Chaînes de Markov régulées et approximation de Poisson pour l'analyse de séquences biologiques

Nicolas Vergne

Université d'Evry Val d'Essonne

École doctorale *Des Génomes aux Organismes*

Chaînes de Markov régulées et approximation de Poisson pour l'analyse de séquences biologiques

THÈSE

présentée et soutenue publiquement le 11 juillet 2008

pour l'obtention du

Doctorat de l'université d'Evry Val d'Essonne

(spécialité Mathématiques Appliquées)

par

Nicolas Vergne

Composition du jury

Président :	Gregory Nuel	Chargé de recherche CNRS, Université Paris Descartes
Rapporteurs :	Michel Termier	Maître de Conférences, Université Paris-Sud
	Benoît Saussol	Professeur, Université de Bretagne Occidentale
Examinateurs :	Dominique Cellier	Maître de Conférences, Université de Rouen
	Jacques van Helden	Chargé de cours, Université Libre de Bruxelles
Directeur de thèse :	Bernard Prum	Professeur, Université d'Evry

Laboratoire Statistique et Génome – UEVE – UMR CNRS 8071 **– UMR INRA** 1152

Mis en page avec la classe thloria.

Liste des publications

Ce travail a donné lieu aux différentes publications suivantes :

Articles

- N. Vergne. Drifting Markov Models with Polynomial Drift and Applications to DNA Sequences. *Statistical Applications in Genetics and Molecular Biology,* **7**(1), http://www.bepress.com/sagmb/vol7/iss1/art6/
- N. Vergne and G. Nuel. Pattern Statistics according to Drifting Markov Models, *in progress.*
- N. Vergne and M. Termier. Drifting Markov Models with Polynomial Splines and Applications to DNA sequences, *in progress.*
- M. Abadi and N. Vergne. Sharp error terms for return time statistics under mixing conditions, *accepted in Journal of Theoretical Probability.*
- M. Abadi and N. Vergne. Sharp error terms for point-wise Poisson approximations under mixing conditions : A new Approach, *submitted in Nonlinearity.*
- N. Vergne and M. Abadi. Poisson approximation for search of rare words in DNA sequences, *accepted in ALEA.*

Conférences

- N. Vergne. Les modèles de Markov régulés pour l'étude des séquences biologiques : dérive polynomiale et applications. *Colloque "Jeunes probabilistes et statisticiens"*, Aussois, 2006.
- N. Vergne. Les modèles de Markov régulés pour l'étude des séquences biologiques : dérive polynomiale et applications. *6ème Journées Jeunes Chercheurs en Biométrie*, Villejuif, 2006.
- N. Vergne. Drifting Markov Models *ECCB poster*, Madrid, 2005.
- N. Vergne. Les chaînes de Markov régulées *JOBIM poster*, Lyon, 2005.

Logiciels

- PANOW (Poisson Approximation for the Number of Occurrences of Words) : http://stat.genopole.cnrs.fr/software/panow
- DRIMM (Drifting Markov Models) : http://stat.genopole.cnrs.fr/software/drimm

Remerciements

Je remercie en premier lieu les gens qui me sont chers : ma famille, ma chérie, mes amis. Qu'ils sachent que tout ce travail ne serait rien sans eux. Merci de m'avoir soutenu, encouragé, enguirlandé, poussé et surtout merci de m'avoir supporté !! Mes parents, mon frère et toute la famille ont toujours été à mes côtés dans les bons et les mauvais moments, me montrant une confiance sans faille. Je dis souvent avoir de la chance et je vais le répéter encore : j'ai une chance démesurée d'avoir une famille aussi géniale. Puis j'ai trouvé le bonheur avec Sylvie, mon ange-fée. Elle est celle qui a permis de me reconstruire et qui me comble de bonheur chaque jour. Ensuite je ne voudrais pas oublier tous mes amis qui m'ont soutenu des années durant et ce n'était pas facile avec mon caractère... Christophe et Fabrice pour de longues discussions interminables, Claire et Denis pour leur fidélité sans faille, Mathieu et Nadège pour les soirées qui font oublier le reste. Je n'oublie pas mon vrai ami Nidal, Emilie, Sebastien, Guillaume, Bao, Isabelle et les savoyards, Xavier et les gars du foot. Et je n'oublie pas non plus mon petit Rémy.

J'aimerais ensuite remercier les gens du labo qui ont été essentiels à mes yeux. Au premier rang de ceux-ci figure évidemment Bernard (mais j'en reparlerai plus tard) et Mark, dont j'apprécie beaucoup la personnalité, l'humour, la disponibilité et le petit bonjour du matin (et je me souviendrai toujours de ce dimanche où tu es venu me voir jouer au foot alors que bon... le foot... ce n'est pas vraiment ta tasse de thé...). Je voudrais remercier aussi Anne-Sophie et Pierre dont les discours un peu différents regonflent le moral en entrouvrant d'autres perspectives. Claudine et sa bonne humeur contagieuse m'a permis bien des fois de retrouver le sourire.

Après ces quatre "coups de coeur", j'aimerais adresser un merci très particulier à Vincent et Catherine pour m'avoir apporté le maximum de leur aide et de leur soutien depuis mon arrivée jusqu'à la fin de ma thèse. Ils ont été essentiels à l'élaboration de ce travail aussi bien sur le plan scientifique que sur le plan technique. Je remercie aussi beaucoup Miguel, avec qui le contact est très facile et dont la collaboration entamée dès mon entrée au labo a abouti à une partie de ce manuscrit. Il est (et donc restera) le premier collègue avec qui j'ai co-écrit et j'en suis fier.

Ensuite, je voudrais préciser que j'ai eu la chance, durant plusieurs années, de faire partie d'un laboratoire d'une qualité humaine et scientifique rare. J'aurais aimé adressé ne serait-ce qu'un petit mot à chacune des personnes croisées dans ce labo tant ils m'ont permis de continuer à avancer, mais je vais me permettre d'en oublier... Je pense notamment à Maurice qui nous permet de travailler dans des conditions parfaites chaque jour. Je pense à Michèle qui a remis de l'ordre dans notre secrétariat. Je pense à Hugues, Pierre-Yves et Eduardo, mes sympathiques collègues de bureau. Je pense à David et je me souviens des encouragements et des parties de tennis. Je pense à Marie-Pierre et aux mots fléchés à l'aéroport de Montréal. Je pense à Etienne avec qui on pouvait parler foot de temps en temps ! Je pense aussi à Ivan, toujours de la gentillesse dans la voix pour entamer une conversation, souvent sportive. Que ceux que j'oublie me pardonnent : Ana, Adeline, Mickaël, Sophie, Elodie, Christophe, Julien, Florence, Cédric, Yolande, Carène. Le travail de chercheur inclue les conférences et au rythme de ces conférences se créent aussi des rencontres et amitiés sympathiques qui permettent d'avancer. Aussi je souhaiterais remercier Sébastien, Catherine et Sabrina. Lors de cette thèse, j'ai enseigné. Je ne veux pas oublier cette partie de mon travail qui a souvent été une bouée de sauvetage à laquelle me raccrocher. Je remercie ainsi mes étudiants, qui, pour la plupart, venaient en cours avec le sourire.

J'en viens maintenant aux membres du jury (et oui, j'ai gardé le meilleur pour la fin !!). Je remercie Benoit Saussol et Michel Termier d'avoir accepté d'être les rapporteurs de ma thèse. Benoit, je viendrai à Brest, merci pour tout. Michel, nos quelques discussions et quelques mails m'ont beaucoup apporté et j'ai hâte de collaborer. Je remercie Dominique Cellier d'avoir bien voulu, sans me connaître, faire partie de mon jury pour finalement être rapporteur ! Toutes tes remarques sont justes et intéressantes. Je remercie Jacques van Helden d'avoir fait le déplacement malgré un calendrier chargé, j'en suis très honoré. Du travail en commun nous attend forcément. J'aimerais maintenant remercier Grégory Nuel, président du jury, pour les années passées dans le même bureau, pour le sourire et le plaisir qu'il a en travaillant, pour sa "tchatche" le midi lors des déjeuners, pour toutes ses idées, pour son aide et pour son esprit critique. C'est un plaisir de travailler avec toi.

Enfin, je remercie Bernard... Je le remercie du fond du coeur. Je ne sais pas si il sait à quel point il a été essentiel. Je ne sais pas si il sait à quel point je l'estime. Je souhaite à tout le monde d'avoir un "chef" comme Bernard. L'utilisation du mot exceptionnel doit rester exceptionnel, mais je l'utilise trois fois dans la même phrase car Bernard est quelqu'un d'exceptionnel à mes yeux. Si cette thèse s'est terminée, c'est d'abord grâce à lui et ensuite pour lui. Il a su me faire confiance dès mon arrivée au labo, puis n'a cessé de de me montrer cette confiance. Jamais il n'a paru découragé dans les moments difficiles. Je ne pourrai jamais le remercier à la hauteur de son mérite. Sa disponibilité, son aide précieuse sur n'importe quelle question mathématique, ses encouragements sans fin, son sourire, son humour, toutes ces qualités qui m'ont poussé chaque jour à donner le maximum de moi-même, en font un directeur de thèse exemplaire.

J'en ai trop dit ou pas assez mais je conclue ces remerciements : MERCI Bernard, merci pour tout.

À ma mémé et à Mathieu...

Les baleines

On me dit que l'insouciance
Est un sacré défaut
Moi j'me sens toujours en vacances
Et j'sens le vent sur ma peau
On me dit qu'j'suis jamais d'attaque
Trop molle pour faire quelque chose
Mais moi j'carbure pas au Prozac
Et l'vent me donne les joues roses
Un jour j'prendrai le large
J'habiterai avec les poissons
Les baleines et les coquillages
Pas de vague, pas d'hameçon
Quand on m'dit qu'la mode est nouvelle
Le monde est semé d'embûches
Moi je cape mes vieilles bretelles et
J'prends mes ballons de baudruche
On m'dit alors qu'y a des épines
Qui vont t'les faire éclater
Ben moi j'm'appellerai Mélusine
Et j'te les regonflerai
Un jour j'prendrai le large
J'habiterai avec les poissons
Les baleines et les coquillages
Pas de vague, pas d'hameçon
Quand on m'parle de politique,
J'fais même pas semblant d'comprendre,
J'bois un coup et pour moi j'explique
Que la Terre n'est pas à vendre
Moi dans ma bulle, y a pas d'misère
Y a pas de gens que j'aime pas
Y a pas d'goudron dans mon air
Puisque y a personne d'autre que moi

Un jour j'prendrai le large
J'habiterai avec les poissons
Les baleines et les coquillages
Pas de vague, pas d'hameçon

En fait on m'dit que j'suis naïve
Que j'devrais faire attention
Mais quand l'âme est à la dérive
Elle s'éloigne des cons

Un jour j'prendrai le large
J'habiterai avec les poissons
Les baleines et les coquillages
Pas de vague, pas d'hameçon

Pas ces salopes de sirènes
Comme ça, pas d'comparaison
Pas de vomi, pas de gangrène
De l'eau claire et pour de bon
De l'eau claire, pour de bon
De l'eau claire, pour de bon
De l'eau claire et pour de bon.

Marie Cherrier

Table des matières

III Approximation de Poisson

Liste des tableaux

Table des figures

Notations

Nous listons ici les principales notations utilisées dans ce document.

- DMM : drifting Markov model (modèle de Markov régulé) ;
- HMM : hidden Markov model (modèle de Markov caché) ;
- $\mathcal{A}$: Alphabet (par exemple $\{a, c, g, t\}$) ;
- k : ordre du modèle de Markov ;
- u : passé markovien (à l'ordre k, $u \in \mathcal{A}^k$) ;
- v : présent ($v \in \mathcal{A}$) ;
- $n+1$: taille de la séquence ;
- t : position dans la séquence (t varie de 0 à n) ;
- X_t : variable aléatoire à valeurs dans $\mathcal{A}$;
- Π : matrice de transition d'un modèle de Markov ;
- Π^t : transposée de Π ;
- μ : loi stationnaire ou distribution de probabilité ;
- L : vraisemblance ;
- ℓ : log-vraisemblance ;
- $\|.\|$: distance en variation totale ;
- $\mathbb{1}$: fonction indicatrice ;
- $\mathbb{1}_u = \mathbb{1}_{\{X_{t-k} \dots X_{t-1} = u\}}$;
- $\mathbb{1}_v = \mathbb{1}_{\{X_t = v\}}$;
- $\mathbb{1}_{uv} = \mathbb{1}_{\{X_{t-k} \dots X_t = uv\}}$;
- $[\![a, b]\!]$: ensemble des entiers de a à b ;
- $\mathbb{P}$: probabilité ;
- $\mathbb{E}$: espérance ;
- $\mathbb{V}$: variance ;
- $\mathbb{C}\mathrm{ov}$: covariance ;
- $\{\alpha_0, \dots, \alpha_N\}$: découpage de la séquence en N segments ;

Introduction

Séquences biologiques et modèles de Markov

Avant toute chose, et bien que matheux de nature, la multidisciplinarité de mon travail m'a induit à rendre cette thèse accessible à tous : aux non-statisticiens comme aux non-biologistes ou non-informaticiens.

*Dans ce manuscrit, je vous présente le résultat de mes recherches sur les "modèles de Markov régulés" (*Drifting Markov Models*). Cette introduction générale à la génétique a pour but de situer le contexte du travail et d'entrevoir les possibilités d'applications de ces nouveaux modèles.*

Je présente aussi plusieurs articles (en collaboration avec Miguel Abadi[1]*) sur un sujet différent : les temps de retour et nombres d'occurrences dans les processus de mélange. Ces articles participent à l'analyse des séquences par la recherche de mots exceptionnels. Ils sont l'objet de la partie III.*

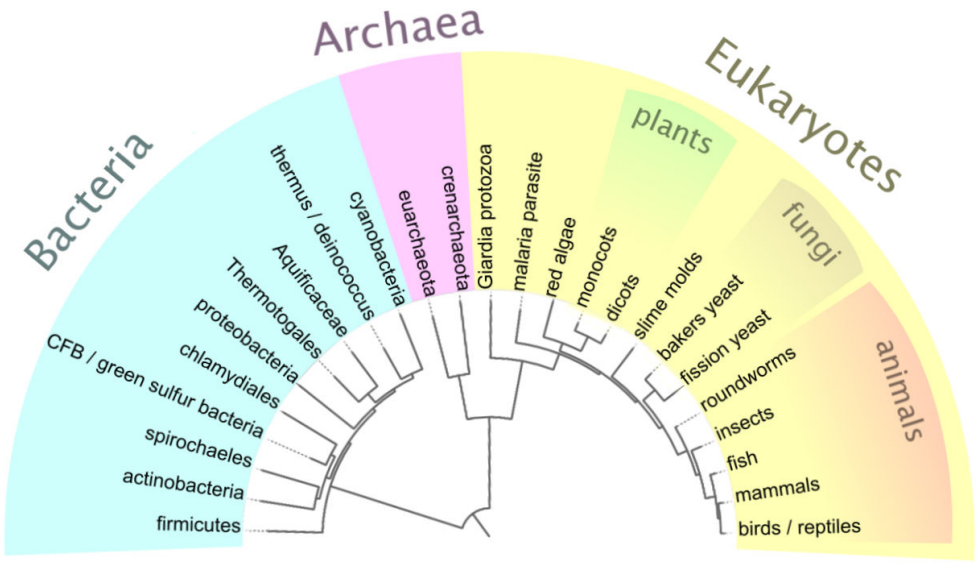
L'analyse statistique des séquences biologiques telles les séquences nucléotidiques (l'ADN et l'ARN) ou d'acides aminés (les protéines) nécessite la conception de différents modèles s'adaptant chacun à un ou plusieurs cas d'étude. Les séquences nucléotidiques d'ADN forment le patrimoine génétique des individus : elles les caractérisent au moins en partie et expliquent leurs ressemblances ainsi que leurs différences. Biologistes, mathématiciens, informaticiens se trouvent donc confrontés à un problème de taille : décrypter (toute) l'information contenue dans ces séquences. L'enjeu scientifique est important. Acquérir de telles connaissances sur le fonctionnement des espèces favorisera la résolution des problèmes telles les maladies. Bien qu'un nombre important de mécanismes biologiques soit déjà connu, un nombre beaucoup plus important reste encore un mystère. L'analyse statistique des séquences offre une vision nouvelle sur ces mécanismes de la biologie et aide à la compréhension de ceux-ci. Étant donnée la dépendance de la succession des nucléotides dans les séquences d'ADN, les modèles généralement utilisés sont des modèles de Markov (voir 0.2). Le problème de ces modèles est de supposer l'homogénéité des séquences. Or, les séquences biologiques ne sont pas homogènes. Un exemple bien connu est la répartition en gc (voir 0.1.1 sur les nucléotides) : le long d'une même séquence, alternent des régions riches en gc et des régions pauvres en gc. Pour rendre compte de l'hétérogénéité des séquences, d'autres modèles sont utilisés : les modèles de Markov cachés (voir 0.3). La séquence est divisée en plusieurs régions homogènes. Les applications sont nombreuses, telle la recherche des régions codantes. Certaines particularités biologiques ne pouvant apparaître suivant ces modèles, nous proposons de nouveaux modèles, les chaînes de Markov régulées. La matrice de transition, constante par morceaux pour des modèles de Markov cachés, est autorisée à varier le long de la séquence. Ces modèles peuvent être vus comme une alternative mais aussi comme un outil complémentaire aux modèles de Markov cachés (voir 0.4).

1 Séquences biologiques

Notre étude pourra porter sur diverses séquences biologiques que nous présentons maintenant. Pour plus d'information sur la biologie et en particulier sur l'informatique en biologie, il est utile de se référer aux récents ouvrages de bioinformatique résumant assez bien les problèmes posés par la biologie aux informaticiens et mathématiciens (voir [DK02] et [Ber01b]). Avant tout, donnons quelques notions de classification des êtres vivants. Le développement de la génétique et le séquençage automatique des êtres vivants a permis de revoir leur mode de classification. À la classification traditionnelle du vivant en cinq règnes (les procaryotes, les protistes, les champignons, les végétaux et les animaux), a succédé la classification phylogénétique. Plutôt que de se fonder sur des

[1] http://www.ime.unicamp.br/~miguel/

FIG. 1 – Classification des êtres vivants ([CDvM+06])



caractères biologiques, phénotypiques ou physiologiques, la classification phylogénétique repose principalement sur la génétique. Les êtres vivants sont regroupés en trois grandes catégories :

- Eubactéries ;
- Archéobactéries ;
- Eucaryotes.

[CDvM+06] propose une classification récente représentée à la Figure 1. Les archéobactéries se différencient des eubactéries (les "vraies" bactéries) par leurs gènes contenant exons et introns. En général, l'ADN des archéobactéries et des eubactéries se compacte en un unique chromosome circulaire, tandis que l'ADN des eucaryotes se compacte en plusieurs chromosomes linéaires (voir Figure 2 et Figure 3 sur la compaction de l'ADN). Les eucaryotes se différencient aussi des archéobactéries et des eubactéries par leur noyau.

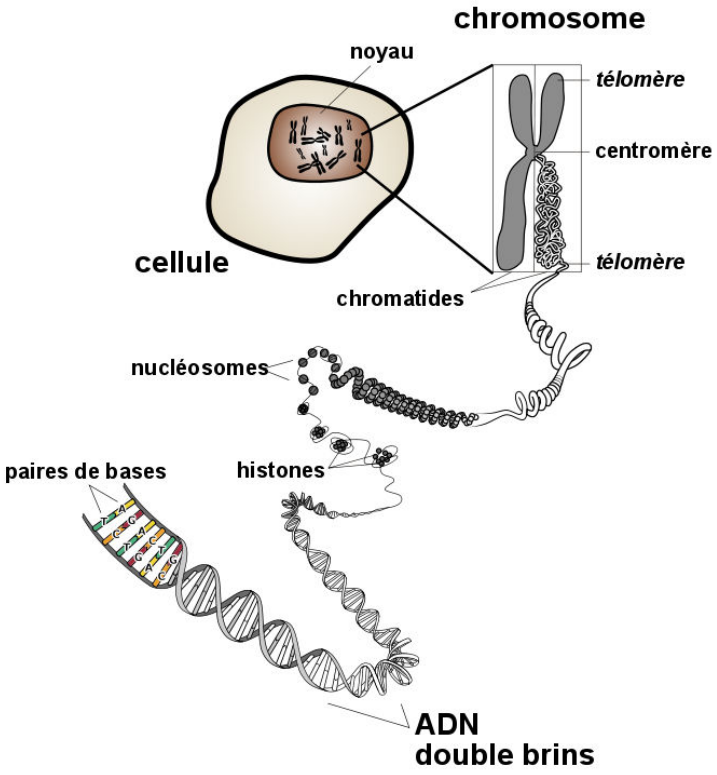
Plusieurs entités sont très proches des êtres vivants mais ne sont pas (encore ?) considérées comme tels, la définition du vivant faisant débat. Il s'agit en particulier des virus. Ces organismes acellulaires sont incapables de se reproduire seuls et sont pour cette raison écartés du monde du vivant. Pourtant, ils ont été récemment soupçonnés d'être à l'origine de la première cellule à ADN. En effet, Didier Raoult et Jean-Michel Claverie ont découvert le mimivirus : un virus géant à ADN (son génome étant deux fois plus long que le plus petit génome bactérien connu). La particularité de ce virus est qu'il peut produire des protéines impliquées dans la traduction de l'ARN en protéines (comme des enzymes chargeant des acides aminés sur des ARNt), il pourrait donc avoir pour ancêtres des virus plus anciens que la première cellule à ADN (voir [LSAR+03]). Nous ne rentrerons pas plus loin dans ce débat sur le vivant.

1.1 Les séquences nucléotidiques

Le patrimoine génétique de chaque individu, communément appelé génome, est entièrement contenu dans chacune des cellules de son corps, plus précisément dans leur noyau chez les eucaryotes, le long de longues chaînes orientées constituées de quatre molécules différentes : les nucléotides `a`, `c`, `g` et `t` (du nom des molécules : adénine, cytosine, guanine et thymine). Ces longues chaînes sont les fameuses séquences d'ADN (Acide DésoxyriboNucléique). Une séquence d'ADN est peut donc être représenté par un texte écrit avec un alphabet constitué des seules quatre lettres précédentes. La molécule d'ADN est structurée de manière assez originale. Elle est formée de deux brins complémentaires et anti-parallèles. En effet, les nucléotides ont la capacité de s'apparier deux à deux (par des liaisons chimiques), le `a` avec le `t` et le `c` avec le `g` : ainsi s'explique le terme complémentaire ; connaissant un brin, l'autre brin est aisément déductible. Ensuite, ces deux brins s'enroulent pour former une hélice. Chez les eucaryotes, cette hélice s'enroule autour d'histones puis s'enroule encore et encore pour finir par former ce que nous nommons les chromosomes. Cet enroulement de l'ADN est schématisé dans la Figure 2[2], pour des organismes eucaryotes. La Figure 3[3] montre plus précisément la compaction de l'ADN. Ainsi, nous pouvons aisément

[2] http://fr.wikipedia.org/wiki/Image:Chromosome_fr.svg
[3] http://www.humans.be/images/chromatine2.jpg

FIG. 2 – Enroulement et compaction de l'ADN dans le chromosome

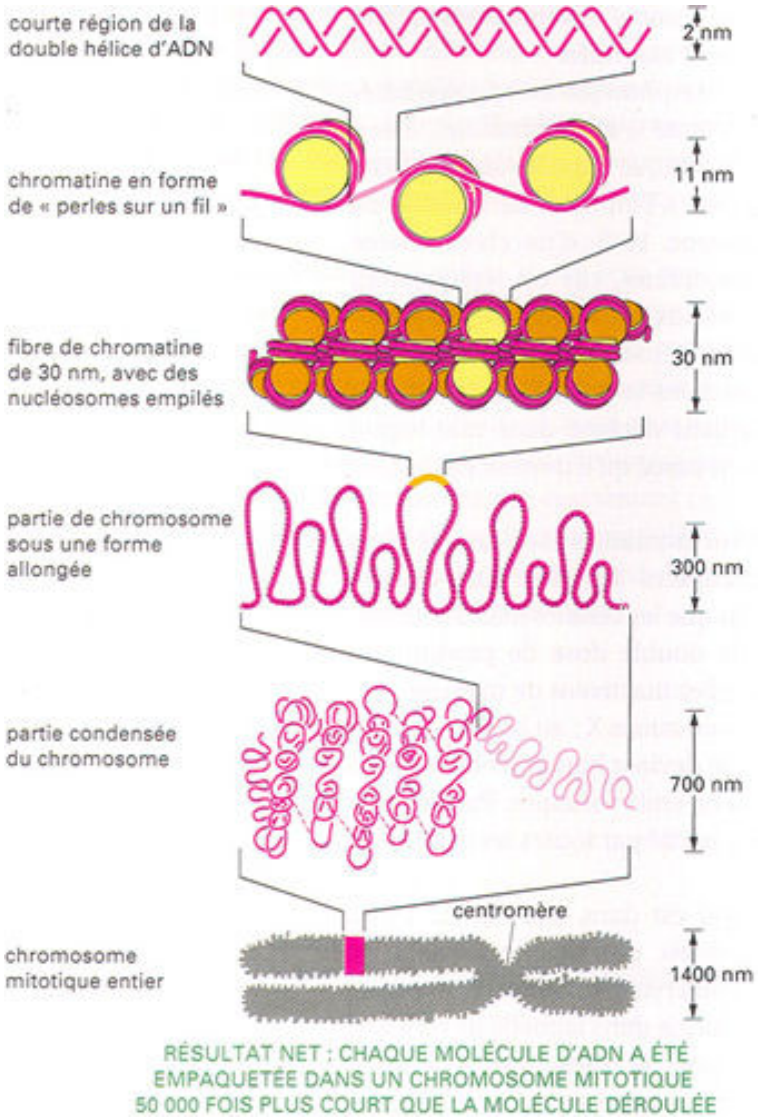


remarquer le peu d'espace occupé par un chromosome malgré la grande longueur des séquences d'ADN. En effet, donnons un exemple simple chez l'homme. Alors qu'un chromosome compacté ne mesure que quelques microns (10^{-6} m), les séquences d'ADN de tous les chromosomes mises bout à bout, totalement dépliées, formeraient un fil de plus de deux mètres de longueur (et ceci dans chaque cellule de notre corps). Pour nous faire une idée des proportions, l'ADN des 6.10^{13} cellules du corps humain, mis bout à bout, pourrait couvrir 300 fois la distance de la Terre à la Lune (300 fois environ 400 000 km), ou bien encore la quasi totalité de la distance de la Terre au Soleil (environ 150 000 000 km). Rappelons toutefois que ce super-enroulement et cette compaction de l'ADN n'a lieu que lors de la division cellulaire, lorsque les chromosomes se forment pour finalement se séparer dans chacune des deux nouvelles cellules ainsi formées.

