longueur de la séquence à générer.

> generatehmmseq <− function(transitionmatrix, emissionmatrix, initialprobs, seqlength)

+ {

+ nucleotides ¡- c("A", "C", "G", "T") # Définir l'alphabet des nucléotides.

+ states <− c("AT-rich", "GC-rich") # Définir les noms des états.

+ mysequence <− character() # Créer un vecteur pour stocker la nouvelle séquence.

+ mystates <− character() # Créer un vecteur pour stocker l'état que chaque position dans

la nouvelle séquence.
+ # a été générée
+ ## Choisissez l'état de la première position dans la séquence :
+ firststate < − sample(states, 1, rep=TRUE, prob=initialprobs) + # Obtenez les probabilités du nucléotide actuel, étant donné que nous sommes dans l'état "firststate" :
+ probabilities < − emissionmatrix[firststate,]
+ # Choisissez le nucléotide pour la première position dans la séquence :
+ firstnucleotide < − sample(nucleotides, 1, rep=TRUE, prob=probabilities)
+ mysequence[1] < − firstnucleotide # Stocker le nucléotide pour la première position de la séquence.
+ mystates[1] < − firststate # Stocke l'état dans lequel la première position de la séquence a été générée par :
+ for (i in 2 :seqlength) + { + prevstate < − mystates[i-1] # btenir l'état dans lequel le nucléotide précédent de la séquence a été généré par :
+ # Obtenir les probabilités de l'état actuel, étant donné que le nucléotide précédent a été généré par état "prevstate"
+ stateprobs < − transitionmatrix[prevstate,] + # Choisissez l'état de la ième position dans la séquence :
+ state < − sample(states, 1, rep=TRUE, prob=stateprobs)
+ # Obtenez les probabilités du nucléotide actuel, étant donné que nous sommes dans l'état "state" :
+ probabilities < − emissionmatrix[state,]
+ # Choisissez le nucléotide pour la ième position dans la séquence :
+ nucleotide < − sample(nucleotides, 1, rep=TRUE, prob=probabilities)
+ mysequence[i] < − nucleotide # Stocker le nucléotide pour la position actuelle de la séquence
+ mystates[i] < − state # Stocke l'état dans lequel la position actuelle dans la séquence a été générée par
+ }
+ for (i in 1 :length(mysequence))
+ {
+ nucleotide < − mysequence[i]
+ state < − mystates[i]
+ print(paste("Position", i, ", State", state, ", Nucleotide = ", nucleotide))

+ }
+ }
> ##### prob init = (0.5, 0.5)
> theinitialprobs < − c(0.5, 0.5)
> generatehmmseq(thetransitionmatrix, theemissionmatrix, theinitialprobs, 100)
> ##### prob init = (0.75,0.25)
> theinitialprobs < − c(0.75,0.25)
> generatehmmseq(thetransitionmatrix, theemissionmatrix, theinitialprobs, 100)
> ##### **Algorithme**
> ##### Définie la chaîne de Markov caché
> library(HMM)
>stat=c(”AT-rich”,”GC-rich”)
> stat
> symbol=c(”A”,”C”,”G”,”T”)
> symbol
> startProb=c(0.75,0.25)
> transProb=matrix(c(0.6,0.4,0.8,0.2),nrow=2,byrow=TRUE)
> emissionProb=matrix(c(0.37, 0.13, 0.14, 0.136, 0.10,0.40,0.40, 0.10),nrow=4,byrow=TRUE)
> cmc=initHMM(stat, symbol, startProb, transProb, emissionProb)
>cmc
> observation=c(”G”, ”T” ,”G” ,”C” ,”G” ,”T”, ”G” ,”G” ,”A”, ”G” ,”C”, ”A”, ”G”, ”C”, ”G” ,”A”, ”C”, ”G” ,”A”, ”C”, ”A” ,”T” ,”G” ,”C” ”C” ,”G” ,”T” ,”G” ,”T” ,”A”)
>observation
> ### **Algorithme de Forward**
> logForwardProbabilities = forward(cmc,observation)
> print(exp(logForwardProbabilities))
> ### **Algorithme de Backward**
> logBackwardProbabilities = backward(cmc,observation)
> print(exp(logBackwardProbabilities))
> ### **Algorithme de Viterbi**
> viterbi < − function(sequence, transitionmatrix, emissionmatrix)
+ {
+ # Obtenez les noms des états dans le HMM :