Donnons l'exemple d'une (courte) séquence ADN, celle du génome complet de *HIV1*, le virus du SIDA.

```
tggaagggct aattcactcc caacgaagac aagatatcct tgatctgtgg atctaccaca cacaaggcta cttccctgat tagcagaact
acacaccagg gccagggatc agatatccac tgacctttgg atggtgctac aagctagtac cagttgagcc agagaagtta gaagaagcca
acaaaggaga gaacaccagc ttgttacacc ctgtgagcct gcatggaatg gatgacccgg agagagaagt gttagagtgg aggtttgaca
gccgcctagc atttcatcac atggcccgag agctgcatcc ggagtacttc aagaactgct gacatcgagc ttgctacaag ggactttccg
ctggggactt tccagggagg cgtggcctgg gcgggactgg ggagtggcga gccctcagat cctgcatata agcagctgct ttttgcctgt
actgggtctc tctggttaga ccagatctga gcctgggagc tctctggcta actagggaac ccactgctta agcctcaata aagcttgcct
tgagtgcttc aagtagtgtg tgcccgtctg ttgtgtgact ctggtaacta gagatccctc agacccttttagtcagtgtg gaaaatctct
agcagtggcg cccgaacagg gacctgaaag cgaaagggaa accagagctc tctcgacgca ggactcggct tgctgaagcg cccgcacggc
aagaggcgag gggcggcgac tggtgagtac gccaaaaatt ttgactagcg gaggctagaa ggagagagat gggtgcgaga gcgtcagtat
taagcggggg agaattagat cgatgggaaa aaattcggtt aaggccaggg ggaaagaaaa aatataaatt aaaacatata gtatgggcaa
gcagggagct agaacgattc gcagttaatc ctggcctgtt agaaacatca gaaggctgta gacaaatact gggacagcta caaccatccc
ttcagacagg atcagaagaa cttagatcat tatataatac agtagcaacc ctctattgtg tgcatcaaag gatagagata aaagacacca
aggaagcttt agacaagata gaggaagagc aaaacaaaag taagaaaaaa gcacagcaag cagcagctga cacaggacac agcaatcagg
tcagccaaaa ttaccctata gtgcagaaca tccaggggca aatggtacat caggccatat cacctagaac tttaaatgca tgggtaaaag
tagtagaaga gaaggctttc agcccagaag tgatacccat gttttcagca ttatcagaag gagccacccc acaagattta aacaccatgc
taaacacagt gggggggacat caagcagcca tgcaaatgtt aaaagagacc atcaatgagg aagctgcaga atgggataga gtgcatccag
tgcatgcagg gcctattgca ccaggccaga tgagagaacc aaggggaagt gacatagcag gaactactag tacccttcag gaacaaatag
gatggatgac aaataatcca cctatcccag taggagaaat ttataaaaga tggataatcc tgggattaaa taaaatagta agaatgtata
gccctaccag cattctggac ataagacaag gaccaaagga acccttttaga gactatgtag accggttcta taaaactcta agagccgagc
aagcttcaca ggaggtaaaa aattggatga cagaaacctt gttggtccaa aatgcgaacc cagattgtaa gactatttta aaagcattgg
gaccagcggc tacactagaa gaaatgatga cagcatgtca gggagtagga ggacccggcc ataaggcaag agttttggct gaagcaatga
gccaagtaac aaattcagct accataatga tgcagagagg caatttttagg aaccaaagaa agattgttaa gtgtttcaat tgtggcaaag
aagggcacac agccagaaat tgcagggccc ctaggaaaaa gggctgttgg aaatgtggaa aggaaggaca ccaaatgaaa gattgtactg
```

FIG. 3 – Structure d'un nucléosome



agagacaggc taatttttta gggaagatct ggccttccta caagggaagg ccagggaatt ttcttcagag cagaccagag ccaacagccc
caccagaaga gagcttcagg tctggggtag agacaacaac tccccctcag aagcaggagc cgatagacaa ggaactgtat cctttaactt
ccctcaggtc actctttggc aacgacccct cgtcacaata aagatagggg ggcaactaaa ggaagctcta ttagatacag gagcagatga
tacagtatta gaagaaatga gtttgccagg aagatggaaa ccaaaaatga taggggaat tggaggtttt atcaaagtaa gacagtatga
tcagatactc atagaaatct gtggacataa agctataggt acagtattag taggacctac acctgtcaac ataattggaa gaaatctgtt
gactcagatt ggttgcactt taaatttcc cattagccct attgagactg taccagtaaa attaaagcca ggaatggatg gcccaaaagt
taaacaatgg ccattgacag aagaaaaaat aaaagcatta gtagaaattt gtacagagat ggaaaaggaa gggaaaattt caaaaattgg
gcctgaaaat ccatacaata ctccagtatt tgccataaag aaaaaagaca gtactaaatg gagaaaatta gtagatttca gagaacttaa
taagagaact caagacttct gggaagttca attaggaata ccacatcccg cagggttaaa aaagaaaaaa tcagtaacag tactggatgt
gggtgatgca tatttttcag ttcccttaga tgaagacttc aggaagtata ctgcatttac catacctagt ataaacaatg agacaccagg
gattagatat cagtacaatg tgcttccaca gggatggaaa ggatcaccag caatattcca aagtagcatg acaaaaatct tagagccttt
tagaaaacaa aatccagaca tagttatcta tcaatacatg gatgatttgt atgtaggatc tgacttagaa atagggcagc atagaacaaa
aatagaggag ctgagacaac atctgttgag gtggggactt accacaccag acaaaaaaca tcagaaagaa cctccattcc tttggatggg
ttatgaactc catcctgata aatggacagt acagcctata gtgctgccag aaaaagacag ctggactgtc aatgacatac agaagttagt
ggggaaattg aattgggcaa gtcagattta cccagggatt aaagtaaggc aattatgtaa actccttaga ggaaccaaag cactaacaga
agtaatacca ctaacagaag aagcagagct agaactggca gaaaacagag agattctaaa agaaccagta catggagtgt attatgaccc
atcaaaagac ttaatagcag aaatacagaa gcaggggcaa ggccaatgga catatcaaat ttatcaagag ccatttaaaa atctgaaaac
aggaaaatat gcaagaatga ggggtgccca cactaatgat gtaaaacaat taacagaggc agtgcaaaaa ataaccacag aaagcatagt
aatatgggga aagactccta aatttaaact gcccatacaa aaggaaacat gggaaacatg gtggacagag tattggcaag ccacctggat
tcctgagtgg gagtttgtta ataccccctcc cttagtgaaa ttatggtacc agttagagaa agaacccata gtaggagcag aaaccttcta
tgtagatggg gcagctaaca gggagactaa attaggaaaa gcaggatatg ttactaatag aggaagacaa aaagttgtca ccctaactga
cacaacaaat cagaagactg agttacaagc aatttatcta gctttgcagg attcgggatt agaagtaaac atagtaacag actcacaata
tgcattagga atcattcaag cacaaccaga tcaaagtgaa tcagagttag tcaatcaaat aatagagcag ttaataaaaa aggaaaaggt
ctatctggca tgggtaccag cacacaaagg aattggagga aatgaacaag tagataaatt agtcagtgct ggaatcagga aagtactatt
tttagatgga atagataagg cccaagatga acatgagaaa tatcacagta attggagagc aatggctagt gattttaacc tgccacctgt
agtagcaaaa gaaatagtag ccagctgtga taaatgtcag ctaaaaggag aagccatgca tggacaagta gactgtagtc caggaatatg
gcaactagat tgtacacatt tagaaggaaa agttatcctg gtagcagttc atgtagccag tggatatata gaagcagaag ttattccagc
agaaacaggg caggaaacag catattttct tttaaaatta gcaggaagat ggccagtaaa aacaatacat actgacaatg gcagcaattt
caccggtgct acggttaggg ccgcctgttg gtgggcggga atcaagcagg aatttggaat tccctacaat ccccaaagtc aaggagtagt
agaatctatg aataaagaat taaagaaaat tataggacag gtaagagatc aggctgaaca tcttaagaca gcagtacaaa tggcagtatt
catccacaat tttaaaagaa aaggggggat tgggggtac agtgcagggg aaagaatagt agacataata gcaacagaca tacaaactaa
agaattacaa aaacaaatta caaaaattca aaattttcgg gtttattaca gggacagcag aaattcactt tggaaaggac cagcaaagct
cctctggaaa ggtgaagggg cagtagtaat acaagataat agtgacataa aagtagtgcc aagaagaaaa gcaaagatca ttagggatta
tggaaaacag atggcaggtg atgattgtgt ggcaagtaga caggatgagg attagaacat ggaaaagttt agtaaaacac catatgtatg
tttcagggaa agctagggga tggttttata gacatcacta tgaaagccct catccaagaa taagttcaga agtacacatc ccactagggg
atgctagatt ggtaataaca acatattggg gtctgcatac aggagaaaga gactggcatt tgggtcaggg agtctccata gaatggagga
aaaagagata tagcacacaa gtagaccctg aactagcaga ccaactaatt catctgtatt actttgactg tttttcagac tctgctataa
gaaaggcctt attaggacac atagttagcc ctaggtgtga atatcaagca ggacataaca aggtaggatc tctacaatac ttggcactag
cagcattaat aacaccaaaa aagataaagc caccttttgcc tagtgttacg aaactgacag aggatagatg gaacaagccc cagaagacca
agggccacag agggagccac acaatgaatg gacactagag cttttagagg agcttaagaa tgaagctgtt agacattttc ctaggatttg
gctccatggc ttagggcaac atatctatga aacttatggg gatacttggg caggagtgga agccataata agaattctgc aacaactgct
gtttatccat tttcagaatt gggtgtcgac atagcagaat aggcgttact cgacagagga gagcaagaaa tggagccagt agatcctaga
ctagagccct ggaagcatcc aggaagtcag cctaaaactg cttgtaccaa ttgctattgt aaaaagtgtt gctttcattg ccaagtttgt
ttcataacaa aagccttagg catctcctat ggcaggaaga agcggagaca gcgacgaaga gctcatcaga acagtcagac tcatcaagct
tctctatcaa agcagtaagt agtacatgta acgcaaccta taccaatagt agcaatagta gcattagtag tagcaataat aatagcaata
gttgtgtggt ccatagtaat catagaatat aggaaaatat taagacaaag aaaaatagac aggttaattg atagactaat agaaagagca
gaagacagtg gcaatgagag tgaaggagaa atatcagcac ttgtggagat gggggtggag atggggcacc atgctccttg ggatgttgat
gatctgtagt gctacagaaa aattgtgggt cacagtctat tatggggtac ctgtgtggaa ggaagcaacc accactctat tttgtgcatc
agatgctaaa gcatatgata cagaggtaca taatgtttgg gccacacatg cctgtgtacc cacagacccc aacccacaag aagtagtatt
ggtaaatgtg acagaaaatt ttgacatgtg gaaaaatgac atggtagaac agatgcatga ggatataatc agtttatggg atcaaagcct
aaagccatgt gtaaaattaa ccccactctg tgttagttta aagtgcactg atttgaagaa tgatactaat accaatagta gtagcgggag
aatgataatg gagaaaggag agataaaaaa ctgctctttc aatatcagca caagcataag aggtaaggtg cagaaagaat atgcattttt
ttataaactt gatataatac caatagataa tgatactacc agctatagct tgacaagttg taacacctca gtcattacac aggcctgtcc
aaaggtatcc tttgagccaa ttcccataca ttattgtgcc ccggctggtt ttgcgattct aaaatgtaat aataagacgt tcaatggaac
aggaccatgt acaaatgtca gcacagtaca atgtacacat ggaattaggc cagtagtatc aactcaactg ctgttaaatg gcagtctagc
agaagaagag gtagtaatta gatctgtcaa tttcacggac aatgctaaaa ccataatagt acagctgaac acatctgtag aaattaattg
tacaagaccc aacaacaata caagaaaaag aatccgtatc cagagaggac cagggagagc atttgttaca ataggaaaaa taggaaatat
gagacaagca cattgtaaca ttagtagagc aaaatggaat aacactttaa aacagataga tagcaaatta agagaacaat tcggaaataa
taaaacaata atctttaagc aatcctcagg aggggaccca gaaattgtaa cgcacagttt taattgtgga ggggaatttt tctactgtaa
ttcaacacaa ctgtttaata gtacttggtt taatagtact tggagtactg aagggtcaaa taacactgaa ggaagtgaca caatcaccct
cccatgcaga ataaaacaaa ttataaacat gtggcagaaa gtaggaaaag caatgtatgc ccctcccatc agtggacaaa ttagatgttc
atcaaatatt acagggctgc tattaacaag agatggtggt aatagcaaca atgagtccga gatcttcaga cttggaggag gagatatgag
ggacaattgg agaagtgaat tatataaata taaagtagta aaaattgaac cattaggagt agcacccacc aaggcaaaga gaagagtggt
gcagagagaa aaaagagcag tgggaatagg agctttgttc cttgggttct tgggagcagc aggaagcact atgggcgcag cctcaatgac
gctgacggta caggccagac aattattgtc tggtatagtg cagcagcaga acaatttgct gagggctatt gaggcgcaac agcatctgtt
gcaactcaca gtctggggca tcaagcagct ccaagcaaga atcctagctg tggaaagata cctaaaggat caacagctcc tagggatttg
gggttgctct ggaaaactca tttgcaccac tgctgtgcct tggaatgcta gttggagtaa taaatctctg gaacagatct ggaatcacac
gacctggatg gagtgggaca gagaaattaa caattacaca agcttaatac actccttaat tgaagaatcg caaaaccagc aagaaaagaa
tgaacaagaa ttattggaat tagataaatg ggcaagtttg tggaattggt ttaacataac aaattggctg tggtatataa aattattcat
aatgatagta ggaggcttgg taggtttaag aatagttttt gctgtacttt ctatagtgaa tagagttagg cagggatatt caccattatc
gtttcagacc cacctcccaa tcccgagggg acccgacagg cccgaaggaa tagaagaaga aggtggagag agagacagag acagatccat
tcgattagtg aacggatcct tggcacttat ctgggacgat ctgcggagcc tgtgcctctt cagctaccac cgcttgagag acttactctt
gattgtaacg aggattgtgg aacttctggg acgcaggggg tgggaagccc tcaaatattg gtggaatctc ctacagtatt ggagtcagga

```
actaaagaat agtgctgtta gcttgctcaa tgccacagcc atagcagtag ctgaggggac agatagggtt atagaagtag tacaaggagc
ttgtagagct attcgccaca tacctagaag aataagacag ggcttggaaa ggattttgct ataagatggg tggcaagtgg tcaaaaagta
gtgtgattgg atggcttact gtaagggaaa gaatgagacg agctgagcca gcagcagatg gggtgggagc agcatctcga gacctggaaa
aacatggagc aatcacaagt agcaacacag cagctaccaa tgctgcttgt gcctggctag aagcacaaga ggaggaggag gtgggttttc
cagtcacacc tcaggtacct ttaagaccaa tgacttacaa ggcagctgta gatcttagcc acttttaaa agaaaagggg ggactggaag
ggctaattca ctcccaaaga agacaagata tccttgatct gtggatctac cacacacaag gctacttccc tgattgacag aactacacac
cagggccagg ggtcagatat ccactgacct ttggatggtg ctacaagcta gtaccagttg agccagataa gatagaagag gccaataaag
gagagaacac cagcttgtta caccctgtga gcctgcatgg gatggatgac ccggagagag aagtgttaga gtggaggttt gacagccgcc
tagcatttca tcacgtggcc cgagagctgc atccggagta cttcaagaac tgctgacatc gagcttgcta caagggactt tccgctgggg
actttccagg gaggcgtggc ctgggcggga ctggggagtg gcgagccctc agatcctgca tataagcagc tgctttttgc ctgtactggg
tctctctggt tagaccagat ctgagcctgg gagctctctg gctaactagg gaacccactg cttaagcctc aataaagctt gccttgagtg
cttcaagtag tgtgtgcccg tctgttgtgt gactctggta actagagatc cctcagaccc ttttagtcag tgtggaaaat ctctagca
```

L'ADN de ce virus ne compte donc que 9718 nucléotides (ça, c'est pour rassurer ceux qui ont compté[4]). À titre de comparaison, le tableau 1 donne les tailles des génomes complets des espèces considérées. Le génome complet d'une espèce ou d'un individu comprend une ou plusieurs séquences d'ADN formant respectivement un ou plusieurs chromosomes. Nous parlons souvent de génome complet d'une espèce : en réalité, chaque individu a bien évidemment son propre ADN, différent de celui de son voisin. Il existe cependant de très grandes similarités entre les génomes d'une même espèce. En effet, deux individus d'une même espèce partagent le même nombre de chromosomes et environ le même nombre de gènes (voir 0.1.3) disposés à peu près aux mêmes endroits dans le génome. Les génomes d'individus de la même espèce sont de même taille, de même que leurs chromosomes. Nous ne sommes donc pas réellement réducteurs en parlant de génome d'une espèce. Nous pouvons nous apercevoir en

TAB. 1 – Tailles des génomes

Espèce	taille (pb)	gènes
Virus du SIDA	9750	9
Mycoplasme génital	580 000	480
Helicobacter pylori (ulcère stomachal)	1 667 867	1 590
Escherichia coli	4 639 221	4 288
Levure de bière	12 067 280	6 200
Plasmodium falciparum (malaria)	25 000 000	5 400
Trypanosome	35 000 000	8000
Nématode	110 000 000	19100
Arabidopsis thaliana (arabette)	125 000 000	25 500
Drosophile	150 000 000	13 600
Tétraodon (poisson-zèbre)	350 000 000	30 000
Tomate	655 000 000	?
Soja	1 115 000 000	?
Poulet	1 200 000 000	?
Boa constrictor	2 100 000 000	?
Homme	3 400 000 000	30 000

lisant ce tableau que ces séquences sont parfois très longues. Relativisons tout de même la position de l'homme dans ce tableau en observant le tableau 2. Au vu de ce second tableau, une idée selon laquelle "Plus il y a d'ADN,

TAB. 2 – Tailles des génomes

Espèce	taille (pb)	gènes
Homme	3 400 000 000	30 000
Salamandre	81 000 000 000	?
Fougère	100 000 000 000	?

plus l'espèce est évoluée" se trouve contredite (par l'idée que la fougère ou la salamandre serait moins "évoluée" que l'homme). Il y a mieux encore, une amibe, l'*Amoeba dubia* (un organisme unicellulaire !), a un génome de 670 milliards de paires de bases ([Li97] page 383). Il existe bon nombre de tels exemples tendant à montrer que l'essentiel n'est pas la taille du génome (voir un tableau étoffé, contenant les tailles des génomes de nombreux organismes, à

[4]Merci Bernard, une touche d'humour de temps en temps ça donne le sourire ! (*"Un sourire ne coûte rien, il enrichit celui qui le reçoit sans appauvrir celui qui le donne."*)

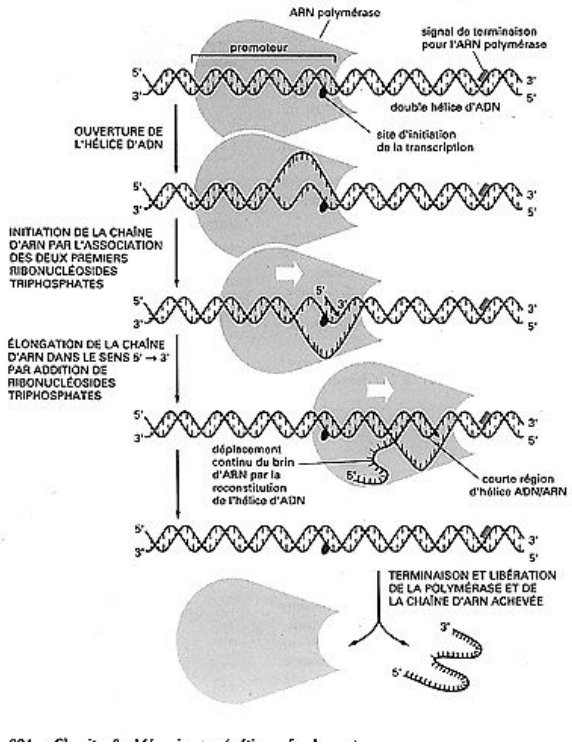
l'adresse suivante www.cbs.dtu.dk/databases/DOGS/abbr_table.bysize.txt). Une explication simple est donnée par les biologistes : il existerait de l'ADN ne servant à rien ! Ce sont les séquences non-codantes dont nous reparlerons plus tard (voir 0.1.2). En réalité, il est improbable que ces séquences soient inutiles, mais leur fonction n'a pas encore été déterminée ou démontrée. À titre d'exemple, chez l'homme, seulement 1.4% des séquences sont codantes (voir http://www.genoscope.cns.fr/externe/Francais/Sequencage/#2.3 pour quelques précisions intéressantes). Un critère plus raisonnable du degré d'évolution d'une espèce semblait être le nombre de gènes (voir Tableau 1). Mais l'homme a été déçu des résultats : à peine deux fois plus de gènes qu'une mouche, autant qu'un poisson et sans doute moins que ces fameuses salamandres et fougères (dont les gènes n'ont pas encore tous été trouvés ou prédits). En réalité, nous recherchons encore un critère scientifique permettant d'affirmer que l'homme est l'être vivant le plus évolué de la planète. Mais en réalité, rien n'est moins sûr... Refermons cette parenthèse philosophique et concluons ce paragraphe en constatant la grande quantité de séquences biologiques provenant de nombreuses espèces différentes et dont la longueur nous pousse à les modéliser afin de pouvoir étudier leurs caractéristiques.

1.2 Les séquences protéiques

De l'ADN à la protéine

Il existe des séquences biologiques autres que les séquences nucléotidiques, ce sont les séquences protéiques. Rappelons brièvement le mécanisme permettant de passer d'une séquence ADN à une séquence protéique. Tout

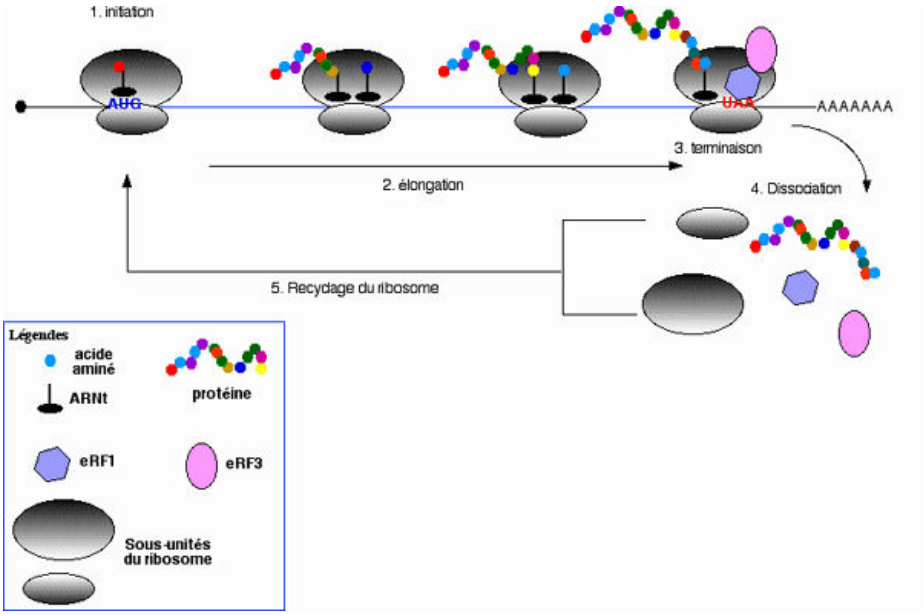
FIG. 4 – Le mécanisme de la transcription



d'abord, la séquence est "recopiée" : les nucléotides complémentaires de la séquence s'apparient pour former l'ARN (Acide RiboNucléique). Il existe toutefois une particularité à ces séquences d'ARN. En effet, le complémentaire d'un `a` sur la séquence d'ADN devient un `u` (un cinquième nucléotide, l'uracile) sur la séquence d'ARN. Nous avons donc une séquence d'ARN formée des seules quatre lettres `a`, `c`, `g`, `u`, et de plus complémentaire à la séquence ADN initiale. Intervient ensuite chez les eucaryotes et archéobactéries le phénomène d'épissage (voir 0.1.3 pour des précisions sur l'épissage). Sur les séquences d'ARN, certaines parties de la séquence codent pour former une protéine, ce sont les exons, et d'autres ne codent pas pour former une protéine, ce sont les introns. L'épissage permet de ne conserver de la séquence ARN que les parties codantes, les exons. Nous obtenons ainsi, une nouvelle séquence, appelée ARNm (ARN messager) contenant le code d'une protéine. Tout ce processus, le passage de l'ADN à l'ARNm est appelé la *transcription* (voir Figure 4[5]). Ensuite, à partir de cette séquence d'ARNm, est

[5] http://www.med.univ-rennes1.fr/wkf/stock/RENNES20080116042653mosserPCEM1_BM_Mosser2-08.pdf

FIG. 5 – Le mécanisme de la traduction



synthétisée une séquence protéique qui est une chaîne d'acides aminés (voir Tableau 5). Il en existe principalement 20 et chacun est codé par un ou plusieurs triplets de nucléotides selon le code fourni par le Tableau 4. Chaque triplet est appelé un codon. Parmi les 64 triplets possibles, plusieurs codent pour le même acide aminé, trois codent pour un codon stop et un seul code pour un codon start. Un codon est reconnu par un ARNt (ARN de transfert) qui est une courte séquence d'ARN, de structure complexe, capable d'apporter (dans la protéine) l'acide aminé correspondant au codon. Nous obtenons ainsi une chaîne d'acides aminés formant la protéine. Ce processus, le passage de l'ARNm à la protéine est appelé la *traduction* (voir Figure 5[6]). Le codon start (`atg`) marque le début de la traduction et le codon stop en marque la fin. Le tableau 3 donne un exemple de ce qui vient d'être dit pour faciliter la compréhension.

TAB. 3 – Le passage de l'ADN à la protéine (le rouge symbolise le codant)

Transcription	
ADN	atgaacagttcacacgtacatcgatttacgccgtaatag
ARN	uacuugucaagugugcauguagcuaaaugcggcauuauc
	Épissage
ARNm	caagauguaaugcgg
Traduction	
Protéine	QDVMR

Les protéines assurent les principales fonctions cellulaires. Elles assurent des fonctions physiologiques essentielles touchant le système digestif, hormonal et immunitaire. Elles interviennent dans l'élaboration de tous nos organes, des muscles, des dents, des os, des nerfs, des cheveux, etc... Leur principal rôle est d'apporter les éléments nécessaires à la croissance, à la production des enzymes et des hormones, et au renouvellement des cellules à tous les âges de la vie.

Sont rassemblées dans le Tableau 5, quelques caractéristiques des acides aminés. Les * correspondent chez l'homme, aux 8 acides aminés essentiels, ceux que nous ne produisons pas naturellement. "Le très lyrique Tristan fait vachement méditer Iseult" est un bon moyen mnémotechnique pour s'en souvenir (Le : leucine, très : thréonine, lyrique : lysine, Tristan : tryptophane, fait : phénylalanine, vachement : valine, méditer : méthionine, Iseult : isoleucine)

Il existe des centaines d'autres acides aminés qui diffèrent selon les espèces, dans la composition et la forme de

[6] http://www.futura-sciences.com/uploads/tx_oxcsfutura/images/figure1_namy.gif

TAB. 4 – Le code général de correspondance codon / acide aminé

AAA	Lys	ACA	Thr	AGA	Arg	ATA	Ile
AAC	Asn	ACC	Thr	AGC	Ser	ATC	Ile
AAG	Lys	ACG	Thr	AGG	Arg	ATG	Met
AAT	Asn	ACT	Thr	AGT	Lys	ATT	Ile
CAA	Gln	CCA	Pro	CGA	Arg	CTA	Leu
CAC	His	CCC	Pro	CGC	Arg	CTC	Leu
CAG	Gln	CCG	Pro	CGG	Arg	CTG	Leu
CAT	His	CCT	Pro	CGT	Arg	CTT	Leu
GAA	Glu	GCA	Ala	GGA	Gly	GTA	Val
GAC	Asp	GCC	Ala	GGC	Gly	GTC	Val
GAG	Glu	GCG	Ala	GGG	Gly	GTG	Val
GAT	Asp	GCT	Ala	GGT	Gly	GTT	Val
TAA	stop	TCA	Ser	TGA	stop	TTA	Leu
TAC	Tyr	TCC	Ser	TGC	Cys	TTC	Phe
TAG	stop	TCG	Ser	TGG	Trp	TTG	Leu
TAT	Tyr	TCT	Ser	TGT	Cys	TTT	Phe

TAB. 5 – Les acides aminés

Acide aminé	Symboles	Lettre	pK - COOH	pK - NH2	pK - R	Polarité
Alanine	Ala	A	2,35	9,87		apolaire
Arginine	Arg	R	1,82	8,99	12,48	chargé
Asparagine	Asn	N	2,14	8,72		polaire
Acide aspartique	Asp	D	1,99	9,9	3,9	chargé
Cystéine	Cys	C	1,92	10,7	8,37	polaire
Glutamine	Gln	Q	2,17	9,13		polaire
Acide glutamique	Glu	E	2,1	9,47	4,07	chargé
Glycine	Gly	G	2,35	9,78		apolaire
Histidine	His	H	1,8	9,33	6,04	chargé
Isoleucine *	Ile	I	2,32	9,76		apolaire
Leucine *	Leu	L	2,33	9,74		apolaire
Lysine *	Lys	K	2,16	9,06	10,54	chargé
Méthionine *	Met	M	2,13	9,28		apolaire
Phénylalanine *	Phe	F	2,2	9,31		apolaire
Proline	Pro	P	1,95	10,64		apolaire
Sérine	Ser	S	2,19	9,21		polaire
Thréonine *	Thr	T	2,09	9,1		polaire
Tryptophane *	Trp	W	2,46	9,41		apolaire
Tyrosine	Tyr	Y	2,2	9,21	10,46	polaire
Valine *	Val	V	2,29	9,74		apolaire

leurs molécules, en voici quelques-uns :

- Bêta-Alanine : Le seul bêta acide aminé naturel ;
- Carnitine : Acide aminé impliqué dans le métabolisme mitochondrial ;
- Citrulline : Acide aminé impliqué dans le métabolisme de l'ammoniaque ;
- Cystine : Produit d'oxydation de la cystéine ;
- Acide Gamma-Aminobutyrique (GABA) : Acide aminé agissant comme neurotransmetteur ;
- Glutathion ;
- Hydroxyproline : Proline hydroxylée ;
- Pyrrolysine ;
- Ornithine : Acide aminé impliqué dans le cycle de l'urée ;
- Sélénocystéine ;
- Taurine ; Acide aminé impliqué dans le métabolisme biliaire.

La variabilité du génome

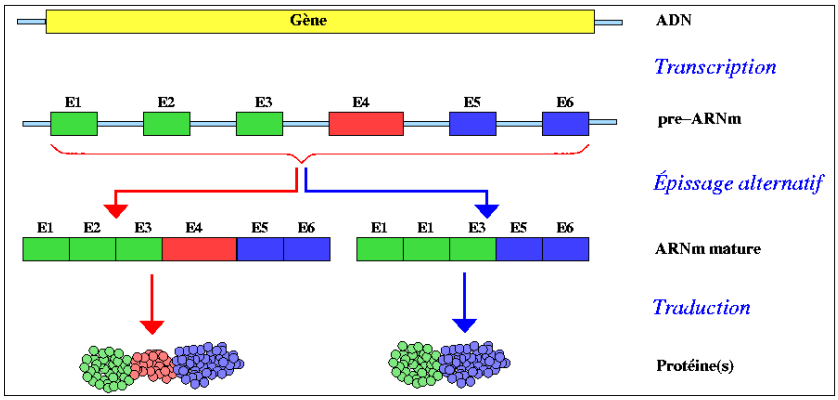
Existence de différents cadres de lecture Lors de la traduction, la lecture de la séquence de nucléotides codant pour une protéine (ARNm) se fait par triplet. Ainsi il existe trois façons de lire une séquence nucléotidique (trois cadres de lecture) :

- +1 : lecture des triplets débutant au premier nucléotide ;
- +2 : lecture des triplets débutant au deuxième nucléotide ;
- +3 : lecture des triplets débutant au troisième nucléotide.

De plus, l'ADN étant formé de deux brins antiparallèles, la protéine synthétisée sera différente selon le brin codant : un brin est lu dans dans un sens, l'autre brin est lu dans l'autre sens. Au total, pour une seule et même séquence, 6 phases de lecture sont possibles. Ce phénomène conjugué à celui de l'épissage alternatif induit une très grande diversité de protéines produites à partir d'une seule séquence ADN (et même d'une seule région codante).

Épissage alternatif Chez les eucaryotes, l'épissage consiste à supprimer les introns ou exons non nécessaires au codage de la protéine à produire. Une même séquence ADN peut produire plusieurs protéines différentes selon que l'épissage supprime certains introns/exons ou bien certains autres. On appelle ce phénomène l'épissage alternatif. Un exemple en est donné à la Figure 6[7]. Constatons qu'ainsi, un nombre très important de produits différents peut

FIG. 6 – Exemple d'épissage alternatif



être synthétisé au moyen de la même séquence initiale. Ceci n'est qu'un seul des multiples aspects qui rendent la biologie si intéressante, tant au niveau combinatoire qu'au niveau des possibilités phénotypiques offertes par ces mécanismes tel l'épissage.

Le cycle cellulaire Le cycle cellulaire offre d'autres mécanismes mettant en jeu le hasard et la combinatoire. La méiose intervient lors de la fécondation. Lors d'une reproduction sexuée, c'est le mélange du génome paternel et du génome maternel. De nombreux croisements entre séquences d'ADN, appelés "crossing-overs", surviennent lors de la méiose, assurant la variabilité génétique (un chromosome d'une cellule fille ne sera pas exactement le chromosome paternel ni exactement le chromosome maternel, mais des morceaux de l'un avec des morceaux de l'autre, pour schématiser simplement). La mitose correspond à la division cellulaire. Elle est nécessaire à la croissance de

[7] http://fr.wikipedia.org/wiki/Image:Altspli.png

l'organisme. Lors de cette division, les deux cellules filles créées sont censées avoir le même patrimoine génétique. Cependant, ici encore, une variabilité est parfois créée par les mutations qui sont des erreurs de recopie. Il en existe de plusieurs sortes :

- les substitutions : des nucléotides sont remplacés par des autres ;
- les insertions : des nouveaux nucléotides sont ajoutés dans la séquence ;
- les délétions : des nucléotides sont supprimés.

L'inversion de courts segments (moins de 1000 bases) est soupçonnée être un mécanisme majeur de l'évolution des génomes (voir [GMSP00]). Ceci est l'objet de la thèse de David Robelin ([Rob05]). [RRP03] présente un logiciel permettant de retrouver ces segments inversés. Lors de la reproduction ce remaniement global est fréquent : l'ADN se brise et un morceau d'ADN (provenant du même brin d'ADN, du brin complémentaire, d'un virus ou autre) vient se greffer à cet endroit ce qui provoque la mutation du génome. [GFM+03] montre les effets de ces retournements d'ADN sur les protéines.

1.3 Les gènes

La véritable définition d'un gène fait l'objet de grandes discussions métaphysiques. Nous pourrions en effet parler de gènes seulement comme les facteurs héréditaires transmis d'une génération à une autre. Plus précisément nous pourrions définir un gène comme une séquence ordonnée de nucléotides située à un endroit précis dans le chromosome. Mais où commence exactement un gène ? Au codon start, au promoteur (petite portion de séquence nucléotidique qui forme un signal pour enclencher la traduction en protéines) ? Comprend-il les signaux de régulations (encore et toujours une petite portion de séquence déterminant la vitesse et l'intensité de la traduction...) ? Comprend-il toute la séquence ADN ou bien seulement les introns ? Nous n'avons rien de vraiment précis pour pouvoir définir un gène. Gène est en réalité un terme très général, que nous utilisons différemment selon que nous voulons parler de son rôle ou bien de sa localisation.

2 Chaînes de Markov

Modéliser l'ADN à l'aide de modèles stochastiques et développer des méthodes statistiques pour analyser la masse de données résultant des multiples projets de séquençages sont des questions difficiles pour les statisticiens et les biologistes. Sachant que dans les séquences d'ADN, l'apparition d'un nucléotide n'est pas indépendante des nucléotides précédents, les modèles les plus populaires dans ce domaine sont les modèles de Markov qui offrent une description générale du comportement des séquences (voir [Alm83, Bla85, PAI87, GKP92] and [SYM05]). Les différentes fréquences des dinucléotides sont l'un des multiples exemples de description possible. Ainsi, à l'aide des propriétés statistiques de ces modèles, différentes propriétés biologiques des séquences d'ADN ou protéiques peuvent être soulignées.

2.1 Généralités

Un modèle de Markov est un processus aléatoire (ou stochastique), permettant de gérer la dépendance des événements. Soit $(X_t)_{t\in\mathbb{N}}$ une suite de variables aléatoires à valeurs dans un alphabet fini $\mathcal{A}$. Une chaîne de Markov homogène d'ordre k suffit pour prédire l'information de toutes les variables qui précèdent par les k dernières :

$$\mathbb{P}(X_t \mid X_u, u < t) = \mathbb{P}(X_t \mid X_{t-1}, ..., X_{t-k}).$$

Une chaîne de Markov est donc uniquement définie par sa *matrice de transition* Π qui réunit les probabilités précédentes pour chacun des cas possibles, et une loi initiale μ_0 qui définit les probabilités d'apparition des k premières lettres.

Exemple 1 *Pour $\mathcal{A} = \{a, c, g, t\}$, nous avons à l'ordre 1 :*

$$\Pi = \begin{pmatrix} \pi_{aa} & \pi_{ac} & \pi_{ag} & \pi_{at} \\ \pi_{ca} & \pi_{cc} & \pi_{cg} & \pi_{ct} \\ \pi_{ga} & \pi_{gc} & \pi_{gg} & \pi_{gt} \\ \pi_{ta} & \pi_{tc} & \pi_{tg} & \pi_{tt} \end{pmatrix}$$

$$\mu_0 = \begin{pmatrix} \mu_0(a) & \mu_0(c) & \mu_0(g) & \mu_0(t) \end{pmatrix}$$

où $\pi_{uv} = \mathbb{P}(X_t = v \mid X_{t-1} = u)$, avec $u \in \mathcal{A}$ et $v \in \mathcal{A}$ et où nous pouvons choisir par exemple 0.25 pour $\mu_0(u)$, $\forall u \in \mathcal{A}$.

Remarque 1 *Il est toujours possible de nous ramener au cas d'une chaîne de Markov d'ordre* 1. *Nous allons seulement présenter le cas d'une chaîne de Markov d'ordre* 2 *se ramenant à une chaîne de Markov d'ordre* 1. *Le cas général en découle directement.*

Soit $X = (X_t)_{t\in\mathbb{N}}$ *une suite de variables aléatoires à valeurs dans l'alphabet* $\mathcal{A} = \{a, c, g, t\}$, *selon un modèle markovien d'ordre* 2. *Nous avons* $\pi_{uvw} = \mathbb{P}(X_t = w \mid X_{t-1} = v, X_{t-2} = u)$ *avec* $u \in \mathcal{A}$, $v \in \mathcal{A}$ *et* $w \in \mathcal{A}$. *Prenons par exemple :*

$$X = gctggtgg.$$

Nous réécrivons cette séquence en groupant les lettres deux à deux (avec chevauchement) :

$$X^\star = (gc)(ct)(tg)(gg)(gt)(tg)(gg).$$

La suite $X^\star$ *est écrite dans un nouvel alphabet, l'alphabet*

$$\mathcal{A}^\star = \mathcal{A} \times \mathcal{A} = \{(uv), u \in \mathcal{A}, v \in \mathcal{A}\}.$$

$X^\star$ *est ainsi une chaîne de Markov d'ordre* 1 *sur* $\mathcal{A}^\star$ *de matrice de transition* $\Pi^\star$ *définie comme suit :*

$$\pi^\star_{(uv)(vw)} = \pi_{uvw} \text{ et } \pi^\star_{(uv)(v'w)} = 0 \text{ si } v \neq v'.$$

Nous gardons la même loi initiale. Remarquons que les matrices de transition ainsi créées sont carrées et très creuses. Ici, par exemple :

$$\Pi^\star = \begin{pmatrix} \pi_{aaa} & \pi_{aac} & \pi_{aag} & \pi_{aat} & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & \pi_{aca} & \pi_{acc} & \pi_{acg} & \pi_{act} & \cdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ \pi_{caa} & \pi_{cac} & \pi_{cag} & \pi_{cat} & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & \pi_{cca} & \pi_{ccc} & \pi_{ccg} & \pi_{cct} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Il ne s'agit en réalité que d'une astuce : nous changeons l'alphabet.

2.2 Définitions

Nous donnons ici quelques définitions portant sur les chaînes de Markov à espaces d'états finis.

Définition 1 *Une distribution* μ *pour la chaîne de Markov* Π *est dite stationnaire si* $\mu = \mu\Pi$.

Définition 2 *Une chaîne de Markov est dite ergodique s'il existe une unique distribution stationnaire* μ *telle que pour toute distribution initiale* μ_0,

$$\lim_{\ell\to\infty} \mu_0\Pi^\ell = \mu.$$

Définition 3 *Une chaîne de Markov* Π *est dite irréductible si*

$$\forall i, \forall j, \exists \ell \text{ tel que } \pi^{(\ell)}_{ij} > 0,$$

où $\pi^{(\ell)}_{ij} = \left(\Pi^\ell\right)_{ij} = \mathbb{P}\left(X_{t+\ell} = j | X_t = i\right)$.

Définition 4 *Une chaîne de Markov* Π *est dite apériodique si*

$$\forall i, \, pgcd\left\{\ell \,|\, \pi^{(\ell)}_{ii} > 0\right\} = 1.$$

Théorème 1 *Une chaîne de Markov irréductible et apériodique est ergodique.*

Définition 5 *Une chaîne de Markov ergodique* Π *de distribution stationnaire* μ *est dite réversible si* $\forall i, \forall j$ *on a* $\mu_i\pi_{ij} = \mu_j\pi_{ji}$.

2.3 Estimateurs classiques

Nous allons présenter l'estimation classique de la matrice de transition d'une chaîne de Markov à partir d'une séquence observée : l'estimation par la méthode du maximum de vraisemblance. Soit $X = (X_t)_{t\in[\![0,n]\!]}$ la séquence observée à valeurs dans $\mathcal{A}$. Supposons que nous voulions modéliser cette séquence par une chaîne de Markov d'ordre k de matrice de transition et de loi initiale la loi stationnaire de cette matrice :

$$\Pi = (\pi_{uv})_{u\in\mathcal{A}^k, v\in\mathcal{A}} \text{ et } \mu = (\mu_u)_{u\in\mathcal{A}^k}.$$

Déterminons la vraisemblance L du modèle :

$$\begin{aligned} L(X,\Pi) &= \mu_{X_0\dots X_{k-1}} \prod_{t=k}^{n} \pi_{X_{t-k}\dots X_{t-1}X_t} \\ L(X,\Pi) &= \mu_{X_0\dots X_{k-1}} \prod_{u\in\mathcal{A}^k,\, v\in\mathcal{A}} {\pi_{uv}}^{N(uv)} \end{aligned}$$

avec $N(uv)$ le nombre observé de uv dans la séquence. Maximisons la log-vraisemblance ℓ sous la contrainte $\sum_{v\in\mathcal{A}} \pi_{uv} = 1$. Pour cela posons $u^\star$ la dernière lettre de u et $\pi_{uu^\star} = 1 - \sum_{v\in\mathcal{A}\setminus\{u^\star\}} \pi_{uv}$. Il ne nous reste donc que $|\mathcal{A}|^k(|\mathcal{A}|-1)$ paramètres à estimer (où $|\mathcal{A}|$ est le cardinal de $\mathcal{A}$). Donnons l'expression de la log-vraisemblance :

$$\ell(X,\Pi) = \ln \mu_{X_0\dots X_{k-1}} + \sum_{u\in\mathcal{A}^k,\, v\in\mathcal{A}} N(uv)\ln\pi_{uv}.$$

Maintenant, annulons la dérivée :

$$\begin{aligned} &\forall u\in\mathcal{A}^k, \forall v\in\mathcal{A}\setminus\{u^\star\},\ \frac{\partial\ell(X,\Pi)}{\partial\pi_{uv}} = 0 \\ \iff &\forall u\in\mathcal{A}^k, \forall v\in\mathcal{A}\setminus\{u^\star\}, \frac{N(uv)}{\pi_{uv}} - \frac{N(uu^\star)}{\pi_{uu^\star}} = 0. \\ \iff &\forall u\in\mathcal{A}^k, \forall v\in\mathcal{A}\setminus\{u^\star\}, N(uv)\pi_{uv} = \frac{N(uu^\star)}{\pi_{uu^\star}}\pi_{uv}. \end{aligned}$$

Nous avons ainsi, en sommant sur v

$$\forall u\in\mathcal{A}^k, \forall v\in\mathcal{A}\setminus\{u^\star\},\ \frac{N(uv)}{\pi_{uv}} = \frac{N(uu^\star)}{\pi_{uu^\star}} = \frac{\sum_{v\in\mathcal{A}} N(uv)}{\sum_{v\in\mathcal{A}} \pi_{uv}} = N(u+)$$

où $N(u+)$ est le nombre de u suivis d'une lettre dans la séquence. Nous obtenons ainsi l'estimateur de Π :

$$\widehat{\Pi} = \left(\widehat{\pi_{uv}} = \frac{N(uv)}{N(u+)}\right)_{u\in\mathcal{A}^k,\, v\in\mathcal{A}}.$$

Notons que la probabilité d'apparition de la lettre u est estimée par

$$\widehat{\mu}_u = \frac{N(u)}{n+1}.$$

La modélisation par simples chaînes de Markov réduit énormément l'importante quantité d'information que peut contenir une séquence. En effet, une chaîne de Markov tend très vite vers sa loi stationnaire. Un résultat connu est le théorème suivant :

Théorème 2 *Si Π est irréductible et apériodique de loi stationnaire μ, alors il existe des constantes C et η, avec $0 < C < +\infty$ et $0 \le \eta < 1$, telles que pour tout i*

$$\left\|e_i\Pi^\ell - \mu\right\| \le C\eta^\ell,$$

où les e_i forment la base orthonormale naturelle de l'espace des vecteurs de dimension $|\mathcal{A}|$ et où $\|.\|$ est la distance en variation totale.

En général, nous pouvons prendre $C = \dfrac{1}{\inf_{j\in\mathcal{A}}\mu(j)}$ et $\eta = \rho$ avec ρ la deuxième valeur propre de la matrice de transition Π. Pour plus de détails, nous pouvons nous reporter aux références bibliographiques suivantes : [Bel98], [Ros95], [Wil99], [Fil91] et [Liu96]. Il existe en effet bon nombre de méthodes permettant le calcul d'un C et d'un η.

3 Chaînes de Markov cachées

D'après le Théorème 2, lors de l'étude d'une séquence, mises à part quelques positions au début, une chaîne de Markov décrivant un génome est en régime stationnaire. Le principal défaut à utiliser de tels modèles est donc qu'il suppose l'homogénéité des séquences, alors qu'il est reconnu que les séquences biologiques ne sont pas homogènes. L'utilisation de chaînes de Markov cachées (CMC ou HMM pour hidden Markov model) est une manière de prendre en compte cette hétérogénéité. Les CMC sont largement utilisées pour la modélisation des séquences biologiques. Par exemple, [Chu89] analyse l'hétérogénéité des séquences d'ADN. Voir aussi [SW03] et [KMH94] pour des applications à la prédiction de gènes. À l'aide des CMC, on peut détecter les régions codantes et régions non-codantes, les exons et introns, les homologies entre séquences ou encore découvrir des transferts horizontaux ([NBM+02]). Les CMC correspondent au fait biologiques que certains signaux se succèdent le long des séquences. Par exemple, sur un brin d'ADN, un gène peut être suivi d'une région non-codante, puis par un promoteur, puis un autre gène et ainsi de suite. Les protéines sont souvent composées de domaines variés. Il est naturel de penser que la façon dont se succèdent les lettres varie selon la région et cela explique le succès des CMC dans la recherche de régions biologiques de différents rôles.

La modélisation par chaînes de Markov cachées implique l'hypothèse préalable qu'il existe un (petit) nombre de modes de succession des lettres dans les séquences. Le long d'une séquence, se succèdent alors des plages plus ou moins longues suivant chacune son propre mode de succession des lettres que nous nommerons indifféremment régime ou état caché (caché car nous ne connaissons a priori ni la position ni la taille de ces plages). Comme pour les modèles de Markov, nous avons $(X_t)_{t\in\mathbb{N}}$ une suite de variables aléatoires à valeurs dans un alphabet $\mathcal{A}$. À chaque position t dans la séquence correspond la valeur de X_t (l'observation), et le régime de X_t, nommé r_t (l'état caché). La succession des R différents régimes est modélisée par une chaîne de Markov classique A d'ordre 1 :

$$\forall t,\ \forall (i,j) \in \{1,2,\ldots,R\}^2,\ A(i,j) = \mathbb{P}(r_t = j | r_{t-1} = i).$$

Augmenter l'ordre de cette chaîne n'a pas grand intérêt puisque nous voulons des plages assez longues. Excepté pour les rares changements de régimes, $r_{i-1} = r_{i-2}$, il est donc inutile de conditionner aussi r_t par r_{t-2}. Chaque régime $i \in \{1,2,\ldots,R\}$ est lui-même modélisé par une chaîne de Markov classique Π_i d'ordre k. La probabilité d'apparition d'une lettre v en position t dépendra non seulement des k lettres précédentes, mais aussi du régime correspondant à la position t :

$$\forall t,\ \forall i \in \{1,2,\ldots,R\},\ \forall u \in \mathcal{A}^k,\ \forall v \in \mathcal{A},\ \Pi_i(u,v) = \mathbb{P}(X_t = v | (X_{t-k},...,X_{t-1}) = u, r_t = i).$$

L'utilisation des chaînes de Markov cachées, notées $M1Mk$ ($M1$ correspondant à la modélisation de la succession des régimes par une chaîne de Markov d'ordre 1, et Mk correspondant à la modélisation de chaque régime par une chaîne de Markov d'ordre k) permet la modélisation d'une séquence en suite de plages homogènes.

Exemple 2 *On veut modéliser une séquence par un modèle de Markov caché à deux états. Ici, les états cachés correspondent au pourcentage en* `gc`*. Se succèdent alors dans la séquence des plages ayant peu de* `g` *et de* `c`*, et des plages ayant beaucoup de* `g` *et de* `c` *(voir Tableau 6).*

TAB. 6 – Exemple de succession de deux régimes

séquence	`aaattgcacaccttaaagaagacgcacgatattagcccatatatatcgactatatgatcgggcaccagtatacgacgcg`
états	`111111111111111111122222222222222222221111111111111111111122222222222222222222`

Étant donnée une séquence biologique, le problème est d'estimer les paramètres (A, Π_i) et de reconstruire la suite des états cachés r_t (c'est-à-dire de déterminer l'état caché de chaque position t). Nous pouvons trouver dans la littérature une grande quantité de variantes à ces modèles de Markov. Le logiciel SHOW (http://genome.jouy.inra.fr/ssb/SHOW/), développé par Pierre Nicolas (voir [Nic03]) permet de créer ses propres modèles de Markov cachés, pour la détection de gènes par exemple. SHOW aide à décrire l'hétérogénéité de *Bascillus subtilis* (voir [NBM+02]) ou à la détection de petits gènes toujours chez *Bascillus subtilis* (voir [Nic03]). Ainsi de nombreuses méthodes informatiques sont disponibles pour estimer les paramètres d'une chaîne de Markov cachée.

4 Chaînes de Markov régulées

Grâce à leur modélisation intuitive, les chaînes de Markov cachées sont largement utilisées pour modéliser les séquences biologiques. Pourtant il est commun d'observer des variations continues le long d'une séquence biologique,

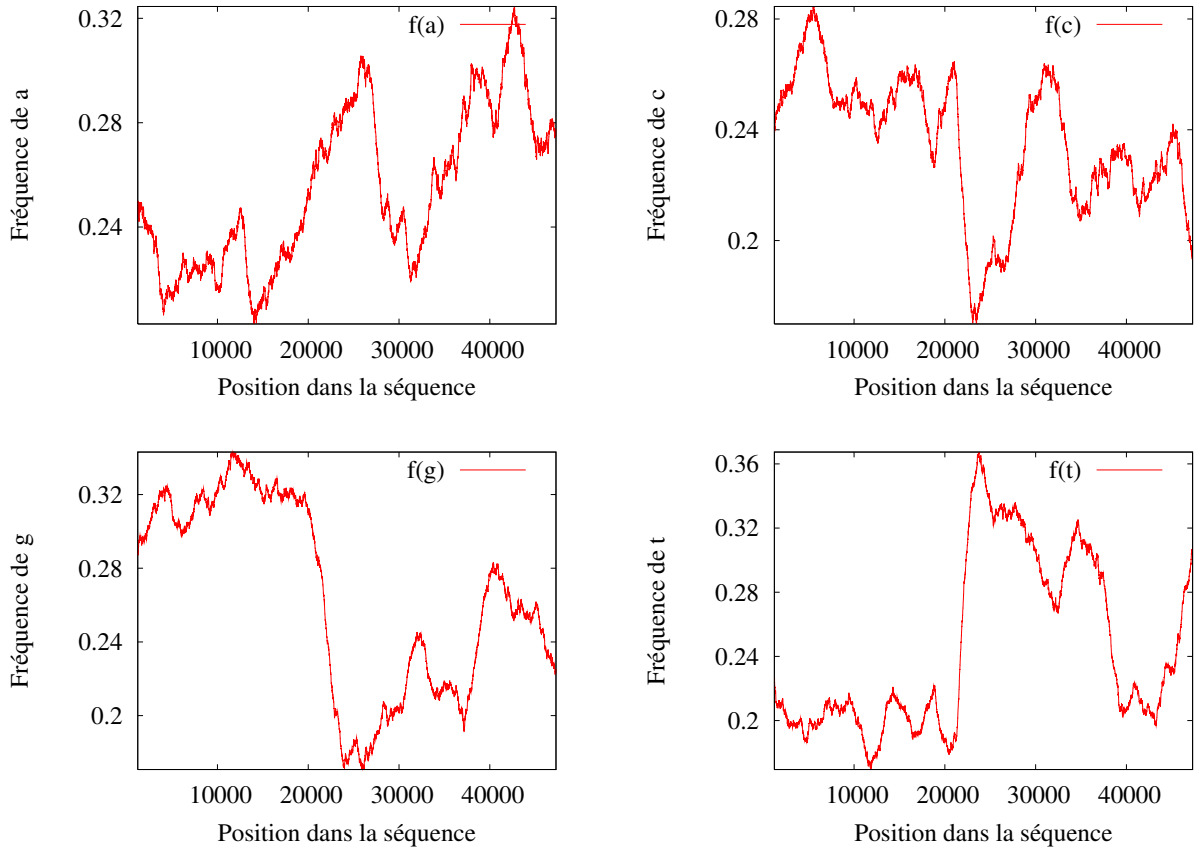
soit à un niveau global soit à l'intérieur d'une des régions précédemment citées (région non codante, gène, etc...). Par exemple la richesse en gc d'une séquence varie selon la position. Un premier modèle distingue deux sortes de comportements : les forts pourcentages en gc (notés H) contre les faibles pourcentages en gc (notés L). Un affinement de ce modèle a été donné, introduisant les régions H1, H2, H3 et L1, L2. On parle d'isochores (voir [Ber93] et Tableau 7). Malgré cet affinement, l'aspect simplificateur de ce modèle est reconnu ([DSP+02], [NL00]).

TAB. 7 – Les isochores

isochore	Pourcentage en gc
L1	gc $<$ 38 %
L2	38 % $<$ gc $<$ 42 %
H1	42 % $<$ gc $<$ 47 %
H2	47 % $<$ gc $<$ 52 %
H3	52 % $<$ gc

Giorgio Bernardi lui-même affirme (ou ré-affirme) dans une réponse à l'*International Human Genome Sequencing Consortium* qu'il n'existe pas d'isochores stricts (voir [Con01] et [Ber01a] pour la réponse). Il est en effet impossible de déterminer précisément les limites de ces cinq régions : une transition douce de la richesse en gc est toujours observée. Même si nous parlons d'isochores principalement chez les mammifères, nous pouvons faire ce constat pour toute séquence d'ADN. Sur le génome complet du phage Lambda ([WT71]), la figure 7 offre un exemple de la variation continue du génome. En effet, les courbes montrent une estimation de la richesse en chacun des quatre nucléotides en fonction de la position dans le génome (cette estimation est obtenue par l'utilisation d'une fenêtre glissante de largeur 2000). La figure montre qu'à chaque position au moins l'une des quatre courbes a une évolution douce. Même autour de la position 22000 où trois courbes semblent avoir une discontinuité (atténuée par l'utilisation d'une fenêtre), la quatrième (correspondant au nucléotide a) a une variation continue. Même à l'intérieur des gènes, ce type de comportement est observé ([NBM+02]).

FIG. 7 – Fréquences des quatre nucléotides chez le phage Lambda



Il est donc nécessaire de développer des outils mathématiques pour prendre en compte de tels changements et nous proposons un modèle, les chaînes de Markov régulées (**CMR** ou **DMM** pour drifting Markov model). Ces modèles peuvent être vus comme une alternative aux modèles de Markov cachés : une CMR peut être ajustée sur une séquence entière. Mais ils peuvent aussi être vus comme un outil complémentaire aux modèles de Markov cachés : les états cachés, habituellement de simples modèles de Markov, peuvent devenir des modèles de Markov régulés.

Les *walking Markov models* (WMM), introduits par [FTW92] ont été les premiers modèles avec une variation continue de la composition en nucléotides. Les auteurs veulent modéliser la composition en `gc` et `at` dans les séquences d'ADN (cela revient à étudier les fameux isochores dont nous venons de parler). Par exemple, ils coupent la séquence en morceaux de 1000 bases puis estiment un modèle de Markov sur tous les morceaux comprenant entre 300 et 400 `at`, et trois autres modèles sur les morceaux comprenant entre 400 et 500, entre 500 et 600 et entre 600 et 700. Ensuite, pour n'importe quelle valeur w (la richesse en `at`), un modèle de Markov M_w est défini par interpolation linéaire entre ces premiers modèles. Enfin, un WMM est défini par une marche aléatoire sur w : ils choisissent une valeur initiale pour w entre 1/3 et 2/3 (cela change en fonction de la séquence étudiée) et pour générer chaque prochaine base, ils ajoutent ou soustraient (avec probabilité 0.5) 0.0015 à w et utilisent M_w.

Nous définissons nos modèles d'une manière totalement différente. Tout d'abord, nous n'utilisons pas de marche aléatoire pour choisir notre matrice de transition : nos modèles sont fondés sur la séquence. Ensuite, nos modèles sont adaptables pour n'importe quel espace d'états sans de nombreux traitements préliminaires telle l'estimation de plusieurs modèles de Markov. Il serait difficile d'adapter les WMM à un espace d'états de taille 20 (comme l'alphabet protéique). Bien sur, les WMM, comme les DMM, ne modélisent pas la structure locale détaillée, comme la structure locale de gènes. Ils ont pour but de modéliser les variations à grande échelle de la composition en bases d'un génome.

Expliquons maintenant le principe des DMM. Au lieu d'ajuster une matrice de transition sur une séquence entière (modèle de Markov homogène classique) ou différentes matrices de transition sur différentes régions de la séquence (modèles de Markov cachés), nous permettons à la matrice de transition de varier (*to drift*) du début à la fin de la séquence. À chaque position, nous avons une matrice de transition (éventuellement) différente. Nos modèles sont donc des modèles de Markov hétérogènes contraints. Dans cette thèse, nous donnerons essentiellement deux manières de contraindre les modèles :

- la modélisation polynomiale ;
- la modélisation par splines.

5 Quelles analyses sur les séquences ?

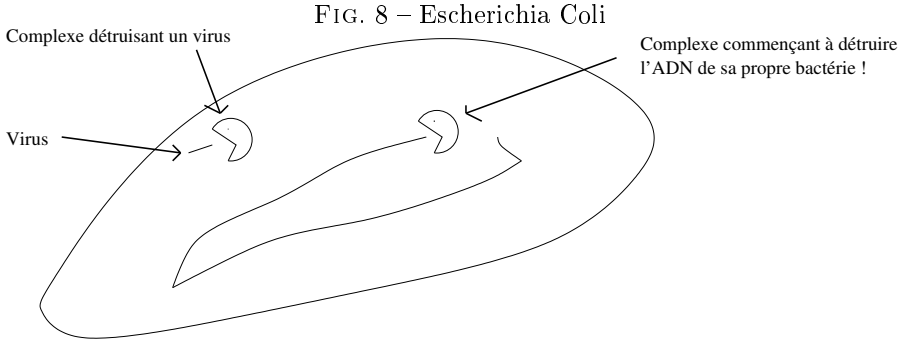
Nous montrons tout d'abord que les distributions de probabilité des nucléotides dans les DMM s'ajustent correctement sur les fréquences des nucléotides calculées sur la vraie séquence. Cela montre que nos modèles offrent un paramétrage plus flexible des données qui nous laisse espérer un meilleur ajustement que les modèles de Markov et les modèles de Markov cachés. En outre, nous validons aussi nos modèles à l'aide d'une application sur les origines de réplication chez les bactéries. Reposant sur les asymétries entre les deux brins de réplication, le programme ORILOC [Lob00] aide à la prédiction de ces origines de réplication. Nous proposons une méthode alternative qui présente l'avantage de pouvoir calculer analytiquement un maximum. En raison d'une difficulté supplémentaire (de modélisation) dans le programme, cette nouvelle méthode est une validation de nos modèles plutôt qu'une méthode concurrente.

La principale application de ces modèles à la biologie est la recherche de mots exceptionnels. Un problème important est de déterminer la significativité statistique de la fréquence d'un mot dans une séquence ADN. [NDV02] discute de cette importance de trouver des mots sur- ou sous-représentés. De nombreux articles parlent de mots exceptionnels dans les séquences mais la plupart sont basés sur les chaînes de Markov et leur homogénéité. [SPdT95] identifie des motifs exceptionnels dans les séquences en utilisant les modèles de Markov. [RS98] propose une autre méthode de détection de mots exceptionnels dans les séquences biologiques tandis que [Nue06b] compare les méthodes les plus utilisées pour découvrir des motifs importants dans les séquences modélisées par des chaînes de Markov. [RRS03] introduit les idées mathématiques et statistiques utilisées pour la résolution de ce problème biologique. Nous introduisons la possibilité d'étudier les mots exceptionnels à l'aide d'un modèle hétérogène correspondant mieux à la séquence réelle.

L'étude des mots biologiquement importants se fait à l'aide de leurs nombres d'occurrences (faibles ou élevés) ou leurs répartitions (beaucoup au début, puis de moins en moins ou bien, beaucoup au début, beaucoup à la fin, peu au milieu...). Ces mots importants sont par exemple les sites de restrictions, les Chi, les séquences uptake, les palindromes, les inverses complémentaires... Les sites de restriction sont des endroits précis sur la séquence, reconnus par des enzymes de restriction qui y coupent la séquence lors de la transcription, la traduction ou bien encore la réplication. Ces endroits sont importants car ils sont parfois les indicateurs de la présence de gènes. Pour

illustrer l'importance des Chi (Crossover Hotspot Investigator), parlons par exemple du Chi d'*Escherichia coli* ([BPB+97]) qui est l'une des bactéries les plus étudiées. Certains virus parviennent à s'introduire à l'intérieur de sa cellule et la détruiraient facilement sans l'existence d'un système de défense. Dans la bactérie, il existe donc un complexe chargé de détruire les virus (en démontant un à un les nucléotides composant le virus)(voir Figure 8). Le tout est de savoir si ce complexe s'attaque bien au virus et non à l'ADN de la bactérie qu'il protège ! Pour cela,

FIG. 8 – Escherichia Coli



il faut au complexe un moyen de reconnaissance, qui lui dirait : "Stop ! C'est moi !" Dans ce cas précis, ce moyen de reconnaissance, appelé Chi, c'est le mot `gctggtgg`. Quand le complexe rencontre ce mot, il s'arrête de détruire. Il est donc nécessaire que dans l'ADN de la bactérie, ce mot apparaisse très souvent afin d'en assurer sa survie en évitant le "suicide" ! Un bon nombre d'organismes possède un Chi et il peut être très utile de connaître ces mots exceptionnels. Les séquences uptake sont d'autres mots importants dans le cadre des transferts horizontaux (un morceau d'ADN d'un organisme est intégré à l'ADN d'un autre). Les palindromes sont souvent des sites de restriction. Par exemple chez *Escherichia coli* les palindromes de taille 6 sont connus pour être des sites de restriction ([KBM92]). Nous nous attendons à en rencontrer assez peu du fait qu'ils fragilisent le génome. Nous cherchons les régions significativement riches en Chi, parallèlement, nous cherchons les régions significativement pauvres en sites de restriction. En effet, ces régions doivent contenir des éléments vitaux pour l'organisme pour être autant protégées.

D'autres analyses pourraient porter sur le pourcentage en `gc`, le degré d'hydrophobicité, la structure de la protéine (hélices α, feuillets β,...). Parmi les principales informations contenues dans les séquences, on peut citer :

- les signaux de maintenance de l'ADN (réplication ou réparation, protection et entretien du génome) ;
- le message génétique (parties codantes et non codantes) ;
- les signaux d'expression (transcription, traduction, épissage, régulation) ;
- les signaux structuraux ;
- les signaux contenus dans les séquences protéiques.

Parmi les objectifs principaux de l'analyse des séquences, on peut citer :

- l'identification des gènes
  - par leurs propriétés intrinsèques (codon start, codon stop, promoteur...),
  - par similarité avec d'autres gènes connus.

  Dans les deux cas, les mathématiques et l'informatique sont utiles.
- La détermination de leur fonction.
- L'acquisition d'une meilleure connaissance des contrôles (signaux de régulation), des interactions entre "objets" biologiques.