+ states <- rownames(theemissionmatrix)
+ # Faire la matrice de Viterbi v :
+ v <- makeViterbimat(sequence, transitionmatrix, emissionmatrix)
+ # Parcourez chacune des lignes de la matrice v (où chaque ligne représente
+ # une position dans la séquence d'ADN), et découvrez quelle colonne a le
+ # valeur maximale pour cette ligne (où chaque colonne représente un état de
+ # le HMM) :
+ mostprobablestatepath <- apply(v, 1, function(x) which.max(x))
+ # afficher le chemin d'état le plus probable :
+ prevnucleotide <- sequence[1]
+ prevmostprobablestate <- mostprobablestatepath[1]
+ prevmostprobablestatename <- states[prevmostprobablestate]
+ startpos <- 1
+ for (i in 2 :length(sequence))
+ {
+ nucleotide <- sequence[i]
+ mostprobablestate <- mostprobablestatepath[i]
+ mostprobablestatename <- states[mostprobablestate]
+ if (mostprobablestatename != prevmostprobablestatename)
+ {
+ print(paste("Positions",startpos,"-",(i-1), "Most probable state = ", prevmostprobablestatename))
+ startpos <- i
+ }
+ prevnucleotide ¡- nucleotide
+ prevmostprobablestatename <- mostprobablestatename
+ }
+ print(paste("Positions",startpos,"-",i, "Most probable state = ", prevmostprobablestatename))
+ }
> makeViterbimat <- function(sequence, transitionmatrix, emissionmatrix)
+ # Cela rend la matrice v en utilisant l'algorithme de Viterbi.
+ {

+ sequence <- toupper(sequence)
+ # Découvrez combien d'états sont dans le HMM
+ numstates <- dim(transitionmatrix)[1]
+ # Faire une matrice avec autant de lignes que de positions dans la séquence, et autant de
+ # colonnes en tant qu'états dans le HMM
+ v <- matrix(NA, nrow = length(sequence), ncol = dim(transitionmatrix)[1])
+ #Définissez les valeurs de la première ligne de la matrice v (représentant la première position
+ de la séquence) à 0
+ v[1,] <- 0
+ #Définissez la valeur dans la première ligne de la matrice v, première colonne à 1.
+ v[1,1] <- 1
+ # Remplissez la matrice v :
+ for (i in 2 :length(sequence)) # For each position in the DNA sequence :
+ {
+ for (l in 1 :numstates) # For each of the states of in the HMM :
+ {
+ # Trouver la probabilité, si on est dans l'état l, de choisir le nucléotide à la position dans
+ la séquence
+ statelprobnucleotidei <- emissionmatrix[l,sequence[i]]

+ # v[(i-1),]donne les valeurs de v pour la (i-1)ème rangée de v, ie. la (i-1)ème position dans la séquence.
+ # Dans v[(i-1),] il y a des valeurs de v à la (i-1)ème rangée de la séquence pour chaque état
+ possible k
+ # v[(i-1),k] donne la valeur de v à la (i-1)ième ligne de la séquence pour un état
+ particulier k.
+ # transitionmatrix[l,] donne les valeurs de la lième ligne de la matrice de transition, .
+ # probabilités de passer d'un état précédent k à un état actuel l.
+ # max(v[(i-1),] * transitionmatrix[l,]) est la probabilité maximale pour le nucléotide observé
+ à la position précédente dans la séquence à l'état k, suivi d'une transition de l'état précédent
+ k à l'état actuel l à la position actuelle du nucléotide.
+ # Définissez la valeur dans la matrice v pour la ligne i (position de nucléotide i), colonne l
+ (état l) comme suit :

```
+ v[i,l] <- statelprobnucleotidei * max(v[(i-1),] * transitionmatrix[,l])
+ }
+ }
+ return(v)
+ }
> myseq <- c("G", "T", "G" ,"C", "G" ,"T", "G", "G", "A" ,"G", "C" ,"A", "G" ,"C",
"G", "A", "C", "G", "A", "C", "A", "T" ,"G", "C" ,"C" ,"G", "T", "G" ,"T" ,"A")
> viterbi(myseq, thetransitionmatrix, theemissionmatrix)
> ### Algorithme de Baum-Welch
> bw = baumWelch(cmc,observation,10)
> bw
> print(bw$hmm)
```