Première partie

Dérive polynomiale et dérive par splines

Chapitre 1

Dérive linéaire

Afin de faciliter la compréhension des modèles de Markov régulés, nous commençons par un cas particulier (qui peut d'ailleurs être dans bien des cas celui que l'on utilisera) : une variation linéaire de la matrice de transition le long de la séquence. Nous considérons ici que la séquence est modélisée par une chaîne de Markov régulée d'ordre k. Soit $X = (X_t)_{t\in[\![0,n]\!]}$ une suite de variables aléatoires à valeurs dans un alphabet fini $\mathcal{A}$. Nous définissons la loi des X_t comme suit :

$$\mathbb{P}\left(X_t = v | X_{t-k}\dots X_{t-1} = u\right) = \Pi_{\frac{t}{n}}(u,v)$$

avec $u = u_1u_2\dots u_k$ définissant le passé markovien et $(u_1,u_2,\dots,u_k,v)\in\mathcal{A}^{k+1}$.

Nous considérons deux modèles pour définir $\Pi_{\frac{t}{n}}(u,v)$:

- **le modèle des points d'appui :** $\Pi_{\frac{t}{n}}(u,v) = \left(1-\frac{t}{n}\right)\Pi_0(u,v) + \left(\frac{t}{n}\right)\Pi_1(u,v)$;
- **le modèle des polynômes :** $\Pi_{\frac{t}{n}}(u,v) = M_0(u,v) + \left(\frac{t}{n}\right)M_1(u,v)$;

où Π_0 et Π_1 sont les matrices de transition de chaînes de Markov classiques d'ordre k (voir 0.2), où M_0 est une matrice stochastique et M_1 est une matrice dont les lignes somment à 0. Constatons simplement que ces deux modèles ne sont que deux écritures différentes qui sont équivalentes. En effet, en prenant $M_0 = \Pi_0$ et $M_1 = \Pi_1 - \Pi_0$ pour le second modèle, nous revenons au premier et en prenant $\Pi_0 = M_0$ et $\Pi_1 = M_1 + M_0$ pour le premier, nous revenons au second. L'utilité du premier modèle réside en sa facilité de compréhension : en position 0 la matrice de transition est $\Pi_{\frac{0}{n}} = \Pi_0$ et en position n la matrice de transition est $\Pi_{\frac{n}{n}} = \Pi_1$. L'utilité du second modèle est sa forme simple qui facilite les calculs. En particulier, ce modèle servira pour l'estimation par splines (voir Chapitre 3)

L'objectif principal étant de trouver la meilleure modélisation de la séquence étudiée, nous souhaitons trouver les matrices Π_0 et Π_1 ou bien M_0 et M_1 s'adaptant le mieux à la séquence. Nous avons au moins trois méthodes pour estimer ces matrices :

- la méthode du maximum de vraisemblance ;
- la méthode de régression matricielle ;
- la méthode point par point.

Malheureusement, la méthode du maximum de vraisemblance est impossible à mettre en oeuvre ici. La méthode de régression matricielle impose le découpage de la séquence en plusieurs morceaux, dont le nombre est à déterminer, et sur chacun desquels nous estimons la matrice de transition d'une chaîne de Markov classique par la méthode habituelle (voir 0.2). Nous minimisons ensuite une somme de distances entre les matrices estimées sur chaque segment et les matrices de notre modèle régulé. Enfin, la méthode point par point est une méthode des moindres carrés. Il s'agit de minimiser de façon globale les carrés des erreurs de prédiction du modèles. Nous présentons dans ce premier chapitre les détails de ces trois méthodes.

1.1 Modèles des points d'appui : estimateurs de Π_0 et Π_1

Sous le modèle des points d'appui, la matrice de transition prend la forme suivante :

$$\Pi_{\frac{t}{n}}(u,v) = \left(1-\frac{t}{n}\right)\Pi_0(u,v) + \left(\frac{t}{n}\right)\Pi_1(u,v).$$

Nous fixons une matrice de transition initiale Π_0 et une matrice de transition finale Π_1 et nous permettons à la matrice $\Pi_{\frac{t}{n}}$ de varier linéairement de l'une à l'autre. Ainsi, la matrice de transition générale repose (prend appui) sur chacune des deux autres matrices. Les polynômes $(1-t/n)$ et t/n sont choisis pour établir la stochasticité de

Π_0 et Π_1. Évidemment, le rôle de ces matrices est artificiel comme n'importe quel paramètre de modèle, mais la stochasticité des matrices rend plus facile la compréhension du modèle. Nous voulons estimer ces deux matrices dans le but de construire notre modèle. Dans le cas d'un modèle de Markov classique, la méthode du maximum de vraisemblance est utilisée avec succès (comme nous l'avons vu dans l'introduction 0.2), mais à cause de complexités numériques, nous ne pouvons l'utiliser ici. Nous la présentons tout de même brièvement avant de mettre en place les deux méthodes d'estimation dont nous disposons.

1.1.1 Estimateur du maximum de vraisemblance

Nous ne donnons qu'un essai de maximisation de la vraisemblance pour des modèles de Markov régulés d'ordre 1 et de degré 1 sur l'alphabet des nucléotides $\mathcal{A} = \{a, c, g, t\}$. Sachant que Π_0 et Π_1 sont des matrices stochastiques, nous pouvons réduire notre estimation aux paramètres $\Pi_0(u, v)$ et $\Pi_1(u, v)$ tels que $u \neq v$. A l'ordre 1, u est bien de même nature que v, c'est-à-dire un état de la chaîne de Markov (une lettre appartenant à l'alphabet $\mathcal{A}$).

Déterminons la vraisemblance L du modèle :

$$L(X, \Pi_0, \Pi_1) = \mu_0(X_0) \prod_{t=1}^{n} \Pi_{\frac{t}{n}}(X_{t-1}, X_t) = \mu_0(X_0) \prod_{t=1}^{n} \left[\left(1 - \frac{t}{n}\right) \Pi_0(X_{t-1}, X_t) + \left(\frac{t}{n}\right) \Pi_1(X_{t-1}, X_t) \right].$$

Travailler avec des sommes étant plus facile qu'avec des produits, nous déterminons la log-vraisemblance ℓ du modèle :

$$\begin{aligned}
\ell(X, \Pi_0, \Pi_1) &= \ln \mu_0(X_0) + \sum_{t=1}^{n} \ln \left[\left(1 - \frac{t}{n}\right) \Pi_0(X_{t-1}, X_t) + \left(\frac{t}{n}\right) \Pi_1(X_{t-1}, X_t) \right] \\
&= \ln \mu_0(X_0) + \sum_{t=1}^{n} \sum_{u \in \mathcal{A}} \mathbb{1}_{\{X_{t-1}=u\}} \sum_{v \in \mathcal{A}} \mathbb{1}_{\{X_t=v\}} \ln \left(\Pi_{\frac{t}{n}}(u, v) \right) \\
&= \ln \mu_0(X_0) + \sum_{t=1}^{n} \sum_{(u,v) \in \mathcal{A}^2} \mathbb{1}_{\{X_{t-1}X_t=uv\}} \ln \left(\Pi_{\frac{t}{n}}(u, v) \right).
\end{aligned}$$

La stochasticité des matrices nous permet de poser

$$\Pi_{\frac{t}{n}}(u, u) = \left(1 - \frac{t}{n}\right) \left(1 - \sum_{v \in \mathcal{A} \setminus \{u\}} \Pi_0(u, v)\right) + \left(\frac{t}{n}\right) \left(1 - \sum_{v \in \mathcal{A} \setminus \{u\}} \Pi_1(u, v)\right).$$

Ainsi,

$$\ell = \ln \mu_0(X_0) + \sum_{t=1}^{n} \sum_{u \in \mathcal{A}} \mathbb{1}_{\{X_{t-1}=u\}} \left(\left(\sum_{v \in \mathcal{A} \setminus \{u\}} \left(\mathbb{1}_{\{X_t=v\}} \ln \left(\Pi_{\frac{t}{n}}(u, v) \right) \right) \right) + \mathbb{1}_{\{X_t=u\}} \ln \left(\Pi_{\frac{t}{n}}(u, u) \right) \right).$$

La maximisation de la vraisemblance s'effectue en annulant la dérivée. Pour $u \in \mathcal{A}$ et $v \in \mathcal{A} \setminus \{u\}$, nous avons

$$\begin{aligned}
&\begin{cases} \dfrac{\partial \ell(X, \Pi_0, \Pi_1)}{\partial \Pi_0(u, v)} &= 0 \\ \dfrac{\partial \ell(X, \Pi_0, \Pi_1)}{\partial \Pi_1(u, v)} &= 0 \end{cases} \\
\iff &\begin{cases} \displaystyle\sum_{t=1}^{n} \mathbb{1}_{\{X_{t-1}=u\}} \left(\left(1 - \frac{t}{n}\right) \frac{\mathbb{1}_{\{X_t=v\}}}{\Pi_{\frac{t}{n}}(u, v)} - \left(1 - \frac{t}{n}\right) \frac{\mathbb{1}_{\{X_t=u\}}}{\Pi_{\frac{t}{n}}(u, u)} \right) &= 0 \\ \displaystyle\sum_{t=1}^{n} \mathbb{1}_{\{X_{t-1}=u\}} \left(\left(\frac{t}{n}\right) \frac{\mathbb{1}_{\{X_t=v\}}}{\Pi_{\frac{t}{n}}(u, v)} - \left(\frac{t}{n}\right) \frac{\mathbb{1}_{\{X_t=u\}}}{\Pi_{\frac{t}{n}}(u, u)} \right) &= 0 \end{cases} \\
\iff &\begin{cases} \displaystyle\sum_{t=1}^{n} \left(1 - \frac{t}{n}\right) \frac{\mathbb{1}_{\{X_{t-1}=u, X_t=v\}}}{\Pi_{\frac{t}{n}}(u, v)} &= \displaystyle\sum_{t=1}^{n} \left(1 - \frac{t}{n}\right) \frac{\mathbb{1}_{\{X_{t-1}=u, X_t=u\}}}{\Pi_{\frac{t}{n}}(u, u)} \\ \displaystyle\sum_{t=1}^{n} \left(\frac{t}{n}\right) \frac{\mathbb{1}_{\{X_{t-1}=u, X_t=v\}}}{\Pi_{\frac{t}{n}}(u, v)} &= \displaystyle\sum_{t=1}^{n} \left(\frac{t}{n}\right) \frac{\mathbb{1}_{\{X_{t-1}=u, X_t=u\}}}{\Pi_{\frac{t}{n}}(u, u)} \end{cases}.
\end{aligned}$$

Nous obtenons ainsi un système de $2|\mathcal{A}|(|\mathcal{A}|-1)$ équations à $2|\mathcal{A}|(|\mathcal{A}|-1)$ inconnues (toutes aux dénominateurs). En fait, il se réduit à $|\mathcal{A}|$ systèmes de $2(|\mathcal{A}|-1)$ équations à $2(|\mathcal{A}|-1)$ inconnues. Ces systèmes ne sont pas linéaires, un simple exemple à l'ordre 1 montre qu'il est impensable de vouloir résoudre de tels systèmes analytiquement ou même numériquement.

Exemple 3 *Pour $\mathcal{A}=\{a,c,g,t\}$, nous devons résoudre 4 systèmes de 6 équations à 6 inconnues. Nous présentons un seul de ces systèmes :*

$$\left\{\begin{array}{l}
\bullet\displaystyle\sum_{t=1}^{n}\frac{\mathbb{1}_{\{X_{t-1}=a,X_t=c\}}\left(\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)\Pi_0(a,c)+\left(\frac{t}{n}\right)\Pi_1(a,c)}= \\
\displaystyle\sum_{t=1}^{n}\frac{\mathbb{1}_{\{X_{t-1}=a,X_t=a\}}\left(\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)(1-\Pi_0(a,c)-\Pi_0(a,g)-\Pi_0(a,t))+\left(\frac{t}{n}\right)(1-\Pi_1(a,c)-\Pi_1(a,g)-\Pi_1(a,t))} \\
\bullet\displaystyle\sum_{t=1}^{n}\frac{\mathbb{1}_{\{X_{t-1}=a,X_t=g\}}\left(\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)\Pi_0(a,g)+\left(\frac{t}{n}\right)\Pi_1(a,g)}= \\
\displaystyle\sum_{t=1}^{n}\frac{\mathbb{1}_{\{X_{t-1}=a,X_t=a\}}\left(\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)(1-\Pi_0(a,c)-\Pi_0(a,g)-\Pi_0(a,t))+\left(\frac{t}{n}\right)(1-\Pi_1(a,c)-\Pi_1(a,g)-\Pi_1(a,t))} \\
\bullet\displaystyle\sum_{t=1}^{n}\frac{\mathbb{1}_{\{X_{t-1}=a,X_t=t\}}\left(\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)\Pi_0(a,t)+\left(\frac{t}{n}\right)\Pi_1(a,t)}= \\
\displaystyle\sum_{t=1}^{n}\frac{\mathbb{1}_{\{X_{t-1}=a,X_t=a\}}\left(\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)(1-\Pi_0(a,c)-\Pi_0(a,g)-\Pi_0(a,t))+\left(\frac{t}{n}\right)(1-\Pi_1(a,c)-\Pi_1(a,g)-\Pi_1(a,t))} \\
\bullet\displaystyle\sum_{t=1}^{n}\frac{\mathbb{1}_{\{X_{t-1}=a,X_t=c\}}\left(1-\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)\Pi_0(a,c)+\left(\frac{t}{n}\right)\Pi_1(a,c)}= \\
\displaystyle\sum_{t=1}^{n}\frac{\mathbb{1}_{\{X_{t-1}=a,X_t=a\}}\left(1-\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)(1-\Pi_0(a,c)-\Pi_0(a,g)-\Pi_0(a,t))+\left(\frac{t}{n}\right)(1-\Pi_1(a,c)-\Pi_1(a,g)-\Pi_1(a,t))} \\
\bullet\displaystyle\sum_{t=1}^{n}\frac{\mathbb{1}_{\{X_{t-1}=a,X_t=g\}}\left(1-\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)\Pi_0(a,g)+\left(\frac{t}{n}\right)\Pi_1(a,g)}= \\
\displaystyle\sum_{t=1}^{n}\frac{\mathbb{1}_{\{X_{t-1}=a,X_t=a\}}\left(1-\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)(1-\Pi_0(a,c)-\Pi_0(a,g)-\Pi_0(a,t))+\left(\frac{t}{n}\right)(1-\Pi_1(a,c)-\Pi_1(a,g)-\Pi_1(a,t))} \\
\bullet\displaystyle\sum_{t=1}^{n}\frac{\mathbb{1}_{\{X_{t-1}=a,X_t=t\}}\left(1-\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)\Pi_0(a,t)+\left(\frac{t}{n}\right)\Pi_1(a,t)}= \\
\displaystyle\sum_{t=1}^{n}\frac{\mathbb{1}_{\{X_{t-1}=a,X_t=a\}}\left(1-\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)(1-\Pi_0(a,c)-\Pi_0(a,g)-\Pi_0(a,t))+\left(\frac{t}{n}\right)(1-\Pi_1(a,c)-\Pi_1(a,g)-\Pi_1(a,t))}
\end{array}\right.$$

Pour des modèles d'ordre k supérieur, le nombre de systèmes s'élève à $|\mathcal{A}|^k$ systèmes. Seul le nombre de systèmes changeant, il suffirait de savoir en résoudre un seul pour pouvoir utiliser cette méthode. Mais le problème est insoluble et au vu de cet exemple, nous abandonnons la méthode du maximum de vraisemblance.

1.1.2 Régression matricielle

Soit une séquence $X_0X_1...X_n$. Une première idée pour obtenir les matrices Π_0 et Π_1 est de diviser la séquence en N segments de taille m (que nous choisirons plus tard) :

$$X_0\cdots X_{m-1}, X_m\cdots X_{2m-1}, \ldots, X_{(N-1)m}\cdots X_n.$$

Nous posons S_ℓ l'ensemble des indices t des X_t appartenant au segment ℓ, et $S_\ell^\star$ l'ensemble S_ℓ privé de ses k premiers éléments, où k est l'ordre du modèle. Tous les segments sont de taille $m \geq k$ excepté éventuellement le dernier (S_N) dont la taille peut être plus grande. Ainsi,

- $S_\ell = [\![(\ell-1)m, \ell m-1]\!]$ pour $\ell \in [\![1, N-1]\!]$;
- $S_N = [\![(N-1)m, n]\!]$;
- $S_\ell^\star = [\![(\ell-1)m+k, \ell m-1]\!]$ pour $\ell \in [\![1, N-1]\!]$;
- $S_N^\star = [\![(N-1)m+k, n]\!]$;

Nous avons

- $|S_\ell| = m$ pour $\ell \in [\![1, N-1]\!]$;
- $|S_N| = n-(N-1)m+1$;
- $|S_\ell^\star| = m-k$ pour $\ell \in [\![1, N-1]\!]$;

– $|S_N^\star| = n - (N-1)m - k + 1$.

Exemple 4 *Considérons une séquence de taille* $n = 1700$ *et un modèle d'ordre* 1. *Choisissons* $m = 400$, *ainsi nous obtenons* $N = 4$ *segments :*
– $S_1 = [\![0, 399]\!]$ *et* $S_1^\star = [\![1, 399]\!]$ *;*
– $S_2 = [\![400, 799]\!]$ *et* $S_2^\star = [\![401, 799]\!]$ *;*
– $S_3 = [\![800, 1199]\!]$ *et* $S_3^\star = [\![801, 1199]\!]$ *;*
– $S_4 = [\![1200, 1699]\!]$ *et* $S_4^\star = [\![1201, 1699]\!]$.

L'idée de cette méthode est d'utiliser une quasi-homogénéité sur chaque segment. Sur chaque segment S_ℓ, nous estimons une matrice de transition homogène avec les estimateurs habituels :

$$\widehat{\Pi_{S_\ell}}(u,v) = \frac{N_{S_\ell}(uv)}{N_{S_\ell}(u+)} = \frac{\sum_{t \in S_\ell^\star} \mathbb{1}\{X_{t-k} \dots X_{t-1} = u, X_t = v\}}{\sum_{j \in S_\ell^\star} \mathbb{1}\{X_{j-k} \dots X_{j-1} = u\}}.$$

Nous nous restreignons aux $u \in \mathcal{A}^k$, $\ell \in [\![1, N]\!]$ tels que $N_{S_\ell}(u+) \neq 0$.

Distances

Dans le but d'ajuster notre modèle hétérogène sur la séquence entière nous choisissons un point dans chaque segment. Nous choisissons les N milieux τ_ℓ des segments S_ℓ car $\mathbb{E}\left(\widehat{\Pi}_{S_\ell}\right)$ tend vers Π_{τ_ℓ} quand m tend vers l'infini. Nous aurions pu choisir plus d'un point par segment mais cela aurait introduit des complexités numériques sans amélioration sensible de l'estimation (voir A.1.3). Nous voulons qu'au milieu de chaque segment S_ℓ, notre matrice $\Pi_{\frac{t}{n}}$ soit la plus proche possible de $\widehat{\Pi}_{S_\ell}$. Ainsi, la régression matricielle revient à minimiser la somme des distances entre les matrices estimées sur chaque segment et la matrice de transition $\Pi_{\frac{t}{n}}$ au milieu τ_ℓ du $\ell^{ème}$ segment, pour une distance matricielle d choisie :

$$\sum_{\ell \in [\![1,N]\!]} d\left(\widehat{\Pi_{S_\ell}}, (1-\tau_\ell)\Pi_0 + \tau_\ell \Pi_1\right).$$

Nous choisissons la distance quadratique suivante :

$$d(N_1, N_2) = \sum_{i,j} (N_1(i,j) - N_2(i,j))^2$$

avec N_1, N_2 deux matrices d'un espace matriciel quelconque. En particulier ici, notre espace matriciel sera celui des matrices $|\mathcal{A}|^k \times |\mathcal{A}|$. Nous devons donc minimiser la fonction SR_1 suivante :

$$SR_1(\Pi_0, \Pi_1) = \sum_{\ell \in [\![1,N]\!]} \sum_{u \in \mathcal{A}^k} \sum_{v \in \mathcal{A}} \left(\widehat{\Pi_{S_\ell}}(u,v) - (1-\tau_\ell)\Pi_0(u,v) - \tau_\ell \Pi_1(u,v)\right)^2.$$

La recherche des zéros de la dérivée nous donne le système suivant pour chaque couple $(u,v) \in \mathcal{A}^k \times \mathcal{A}$:

$$\begin{cases} \dfrac{\partial SR_1(\Pi_0,\Pi_1)}{\partial \Pi_0(u,v)} &= 0 \\ \dfrac{\partial SR_1(\Pi_0,\Pi_1)}{\partial \Pi_1(u,v)} &= 0 \end{cases}$$

$$\iff \begin{cases} \displaystyle\sum_{\ell=1}^N 2(1-\tau_\ell)\left(\tau_\ell \widehat{\Pi_1}(u,v) + (1-\tau_\ell)\widehat{\Pi_0}(u,v) - \widehat{\Pi_{S_\ell}}(u,v)\right) &= 0 \\ \displaystyle\sum_{\ell=1}^N 2\tau_\ell\left(\tau_\ell \widehat{\Pi_1}(u,v) + (1-\tau_\ell)\widehat{\Pi_0}(u,v) - \widehat{\Pi_{S_\ell}}(u,v)\right) &= 0 \end{cases}$$

$$\iff \begin{cases} \displaystyle\sum_{\ell=1}^N \left(\tau_\ell \widehat{\Pi_1}(u,v) + (1-\tau_\ell)\widehat{\Pi_0}(u,v) - \widehat{\Pi_{S_\ell}}(u,v)\right) &= 0 \\ \displaystyle\sum_{\ell=1}^N \tau_\ell\left(\tau_\ell \widehat{\Pi_1}(u,v) + (1-\tau_\ell)\widehat{\Pi_0}(u,v) - \widehat{\Pi_{S_\ell}}(u,v)\right) &= 0 \end{cases}$$

$$\Longleftrightarrow \left\{ \begin{array}{rcl} \widehat{\Pi_0}(u,v)\left(\sum_{\ell=1}^{N} 1-\tau_\ell\right) + \widehat{\Pi_1}(u,v)\left(\sum_{\ell=1}^{N}\tau_\ell\right) - \sum_{\ell=1}^{N}\widehat{\Pi_{S_\ell}}(u,v) & = & 0 \\ \\ \widehat{\Pi_0}(u,v)\left(\sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell)\right) + \widehat{\Pi_1}(u,v)\left(\sum_{\ell=1}^{N}{\tau_\ell}^2\right) - \sum_{\ell=1}^{N}\tau_\ell\widehat{\Pi_{S_\ell}}(u,v) & = & 0 \end{array} \right.$$

$$\Longleftrightarrow \left\{ \begin{array}{rcl} R_1\widehat{\Pi_0}(u,v) + R_2\widehat{\Pi_1}(u,v) - R_3(u,v) & = & 0 \\ \\ R_4\widehat{\Pi_0}(u,v) + R_5\widehat{\Pi_1}(u,v) - R_6(u,v) & = & 0 \end{array} \right.$$

$$\Longleftrightarrow \left\{ \begin{array}{rcl} \widehat{\Pi_0}(u,v) & = & \dfrac{R_2R_6(u,v) - R_5R_3(u,v)}{R_4R_2 - R_1R_5} \\ \\ \widehat{\Pi_1}(u,v) & = & \dfrac{R_4R_3(u,v) - R_1R_6(u,v)}{R_4R_2 - R_1R_5} \end{array} \right.$$

avec

$$\begin{array}{rclrclrcl} R_1 & = & \sum_{\ell=1}^{N} 1-\tau_\ell, & R_2 & = & \sum_{\ell=1}^{N}\tau_\ell, & R_3(u,v) & = & \sum_{\ell=1}^{N}\widehat{\Pi_{S_\ell}}(u,v), \\ R_4 & = & \sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell), & R_5 & = & \sum_{\ell=1}^{N}{\tau_\ell}^2, & R_6(u,v) & = & \sum_{\ell=1}^{N}\tau_\ell\widehat{\Pi_{S_\ell}}(u,v). \end{array}$$

Pour un alphabet de 4 lettres, nous avons ainsi à l'ordre 1, 16 systèmes de 2 équations à 2 inconnues. Plus généralement, pour un alphabet $\mathcal{A}$ donné, nous obtenons $|\mathcal{A}|^{k+1}$ systèmes de 2 équations à 2 inconnues. Nous avons alors à m fixé, les estimateurs de Π_0 et Π_1. Il est important de noter que nous n'obtenons pas un modèle homogène sur les segments et que cette hypothèse est seulement utilisée pour l'estimation préliminaire des $\widehat{\Pi}_{S_\ell}$.

Théorème 3 *Les matrices $\widehat{\Pi_0}$ et $\widehat{\Pi_1}$ sont stochastiques (voir A.1.1). Dans de rares cas, il est possible d'obtenir des termes négatifs dans les matrices estimées. Ce problème est résolu par un réajustement proportionnel des valeurs (voir A.1.2).*

Espérances et variances de $\widehat{\Pi_{S_\ell}}(u,v)$ pour le modèle des chaînes de Markov régulées

Les estimateurs $\widehat{\Pi_{S_\ell}}(u,v)$ pour le modèle des chaînes de Markov classiques sont sans biais et de variance qui tend vers zéro quand $|S_\ell| \to +\infty$, et par conséquent convergents (voir A.1.4). Nous nous servons de ces résultats pour déterminer les espérances et variances des estimateurs $\widehat{\Pi_{S_\ell}}(u,v)$ pour le modèle régulé. Alors que pour un modèle classique, la probabilité d'apparition d'un v derrière un u est la même sur tout le segment (égale à $\Pi_{S_\ell}(u,v)$), elle dépend de la position dans le segment pour un modèle régulé.

Nous obtenons les espérances suivantes (voir A.1.5 pour les détails) :

$$\begin{array}{rcl} \mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right) & \approx & (1-(1-\mu_\ell(u))^{m-k})\Pi_{\frac{2\ell m-m+k-1}{2n}}(u,v) \\ \mathbb{E}\left(\widehat{\Pi_{S_N}}(u,v)\right) & \approx & (1-(1-\mu_\ell(u))^{n-(N-1)m-k+1})\Pi_{\frac{(N-1)m+k+n}{2n}}(u,v). \end{array}$$

Théorème 4 *Les estimateurs des $\widehat{\Pi_{S_\ell}}(u,v)$ pour le modèle des chaînes de Markov régulées sont asymptotiquement sans biais.*

Preuve *En effet, nous avons*

$$\lim_{m\to+\infty}\mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right) = \Pi_{\tau_\ell}(u,v) \text{ et } \lim_{m\to+\infty}\mathbb{E}\left(\widehat{\Pi_{S_N}}(u,v)\right) = \Pi_{\tau_N}(u,v)$$

où les τ_ℓ sont (presque) les milieux des segments S_ℓ :

- $\tau_\ell = \dfrac{2\ell m - m + k - 1}{2n}$;
- $\tau_N = \dfrac{(N-1)m + k + n}{2n}$.

Théorème 5 *Les estimateurs $\widehat{\Pi_{S_\ell}}(u,v)$ ont une variance qui tend vers zéro quand $|S_\ell| \to +\infty$.*

Preuve *Introduisons quelques notations pour le calcul des variances de ces estimateurs. Nous posons*

$$\begin{aligned} A_- &= \frac{1-(1-\mu_\ell(u))^{|S_\ell^\star|+1}-(|S_\ell^\star|+1)\mu_\ell(u)(1-\mu_\ell(u))^{|S_\ell^\star|}}{|S_\ell^\star|\mu_\ell(u)}, \\ A_+ &= \frac{2\left(1-(1-\mu_\ell(u))^{|S_\ell^\star|+1}-(|S_\ell^\star|+1)\mu_\ell(u)(1-\mu_\ell(u))^{|S_\ell^\star|}\right)}{|S_\ell^\star|\mu_\ell(u)}, \end{aligned}$$

$$\begin{aligned} B_- &= I - II_+, \\ B_+ &= I - II_-, \\ I &= \left(1-(1-\mu_\ell(u))^{|S_\ell^\star|}-|S_\ell^\star|\mu_\ell(u)(1-\mu_\ell(u))^{|S_\ell^\star|-1}\right), \\ II_- &= \frac{1-(1-\mu_\ell(u))^{|S_\ell^\star|+1}-(|S_\ell^\star|+1)(1-\mu_\ell(u))^{|S_\ell^\star|}\mu_\ell(u)-\frac{|S_\ell^\star|(|S_\ell^\star|+1)}{2}(1-\mu_\ell(u))^{|S_\ell^\star|-1}\mu_\ell(u)^2}{\mu_\ell(u)(|S_\ell^\star|+1)}, \\ II_+ &= \frac{2\left(1-(1-\mu_\ell(u))^{|S_\ell^\star|+1}-(|S_\ell^\star|+1)(1-\mu_\ell(u))^{|S_\ell^\star|}\mu_\ell(u)-\frac{|S_\ell^\star|(|S_\ell^\star|+1)}{2}(1-\mu_\ell(u))^{|S_\ell^\star|-1}\mu_\ell(u)^2\right)}{\mu_\ell(u)(|S_\ell^\star|+1)} \end{aligned}$$

et $K=\dfrac{a-b-2}{12n^2}$ *avec* a *et* b *premier et dernier élément de* S_ℓ*. Nous obtenons un encadrement de la variance :*

$$\begin{aligned} &A_-\Pi_{\tau_\ell}(u,v)+B_-\left(\Pi_{\tau_\ell}(u,v)^2+K(\Pi_1(u,v)-\Pi_0(u,v))^2\right)-\Pi_{\tau_\ell}(u,v)^2\left(1-(1-\mu_\ell(u))^{|S_\ell^\star|}\right)^2 \\ &\qquad\qquad \leq \mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)\leq \\ &A_+\Pi_{\tau_\ell}(u,v)+B_+\left(\Pi_{\tau_\ell}(u,v)^2+K(\Pi_1(u,v)-\Pi_0(u,v))^2\right)-\Pi_{\tau_\ell}(u,v)^2\left(1-(1-\mu_\ell(u))^{|S_\ell^\star|}\right)^2 \end{aligned}$$

Nous montrons (voir A.1.5) que minorant et majorant tendent vers 0 *quand* $|S_\ell|\to+\infty$*. Ainsi, la variance de nos estimateurs tend vers zéro.*

Théorème 6 *Les estimateurs* $\widehat{\Pi_{S_\ell}}(u,v)$ *sont convergents.*

Preuve *La convergence découle directement des Théorèmes 4 et 5.*

Espérances et variances de $\widehat{\Pi_0}(u,v)$ et $\widehat{\Pi_1}(u,v)$

Nous avons calculé les espérances et variances des estimateurs pour chaque segment ce qui va nous permettre de calculer les espérances et variances des estimateurs de $\widehat{\Pi_0}(u,v)$ et $\widehat{\Pi_1}(u,v)$. Calculons tout d'abord l'espérance de ces estimateurs :

$$\left\{\begin{array}{rcl} \widehat{\Pi_0}(u,v) &=& \dfrac{R_2R_6(u,v)-R_5R_3(u,v)}{R_4R_2-R_1R_5} \\ \widehat{\Pi_1}(u,v) &=& \dfrac{R_4R_3(u,v)-R_1R_6(u,v)}{R_4R_2-R_1R_5} \end{array}\right. \Longleftrightarrow \left\{\begin{array}{rcl} \mathbb{E}\left(\widehat{\Pi_0}(u,v)\right) &=& \dfrac{R_2\mathbb{E}\left(R_6(u,v)\right)-R_5\mathbb{E}\left(R_3(u,v)\right)}{R_4R_2-R_1R_5} \\ \mathbb{E}\left(\widehat{\Pi_1}(u,v)\right) &=& \dfrac{R_4\mathbb{E}\left(R_3(u,v)\right)-R_1\mathbb{E}\left(R_6(u,v)\right)}{R_4R_2-R_1R_5} \end{array}\right.$$

Ainsi

$$\begin{aligned} \mathbb{E}\left(\widehat{\Pi_0}(u,v)\right) &= \frac{\left(\sum_{\ell=1}^N\tau_\ell\right)\left(\sum_{\ell=1}^N\tau_\ell\mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)\right)-\left(\sum_{\ell=1}^N{\tau_\ell}^2\right)\left(\sum_{\ell=1}^N\mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)\right)}{\left(\sum_{\ell=1}^N\tau_\ell(1-\tau_\ell)\right)\left(\sum_{\ell=1}^N\tau_\ell\right)-\left(\sum_{\ell=1}^N1-\tau_\ell\right)\left(\sum_{\ell=1}^N{\tau_\ell}^2\right)} \\ \mathbb{E}\left(\widehat{\Pi_1}(u,v)\right) &= \frac{\left(\sum_{\ell=1}^N\tau_\ell(1-\tau_\ell)\right)\left(\sum_{\ell=1}^N\mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)\right)-\left(\sum_{\ell=1}^N1-\tau_\ell\right)\left(\sum_{\ell=1}^N\tau_\ell\mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)\right)}{\left(\sum_{\ell=1}^N\tau_\ell(1-\tau_\ell)\right)\left(\sum_{\ell=1}^N\tau_\ell\right)-\left(\sum_{\ell=1}^N1-\tau_\ell\right)\left(\sum_{\ell=1}^N{\tau_\ell}^2\right)}. \end{aligned}$$

Théorème 7 *Les estimateurs* $\widehat{\Pi_0}(u,v)$ *et* $\widehat{\Pi_1}(u,v)$ *sont asymptotiquement sans biais (voir A.1.6).*

Calculons ensuite la variance de ces estimateurs.

$$\left\{\begin{array}{lcl}\widehat{\Pi_0}(u,v) &=& \dfrac{R_2R_6(u,v)-R_5R_3(u,v)}{R_4R_2-R_1R_5}\\[2ex] \widehat{\Pi_1}(u,v) &=& \dfrac{R_4R_3(u,v)-R_1R_6(u,v)}{R_4R_2-R_1R_5}\end{array}\right.$$

$$\Longleftrightarrow \left\{\begin{array}{lcl}\mathbb{V}\left(\widehat{\Pi_0}(u,v)\right) &=& \dfrac{{R_2}^2\mathbb{V}\left(R_6(u,v)\right)+{R_5}^2\mathbb{V}\left(R_3(u,v)\right)-2R_2R_5\mathbb{C}\mathrm{ov}\left(R_3(u,v),R_6(u,v)\right)}{\left(R_4R_2-R_1R_5\right)^2}\\[2ex] \mathbb{V}\left(\widehat{\Pi_1}(u,v)\right) &=& \dfrac{{R_4}^2\mathbb{V}\left(R_3(u,v)\right)+{R_1}^2\mathbb{V}\left(R_6(u,v)\right)-2R_1R_4\mathbb{C}\mathrm{ov}\left(R_3(u,v),R_6(u,v)\right)}{\left(R_4R_2-R_1R_5\right)^2}\end{array}\right.$$

Nous avons l'indépendance des estimateurs $\widehat{\Pi_{S_\ell}}(u,v)$ des segments. Ainsi :

$$\begin{array}{lcl}\mathbb{V}\left(R_3(u,v)\right) &=& \displaystyle\sum_{\ell=1}^{N}\mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)\\ \mathbb{V}\left(R_6(u,v)\right) &=& \displaystyle\sum_{\ell=1}^{N}{\tau_\ell}^2\mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)\\ \mathbb{C}\mathrm{ov}\left(R_3(u,v),R_6(u,v)\right) &=& \mathbb{E}\left(R_3(u,v)R_6(u,v)\right)-\mathbb{E}\left(R_3(u,v)\right)\mathbb{E}\left(R_6(u,v)\right)\\ &=& \displaystyle\sum_{(\ell 1,\ell 2)\in[\![1,N]\!]^2}\tau_{\ell 2}\mathbb{E}\left(\widehat{\Pi_{S_{\ell 1}}}(u,v)\widehat{\Pi_{S_{\ell 2}}}(u,v)\right)\\ &-& \left(\displaystyle\sum_{\ell_1=0}^{N-1}\mathbb{E}\left(\widehat{\Pi_{S_{\ell_1}}}(u,v)\right)\right)\left(\displaystyle\sum_{\ell_2=0}^{N-1}\tau_{\ell_2}\mathbb{E}\left(\widehat{\Pi_{S_{\ell_2}}}(u,v)\right)\right)\\ &=& \displaystyle\sum_{\ell=1}^{N}\tau_\ell\mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)\end{array}$$

En effet les termes pour lesquels $\ell_1\neq\ell_2$ s'annulent du fait de l'indépendance des estimateurs des segments. Ainsi :

$$\begin{array}{lcl}\mathbb{V}\left(\widehat{\Pi_0}(u,v)\right) &=& \dfrac{\left(\displaystyle\sum_{\ell=1}^{N}\tau_\ell\right)^2\left(\displaystyle\sum_{\ell=1}^{N}{\tau_\ell}^2\mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)\right)+\left(\displaystyle\sum_{\ell=1}^{N}{\tau_\ell}^2\right)^2\left(\displaystyle\sum_{\ell=1}^{N}\mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)\right)}{\left[\left(\displaystyle\sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell)\right)\left(\displaystyle\sum_{\ell=1}^{N}\tau_\ell\right)-\left(\displaystyle\sum_{\ell=1}^{N}1-\tau_\ell\right)\left(\displaystyle\sum_{\ell=1}^{N}{\tau_\ell}^2\right)\right]^2}\\ &-& \dfrac{2\left(\displaystyle\sum_{\ell=1}^{N}\tau_\ell\right)\left(\displaystyle\sum_{\ell=1}^{N}{\tau_\ell}^2\right)\left(\displaystyle\sum_{\ell=1}^{N}\tau_\ell\mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)\right)}{\left[\left(\displaystyle\sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell)\right)\left(\displaystyle\sum_{\ell=1}^{N}\tau_\ell\right)-\left(\displaystyle\sum_{\ell=1}^{N}1-\tau_\ell\right)\left(\displaystyle\sum_{\ell=1}^{N}{\tau_\ell}^2\right)\right]^2}\end{array}$$

$$\begin{array}{lcl}\mathbb{V}\left(\widehat{\Pi_1}(u,v)\right) &=& \dfrac{\left(\displaystyle\sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell)\right)^2\left(\displaystyle\sum_{\ell=1}^{N}\mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)\right)+\left(\displaystyle\sum_{\ell=1}^{N}1-\tau_\ell\right)^2\left(\displaystyle\sum_{\ell=1}^{N}{\tau_\ell}^2\mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)\right)}{\left[\left(\displaystyle\sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell)\right)\left(\displaystyle\sum_{\ell=1}^{N}\tau_\ell\right)-\left(\displaystyle\sum_{\ell=1}^{N}1-\tau_\ell\right)\left(\displaystyle\sum_{\ell=1}^{N}{\tau_\ell}^2\right)\right]^2}\\ &-& \dfrac{2\left(\displaystyle\sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell)\right)\left(\displaystyle\sum_{\ell=1}^{N}1-\tau_\ell\right)\left(\displaystyle\sum_{\ell=1}^{N}\tau_\ell\mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)\right)}{\left[\left(\displaystyle\sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell)\right)\left(\displaystyle\sum_{\ell=1}^{N}\tau_\ell\right)-\left(\displaystyle\sum_{\ell=1}^{N}1-\tau_\ell\right)\left(\displaystyle\sum_{\ell=1}^{N}{\tau_\ell}^2\right)\right]^2}\end{array}$$

Nous expliquons en annexe (voir A.1.6) notre choix $m=\sqrt{n}$ pour la longueur des segments.

1.1.3 Méthode point par point

Rappelons le modèle des points d'appui :

$$\Pi_{\frac{t}{n}}(u,v) = \Pi_0(u,v)\left(1-\frac{t}{n}\right) + \Pi_1(u,v)\left(\frac{t}{n}\right).$$

Une autre idée pour obtenir les matrices Π_0 et Π_1 est une méthode des moindres carrés. Nous minimisons une forme quadratique des différents paramètres qui est la somme des erreurs de prédictions. À chaque position t dans la séquence, connaissant le mot $u = X_{t-k}\dots X_{t-1}$ de taille k précédent X_t, nous voulons que le coefficient $\Pi_{\frac{t}{n}}$ de la matrice de transition soit le plus proche possible de 1 lorsque $X_t = v$ et le plus proche possible de 0 lorsque $X_t \neq v$. L'erreur de prédiction en t se définit donc par

$$\mathbb{1}_{\{X_{t-k}\dots X_{t-1}=u\}}\Pi_{\frac{t}{n}}(u,v) - \mathbb{1}_{\{X_{t-k}\dots X_t=uv\}}.$$

Nous choisissons une distance quadratique, ainsi nous devons minimiser la fonction SP_1 suivante :

$$SP_1\left(\Pi_0,\Pi_1\right) = \sum_{t=k}^{n}\left[\sum_{u\in\mathcal{A}^k}\mathbb{1}\{X_{t-k}\dots X_{t-1}=u\}\left[\sum_{v\in\mathcal{A}}\left(\Pi_{\frac{t}{n}}(u,v) - \mathbb{1}_{\{X_t=v\}}\right)^2\right]\right].$$

Nous notons $\mathbb{1}_u = \mathbb{1}_{\{X_{t-k}\dots X_{t-1}=u\}}$, $\mathbb{1}_v = \mathbb{1}_{\{X_t=v\}}$ et $\mathbb{1}_{uv} = \mathbb{1}_{\{X_{t-k}\dots X_t=uv\}}$.

Déterminons la dérivée de SP_1.

$$\begin{aligned}\frac{\partial SP_1}{\partial\Pi_0(u,v)}\left(\Pi_0,\Pi_1\right) &= \sum_{t=k}^{n}\mathbb{1}_u 2\left(1-\frac{t}{n}\right)\left(\Pi_0(u,v)\left(1-\frac{t}{n}\right)+\Pi_1(u,v)\frac{t}{n}-\mathbb{1}_v\right)\\ &= P_1(u)\Pi_0(u,v)+P_2(u)\Pi_1(u,v)-P_3(u,v)\end{aligned}$$

avec

$$P_1(u) = 2\sum_{t=k}^{n}\mathbb{1}_u\left(1-\frac{t}{n}\right)^2,\ P_2(u) = 2\sum_{t=k}^{n}\mathbb{1}_u\left(1-\frac{t}{n}\right)\left(\frac{t}{n}\right)\ \text{et}\ P_3(u,v) = 2\sum_{t=k}^{n}\mathbb{1}_{uv}\left(1-\frac{t}{n}\right).$$

$$\begin{aligned}\frac{\partial SP_1}{\partial\Pi_1(u,v)}\left(\Pi_0,\Pi_1\right) &= \sum_{t=k}^{n}\mathbb{1}_u 2\left(\frac{t}{n}\right)\left(\Pi_0(u,v)\left(1-\frac{t}{n}\right)+\Pi_1(u,v)\frac{t}{n}-\mathbb{1}_v\right)\\ &= P_4(u)\Pi_0(u,v)+P_5(u)\Pi_1(u,v)-P_6(u,v)\end{aligned}$$

avec

$$P_4(u) = 2\sum_{t=k}^{n}\mathbb{1}_u\left(1-\frac{t}{n}\right)\left(\frac{t}{n}\right),\ P_5(u) = 2\sum_{t=k}^{n}\mathbb{1}_u\left(\frac{t}{n}\right)^2\ \text{et}\ P_6(u,v) = 2\sum_{t=k}^{n}\mathbb{1}_{uv}\left(\frac{t}{n}\right).$$

Remarque 2 *Nous avons les affirmations suivantes :*
- $P_4(u) = P_2(u)$ *;*
- $P_1(u), P_4(u), P_2(u), P_5(u) \neq 0$ *si u apparaît dans la séquence ;*
- $P_3(u,v), P_6(u,v) \neq 0$ *si uv apparaît dans la séquence.*

Ces coefficients sont presque toujours non nuls pour une séquence assez longue (un exemple : à l'ordre 1*, il suffit que tous les mots de deux lettres apparaissent dans la séquence).*

Déterminons le minimum.

$$\begin{aligned}&\begin{cases} P_1(u)\widehat{\Pi_0}(u,v)+P_2(u)\widehat{\Pi_1}(u,v)-P_3(u,v) &= 0\\ P_4(u)\widehat{\Pi_0}(u,v)+P_5(u)\widehat{\Pi_1}(u,v)-P_6(u,v) &= 0\end{cases}\\ \iff &\begin{cases}\widehat{\Pi_0}(u,v) &= \dfrac{P_5(u)P_3(u,v)-P_2(u)P_6(u,v)}{P_1(u)P_5(u)-P_4(u)P_2(u)}\\ \widehat{\Pi_1}(u,v) &= \dfrac{P_1(u)P_6(u,v)-P_4(u)P_3(u,v)}{P_1(u)P_5(u)-P_4(u)P_2(u)}\end{cases}.\end{aligned}$$

Donc

$$\widehat{\Pi_0}(u,v) = \frac{\left(\begin{array}{c} \left(\sum_{t=k}^{n} \mathbb{1}_u \left(\frac{t}{n}\right)^2\right)\left(\sum_{t=k}^{n} \mathbb{1}_{uv}\left(1-\frac{t}{n}\right)\right) - \\ \left(\sum_{t=k}^{n} \mathbb{1}_u \left(1-\frac{t}{n}\right)\left(\frac{t}{n}\right)\right)\left(\sum_{t=k}^{n} \mathbb{1}_{uv}\left(\frac{t}{n}\right)\right) \end{array} \right)}{\left(\begin{array}{c} \left(\sum_{t=k}^{n} \mathbb{1}_u \left(1-\frac{t}{n}\right)^2\right)\left(\sum_{t=k}^{n} \mathbb{1}_u\left(\frac{t}{n}\right)^2\right) - \\ \left(\sum_{t=k}^{n} \mathbb{1}_u \left(1-\frac{t}{n}\right)\left(\frac{t}{n}\right)\right)\left(\sum_{t=k}^{n} \mathbb{1}_u\left(1-\frac{t}{n}\right)\left(\frac{t}{n}\right)\right) \end{array} \right)},$$

$$\widehat{\Pi_1}(u,v) = \frac{\left(\begin{array}{c} \left(\sum_{t=k}^{n} \mathbb{1}_u \left(1-\frac{t}{n}\right)^2\right)\left(\sum_{t=k}^{n} \mathbb{1}_{uv}\left(\frac{t}{n}\right)\right) - \\ \left(\sum_{t=k}^{n} \mathbb{1}_u \left(1-\frac{t}{n}\right)\left(\frac{t}{n}\right)\right)\left(\sum_{t=k}^{n} \mathbb{1}_{uv}\left(1-\frac{t}{n}\right)\right) \end{array} \right)}{\left(\begin{array}{c} \left(\sum_{t=k}^{n} \mathbb{1}_u \left(1-\frac{t}{n}\right)^2\right)\left(\sum_{t=k}^{n} \mathbb{1}_u\left(\frac{t}{n}\right)^2\right) - \\ \left(\sum_{t=k}^{n} \mathbb{1}_u \left(1-\frac{t}{n}\right)\left(\frac{t}{n}\right)\right)\left(\sum_{t=k}^{n} \mathbb{1}_u\left(1-\frac{t}{n}\right)\left(\frac{t}{n}\right)\right) \end{array} \right)}.$$

Théorème 8 *Les matrices $\widehat{\Pi_0}$ et $\widehat{\Pi_1}$ sont stochastiques (voir A.2.1). Nous réajustons les valeurs négatives le cas échéant (voir A.1.2).*

Espérances et variances de Π_0 et Π_1

Nous montrons à l'aide la delta-méthode que nos estimateurs sont convergents (voir A.2.2).

1.2 Modèles des polynômes : Estimateurs de M_0 et M_1

Rappelons la forme de ce second modèle :

$$\Pi_{\frac{t}{n}}(u,v) = M_0(u,v) + \frac{t}{n} M_1(u,v).$$

Ce n'est que la forme générale d'un polynôme de degré 1 et c'est pourquoi nous parlons de modèle des polynômes. L'avantage de cette écriture sur celle des points d'appui est de ne plus avoir de coefficients "compliqués" devant les matrices à estimer M_0 et M_1. En effet, pour le modèle des points d'appui, les coefficients de Π_0 et Π_1, intuitifs au degré 1, ne le sont plus pour les degrés supérieurs (voir Chapitre 2). Cela simplifie les calculs de changer le modèle dans certains cas, en particulier lors de l'estimation des modèles de Markov régulés par splines (voir Chapitre 3). Les détails de l'estimation des paramètres de ce modèle est donné en Annexe B.

Chapitre 2

Dérive de degré d

Dans le chapitre précédent, nous avons étudié la variation linéaire de la matrice de transition. Nous généralisons à des modèles de Markov régulés par un polynôme de degré d. Ainsi les DMM ont deux paramètres d'ordre : l'ordre k du modèle de Markov et le degré d de la dérive polynomiale. Pour décrire de tels modèles nous avons besoin de $d+1$ matrices de paramètres. Nous généralisons les deux modèles présentés au chapitre précédent (le modèle des points d'appui et le modèle des polynômes) ainsi que les deux méthodes d'estimation (la régression matricielle et la méthode point par point).

2.1 Modèle des points d'appui

Pour une dérive de degré d, nous prenons $d+1$ points d'appui. Ainsi au degré 1 (voir chapitre 1), nous avons pris les deux matrices Π_0 et Π_1 comme points d'appui. Désormais, nous fondons notre modèle sur $d+1$ matrices $\Pi_{\frac{i}{d}}$ pour $0 \leq i \leq d$. Nous choisissons les $\Pi_{\frac{i}{d}}$ uniformément réparties sur la séquence.

Remarque 3 *Tout autre choix de répartition des $d+1$ matrices points d'appui est équivalent. La position de ces matrices dans la séquence n'a aucune influence sur la matrice de transition $\Pi_{\frac{t}{n}}$. Un DMM de degré d s'identifie à l'aide de $d+1$ matrices de la même façon qu'en dimension 1, un polynôme de degré d s'identifie à l'aide de $d+1$ points. Ce choix d'une répartition uniforme n'a pour but qu'une meilleure visualisation de la séquence.*

La matrice de transition dérivante a la forme suivante :

$$\begin{aligned}\Pi_{\frac{t}{n}}(u,v) &= A_0(t)\Pi_0(u,v) + A_1(t)\Pi_{\frac{1}{d}}(u,v) + \ldots + A_{d-1}(t)\Pi_{\frac{d-1}{d}}(u,v) + A_d(t)\Pi_1(u,v)\\ \Pi_{\frac{t}{n}}(u,v) &= \sum_{i=0}^{d} A_i(t)\Pi_{\frac{i}{d}}(u,v)\end{aligned}$$

avec A_i des polynômes de degré d :

$$A_i(t) = a_i^0 + a_i^1\frac{t}{n} + a_i^2\frac{t^2}{n^2} + \ldots + a_i^d\frac{t^d}{n^d} = \sum_{j=0}^{d} a_i^j\frac{t^j}{n^j}.$$

Les polynômes A_i sont tels que :

$$\forall (i,j) \in [\![0,d]\!]^2,\ A_i\left(\frac{nj}{d}\right) = \mathbb{1}_{\{i=j\}}.$$

Remarque 4 *Pour $t = ni/d$, nous avons $\Pi_{\frac{t}{n}} = \Pi_{\frac{i}{d}}$.*

Remarque 5 *Les polynômes A_i sont choisis dans le but d'avoir des matrices $\Pi_{\frac{t}{n}}$ stochastiques pour tout entier $0 \leq t \leq n$. En effet,*

$$\sum_{v\in\mathcal{A}} \Pi_{\frac{t}{n}}(u,v) = \sum_{i=0}^{d} A_i(t) = A(t)$$

où $A(t)$ est un polynôme de degré d valant 1 en $d+1$ points, donc constant et égal à 1.

2.1.1 Calcul des A_i

Pour $d = 1$, nous avons obtenu intuitivement $A_0(t) = 1 - t/n$ et $A_1(t) = t/n$ de manière à obtenir $\Pi_{\frac{t}{n}} = (1 - t/n)\Pi_0 + t/n\Pi_1$. Nous explicitons le calcul des A_i pour un degré $d = 2$ afin de montrer que ces polynômes n'ont pas une expression aussi simple que dans le cas d'une dérive linéraire.

Notons Π_0, $\Pi_{\frac{1}{2}}$ et Π_1, nos matrices points d'appui. Elles réunissent les probabilités de transition respectivement du début, du milieu et de la fin de la séquence. La matrice dérivante $\Pi_{\frac{t}{n}}$ est donc de la forme :

$$\Pi_{\frac{t}{n}}(u,v) = A_0(t)\Pi_0(u,v) + A_1(t)\Pi_{\frac{1}{2}}(u,v) + A_2(t)\Pi_1(u,v)$$

avec A_0, A_1 et A_2 des polynômes de degré 2 :

$$A_0(t) = a_0^2\frac{t^2}{n^2} + a_0^1\frac{t}{n} + a_0^0, A_1(t) = a_1^2\frac{t^2}{n^2} + a_1^1\frac{t}{n} + a_1^0 \text{ et } A_2(t) = a_2^2\frac{t^2}{n^2} + a_2^1\frac{t}{n} + a_2^0.$$

Nous voulons $\Pi_{\frac{t}{n}}$ stochastique et
- pour $t = 0$, $\Pi_{\frac{t}{n}} = \Pi_0$;
- pour $t = n/2$, $\Pi_{\frac{t}{n}} = \Pi_{\frac{1}{2}}$;
- pour $t = n$, $\Pi_{\frac{t}{n}} = \Pi_1$.

Nous en déduisons les A_i :

$$\left\{\begin{array}{lcl} A_0(0) & = & 1 \\ A_0\left(\frac{n}{2}\right) & = & 0 \\ A_0(n) & = & 0 \end{array}\right. \iff \left\{\begin{array}{lcl} a_0^0 & = & 1 \\ \frac{a_0^2}{4} + \frac{a_0^1}{2} + a_0^0 & = & 0 \\ a_0^2 + a_0^1 + a_0^0 & = & 0 \end{array}\right. \iff \left\{\begin{array}{lcl} a_0^2 & = & 2 \\ a_0^1 & = & -3 \\ a_0^0 & = & 1 \end{array}\right.$$

$$\left\{\begin{array}{lcl} A_1(0) & = & 0 \\ A_1\left(\frac{n}{2}\right) & = & 1 \\ A_1(n) & = & 0 \end{array}\right. \iff \left\{\begin{array}{lcl} a_1^0 & = & 0 \\ \frac{a_1^2}{4} + \frac{a_1^1}{2} & = & 1 \\ a_1^2 + a_1^1 & = & 0 \end{array}\right. \iff \left\{\begin{array}{lcl} a_1^2 & = & -4 \\ a_1^1 & = & 4 \\ a_1^0 & = & 0 \end{array}\right.$$

$$\left\{\begin{array}{lcl} A_2(0) & = & 0 \\ A_2\left(\frac{n}{2}\right) & = & 0 \\ A_2(n) & = & 1 \end{array}\right. \iff \left\{\begin{array}{lcl} a_2^0 & = & 0 \\ \frac{a_2^2}{4} + \frac{a_2^1}{2} & = & 0 \\ a_2^2 + a_2^1 & = & 1 \end{array}\right. \iff \left\{\begin{array}{lcl} a_1^2 & = & 2 \\ a_1^1 & = & -1 \\ a_1^0 & = & 0 \end{array}\right.$$

Ainsi, nous obtenons $\Pi_{\frac{t}{n}}$ de la forme suivante :

$$\Pi_{\frac{t}{n}}(u,v) = \left(2\frac{t^2}{n^2} - 3\frac{t}{n} + 1\right)\Pi_0(u,v) + \left(-4\frac{t^2}{n^2} + 4\frac{t}{n}\right)\Pi_{\frac{1}{2}}(u,v) + \left(2\frac{t^2}{n^2} - \frac{t}{n}\right)\Pi_1(u,v).$$

Notons qu'un tel système est facile à résoudre pour n'importe quel degré car c'est un simple système linéaire de $(d+1)(d+1)$ équations indépendantes avec $(d+1)(d+1)$ variables. Néanmoins, nous ne pouvons pas donner une expression générale explicite des A_i pour un degré d quelconque.

Exemple 5 *Au degré $d = 3$, nous avons :*

$$\begin{array}{rcl} \Pi_{\frac{t}{n}} & = & \left(-\frac{9}{2}\frac{t^3}{n^3} + 9\frac{t^2}{n^2} - \frac{11}{2}\frac{t}{n} + 1\right)\Pi_0 + \left(\frac{27}{2}\frac{t^3}{n^3} - \frac{45}{2}\frac{t^2}{n^2} + 9\frac{t}{n}\right)\Pi_{\frac{1}{3}} + \left(-\frac{27}{2}\frac{t^3}{n^3} + 18\frac{t^2}{n^2} - \frac{9}{2}\frac{t}{n}\right)\Pi_{\frac{2}{3}} \\ & + & \left(\frac{9}{2}\frac{t^3}{n^3} - \frac{9}{2}\frac{t^2}{n^2} + \frac{t}{n}\right)\Pi_1. \end{array}$$

2.1.2 Régression matricielle

Distances

L'estimation les $d+1$ matrices $\Pi_{\frac{i}{d}}$ peut se faire par régression matricielle en minimisant la fonction SR_d suivante :

$$SR_d(\Pi_0, \ldots, \Pi_1) = \sum_{\ell \in [\![1,N]\!]} \sum_{u \in \mathcal{A}^k} \sum_{v \in \mathcal{A}} \left(\widehat{\Pi_{S_\ell}}(u,v) - \sum_{i=0}^{d} A_i(n\tau_\ell)\Pi_{\frac{i}{d}}(u,v)\right)^2.$$

La recherche des zéros de la dérivée nous donne le système suivant pour chaque couple $(u,v) \in \mathcal{A}^k \times \mathcal{A}$:

$$\left\{\begin{array}{rcl} \dfrac{\partial SR_d(\Pi_0,\ldots,\Pi_1)}{\partial \Pi_0(u,v)} & = & 0 \\ & \vdots & \\ \dfrac{\partial SR_d(\Pi_0,\ldots,\Pi_1)}{\partial \Pi_{\frac{j}{d}}(u,v)} & = & 0 \\ & \vdots & \\ \dfrac{\partial SR_d(\Pi_0,\ldots,\Pi_1)}{\partial \Pi_1(u,v)} & = & 0 \end{array}\right.$$

$$\iff \left\{\begin{array}{rcl} \displaystyle\sum_{\ell=1}^{N} A_0(n\tau_\ell)\left(\widehat{\Pi_{S_\ell}}(u,v) - \sum_{i=0}^{d} A_i(n\tau_\ell)\widehat{\Pi_{\frac{i}{d}}}(u,v)\right) & = & 0 \\ & \vdots & \\ \displaystyle\sum_{\ell=1}^{N} A_j(n\tau_\ell)\left(\widehat{\Pi_{S_\ell}}(u,v) - \sum_{i=0}^{d} A_i(n\tau_\ell)\widehat{\Pi_{\frac{i}{d}}}(u,v)\right) & = & 0 \\ & \vdots & \\ \displaystyle\sum_{\ell=1}^{N} A_d(n\tau_\ell)\left(\widehat{\Pi_{S_\ell}}(u,v) - \sum_{i=0}^{d} A_i(n\tau_\ell)\widehat{\Pi_{\frac{i}{d}}}(u,v)\right) & = & 0 \end{array}\right.$$

$$\iff \left\{\begin{array}{rcl} \displaystyle\sum_{i=0}^{d} \widehat{\Pi_{\frac{i}{d}}}(u,v)\left(\sum_{\ell=1}^{N} A_0(n\tau_\ell)A_i(n\tau_\ell)\right) - \sum_{\ell=1}^{N} A_0(n\tau_\ell)\widehat{\Pi_{S_\ell}}(u,v) & = & 0 \\ & \vdots & \\ \displaystyle\sum_{i=0}^{d} \widehat{\Pi_{\frac{i}{d}}}(u,v)\left(\sum_{\ell=1}^{N} A_j(n\tau_\ell)A_i(n\tau_\ell)\right) - \sum_{\ell=1}^{N} A_j(n\tau_\ell)\widehat{\Pi_{S_\ell}}(u,v) & = & 0 \\ & \vdots & \\ \displaystyle\sum_{i=0}^{d} \widehat{\Pi_{\frac{i}{d}}}(u,v)\left(\sum_{\ell=1}^{N} A_d(n\tau_\ell)A_i(n\tau_\ell)\right) - \sum_{\ell=1}^{N} A_d(n\tau_\ell)\widehat{\Pi_{S_\ell}}(u,v) & = & 0 \end{array}\right.$$

$$\iff \left\{\begin{array}{rcl} \left(\displaystyle\sum_{\ell=1}^{N} A_0(n\tau_\ell)A_0(n\tau_\ell)\right)\widehat{\Pi_0}(u,v) + \ldots + \left(\displaystyle\sum_{\ell=1}^{N} A_0(n\tau_\ell)A_d(n\tau_\ell)\right)\widehat{\Pi_1}(u,v) - \displaystyle\sum_{\ell=1}^{N} A_0(n\tau_\ell)\widehat{\Pi_{S_\ell}}(u,v) & = & 0 \\ & \vdots & \\ \left(\displaystyle\sum_{\ell=1}^{N} A_j(n\tau_\ell)A_0(n\tau_\ell)\right)\widehat{\Pi_0}(u,v) + \ldots + \left(\displaystyle\sum_{\ell=1}^{N} A_j(n\tau_\ell)A_d(n\tau_\ell)\right)\widehat{\Pi_1}(u,v) - \displaystyle\sum_{\ell=1}^{N} A_j(n\tau_\ell)\widehat{\Pi_{S_\ell}}(u,v) & = & 0 \\ & \vdots & \\ \left(\displaystyle\sum_{\ell=1}^{N} A_d(n\tau_\ell)A_0(n\tau_\ell)\right)\widehat{\Pi_0}(u,v) + \ldots + \left(\displaystyle\sum_{\ell=1}^{N} A_d(n\tau_\ell)A_d(n\tau_\ell)\right)\widehat{\Pi_1}(u,v) - \displaystyle\sum_{\ell=1}^{N} A_d(n\tau_\ell)\widehat{\Pi_{S_\ell}}(u,v) & = & 0 \end{array}\right.$$

Pour un alphabet $\mathcal{A}$ donné, nous obtenons ainsi $(|\mathcal{A}|)^{k+1}$ systèmes de $d+1$ équations à $d+1$ inconnues. Nous avons alors à m fixé, les estimateurs $\widehat{\Pi_{\frac{i}{d}}}$.

Nous réécrivons le système sous la forme matricielle suivante :

$$\begin{pmatrix} \displaystyle\sum_{\ell=1}^{N} A_0(n\tau_\ell)A_0(n\tau_\ell) & \cdots & \displaystyle\sum_{\ell=1}^{N} A_0(n\tau_\ell)A_d(n\tau_\ell) \\ \vdots & & \vdots \\ \displaystyle\sum_{\ell=1}^{N} A_d(n\tau_\ell)A_0(n\tau_\ell) & \cdots & \displaystyle\sum_{\ell=1}^{N} A_d(n\tau_\ell)A_d(n\tau_\ell) \end{pmatrix} \begin{pmatrix} \widehat{\Pi_0}(u,v) \\ \vdots \\ \widehat{\Pi_1}(u,v) \end{pmatrix} = \begin{pmatrix} \displaystyle\sum_{\ell=1}^{N} A_0(n\tau_\ell)\widehat{\Pi_{S_\ell}}(u,v) \\ \vdots \\ \displaystyle\sum_{\ell=1}^{N} A_d(n\tau_\ell)\widehat{\Pi_{S_\ell}}(u,v) \end{pmatrix}$$

Nous notons ce système $MX = P$ avec

$$M_{ij} = \sum_{\ell=1}^{N} A_i(n\tau_\ell)A_j(n\tau_\ell), X_i = \widehat{\Pi_{\frac{i}{d}}}(u,v) \text{ et } P_i = \sum_{\ell=1}^{N} A_i(n\tau_\ell)\widehat{\Pi_{S_\ell}}(u,v).$$

Théorème 9 *Les matrices $\widehat{\Pi_{\frac{i}{d}}}$ sont stochastiques (voir C.1.1).*

Les espérances et variances de $\widehat{\Pi}_{\frac{i}{d}}$

•*Espérance*
Nous avons $\mathbb{E}(X) = M^{-1}\mathbb{E}(P)$. Ainsi

$$\mathbb{E}(P) = \begin{pmatrix} \sum_{\ell=1}^{N} A_0(n\tau_\ell)\mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right) \\ \vdots \\ \sum_{\ell=1}^{N} A_d(n\tau_\ell)\mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right) \end{pmatrix}.$$

Ayant calculé $\mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)$ (voir C.1.3), nous déduisons $\mathbb{E}(X)$.

•*Variance*
Nous avons $\mathbb{V}(X) = M^{-1}\mathbb{V}(P)(M^t)^{-1}$. Le calcul de $\mathbb{V}(P)$ nécessite celui des deux termes suivants :

- $\mathbb{V}\left(\sum_{\ell=1}^{N} A_i(n\tau_\ell)\widehat{\Pi_{S_\ell}}(u,v)\right) = \sum_{\ell=1}^{N} A_i(n\tau_\ell)^2\mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)$
- $\mathbb{C}\text{ov}\left(\sum_{\ell=1}^{N} A_i(n\tau_\ell)\widehat{\Pi_{S_\ell}}(u,v), \sum_{\ell=1}^{N} A_j(n\tau_\ell)\widehat{\Pi_{S_\ell}}(u,v)\right) = \sum_{\ell=1}^{N} A_i(n\tau_\ell)A_j(n\tau_\ell)\mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)$

Nous nous servons de l'indépendance des segments pour la seconde égalité. Ayant calculé $\mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)$, nous déduisons $\mathbb{V}(X)$.

2.1.3 Méthode point par point

Rappelons le modèle des points d'appui :

$$\Pi_{\frac{t}{n}}(u,v) = \sum_{i=0}^{d} A_i(t)\Pi_{\frac{i}{d}}(u,v).$$

De la même façon que précédemment (voir 1.1.3), nous devons minimiser la fonction SP_d suivante :

$$SP_d(\Pi_0,\ldots,\Pi_1) = \sum_{t=k}^{n}\left[\sum_{u\in\mathcal{A}^k}\mathbb{1}\{X_{t-k}\ldots X_{t-1}=u\}\left[\sum_{v\in\mathcal{A}}\left(\Pi_{\frac{t}{n}}(u,v) - \mathbb{1}_{\{X_t=v\}}\right)^2\right]\right] \tag{2.1}$$

La recherche des zéros de la dérivée nous donne le système suivant pour chaque couple $(u,v)\in\mathcal{A}^k\times\mathcal{A}$:

$$\left\{\begin{array}{rcl} \frac{\partial SP_d(\Pi_0,\ldots,\Pi_1)}{\partial\Pi_0(u,v)} &=& 0 \\ &\vdots& \\ \frac{\partial SP_d(\Pi_0,\ldots,\Pi_1)}{\partial\Pi_{\frac{i}{d}}(u,v)} &=& 0 \\ &\vdots& \\ \frac{\partial SP_d(\Pi_0,\ldots,\Pi_1)}{\partial\Pi_1(u,v)} &=& 0 \end{array}\right.$$

$$\Longleftrightarrow \left\{\begin{array}{rcl} \sum_{t=k}^{n}\mathbb{1}_{\{X_{t-k}\ldots X_{t-1}=u\}}A_0(t)\left(\sum_{i=0}^{d}A_i(t)\widehat{\Pi_{\frac{i}{d}}}(u,v) - \mathbb{1}_{\{X_t=v\}}\right) &=& 0 \\ &\vdots& \\ \sum_{t=k}^{n}\mathbb{1}_{\{X_{t-k}\ldots X_{t-1}=u\}}A_j(t)\left(\sum_{i=0}^{d}A_i(t)\widehat{\Pi_{\frac{i}{d}}}(u,v) - \mathbb{1}_{\{X_t=v\}}\right) &=& 0 \\ &\vdots& \\ \sum_{t=k}^{n}\mathbb{1}_{\{X_{t-k}\ldots X_{t-1}=u\}}A_d(t)\left(\sum_{i=0}^{d}A_i(t)\widehat{\Pi_{\frac{i}{d}}}(u,v) - \mathbb{1}_{\{X_t=v\}}\right) &=& 0 \end{array}\right.$$

$$\iff \begin{cases} \sum_{i=0}^{d} \widehat{\Pi_{\frac{i}{d}}}(u,v) \left(\sum_{t=k}^{n} \mathbb{1}_{\{X_{t-k}\dots X_{t-1}=u\}} A_0(t)A_i(t) \right) - \sum_{t=k}^{n} A_0(t)\mathbb{1}_{\{X_{t-k}\dots X_{t-1}=u, X_t=v\}} &= 0 \\ \vdots & \\ \sum_{i=0}^{d} \widehat{\Pi_{\frac{i}{d}}}(u,v) \left(\sum_{t=k}^{n} \mathbb{1}_{\{X_{t-k}\dots X_{t-1}=u\}} A_j(t)A_i(t) \right) - \sum_{t=k}^{n} A_j(t)\mathbb{1}_{\{X_{t-k}\dots X_{t-1}=u, X_t=v\}} &= 0 \\ \vdots & \\ \sum_{i=0}^{d} \widehat{\Pi_{\frac{i}{d}}}(u,v) \left(\sum_{t=k}^{n} \mathbb{1}_{\{X_{t-k}\dots X_{t-1}=u\}} A_d(t)A_i(t) \right) - \sum_{t=k}^{n} A_d(t)\mathbb{1}_{\{X_{t-k}\dots X_{t-1}=u, X_t=v\}} &= 0 \end{cases}$$

Pour un alphabet $\mathcal{A}$ donné, nous obtenons ainsi $(|\mathcal{A}|)^{k+1}$ systèmes de $d+1$ équations à $d+1$ inconnues. Nous avons alors les estimateurs $\widehat{\Pi_{\frac{i}{d}}}$.

En notant $\mathbb{1}_u = \mathbb{1}_{\{X_{t-k}\dots X_{t-1}=u\}}$ et $\mathbb{1}_{uv} = \mathbb{1}_{\{X_{t-k}\dots X_{t-1}=u, X_t=v\}}$, nous réécrivons le système sous la forme matricielle suivante :

$$\begin{pmatrix} \sum_{t=k}^{n} \mathbb{1}_u A_0(t)A_0(t) & \cdots & \sum_{t=k}^{n} \mathbb{1}_u A_0(t)A_d(t) \\ \vdots & & \vdots \\ \sum_{t=k}^{n} \mathbb{1}_u A_d(t)A_0(t) & \cdots & \sum_{t=k}^{n} \mathbb{1}_u A_d(t)A_d(t) \end{pmatrix} \begin{pmatrix} \widehat{\Pi_0}(u,v) \\ \vdots \\ \widehat{\Pi_1}(u,v) \end{pmatrix} = \begin{pmatrix} \sum_{t=k}^{n} A_0(t)\mathbb{1}_{uv} \\ \vdots \\ \sum_{t=k}^{n} A_d(t)\mathbb{1}_{uv} \end{pmatrix} \tag{2.2}$$

Nous notons ce système $MX = P$ avec

$$M_{ij} = \sum_{t=k}^{n} \mathbb{1}_u A_i(t)A_j(t), X_i = \widehat{\Pi_{\frac{i}{d}}}(u,v) \text{ et } P_i = \sum_{t=k}^{n} A_i(t)\mathbb{1}_{uv}$$

Théorème 10 *Les matrices $\widehat{\Pi_{\frac{i}{d}}}$ sont stochastiques (voir C.2.1).*

Les espérances et variances de $\widehat{\Pi}_{\frac{i}{d}}$

Nous montrons à l'aide la delta-méthode que nos estimateurs sont convergents (voir C.2.2).

2.2 Modèle des polynômes

Donnons la forme de ce modèle pour un DMM de degré d :

$$\Pi_{\frac{t}{n}}(u,v) = M_0(u,v) + M_1(u,v)\frac{t}{n} + M_2(u,v)\frac{t^2}{n^2} + \dots + M_d(u,v)\frac{t^d}{n^d} = \sum_{i=0}^{d} M_i(u,v)\frac{t^i}{n^i}$$

L'exemple 5 p36 nous montre que les coefficients des $\Pi_{\frac{i}{d}}$, intuitifs au degré 1, ne le sont plus pour les degrés supérieurs. Ainsi ce modèle des polynômes (d'apparence plus simple) sera particulièrement utile lors du chapitre suivant : l'estimation par splines. Les détails de l'estimation des paramètres de ce modèle est donné en Annexe D.

Chapitre 3

Splines

3.1 Introduction

Développées par les numériciens, les fonctions splines forment un outil très utilisé pour les problèmes d'interpolation ou de régression polynomiale. Un des principaux problèmes de l'interpolation polynomiale est la grande variabilité imposée par la contrainte de passer par tous les points, sans aucune souplesse dans le tracé. La régression polynomiale, bien que passant entre les points, n'échappe pas à ce manque de souplesse des polynômes. Les fonctions splines tentent de remédier à ce problème en proposant une interpolation ou une régression non pas globale mais par morceaux. L'ajustement d'une chaîne de Markov régulée de façon polynomiale présentant le même désavantage que la régression polynomiale (à savoir la variabilité des polynômes), nous proposons d'adapter la régression par splines, plus flexible, afin de construire notre DMM. Il s'agit donc développer une nouvelle méthode d'estimation de nos modèles : **les DMM par splines polynomiales**.

3.1.1 Les splines

Une abondante littérature existe sur les splines, la plupart des articles concernant leurs propriétés numériques et analytiques plutôt que leurs propriétés statistiques. Par exemple, nous pouvons citer les livres suivants : [dB78], [Eub88], [Sch07]. L'usage des fonctions splines a été popularisé en statistique par G. Wahba ([Wah90]) dans les années 70. Les splines polynomiales sont sans doute les plus utilisées. Construites à partir de polynômes, elles possèdent de bien meilleures propriétés d'approximation. En particulier la sensibilité d'une approximation spline à un bruitage localisé est bien moindre que celle d'une approximation polynomiale. Les fonctions splines ont dernièrement trouvé une grande utilité comme fonctions d'approximations en mathématique et analyse numérique. Ce sont les fonctions qui ont le plus de succès parmi les fonctions d'approximation. Les fonctions qui expriment des relations physiques sont fréquemment disjointes ou de nature désunies, c'est à dire que leur comportement dans une région peut n'être sans aucun lien avec leur comportement dans une autre région. Les polynômes, de même que la plupart des autres fonctions mathématiques, ont exactement le comportement inverse. Leur comportement dans une seule petite région a des conséquences sur leur comportement entier. Les splines font face à ce problème puisqu'elles sont définies par morceaux. Leurs bonnes propriétés font des fonctions splines un excellent outil pour l'ajustement de courbes dans l'analyse de données. Pour quelques applications de l'approximation par splines, nous pouvons nous référer à [MDD01, MMCD02].

Les fonctions splines sont définies comme des polynômes par morceaux, de degré d et réalisent, en gros, une régression polynomiale. En général, le degré est 3 (et on parle de splines cubiques), ou 4, mais rien ne s'oppose à l'utilisation des degrés 2 ou 12. Le choix de polynômes de degré 3 permet d'obtenir une grande flexibilité en évitant les oscillations. Les morceaux se relient en ce qu'on appelle les *noeuds* en suivant des conditions de continuité pour la fonction elle-même et ses $d-1$ fonctions dérivées. Ainsi, une fonction spline de degré d est une fonction continue avec $d-1$ dérivées continues. Le problème du nombre de noeuds ainsi que de leur position est un problème encore ouvert. Il est très difficile d'estimer à la fois le nombre optimal de noeud ainsi que leur position optimale. Même empiriquement, les algorithmes existants sont non triviaux que ce soit pour trouver la position optimale à partir d'un nombre de noeuds donnés ou pour trouver le nombre optimal de noeuds sachant leur répartition. Une première stratégie est de choisir un petit nombre de noeuds puis d'estimer la fonction par moindres carrés ([SS95, MNSS89, Die95, Jup78, FS89, Pit02]). Le choix des noeuds est alors important. Une seconde stratégie est d'utiliser un nombre relativement large nombre de noeuds ([DMS98, DGK01, EM05, Lin99, Rup02]). L'importance des noeuds est moindre dans ce cas, l'essentiel étant la manière d'estimer la fonction. [Wan00] et [Lee02] proposent une comparaison des différents algorithmes d'estimation de fonctions splines.

3.1.2 Les splines et les chaînes de Markov régulées

L'estimation de nos DMM par splines polynomiales nécessite l'adaptation à l'espace des matrices des méthodes de splines. L'utilisation de splines de matrices est inédite dans la littérature excepté pour la résolution numérique d'équations différentielles ([DSHS05]). Mais l'estimation d'une matrice de transition d'un modèle de Markov par des splines polynomiales de matrices est entièrement nouvelle. Présentons-en le principe ainsi que quelques notations que nous utiliserons.

Il s'agit de découper la séquence en plusieurs morceaux et d'ajuster sur chacun des morceaux un polynôme de matrices tout en respectant des conditions de continuité entre les morceaux. Ainsi, pour une estimation par splines polynomiales de degré d, nous aurons sur chaque morceau i de la séquence, une matrice $\Pi^i_{\frac{t}{n}}$ définie comme matrice de transition d'un DMM de degré d. Le nombre de contraintes augmente avec le degré de la fonction spline. Nous allons considérer trois manières de voir les choses :

- Estimation naïve des estimateurs des matrices de tous les morceaux simultanément à l'aide d'un unique système linéaire (voir 3.2). Après avoir présenté le principe au degré 1 pour nos modèles sous forme de points d'appuis (voir 2.1), nous considérerons nos modèles des polynômes (voir 2.2) aux degrés supérieurs pour plus de simplicité dans les calculs. Rappelons que ces modèles sont équivalents (voir Annexe B).
- Estimation par morceaux, en combinant des polynômes de bases sur chaque morceau (voir 3.3.2). Cela revient à résoudre plusieurs systèmes linéaires bien plus petits que le précédent.
- Estimation par morceaux, sans polynômes de base (voir 3.3.1).

Dans les deux derniers cas, nous ne traitons que des splines cubiques, les plus courantes, contrairement au premier cas, plus général. Ces deux dernières méthodes font aussi l'objet d'une extension qui permet d'estimer nos matrices de transition récursivement, par allers et retours sur la séquence.

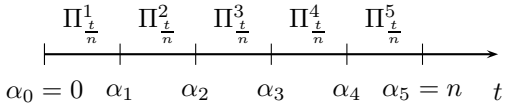
Remarque 6 *Nous ne présentons ici que l'approche théorique de ces trois méthodes. La comparaison des méthodes ainsi que les résultats obtenus seront présentés en Partie II.*

Nos programmes sont développés de sorte de prendre en compte n'importe quel découpage de la séquence. Sachant que l'optimisation du nombre de noeuds est un problème difficile, nous avons choisi de laisser ce choix en paramètre. Concernant la localisation des noeuds, étant donné le grand nombre de données, un découpage uniforme ne présente pas de désavantage particulier.

Notons N le nombre de morceaux et $\{\alpha_0; \ldots; \alpha_N\}$ le découpage (voir Figure 3.1 pour un exemple). Comme auparavant notons aussi $\mathbb{1}_u = \mathbb{1}_{\{X_{t-k}\ldots X_{t-1}=u\}}$, $\mathbb{1}_v = \mathbb{1}_{\{X_t=v\}}$ et $\mathbb{1}_{uv} = \mathbb{1}_{\{X_{t-k}\ldots X_t=uv\}}$. Tout au long de ce chapitre, nous nous efforçons d'estimer les N matrices de transition dérivantes $\Pi^i_{\frac{t}{n}}$ satisfaisant aux conditions de continuité au noeuds. La matrice de transition $\Pi_{\frac{t}{n}}$ est donc définie de la façon suivante :

$$\Pi_{\frac{t}{n}} = \Pi^i_{\frac{t}{n}} \text{ pour } t \in [\alpha_{i-1}, \alpha_i].$$

FIG. 3.1 – Exemple de découpage d'une séquence en 5 morceaux



3.2 Estimation globale

3.2.1 degré 0

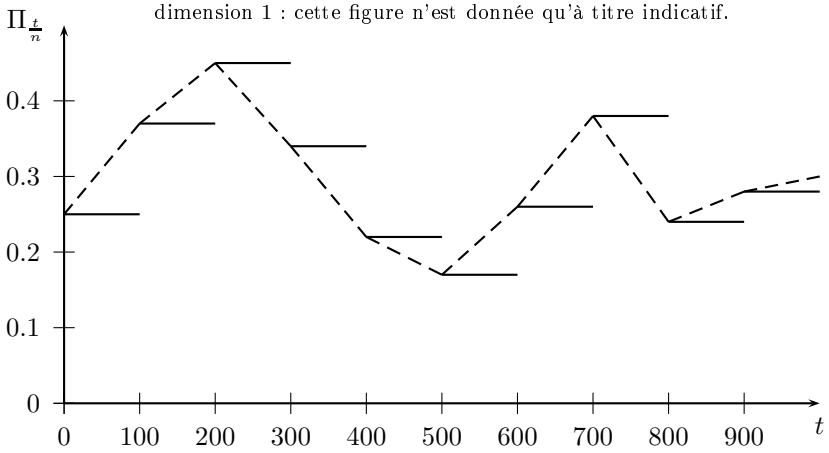
Dans le cadre de la modélisation par chaînes de Markov régulées, le degré 0 n'apporte aucun intérêt nouveau. En effet, une estimation par splines polynomiales de degré 0 revient tout simplement à estimer un modèle de Markov classique (voir 0.2) sur chaque segment (voir Figure 3.2).

Optimiser le nombre et la position des noeuds pourrait être une alternative à la modélisation par chaînes de Markov cachées (voir 0.3).

3.2.2 degré 1, N morceaux

La modélisation par splines polynomiales de degré 1 revient à estimer un DMM de degré 1 sur chaque morceau en imposant la continuité aux noeuds (voir Figure 3.2). Pour chaque segment, nous utilisons deux matrices points d'appui (voir 1.1), une au début et une à la fin du segment. La continuité entre deux segments impose l'égalité de

FIG. 3.2 – Exemple de splines polynomiales de degré 0 et 1 (en pointillés)
Nous considérons une séquence de longueur 1000 découpée en 10 morceaux. L'axe des ordonnées "représente" l'espace des matrices en dimension 1 : cette figure n'est donnée qu'à titre indicatif.



la matrice initiale d'un segment et de la matrice finale du segment précédent. Ainsi sur chaque segment i, pour i allant de 1 à N, une matrice de transition dérivante $\Pi^i_{\frac{t}{n}}$ est définie :

$$\begin{aligned}
\Pi^1_{\frac{t}{n}}(u,v) &= \frac{\alpha_1 - t}{\alpha_1}\Pi_0(u,v) + \frac{t}{\alpha_1}\Pi_{\frac{\alpha_1}{n}}(u,v) \\
&\vdots \\
\Pi^i_{\frac{t}{n}}(u,v) &= \frac{\alpha_i - t}{\alpha_i - \alpha_{i-1}}\Pi_{\frac{\alpha_{i-1}}{n}}(u,v) + \frac{t - \alpha_{i-1}}{\alpha_i - \alpha_{i-1}}\Pi_{\frac{\alpha_i}{n}}(u,v) \\
&\vdots \\
\Pi^N_{\frac{t}{n}}(u,v) &= \frac{n - t}{n - \alpha_{N-1}}\Pi_{\frac{\alpha_{N-1}}{n}}(u,v) + \frac{t - \alpha_{N-1}}{n - \alpha_{N-1}}\Pi_1(u,v)
\end{aligned}$$

Exemple 6 *Nous présentons le modèle pour une séquence découpée en* 2 *segments :*

$$\begin{aligned}
\Pi^1_{\frac{t}{n}}(u,v) &= \frac{\alpha - t}{\alpha}\Pi_0(u,v) + \frac{t}{\alpha}\Pi_{\frac{\alpha}{n}}(u,v) \\
\Pi^2_{\frac{t}{n}}(u,v) &= \frac{n - t}{n - \alpha}\Pi_{\frac{\alpha}{n}}(u,v) + \frac{t - \alpha}{n - \alpha}\Pi_1(u,v).
\end{aligned}$$

Les matrices Π_0, $\Pi_{\frac{\alpha}{n}}$ *et* Π_1 *sont à estimer (voir E.1.1).*

Exemple 7 *Nous présentons le modèle pour une séquence découpée en* 3 *segments :*

$$\begin{aligned}
\Pi^1_{\frac{t}{n}}(u,v) &= \frac{\alpha_1 - t}{\alpha_1}\Pi_0(u,v) + \frac{t}{\alpha_1}\Pi_{\frac{\alpha_1}{n}}(u,v) \\
\Pi^2_{\frac{t}{n}}(u,v) &= \frac{\alpha_2 - t}{\alpha_2 - \alpha_1}\Pi_{\frac{\alpha_1}{n}}(u,v) + \frac{t - \alpha_1}{\alpha_2 - \alpha_1}\Pi_{\frac{\alpha_2}{n}}(u,v) \\
\Pi^3_{\frac{t}{n}}(u,v) &= \frac{n - t}{n - \alpha_2}\Pi_{\frac{\alpha_2}{n}}(u,v) + \frac{t - \alpha_2}{n - \alpha_2}\Pi_1(u,v).
\end{aligned}$$

Les matrices Π_0, $\Pi_{\frac{\alpha_1}{n}}$, $\Pi_{\frac{\alpha_2}{n}}$ *et* Π_1 *sont à estimer (voir E.1.1).*

Comme précédemment (voir 1.1.3), nous utilisons la méthode point par point pour estimer les paramètres du modèle. Nous cherchons toutes les matrices de paramètres simultanément ainsi nous minimisons la fonction f_1 suivante :

$$\begin{aligned}
& f_1(\Pi_0, \ldots, \Pi_{\frac{\alpha_i}{n}}, \ldots, \Pi_1) \\
= & \sum_{t=k}^{n} \left(\sum_{u \in \mathcal{A}^k} \mathbb{1}_u \left(\sum_{v \in \mathcal{A}} \left(\Pi_{\frac{t}{n}}(u,v) - \mathbb{1}_v \right)^2 \right) \right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{t=k}^{\alpha_1}\left(\sum_{u\in\mathcal{A}^k}\mathbb{1}_u\left(\sum_{v\in\mathcal{A}}\left(\frac{\alpha_1-t}{\alpha_1}\Pi_0(u,v)+\frac{t}{\alpha_1}\Pi_{\frac{\alpha_1}{n}}(u,v)-\mathbb{1}_v\right)^2\right)\right)\\
&+ \dots\\
&+ \sum_{t=\alpha_{i-1}+1}^{\alpha_i}\left(\sum_{u\in\mathcal{A}^k}\mathbb{1}_u\left(\sum_{v\in\mathcal{A}}\left(\frac{\alpha_i-t}{\alpha_i-\alpha_{i-1}}\Pi_{\frac{\alpha_{i-1}}{n}}(u,v)+\frac{t-\alpha_{i-1}}{\alpha_i-\alpha_{i-1}}\Pi_{\frac{\alpha_i}{n}}(u,v)-\mathbb{1}_v\right)^2\right)\right)\\
&+ \dots\\
&+ \sum_{t=\alpha_{N-1}+1}^{n}\left(\sum_{u\in\mathcal{A}^k}\mathbb{1}_u\left(\sum_{v\in\mathcal{A}}\left(\frac{n-t}{n-\alpha_{N-1}}\Pi_{\frac{\alpha_{N-1}}{n}}(u,v)+\frac{t-\alpha_{N-1}}{n-\alpha_{N-1}}\Pi_1(u,v)-\mathbb{1}_v\right)^2\right)\right).
\end{aligned}
$$

Explicitons les dérivées :

$$
\begin{aligned}
\frac{\partial f_1}{\partial\Pi_0(u,v)} &= \sum_{t=k}^{\alpha_1}\left(\mathbb{1}_u 2\left(\frac{\alpha_1-t}{\alpha_1}\right)\left(\frac{\alpha_1-t}{\alpha_1}\Pi_0(u,v)+\frac{t}{\alpha_1}\Pi_{\frac{\alpha_1}{n}}(u,v)-\mathbb{1}_v\right)\right)\\
&\vdots\\
\frac{\partial f_1}{\partial\Pi_{\frac{\alpha_i}{n}}(u,v)} &= \sum_{t=\alpha_{i-1}+1}^{\alpha_i}\left(\mathbb{1}_u 2\left(\frac{t-\alpha_{i-1}}{\alpha_i-\alpha_{i-1}}\right)\left(\frac{\alpha_i-t}{\alpha_i-\alpha_{i-1}}\Pi_{\frac{\alpha_{i-1}}{n}}(u,v)+\frac{t-\alpha_{i-1}}{\alpha_i-\alpha_{i-1}}\Pi_{\frac{\alpha_i}{n}}(u,v)-\mathbb{1}_v\right)\right)\\
&+ \sum_{t=\alpha_i+1}^{n}\left(\mathbb{1}_u 2\left(\frac{\alpha_{i+1}-t}{\alpha_{i+1}-\alpha_i}\right)\left(\frac{\alpha_{i+1}-t}{\alpha_{i+1}-\alpha_i}\Pi_{\frac{\alpha_i}{n}}(u,v)+\frac{t-\alpha_i}{\alpha_{i+1}-\alpha_i}\Pi_{\frac{\alpha_{i+1}}{n}}(u,v)-\mathbb{1}_v\right)\right)\\
&\vdots\\
\frac{\partial f_1}{\partial\Pi_1(u,v)} &= \sum_{t=\alpha_{N-1}+1}^{n}\left(\mathbb{1}_u 2\left(\frac{t-\alpha_{N-1}}{n-\alpha_{N-1}}\right)\left(\frac{n-t}{n-\alpha_{N-1}}\Pi_{\frac{\alpha_{N-1}}{n}}(u,v)+\frac{t-\alpha_{N-1}}{n-\alpha_{N-1}}\Pi_1(u,v)-\mathbb{1}_v\right)\right).
\end{aligned}
$$

La minimisation nous donne le système suivant :

$$
\begin{pmatrix}
a_{11} & a_{12} & 0 & \dots & \dots & \dots & 0\\
a_{21} & a_{22} & a_{23} & \ddots & & & \vdots\\
0 & a_{32} & a_{33} & \ddots & \ddots & & \vdots\\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots\\
\vdots & & \ddots & \ddots & \ddots & \ddots & 0\\
\vdots & & & \ddots & \ddots & a_{NN} & a_{N(N+1)}\\
0 & \dots & \dots & \dots & 0 & a_{(N+1)N} & a_{(N+1)(N+1)}
\end{pmatrix}
\begin{pmatrix}
\Pi_0\\ \Pi_{\frac{\alpha_1}{n}}\\ \vdots\\ \Pi_{\frac{\alpha_i}{n}}\\ \vdots\\ \Pi_{\frac{\alpha_{N-1}}{n}}\\ \Pi_1
\end{pmatrix}
=
\begin{pmatrix}
b_1\\ b_2\\ \vdots\\ b_i\\ \vdots\\ b_N\\ b_{N+1}
\end{pmatrix}.
$$

Voici les valeurs des termes de la matrice et du du second membre :

– 1-ère ligne

$$
a_{11}=\sum_{t=k}^{\alpha_1}\mathbb{1}_u 2\left(\frac{\alpha_1-t}{\alpha_1}\right)^2,\quad a_{12}=\sum_{t=k}^{\alpha_1}\mathbb{1}_u 2\left(\frac{\alpha_1-t}{\alpha_1}\right)\frac{t}{\alpha_1},\quad b_1=\sum_{t=k}^{\alpha_1}\mathbb{1}_{uv} 2\left(\frac{\alpha_1-t}{\alpha_1}\right);
$$

– 2-ème ligne

$$
a_{21}=\sum_{t=k}^{\alpha_1}\mathbb{1}_u 2\left(\frac{\alpha_1-t}{\alpha_1}\right)\frac{t}{\alpha_1},\quad a_{22}=\sum_{t=k}^{\alpha_1}\mathbb{1}_u 2\left(\frac{t}{\alpha_1}\right)^2+\sum_{t=\alpha_1+1}^{\alpha_2}\mathbb{1}_u 2\left(\frac{\alpha_2-t}{\alpha_2-\alpha_1}\right)^2,
$$

$$
a_{23}=\sum_{t=\alpha_1+1}^{\alpha_2}\mathbb{1}_u 2\left(\frac{\alpha_2-t}{\alpha_2-\alpha_1}\right)\left(\frac{t-\alpha_1}{\alpha_2-\alpha_1}\right),\quad b_2=\sum_{t=k}^{\alpha_1}\mathbb{1}_{uv} 2\left(\frac{t}{\alpha_1}\right)+\sum_{t=\alpha_1+1}^{\alpha_2}\mathbb{1}_{uv} 2\left(\frac{\alpha_2-t}{\alpha_2-\alpha_1}\right);
$$

– i-ème ligne

$$
a_{(i+1)i} = \sum_{t=\alpha_{i-1}+1}^{\alpha_i}\mathbb{1}_u 2\left(\frac{\alpha_i-t}{\alpha_i-\alpha_{i-1}}\right)\left(\frac{t-\alpha_{i-1}}{\alpha_i-\alpha_{i-1}}\right),
$$

$$
\begin{aligned}
a_{(i+1)(i+1)} &= \sum_{t=\alpha_{i-1}+1}^{\alpha_i} \mathbb{1}_u 2\left(\frac{t-\alpha_{i-1}}{\alpha_i-\alpha_{i-1}}\right)^2 + \sum_{t=\alpha_i+1}^{n} \mathbb{1}_u 2\left(\frac{\alpha_{i+1}-t}{\alpha_{i+1}-\alpha_i}\right)^2,\\
a_{(i+1)(i+2)} &= \sum_{t=\alpha_i+1}^{\alpha_{i+1}} \mathbb{1}_u 2\left(\frac{\alpha_{i+1}-t}{\alpha_{i+1}-\alpha_i}\right)\left(\frac{t-\alpha_i}{\alpha_{i+1}-\alpha_i}\right),\\
b_{(i+1)} &= \sum_{t=\alpha_{i-1}+1}^{\alpha_i} \mathbb{1}_{uv} 2\left(\frac{t-\alpha_{i-1}}{\alpha_i-\alpha_{i-1}}\right) + \sum_{t=\alpha_i+1}^{\alpha_{i+1}} \mathbb{1}_{uv} 2\left(\frac{\alpha_{i+1}-t}{\alpha_{i+1}-\alpha_i}\right);
\end{aligned}
$$

– $(N+1)$-ème ligne

$$
a_{(N+1)N} = \sum_{t=\alpha_{N-1}+1}^{n} \mathbb{1}_u 2\left(\frac{n-t}{n-\alpha_{N-1}}\right)\left(\frac{t-\alpha_{N-1}}{n-\alpha_{N-1}}\right), \quad a_{(N+1)(N+1)} = \sum_{t=\alpha_{N-1}+1}^{n} \mathbb{1}_u 2\left(\frac{t-\alpha_{N-1}}{n-\alpha_{N-1}}\right)^2,
$$

$$
b_{(N+1)} = \sum_{t=\alpha_{N-1}+1}^{n} \mathbb{1}_{uv} 2\left(\frac{t-\alpha_{N-1}}{n-\alpha_{N-1}}\right).
$$

Théorème 11 *Les matrices $\Pi^i_{\frac{t}{n}}$ sont stochastiques pour i allant de 1 à N.*

Preuve *Nous sommons sur $v \in \mathcal{A}$ les équations de notre système. Nous obtenons le système suivant :*

$$
\begin{pmatrix}
a_{11} & a_{12} & 0 & \dots & \dots & \dots & 0\\
a_{21} & a_{22} & a_{23} & \ddots & & & \vdots\\
0 & a_{32} & a_{33} & \ddots & \ddots & & \vdots\\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots\\
\vdots & & \ddots & \ddots & \ddots & \ddots & 0\\
\vdots & & & \ddots & \ddots & a_{NN} & a_{N(N+1)}\\
0 & \dots & \dots & \dots & 0 & a_{(N+1)N} & a_{(N+1)(N+1)}
\end{pmatrix}
\begin{pmatrix}
\sum_{v\in\mathcal{A}} \Pi_0\\
\sum_{v\in\mathcal{A}} \Pi_{\frac{\alpha_1}{n}}\\
\vdots\\
\sum_{v\in\mathcal{A}} \Pi_{\frac{\alpha_i}{n}}\\
\vdots\\
\sum_{v\in\mathcal{A}} \Pi_{\frac{\alpha_{N-1}}{n}}\\
\sum_{v\in\mathcal{A}} \Pi_1
\end{pmatrix}
=
\begin{pmatrix}
\sum_{v\in\mathcal{A}} b_1\\
\sum_{v\in\mathcal{A}} b_2\\
\vdots\\
\sum_{v\in\mathcal{A}} b_i\\
\vdots\\
\sum_{v\in\mathcal{A}} b_N\\
\sum_{v\in\mathcal{A}} b_{N+1}
\end{pmatrix}.
$$

Le vecteur $(1\cdots1)^t$ est solution. Ainsi on montre que les lignes des matrices $\Pi_{\frac{\alpha_i}{n}}$ somment toutes à 1. Par conséquent, la stochasticité des $\Pi^i_{\frac{t}{n}}$ est vérifiée. La même remarque que précédemment (voir A.1.2) peut s'appliquer : il est théoriquement possible d'obtenir des termes négatifs.

3.2.3 degré d, N morceaux

La modélisation par splines polynomiales de degré d revient à estimer un DMM de degré d sur chaque morceau en imposant la continuité aux noeuds ainsi que la continuité des dérivées d'ordre 1 à $d-1$. L'utilisation du modèle des points d'appuis va rendre les calculs complexes. Aussi, nous utilisons désormais le modèle des polynômes (voir Annexe B). Rappelons que ces deux modèles sont équivalents. Sur chaque segment i, pour i allant de 1 à N, nous utilisons $d+1$ matrices de paramètres (voir Annexe B) pour définir notre matrice de transition dérivante $\Pi^i_{\frac{t}{n}}$:

$$
\Pi^i_{\frac{t}{n}}(u,v) = M^i_0(u,v) + \frac{t}{n}M^i_1(u,v) + \ldots + \frac{t^d}{n^d}M^i_d(u,v).
$$

Ce modèle comprend $(d+1)N$ matrices de paramètres.

Exemple 8 *Nous présentons le modèle par splines polynomiales de degré 2 pour une séquence découpée en 2 segments :*

$$
\begin{aligned}
\Pi^1_{\frac{t}{n}}(u,v) &= M^1_0(u,v) + \frac{t}{n}M^1_1(u,v) + \frac{t^2}{n^2}M^1_2(u,v)\\
\Pi^2_{\frac{t}{n}}(u,v) &= M^2_0(u,v) + \frac{t}{n}M^2_1(u,v) + \frac{t^2}{n^2}M^2_2(u,v).
\end{aligned}
$$

Les matrices M^i_j pour $i=1,2$ et $j=0,1,2$ sont à estimer (voir E.1.2).

Nous souhaitons minimiser la fonction f_d suivante sous les contraintes de continuité précitées :

$$\begin{aligned} f_d(M_j^i) &= \sum_{t=k}^{\alpha_1}\left(\sum_{u\in\mathcal{A}^k}\mathbb{1}_u\left(\sum_{v\in\mathcal{A}}\left(\Pi^1_{\frac{t}{n}}(u,v)-\mathbb{1}_v\right)^2\right)\right) \\ &+ \sum_{j=1}^{N-2}\left(\sum_{t=\alpha_j+1}^{\alpha_{j+1}}\left(\sum_{u\in\mathcal{A}^k}\mathbb{1}_u\left(\sum_{v\in\mathcal{A}}\left(\Pi^{j+1}_{\frac{t}{n}}(u,v)-\mathbb{1}_v\right)^2\right)\right)\right) \\ &+ \sum_{t=\alpha_{N-1}+1}^{n}\left(\sum_{u\in\mathcal{A}^k}\mathbb{1}_u\left(\sum_{v\in\mathcal{A}}\left(\Pi^N_{\frac{t}{n}}(u,v)-\mathbb{1}_v\right)^2\right)\right). \end{aligned}$$

Notons les contraintes :
- $\Pi^i_{\frac{\alpha_i}{n}}(u,v)=\Pi^{i+1}_{\frac{\alpha_i}{n}}(u,v)$ pour i allant de 1 à $N-1$ ($N-1$ contraintes)
- $\left(\Pi^i_{\frac{\alpha_i}{n}}\right)'(u,v)=\left(\Pi^{i+1}_{\frac{\alpha_i}{n}}\right)'(u,v)$ pour i allant de 1 à $N-1$ ($N-1$ contraintes)
- $\left(\Pi^i_{\frac{\alpha_i}{n}}\right)^{(j)}(u,v)=\left(\Pi^{i+1}_{\frac{\alpha_i}{n}}\right)^{(j)}(u,v)$ pour i allant de 1 à $N-1$ ($N-1$ contraintes)
- $\left(\Pi^i_{\frac{\alpha_i}{n}}\right)^{(d-1)}(u,v)=\left(\Pi^{i+1}_{\frac{\alpha_i}{n}}\right)^{(d-1)}(u,v)$ pour i allant de 1 à $N-1$ ($N-1$ contraintes)

$\Longleftrightarrow$
- $M_0^i(u,v)+\ldots+\frac{\alpha_i^k}{n^k}M_k^i(u,v)+\ldots+\frac{\alpha_i^d}{n^d}M_d^i(u,v)=M_0^{i+1}(u,v)+\ldots+\frac{\alpha_i^k}{n^k}M_k^{i+1}(u,v)+\ldots+\frac{\alpha_i^d}{n^d}M_d^{i+1}(u,v)$
- $M_1^i(u,v)+2\frac{\alpha_i}{n}M_2^i(u,v)+\ldots+d\frac{\alpha_i^{d-1}}{n^{d-1}}M_d^i(u,v)=M_1^{i+1}(u,v)+2\frac{\alpha_i}{n}M_2^{i+1}(u,v)+\ldots+d\frac{\alpha_i^{d-1}}{n^{d-1}}M_d^{i+1}(u,v)$
- $\sum_{k=j}^d\frac{k!}{(k-j)!}\left(\frac{\alpha_i}{n}\right)^{k-j}M_k^i(u,v)=\sum_{k=j}^d\frac{k!}{(k-j)!}\left(\frac{\alpha_i}{n}\right)^{k-j}M_k^{i+1}(u,v)$
- $(d-1)!M_{d-1}^i(u,v)+d!\frac{\alpha_i}{n}M_d^i(u,v)=(d-1)!M_{d-1}^{i+1}(u,v)+d!\frac{\alpha_i}{n}M_d^{i+1}(u,v)$.

Nous avons donc au total $(N-1)d$ contraintes et $N(d+1)$ paramètres. Utilisons la méthode lagrangienne pour la minimisation sous contraintes. Ajoutons à la fonction à minimiser nos contraintes avec des coefficients $\lambda_{i,j}$ avec i allant de 1 à $N-1$ et j allant de 0 à $d-1$.

Remarque 7 *Il est possible de donner plus de liberté au modèles en supprimant les contraintes de continuité des dérivées d'ordre 2 à $d+1$. Pour cela, il suffit de faire varier les coefficients $\lambda_{i,j}$ pour j allant seulement de 0 à 1. Les splines obtenues seront moins lisses mais nous aurons tout de même une estimation de nos DMM.*

La fonction à minimiser devient donc la fonction F_d suivante :

$$\begin{aligned} F_d(M_j^i,\lambda_{i,j}) &= \sum_{t=k}^{\alpha_1}\left(\sum_{u\in\mathcal{A}^k}\mathbb{1}_u\left(\sum_{v\in\mathcal{A}}\left(\Pi^1_{\frac{t}{n}}(u,v)-\mathbb{1}_v\right)^2\right)\right) \\ &+ \sum_{j=1}^{N-2}\left(\sum_{t=\alpha_j+1}^{\alpha_{j+1}}\left(\sum_{u\in\mathcal{A}^k}\mathbb{1}_u\left(\sum_{v\in\mathcal{A}}\left(\Pi^{j+1}_{\frac{t}{n}}(u,v)-\mathbb{1}_v\right)^2\right)\right)\right) \\ &+ \sum_{t=\alpha_{N-1}+1}^{n}\left(\sum_{u\in\mathcal{A}^k}\mathbb{1}_u\left(\sum_{v\in\mathcal{A}}\left(\Pi^N_{\frac{t}{n}}(u,v)-\mathbb{1}_v\right)^2\right)\right) \\ &+ \sum_{i=1}^{N-1}\sum_{j=0}^{d-1}\lambda_{i,j}\sum_{k=j}^{d}\frac{k!}{(k-j)!}\left(\frac{\alpha_i}{n}\right)^{k-j}\left[M_k^i(u,v)-M_k^{i+1}(u,v)\right]. \end{aligned}$$

Explicitons les dérivées selon les M_j^i (pour i allant de 1 à N et j allant de 0 à d) et selon les $\lambda_{i,j}$ (pour i allant de 1 à $N-1$ et j allant de 0 à $d-1$) :

$$\begin{aligned} \frac{\partial F_d(M_j^i,\lambda_i,\gamma_i)}{\partial M_j^1(u,v)} &= \sum_{t=k}^{\alpha_1}\mathbb{1}_u 2\frac{t^j}{n^j}\left(M_0^1(u,v)+\ldots+\frac{t^d}{n^d}M_d^1(u,v)-\mathbb{1}_v\right)+\sum_{k=0}^{j}\lambda_{1,k}\frac{j!}{(j-k)!}\frac{\alpha_1^{j-k}}{n^{j-k}}; \\ \frac{\partial F_d(M_j^i,\lambda_i,\gamma_i)}{\partial M_j^i(u,v)} &= \sum_{t=\alpha_{i-1}+1}^{\alpha_i}\mathbb{1}_u 2\frac{t^j}{n^j}\left(M_0^i(u,v)+\ldots+\frac{t^d}{n^d}M_d^i(u,v)-\mathbb{1}_v\right) \\ &\quad +\sum_{k=0}^{j}\lambda_{i,k}\frac{\alpha_i^{j-k}}{n^{j-k}}-\sum_{k=0}^{j}\lambda_{i-1,k}\frac{j!}{(j-k)!}\frac{\alpha_{i-1}^{j-k}}{n^{j-k}}; \end{aligned}$$

$$\frac{\partial F_d(M_j^i,\lambda_i,\gamma_i)}{\partial M_j^N(u,v)} = \sum_{t=\alpha_{N-1}}^{n} \mathbb{1}_u 2\frac{t^j}{n^j}\left(M_0^N(u,v)+\ldots+\frac{t^d}{n^d}M_d^N(u,v)-\mathbb{1}_v\right)-\sum_{k=0}^{j}\lambda_{N-1,k}\frac{j!}{(j-k)!}\frac{\alpha_{N-1}^{j-k}}{n^{j-k}};$$

$$\frac{\partial F_d(M_j^i,\lambda_{i,j})}{\partial \lambda_{i,j}} = \sum_{k=j}^{d}\frac{k!}{(k-j)!}\left(\frac{\alpha_i}{n}\right)^{k-j}\left[M_k^i(u,v)-M_k^{i+1}(u,v)\right].$$

Nous obtenons ainsi un système de $(d+1)N+(N-1)d$ équations à $(d+1)N+(N-1)d$ inconnues. Pour écrire ce système nous décomposons sa matrice en quatre blocs. Explicitons les trois premiers blocs (sachant que le dernier bloc est une matrice nulle) :

- $\Sigma_1 = \begin{pmatrix} \Sigma_{1,1} & & 0 \\ & \ddots & \\ 0 & & \Sigma_{1,N} \end{pmatrix}$ avec

$$\Sigma_{1,1} = \begin{pmatrix} 2\sum_{t=k}^{\alpha_1}\mathbb{1}_u & \cdots & 2\sum_{t=k}^{\alpha_1}\mathbb{1}_u\frac{t^d}{n^d} \\ \vdots & \ddots & \vdots \\ 2\sum_{t=k}^{\alpha_1}\mathbb{1}_u\frac{t^d}{n^d} & \cdots & 2\sum_{t=k}^{\alpha_1}\mathbb{1}_u\frac{t^{2d}}{n^{2d}} \end{pmatrix} \text{ et } \Sigma_{1,N} = \begin{pmatrix} 2\sum_{t=\alpha_{N-1}+1}^{n}\mathbb{1}_u & \cdots & 2\sum_{t=\alpha_{N-1}+1}^{n}\mathbb{1}_u\frac{t^d}{n^d} \\ \vdots & \ddots & \vdots \\ 2\sum_{t=\alpha+1}^{n}\mathbb{1}_u\frac{t^d}{n^d} & \cdots & 2\sum_{t=\alpha+1}^{n}\mathbb{1}_u\frac{t^{2d}}{n^{2d}} \end{pmatrix}.$$

  Ainsi $\Sigma_1 \in M_{N(d+1)}(\mathbb{R})$ et $\Sigma_{1,i} \in M_{d+1}(\mathbb{R})$.

- $\Sigma_2 = \begin{pmatrix} \Sigma_{2,1} & & 0 \\ & \ddots & \\ 0 & & \Sigma_{2,N-1} \end{pmatrix}$ avec

$$\Sigma_{2,1} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ \frac{\alpha_1}{n} & 1 & \ddots & \vdots \\ \vdots & \frac{2\alpha_1}{n} & \ddots & 0 \\ \vdots & \vdots & \ddots & (d-1)! \\ \frac{\alpha_1^d}{n^d} & d\frac{\alpha_1^{d-1}}{n^{d-1}} & \ldots & d!\frac{\alpha_1}{n} \\ -1 & 0 & \ldots & 0 \\ -\frac{\alpha_1}{n} & -1 & \ddots & \vdots \\ \vdots & -\frac{2\alpha_1}{n} & \ddots & 0 \\ \vdots & \vdots & \ddots & -(d-1)! \\ -\frac{\alpha_1^d}{n^d} & -d\frac{\alpha_1^{d-1}}{n^{d-1}} & \ldots & -d!\frac{\alpha_1}{n} \end{pmatrix} \text{ et } \Sigma_{2,N-1} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ \frac{\alpha_{N-1}}{n} & 1 & \ddots & \vdots \\ \vdots & \frac{2\alpha_{N-1}}{n} & \ddots & 0 \\ \vdots & \vdots & \ddots & (d-1)! \\ \frac{\alpha_{N-1}^d}{n^d} & d\frac{\alpha_{N-1}^{d-1}}{n^{d-1}} & \ldots & d!\frac{\alpha_{N-1}}{n} \\ -1 & 0 & \ldots & 0 \\ -\frac{\alpha_{N-1}}{n} & -1 & \ddots & \vdots \\ \vdots & -\frac{2\alpha_{N-1}}{n} & \ddots & 0 \\ \vdots & \vdots & \ddots & -(d-1)! \\ -\frac{\alpha_{N-1}^d}{n^d} & -d\frac{\alpha_{N-1}^{d-1}}{n^{d-1}} & \ldots & -d!\frac{\alpha_{N-1}}{n} \end{pmatrix}.$$

  Ainsi $\Sigma_2 \in M_{N(d+1),d(N-1)}(\mathbb{R})$ et $\Sigma_{2,i} \in M_{2(d+1),d}(\mathbb{R})$.

- $\Sigma_3 = \begin{pmatrix} \Sigma_{3,1} & & 0 \\ & \ddots & \\ 0 & & \Sigma_{3,N-1} \end{pmatrix}$ avec

$$\Sigma_{3,1} = \begin{pmatrix} 1 & \frac{\alpha_1}{n} & \ldots & \frac{\alpha_1^{d-1}}{n^{d-1}} & \frac{\alpha_1^d}{n^d} & -1 & -\frac{\alpha_1}{n} & \ldots & -\frac{\alpha_1^{d-1}}{n^{d-1}} & -\frac{\alpha_1^d}{n^d} \\ 0 & 1 & \ldots & (d-1)\frac{\alpha_1^{d-2}}{n^{d-2}} & d\frac{\alpha_1^{d-1}}{n^{d-1}} & 0 & -1 & -\ldots & -(d-1)\frac{\alpha_1^{d-2}}{n^{d-2}} & -d\frac{\alpha_1^{d-1}}{n^{d-1}} \\ \vdots & \ddots & \ddots & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \ldots & 0 & (d-1)! & d!\frac{\alpha_1}{n} & 0 & \ldots & 0 & -(d-1)! & -d!\frac{\alpha_1}{n} \end{pmatrix}$$

  et

$$\Sigma_{3,N-1} = \begin{pmatrix} 1 & \frac{\alpha_{N-1}}{n} & \ldots & \frac{\alpha_{N-1}^{d-1}}{n^{d-1}} & \frac{\alpha_{N-1}^d}{n^d} & -1 & -\frac{\alpha_{N-1}}{n} & \ldots & -\frac{\alpha_{N-1}^{d-1}}{n^{d-1}} & -\frac{\alpha_{N-1}^d}{n^d} \\ 0 & 1 & \ldots & (d-1)\frac{\alpha_{N-1}^{d-2}}{n^{d-2}} & d\frac{\alpha_{N-1}^{d-1}}{n^{d-1}} & 0 & -1 & -\ldots & -(d-1)\frac{\alpha_{N-1}^{d-2}}{n^{d-2}} & -d\frac{\alpha_{N-1}^{d-1}}{n^{d-1}} \\ \vdots & \ddots & \ddots & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \ldots & 0 & (d-1)! & d!\frac{\alpha_{N-1}}{n} & 0 & \ldots & 0 & -(d-1)! & -d!\frac{\alpha_{N-1}}{n} \end{pmatrix}.$$

Ainsi $\Sigma_3 \in M_{(N-1)d,N(d+1)}(\mathbb{R})$ et $\Sigma_{3,i} \in M_{d,2(d+1)}(\mathbb{R})$.
Le système s'écrit finalement :

$$\begin{pmatrix} \Sigma_1 & \Sigma_2 \\ \Sigma_3 & 0 \end{pmatrix} \begin{pmatrix} M_0^1 \\ \vdots \\ M_d^1 \\ \vdots \\ M_0^N \\ \vdots \\ M_d^N \\ \lambda_{1,0} \\ \vdots \\ \lambda_{1,d-1} \\ \vdots \\ \lambda_{N-1,0} \\ \vdots \\ \lambda_{N-1,d-1} \end{pmatrix} = \begin{pmatrix} 2\sum_{t=k}^{\alpha_1} \mathbb{1}_{uv} \\ \vdots \\ 2\sum_{t=k}^{\alpha_1} \mathbb{1}_{uv} \frac{t^d}{n^d} \\ \vdots \\ 2\sum_{t=\alpha_{N-1}+1}^{n} \mathbb{1}_{uv} \\ \vdots \\ 2\sum_{t=\alpha_{N-1}+1}^{n} \mathbb{1}_{uv} \frac{t^d}{n^d} \\ 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Théorème 12 *Les matrices $\Pi^i_{\frac{t}{n}}$ sont stochastiques pour i allant de 1 à N.*

Preuve *Comme précédemment on vérifie que* $(\ 1 \ \cdots \ 0 \ \cdots \ 1 \ \cdots \ 0 \ 0 \ \cdots \ 0 \ \cdots \ 0 \ \cdots \ 0\ )^t$ *est solution du système dans lequel on somme sur tous les $v \in \mathcal{A}$. Ainsi on montre que les lignes de ces matrices somment toutes à 0 sauf pour M_0^i avec $i = 1, 2, \ldots, N$ qui somment à 1. Ainsi la stochasticité des $\Pi^i_{\frac{t}{n}}$ est vérifiée. La même remarque que précédemment (voir A.1.2) peut s'appliquer : il est théoriquement possible d'obtenir des termes négatifs.*

3.3 Estimation par morceaux, allers et retours

La naïveté de la première approche nous incite à mettre en oeuvre d'autres méthodes d'estimation d'un DMM par splines polynomiales. Ainsi nous est venue l'idée de définir les matrices dérivantes récursivement d'un segment à un autre. Nous présentons cette méthode directement avec son extension : les allers et retours. Rappelons nos notations : N le nombre de morceaux, $\{\alpha_0; \ldots; \alpha_N\}$ le découpage (voir Figure 3.1 pour un exemple) et $\Pi^i_{\frac{t}{n}}$ la matrice de transition pour le segment i (i allant de 1 à N). Il s'agit d'estimer notre matrice $\Pi^1_{\frac{t}{n}}$ sur le premier segment, puis à l'aide des conditions de continuité on estime la matrice $\Pi^2_{\frac{t}{n}}$ du deuxième segment et ainsi de suite. Arrivés au dernier segment, nous revenons en arrière, estimant la matrice $\Pi^{N-1}_{\frac{t}{n}}$ toujours à l'aide des conditions de continuité. Le procédé se répète autant que nous voulons.

La littérature reconnaissant unanimement l'emploi des splines de degré 3, nous nous bornerons à ce degré dans la construction de cette méthode. Nous avons envisagé deux modèles sur lesquels utiliser cette méthode :

- un modèle sans fonction de base. Les matrices $\Pi^i_{\frac{t}{n}}$ sont exprimées sous la forme canonique d'un polynôme de matrices, à l'aide des matrices de paramètres H^i_j pour $j = 0, 1, 2, 3$ et i allant de 1 à N :

$$\Pi^i_{\frac{t}{n}} = H^i_0 + \frac{t}{n} H^i_1 + \frac{t^2}{n^2} H^i_2 + \frac{t^3}{n^3} H^i_3.$$

  Les contraintes sont satisfaites dans la résolution de systèmes (voir 3.3.1).
- un modèle avec fonctions de base. Les matrices $\Pi^i_{\frac{t}{n}}$ sont exprimées à l'aide de fonctions de base de splines a, b, c et d :

$$\Pi^i_{\frac{t}{n}} = A_i a\left(\frac{t-\alpha_{i-1}}{\alpha_i - \alpha_{i-1}}\right) + B_i b\left(\frac{t-\alpha_{i-1}}{\alpha_i - \alpha_{i-1}}\right) + C_i c\left(\frac{t-\alpha_{i-1}}{\alpha_i - \alpha_{i-1}}\right) + D_i d\left(\frac{t-\alpha_{i-1}}{\alpha_i - \alpha_{i-1}}\right)$$

Ces fonctions de base sont choisies afin de satisfaire les contraintes au préalable. Ainsi l'estimation se réduit à la résolution de systèmes 2×2 (voir 3.3.2).

3.3.1 Sans fonctions de base

Nous allons construire notre méthode d'estimation à partir du modèle suivant. Pour i allant de 1 à N :

$$\Pi^i_{\frac{t}{n}} = H^i_0 + \frac{t}{n}H^i_1 + \frac{t^2}{n^2}H^i_2 + \frac{t^3}{n^3}H^i_3.$$

Nous posons les contraintes de continuité, pour $i = 1 \dots N-1$:
- $\Pi^i_{\frac{\alpha_i}{n}} = \Pi^{i+1}_{\frac{\alpha_i}{n}}$;
- ${\Pi^i_{\frac{\alpha_i}{n}}}' = {\Pi^{i+1}_{\frac{\alpha_i}{n}}}'$.

L'initialisation de la récurrence s'effectue sur les deux premiers segments en même temps à l'aide des contraintes de continuité. Puis la récurrence se poursuit segment par segment. Ainsi, la méthode lagrangienne nous propose la minimisation des fonctions g_1, g^a_i et g^r_i (pour i allant de 1 à $N-1$) pour l'initialisation, l'aller et le retour :

$$\begin{aligned}
g_1(H^1_j(u,v), H^2_j(u,v), \lambda_1, \gamma_1) &= \sum_{t=k}^{\alpha_1}\left[\sum_{u\in\mathcal{A}^k}\mathbb{1}_u\sum_{v\in\mathcal{A}}\left(\Pi^1_{\frac{t}{n}} - \mathbb{1}_{uv}\right)^2\right] + \sum_{t=\alpha_1}^{\alpha_2}\left[\sum_{u\in\mathcal{A}^k}\mathbb{1}_u\sum_{v\in\mathcal{A}}\left(\Pi^2_{\frac{t}{n}} - \mathbb{1}_{uv}\right)^2\right] \\
&\quad -\lambda_1\left(\Pi^1_{\frac{t}{n}}(\alpha_1) - \Pi^2_{\frac{t}{n}}(\alpha_1)\right) - \gamma_1\left({\Pi^1_{\frac{t}{n}}}'(\alpha_1) - {\Pi^2_{\frac{t}{n}}}'(\alpha_1)\right) \\
g^a_i(H^{i+1}_j(u,v), \lambda_i, \gamma_i) &= \sum_{t=\alpha_i}^{\alpha_{i+1}}\left[\sum_{u\in\mathcal{A}^k}\mathbb{1}_u\sum_{v\in\mathcal{A}}\left(\Pi^{i+1}_{\frac{t}{n}} - \mathbb{1}_{uv}\right)^2\right] \\
&\quad -\lambda_i\left(\Pi^i_{\frac{t}{n}}(\alpha_i) - \Pi^{i+1}_{\frac{t}{n}}(\alpha_i)\right) - \gamma_i\left({\Pi^i_{\frac{t}{n}}}'(\alpha_i) - {\Pi^{i+1}_{\frac{t}{n}}}'(\alpha_i)\right) \\
g^r_i(H^i_j, \lambda_i, \gamma_i) &= \sum_{t=\alpha_{i-1}}^{\alpha_i}\left[\sum_{u\in\mathcal{A}^k}\mathbb{1}_u\sum_{v\in\mathcal{A}}\left(\Pi^i_{\frac{t}{n}} - \mathbb{1}_{uv}\right)^2\right] \\
&\quad -\lambda_i\left(\Pi^i_{\frac{t}{n}}(\alpha_i) - \Pi^{i+1}_{\frac{t}{n}}(\alpha_i)\right) - \gamma_i\left({\Pi^i_{\frac{t}{n}}}'(\alpha_i) - {\Pi^{i+1}_{\frac{t}{n}}}'(\alpha_i)\right)
\end{aligned}$$

La minimisation nous donne pour chaque couple (u,v) un système 6×6 d'initialisation, N systèmes 4×4 pour l'aller et N systèmes 4×4 pour le retour :

INITIALISATION : Segment 1.
Estimation des H^1_j, H^2_j pour j allant de 0 à 3 et de λ_1, γ_1 :

$$\begin{pmatrix} \Sigma_A & 0 & -\sigma^t \\ 0 & \Sigma_B & \sigma^t \\ -\sigma & \sigma & 0 \end{pmatrix} \begin{pmatrix} H^1_0(u,v) \\ H^1_1(u,v) \\ H^1_2(u,v) \\ H^1_3(u,v) \\ H^2_0(u,v) \\ H^2_1(u,v) \\ H^2_2(u,v) \\ H^2_3(u,v) \\ \lambda_1 \\ \gamma_1 \end{pmatrix} = B_0$$

avec

$$\Sigma_A = \begin{pmatrix} 2\sum_{t=k}^{\alpha_1}\mathbb{1}_u & 2\sum_{t=k}^{\alpha_1}\mathbb{1}_u\frac{t}{n} & 2\sum_{t=k}^{\alpha_1}\mathbb{1}_u\frac{t^2}{n^2} & 2\sum_{t=k}^{\alpha_1}\mathbb{1}_u\frac{t^3}{n^3} \\ 2\sum_{t=k}^{\alpha_1}\mathbb{1}_u\frac{t}{n} & 2\sum_{t=k}^{\alpha_1}\mathbb{1}_u\frac{t^2}{n^2} & 2\sum_{t=k}^{\alpha_1}\mathbb{1}_u\frac{t^3}{n^3} & 2\sum_{t=k}^{\alpha_1}\mathbb{1}_u\frac{t^4}{n^4} \\ 2\sum_{t=k}^{\alpha_1}\mathbb{1}_u\frac{t^2}{n^2} & 2\sum_{t=k}^{\alpha_1}\mathbb{1}_u\frac{t^3}{n^3} & 2\sum_{t=k}^{\alpha_1}\mathbb{1}_u\frac{t^4}{n^4} & 2\sum_{t=k}^{\alpha_1}\mathbb{1}_u\frac{t^5}{n^5} \\ 2\sum_{t=k}^{\alpha_1}\mathbb{1}_u\frac{t^3}{n^3} & 2\sum_{t=k}^{\alpha_1}\mathbb{1}_u\frac{t^4}{n^4} & 2\sum_{t=k}^{\alpha_1}\mathbb{1}_u\frac{t^5}{n^5} & 2\sum_{t=k}^{\alpha_1}\mathbb{1}_u\frac{t^6}{n^6} \end{pmatrix}$$

$$\Sigma_B = \begin{pmatrix} 2\sum_{t=\alpha_1}^{\alpha_2} \mathbb{1}_u & 2\sum_{t=\alpha_1}^{\alpha_2} \mathbb{1}_u \frac{t}{n} & 2\sum_{t=\alpha_1}^{\alpha_2} \mathbb{1}_u \frac{t^2}{n^2} & 2\sum_{t=\alpha_1}^{\alpha_2} \mathbb{1}_u \frac{t^3}{n^3} \\ 2\sum_{t=\alpha_1}^{\alpha_2} \mathbb{1}_u \frac{t}{n} & 2\sum_{t=\alpha_1}^{\alpha_2} \mathbb{1}_u \frac{t^2}{n^2} & 2\sum_{t=\alpha_1}^{\alpha_2} \mathbb{1}_u \frac{t^3}{n^3} & 2\sum_{t=\alpha_1}^{\alpha_2} \mathbb{1}_u \frac{t^4}{n^4} \\ 2\sum_{t=\alpha_1}^{\alpha_2} \mathbb{1}_u \frac{t^2}{n^2} & 2\sum_{t=\alpha_1}^{\alpha_2} \mathbb{1}_u \frac{t^3}{n^3} & 2\sum_{t=\alpha_1}^{\alpha_2} \mathbb{1}_u \frac{t^4}{n^4} & 2\sum_{t=\alpha_1}^{\alpha_2} \mathbb{1}_u \frac{t^5}{n^5} \\ 2\sum_{t=\alpha_1}^{\alpha_2} \mathbb{1}_u \frac{t^3}{n^3} & 2\sum_{t=\alpha_1}^{\alpha_2} \mathbb{1}_u \frac{t^4}{n^4} & 2\sum_{t=\alpha_1}^{\alpha_2} \mathbb{1}_u \frac{t^5}{n^5} & 2\sum_{t=\alpha_1}^{\alpha_2} \mathbb{1}_u \frac{t^6}{n^6} \end{pmatrix}$$

$$\sigma = \begin{pmatrix} 1 & \frac{\alpha_1}{n} & \frac{\alpha_1^2}{n^2} & \frac{\alpha_1^3}{n^3} \\ 0 & \frac{1}{n} & 2\frac{\alpha_1}{n^2} & 3\frac{\alpha_1^2}{n^3} \end{pmatrix}$$

$$B_0^t = 2\begin{pmatrix} \sum_{t=k}^{\alpha_1} \mathbb{1}_{uv} & \sum_{t=k}^{\alpha_1} \mathbb{1}_{uv}\frac{t}{n} & \sum_{t=k}^{\alpha_1} \mathbb{1}_{uv}\frac{t^2}{n^2} & \sum_{t=k}^{\alpha_1} \mathbb{1}_{uv}\frac{t^3}{n^3} & \sum_{t=\alpha_1}^{\alpha_2} \mathbb{1}_{uv} & \sum_{t=\alpha_1}^{\alpha_2} \mathbb{1}_{uv}\frac{t}{n} & \sum_{t=\alpha_1}^{\alpha_2} \mathbb{1}_{uv}\frac{t^2}{n^2} & \sum_{t=\alpha_1}^{\alpha_2} \mathbb{1}_{uv}\frac{t^3}{n^3} & 0 & 0 \end{pmatrix}.$$

ALLER : Segment i pour i allant de 1 à $N-1$.
Estimation des H_j^{i+1} pour j allant de 0 à 3 et de λ_i, γ_i :

$$\begin{pmatrix} 2\sum_{t=\alpha_i}^{\alpha_{i+1}} \mathbb{1}_u & 2\sum_{t=\alpha_i}^{\alpha_{i+1}} \mathbb{1}_u \frac{t}{n} & 2\sum_{t=\alpha_i}^{\alpha_{i+1}} \mathbb{1}_u \frac{t^2}{n^2} & 2\sum_{t=\alpha_i}^{\alpha_{i+1}} \mathbb{1}_u \frac{t^2}{n^2} & 1 & 0 \\ 2\sum_{t=\alpha_i}^{\alpha_{i+1}} \mathbb{1}_u \frac{t}{n} & 2\sum_{t=\alpha_i}^{\alpha_{i+1}} \mathbb{1}_u \frac{t^2}{n^2} & 2\sum_{t=\alpha_i}^{\alpha_{i+1}} \mathbb{1}_u \frac{t^3}{n^3} & 2\sum_{t=\alpha_i}^{\alpha_{i+1}} \mathbb{1}_u \frac{t^4}{n^4} & \frac{\alpha_i}{n} & \frac{1}{n} \\ 2\sum_{t=\alpha_i}^{\alpha_{i+1}} \mathbb{1}_u \frac{t^2}{n^2} & 2\sum_{t=\alpha_i}^{\alpha_{i+1}} \mathbb{1}_u \frac{t^3}{n^3} & 2\sum_{t=\alpha_i}^{\alpha_{i+1}} \mathbb{1}_u \frac{t^4}{n^4} & 2\sum_{t=\alpha_i}^{\alpha_{i+1}} \mathbb{1}_u \frac{t^5}{n^5} & \frac{\alpha_i^2}{n^2} & 2\frac{\alpha_i}{n^2} \\ 2\sum_{t=\alpha_i}^{\alpha_{i+1}} \mathbb{1}_u \frac{t^3}{n^3} & 2\sum_{t=\alpha_i}^{\alpha_{i+1}} \mathbb{1}_u \frac{t^4}{n^4} & 2\sum_{t=\alpha_i}^{\alpha_{i+1}} \mathbb{1}_u \frac{t^5}{n^5} & 2\sum_{t=\alpha_i}^{\alpha_{i+1}} \mathbb{1}_u \frac{t^6}{n^6} & \frac{\alpha_i^3}{n^3} & 3\frac{\alpha_i^2}{n^3} \\ 1 & \frac{\alpha_i}{n} & \frac{\alpha_i^2}{n^2} & \frac{\alpha_i^3}{n^3} & 0 & 0 \\ 0 & \frac{i}{n} & 2\frac{\alpha_i}{n^2} & 3\frac{\alpha_i^2}{n^3} & 0 & 0 \end{pmatrix} \begin{pmatrix} H_0^{i+1}(u,v) \\ H_1^{i+1}(u,v) \\ H_2^{i+1}(u,v) \\ H_3^{i+1}(u,v) \\ \lambda_i \\ \gamma_i \end{pmatrix} = \begin{pmatrix} 2\sum_{t=\alpha_i}^{\alpha_{i+1}} \mathbb{1}_{uv} \\ 2\sum_{t=\alpha_i}^{\alpha_{i+1}} \mathbb{1}_{uv}\frac{t}{n} \\ 2\sum_{t=\alpha_i}^{\alpha_{i+1}} \mathbb{1}_{uv}\frac{t^2}{n^2} \\ 2\sum_{t=\alpha_i}^{\alpha_{i+1}} \mathbb{1}_{uv}\frac{t^3}{n^3} \\ \Pi_{\frac{t}{n}}^i(\alpha_i) \\ \Pi_{\frac{t}{n}}^{i\,\prime}(\alpha_i) \end{pmatrix}.$$

RETOUR : Segment i pour i allant de $N-1$ à 1.
Estimation des H_j^i pour j allant de 0 à 3 et de λ_i, γ_i :

$$\begin{pmatrix} 2\sum_{t=\alpha_{i-1}}^{\alpha_i} \mathbb{1}_u & 2\sum_{t=\alpha_{i-1}}^{\alpha_i} \mathbb{1}_u \frac{t}{n} & 2\sum_{t=\alpha_{i-1}}^{\alpha_i} \mathbb{1}_u \frac{t^2}{n^2} & 2\sum_{t=\alpha_{i-1}}^{\alpha_i} \mathbb{1}_u \frac{t^3}{n^3} & -1 & 0 \\ 2\sum_{t=\alpha_{i-1}}^{\alpha_i} \mathbb{1}_u \frac{t}{n} & 2\sum_{t=\alpha_{i-1}}^{\alpha_i} \mathbb{1}_u \frac{t^2}{n^2} & 2\sum_{t=\alpha_{i-1}}^{\alpha_i} \mathbb{1}_u \frac{t^3}{n^3} & 2\sum_{t=\alpha_{i-1}}^{\alpha_i} \mathbb{1}_u \frac{t^4}{n^4} & -\frac{\alpha_i}{n} & -\frac{1}{n} \\ 2\sum_{t=\alpha_{i-1}}^{\alpha_i} \mathbb{1}_u \frac{t^2}{n^2} & 2\sum_{t=\alpha_{i-1}}^{\alpha_i} \mathbb{1}_u \frac{t^3}{n^3} & 2\sum_{t=\alpha_{i-1}}^{\alpha_i} \mathbb{1}_u \frac{t^4}{n^4} & 2\sum_{t=\alpha_{i-1}}^{\alpha_i} \mathbb{1}_u \frac{t^5}{n^5} & -\frac{\alpha_i^2}{n^2} & -2\frac{\alpha_i}{n^2} \\ 2\sum_{t=\alpha_{i-1}}^{\alpha_i} \mathbb{1}_u \frac{t^3}{n^3} & 2\sum_{t=\alpha_{i-1}}^{\alpha_i} \mathbb{1}_u \frac{t^4}{n^4} & 2\sum_{t=\alpha_{i-1}}^{\alpha_i} \mathbb{1}_u \frac{t^5}{n^5} & 2\sum_{t=\alpha_{i-1}}^{\alpha_i} \mathbb{1}_u \frac{t^6}{n^6} & -\frac{\alpha_i^3}{n^3} & -3\frac{\alpha_i^2}{n^3} \\ -1 & -\frac{\alpha_i}{n} & -\frac{\alpha_i^2}{n^2} & -\frac{\alpha_i^3}{n^3} & 0 & 0 \\ 0 & -\frac{1}{n} & -2\frac{\alpha_i}{n^2} & -3\frac{\alpha_i^2}{n^3} & 0 & 0 \end{pmatrix} \begin{pmatrix} H_0^i(u,v) \\ H_1^i(u,v) \\ H_2^i(u,v) \\ H_3^i(u,v) \\ \lambda_i \\ \gamma_i \end{pmatrix} = B_r$$

avec $B_r^t = \begin{pmatrix} 2\sum_{t=\alpha_{i-1}}^{\alpha_i} \mathbb{1}_{uv} & 2\sum_{t=\alpha_{i-1}}^{\alpha_i} \mathbb{1}_{uv}\frac{t}{n} & 2\sum_{t=\alpha_{i-1}}^{\alpha_i} \mathbb{1}_{uv}\frac{t^2}{n^2} & 2\sum_{t=\alpha_{i-1}}^{\alpha_i} \mathbb{1}_{uv}\frac{t^3}{n^3} & -\Pi_{\frac{t}{n}}^{i+1}(\alpha_i) & -\Pi_{\frac{t}{n}}^{i+1\,\prime}(\alpha_i) \end{pmatrix}$.

La stochasticité des matrices $\Pi_{\frac{t}{n}}^i(u,v)$ est vérifiée. La même remarque que précédemment (voir A.1.2) peut s'appliquer : il est théoriquement possible d'obtenir des termes négatifs.

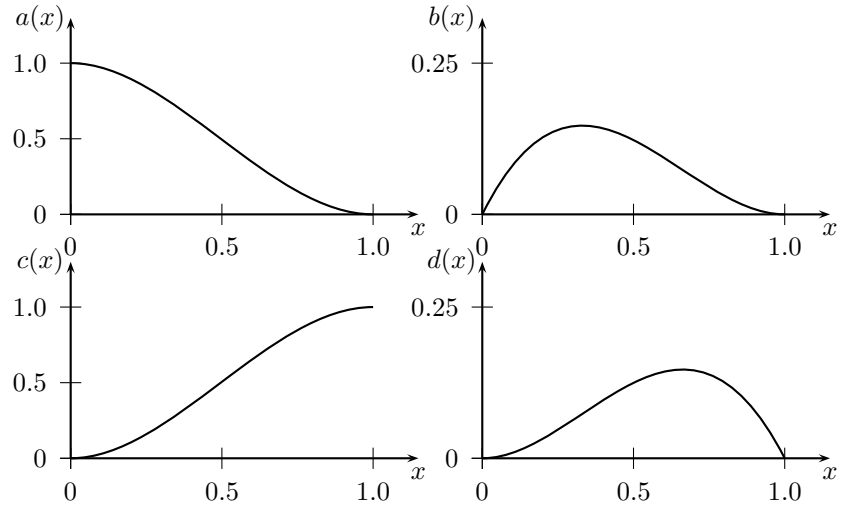
3.3.2 Fonctions de base

Tout polynôme de degré 3 s'exprime comme combinaison linéaire des 4 fonctions de base a, b, c et d définies dans le Tableau 3.1 et choisies de façon à vérifier les propriétés de ce même tableau. Elles sont représentées à la Figure 3.3).

TAB. 3.1 – Les fonctions de bases et leurs propriétés

$a(x) = 2x^3 - 3x^2 + 1$	telle que	$a(0) = 1$, $a'(0) = 0$, $a(1) = 0$ et $a'(1) = 0$;
$b(x) = x^3 - 2x^2 + x$	telle que	$b(0) = 0$, $b'(0) = 1$, $b(1) = 0$ et $b'(1) = 0$;
$c(x) = -2x^3 + 3x^2$	telle que	$c(0) = 0$, $c'(0) = 0$, $c(1) = 1$ et $c'(1) = 0$;
$d(x) = -x^3 + x^2$	telle que	$d(0) = 0$, $d'(0) = 0$, $d(1) = 0$ et $d'(1) = -1$.

FIG. 3.3 – Les fonctions de base des polynômes de degré 3



Ainsi, à l'aide de ces fonctions de base, les matrices $\Pi^i_{\frac{t}{n}}$ de notre modèle s'expriment sous la forme suivante :

$$\Pi^i_{\frac{t}{n}} = A_i a\left(\frac{t-\alpha_{i-1}}{\alpha_i-\alpha_{i-1}}\right) + B_i b\left(\frac{t-\alpha_{i-1}}{\alpha_i-\alpha_{i-1}}\right) + C_i c\left(\frac{t-\alpha_{i-1}}{\alpha_i-\alpha_{i-1}}\right) + D_i d\left(\frac{t-\alpha_{i-1}}{\alpha_i-\alpha_{i-1}}\right).$$

Nous posons les contraintes de continuité, pour $i = 1 \dots N-1$:
- $\Pi^i_{\frac{\alpha_i}{n}} = \Pi^{i+1}_{\frac{\alpha_i}{n}}$;
- ${\Pi^i_{\frac{\alpha_i}{n}}}' = {\Pi^{i+1}_{\frac{\alpha_i}{n}}}'$.

Ces contraintes permettent les simplifications suivantes :
- $C_i = A_{i+1}$ pour chaque i entre 1 et $N-1$,
- $D_i = \frac{\alpha_{i-1}-\alpha_i}{\alpha_{i+1}-\alpha_i} B_{i+1}$ pour chaque i entre 1 et $N-1$,

En effet, au regard du tableau 3.1, nous avons

$$\begin{aligned}
& \Pi^i_{\frac{\alpha_i}{n}} = \Pi^{i+1}_{\frac{\alpha_i}{n}} \\
\iff\ & A_i a(1) + B_i b(1) + C_i c(1) + D_i d(1) = A_{i+1} a(0) + B_{i+1} b(0) + C_{i+1} c(0) + D_{i+1} d(0) \\
\iff\ & C_i = A_{i+1}, \\
& {\Pi^i_{\frac{\alpha_i}{n}}}' = {\Pi^{i+1}_{\frac{\alpha_i}{n}}}' \\
\iff\ & A_i a'(1) + B_i b'(1) + C_i c'(1) + D_i d'(1) = \frac{\alpha_i - \alpha_{i-1}}{\alpha_{i+1} - \alpha_i} \left(A_{i+1} a'(0) + B_{i+1} b'(0) + C_{i+1} c'(0) + D_{i+1} d'(0)\right) \\
\iff\ & D_i = \frac{\alpha_{i-1} - \alpha_i}{\alpha_{i+1} - \alpha_i} B_{i+1}.
\end{aligned}$$

Remarque 8 *Si α_i sont choisis équirépartis, alors $D_i = -B_{i+1}$.*

Comme de plus nous pouvons réduire les 4 fonctions de bases à seulement 2 (car $c(x) = a(1-x)$ et $d(x) = b(1-x)$), le modèle se simplifie. Pour $i = 1 \dots N-1$:

$$\begin{aligned}
\Pi^i_{\frac{t}{n}} &= A_i a\left(\frac{t-\alpha_{i-1}}{\alpha_i - \alpha_{i-1}}\right) + B_i b\left(\frac{t-\alpha_{i-1}}{\alpha_i - \alpha_{i-1}}\right) + A_{i+1} a\left(\frac{\alpha_i - t}{\alpha_i - \alpha_{i-1}}\right) + \frac{\alpha_{i-1}-\alpha_i}{\alpha_{i+1}-\alpha_i} B_{i+1} b\left(\frac{\alpha_i - t}{\alpha_i - \alpha_{i-1}}\right) \\
\Pi^N_{\frac{t}{n}} &= A_N a\left(\frac{t-\alpha_{N-1}}{\alpha_N - \alpha_{N-1}}\right) + B_N b\left(\frac{t-\alpha_{N-1}}{\alpha_N - \alpha_{N-1}}\right) + C_N a\left(\frac{\alpha_N - t}{\alpha_N - \alpha_{N-1}}\right) + D_N b\left(\frac{\alpha_N - t}{\alpha_N - \alpha_{N-1}}\right)
\end{aligned}$$

Dans le but de simplifier les expressions, nous posons :
- $\kappa_i = \dfrac{t-\alpha_{i-1}}{\alpha_i - \alpha_{i-1}}$
- $\delta_i = \dfrac{\alpha_i - t}{\alpha_i - \alpha_{i-1}}$
- $\nu_i = \left(\dfrac{\alpha_{i-1}-\alpha_i}{\alpha_{i+1}-\alpha_i}\right)$
- $A_{N+1} = C_N$
- $B_{N+1} = D_N$.

Les fonctions à minimiser f_i pour i allant de 1 à $N-1$ sont :

$$\begin{aligned}
f_1(A_1(u,v), B_1(u,v), A_2(u,v), B_2(u,v)) &= \sum_{t=k}^{\alpha_1} \left[\sum_{u \in \mathcal{A}^k} \mathbb{1}_u \sum_{v \in \mathcal{A}} \left(\Pi^1_{\frac{t}{n}}(u,v) - \mathbb{1}_{uv} \right)^2 \right] \\
f_i(A_{i+1}(u,v), B_{i+1}(u,v)) &= \sum_{t=\alpha_{i-1}}^{\alpha_i} \left[\sum_{u \in \mathcal{A}^k} \mathbb{1}_u \sum_{v \in \mathcal{A}} \left(\Pi^i_{\frac{t}{n}}(u,v) - \mathbb{1}_{uv} \right)^2 \right]
\end{aligned}$$

Explicitons les dérivées pour i allant de 1 à $N-1$:

$$\begin{aligned}
\frac{\partial f_1(A_1(u,v), B_1(u,v), A_2(u,v), B_2(u,v))}{\partial A_1(u,v)} = 0 &\iff \sum_{t=k}^{m_0} \mathbb{1}_u a(\kappa_1) \left(\Pi^1_{\frac{t}{n}}(u,v) - \mathbb{1}_{uv} \right) = 0 \\
\frac{\partial f_1(A_1(u,v), B_1(u,v), A_2(u,v), B_2(u,v))}{\partial B_1(u,v)} = 0 &\iff \sum_{t=k}^{m_0} \mathbb{1}_u b(\kappa_1) \left(\Pi^1_{\frac{t}{n}}(u,v) - \mathbb{1}_{uv} \right) = 0 \\
\frac{\partial f_1(A_1(u,v), B_1(u,v), A_2(u,v), B_2(u,v))}{\partial A_2(u,v)} = 0 &\iff \sum_{t=k}^{m_0} \mathbb{1}_u a(\delta_1) \left(\Pi^1_{\frac{t}{n}}(u,v) - \mathbb{1}_{uv} \right) = 0 \\
\frac{\partial f_1(A_1(u,v), B_1(u,v), A_2(u,v), B_2(u,v))}{\partial B_2(u,v)} = 0 &\iff \sum_{t=k}^{m_0} \mathbb{1}_u b(\delta_1) \left(\Pi^1_{\frac{t}{n}}(u,v) - \mathbb{1}_{uv} \right) = 0 \\
\frac{\partial f_i(A_{i+1}(u,v), B_{i+1}(u,v))}{\partial A_{i+1}(u,v)} = 0 &\iff \sum_{t=\alpha_{i-1}+1}^{\alpha_i} \mathbb{1}_u a(\delta_i) \left(\Pi^1_{\frac{t}{n}}(u,v) - \mathbb{1}_{uv} \right) = 0 \\
\frac{\partial f_i(A_{i+1}(u,v), B_{i+1}(u,v))}{\partial B_{i+1}(u,v)} = 0 &\iff \sum_{t=\alpha_{i-1}+1}^{\alpha_i} \mathbb{1}_u b(\delta_i) \left(\Pi^1_{\frac{t}{n}}(u,v) - \mathbb{1}_{uv} \right) = 0
\end{aligned}$$

Comme précédemment, l'initialisation de la récurrence s'effectue sur les deux premiers segments en même temps. Puis la récurrence se poursuit segment par segment. La minimisation nous donne pour chaque couple (u,v) un système 4×4 d'initialisation, N systèmes 2×2 pour l'aller et N systèmes 4×4 pour le retour :

INITIALISATION : Segment 1.
Estimation de A_1, B_1, A_2 et B_2 :

$$\begin{pmatrix} \sum_{t=k}^{\alpha_1} \mathbb{1}_u a(\kappa_1)a(\kappa_1) & \sum_{t=k}^{\alpha_1} \mathbb{1}_u a(\kappa_1)b(\kappa_1) & \sum_{t=k}^{\alpha_1} \mathbb{1}_u a(\kappa_1)a(\delta_1) & \nu_1 \sum_{t=k}^{\alpha_1} \mathbb{1}_u a(\kappa_1)b(\delta_1) \\ \sum_{t=k}^{\alpha_1} \mathbb{1}_u b(\kappa_1)a(\kappa_1) & \sum_{t=k}^{\alpha_1} \mathbb{1}_u b(\kappa_1)b(\kappa_1) & \sum_{t=k}^{\alpha_1} \mathbb{1}_u b(\kappa_1)a(\delta_1) & \nu_1 \sum_{t=k}^{\alpha_1} \mathbb{1}_u b(\kappa_1)b(\delta_1) \\ \sum_{t=k}^{\alpha_1} \mathbb{1}_u a(\delta_1)a(\kappa_1) & \sum_{t=k}^{\alpha_1} \mathbb{1}_u a(\delta_1)b(\kappa_1) & \sum_{t=k}^{\alpha_1} \mathbb{1}_u a(\delta_1)a(\delta_1) & \nu_1 \sum_{t=k}^{\alpha_1} \mathbb{1}_u a(\delta_1)b(\delta_1) \\ \sum_{t=k}^{\alpha_1} \mathbb{1}_u b(\delta_1)a(\kappa_1) & \sum_{t=k}^{\alpha_1} \mathbb{1}_u b(\delta_1)b(\kappa_1) & \sum_{t=k}^{\alpha_1} \mathbb{1}_u b(\delta_1)a(\delta_1) & \nu_1 \sum_{t=k}^{\alpha_1} \mathbb{1}_u b(\delta_1)b(\delta_1) \end{pmatrix} \begin{pmatrix} A_1(u,v) \\ B_1(u,v) \\ A_2(u,v) \\ B_2(u,v) \end{pmatrix} =$$

$$\begin{pmatrix} \sum_{t=k}^{\alpha_1} \mathbb{1}_{uv} a(\kappa_1) \\ \sum_{t=k}^{\alpha_1} \mathbb{1}_{uv} b(\kappa_1) \\ \sum_{t=k}^{\alpha_1} \mathbb{1}_{uv} a(\delta_1) \\ \sum_{t=k}^{\alpha_1} \mathbb{1}_{uv} b(\delta_1) \end{pmatrix}.$$

ALLER : Segment i pour i allant de 2 à N : de A_2, B_2 à C_N, D_N.

Estimation de A_2 et B_2 :

$$\begin{pmatrix} \sum_{t=k}^{\alpha_1} \mathbb{1}_u a(\delta_1)a(\delta_1) & \nu_1 \sum_{t=k}^{\alpha_1} \mathbb{1}_u a(\delta_1)b(\delta_1) \\ \sum_{t=k}^{\alpha_1} \mathbb{1}_u b(\delta_1)a(\delta_1) & \nu_1 \sum_{t=k}^{\alpha_1} \mathbb{1}_u b(\delta_1)b(\delta_1) \end{pmatrix} \begin{pmatrix} A_2(u,v) \\ B_2(u,v) \end{pmatrix} =$$

$$\begin{pmatrix} \sum_{t=k}^{\alpha_1} \mathbb{1}_{uv} a(\delta_1) - \sum_{t=k}^{\alpha_1} \mathbb{1}_u a(\delta_1)a(\kappa_1)A_1(u,v) - \sum_{t=k}^{\alpha_1} \mathbb{1}_u a(\delta_1)b(\kappa_1)B_1(u,v) \\ \sum_{t=k}^{\alpha_1} \mathbb{1}_{uv} b(\delta_1) - \sum_{t=k}^{\alpha_1} \mathbb{1}_u b(\delta_1)a(\kappa_1)A_1(u,v) - \sum_{t=k}^{\alpha_1} \mathbb{1}_u b(\delta_1)b(\kappa_1)B_1(u,v) \end{pmatrix}.$$

Estimation des A_i et B_i pour i allant de 3 à N :

$$\begin{pmatrix} \sum_{t=\alpha_{i-1}+1}^{\alpha_i} \mathbb{1}_u a(\delta_i)a(\delta_i) & \nu_i \sum_{t=\alpha_{i-1}+1}^{\alpha_i} \mathbb{1}_u a(\delta_i)b(\delta_i) \\ \sum_{t=\alpha_{i-1}+1}^{\alpha_i} \mathbb{1}_u b(\delta_i)a(\delta_i) & \nu_i \sum_{t=\alpha_{i-1}+1}^{\alpha_i} \mathbb{1}_u b(\delta_i)b(\delta_i) \end{pmatrix} \begin{pmatrix} A_{i+1}(u,v) \\ B_{i+1}(u,v) \end{pmatrix} =$$

$$\begin{pmatrix} \sum_{t=\alpha_{i-1}+1}^{\alpha_i} \mathbb{1}_{uv} a(\delta_i) - \sum_{t=\alpha_{i-1}+1}^{\alpha_i} \mathbb{1}_u a(\delta_i)a(\kappa_i)A_i(u,v) - \sum_{t=\alpha_{i-1}+1}^{\alpha_i} \mathbb{1}_u a(\delta_i)b(\kappa_i)B_i(u,v) \\ \sum_{t=\alpha_{i-1}+1}^{\alpha_i} \mathbb{1}_{uv} b(\delta_i) - \sum_{t=\alpha_{i-1}+1}^{\alpha_i} \mathbb{1}_u b(\delta_i)a(\kappa_i)A_i(u,v) - \sum_{t=\alpha_{i-1}+1}^{\alpha_i} \mathbb{1}_u b(\delta_i)b(\kappa_i)B_i(u,v) \end{pmatrix}.$$

Estimation de C_N et D_N :

$$\begin{pmatrix} \sum_{t=\alpha_{N-1}+1}^{\alpha_N} \mathbb{1}_u a(\delta_N)a(\delta_N) & \sum_{t=\alpha_{N-1}+1}^{\alpha_N} \mathbb{1}_u a(\delta_N)b(\delta_N) \\ \sum_{t=\alpha_{N-1}+1}^{\alpha_N} \mathbb{1}_u b(\delta_N)a(\delta_N) & \sum_{t=\alpha_{N-1}+1}^{\alpha_N} \mathbb{1}_u b(\delta_N)b(\delta_N) \end{pmatrix} \begin{pmatrix} C_N(u,v) \\ D_N(u,v) \end{pmatrix} =$$

$$\begin{pmatrix} \sum_{t=\alpha_{N-1}+1}^{\alpha_N} \mathbb{1}_{uv} a(\delta_N) - \sum_{t=\alpha_{N-1}+1}^{\alpha_N} \mathbb{1}_u a(\delta_N) a(\kappa_N) A_N(u,v) - \sum_{t=\alpha_{N-1}+1}^{\alpha_N} \mathbb{1}_u a(\delta_N) b(\kappa_N) B_N(u,v) \\ \sum_{t=\alpha_{N-1}+1}^{\alpha_N} \mathbb{1}_{uv} b(\delta_N) - \sum_{t=\alpha_{N-1}+1}^{\alpha_N} \mathbb{1}_u b(\delta_N) a(\kappa_N) A_N(u,v) - \sum_{t=\alpha_{N-1}+1}^{\alpha_N} \mathbb{1}_u b(\delta_N) b(\kappa_N) B_N(u,v) \end{pmatrix}.$$

RETOUR : Segment i pour i allant de N à 1 : de A_N, B_N à A_1, B_1.

Estimation de A_N et B_N :

$$\begin{pmatrix} \sum_{t=\alpha_{N-1}+1}^{\alpha_N} \mathbb{1}_u a(\kappa_N) a(\kappa_N) & \sum_{t=\alpha_{N-1}+1}^{\alpha_N} \mathbb{1}_u a(\kappa_N) b(\kappa_N) \\ \sum_{t=\alpha_{N-1}+1}^{\alpha_N} \mathbb{1}_u b(\kappa_N) a(\kappa_N) & \sum_{t=\alpha_{N-1}+1}^{\alpha_N} \mathbb{1}_u b(\kappa_N) b(\kappa_N) \end{pmatrix} \begin{pmatrix} A_N(u,v) \\ B_N(u,v) \end{pmatrix} =$$

$$\begin{pmatrix} \sum_{t=\alpha_{N-1}+1}^{\alpha_N} \mathbb{1}_{uv} a(\kappa_N) - \sum_{t=\alpha_{N-1}+1}^{\alpha_N} \mathbb{1}_u a(\kappa_N) a(\delta_N) C_N(u,v) - \sum_{t=\alpha_{N-1}+1}^{\alpha_N} \mathbb{1}_u a(\kappa_N) b(\delta_N) D_N(u,v) \\ \sum_{t=\alpha_{N-1}+1}^{\alpha_N} \mathbb{1}_{uv} b(\kappa_N) - \sum_{t=\alpha_{N-1}+1}^{\alpha_N} \mathbb{1}_u b(\kappa_N) a(\delta_N) C_N(u,v) - \sum_{t=\alpha_{N-1}+1}^{\alpha_N} \mathbb{1}_u b(\kappa_N) b(\delta_N) D_N(u,v) \end{pmatrix}.$$

Estimation des A_i et B_i pour i allant de $N-1$ à 2 :

$$\begin{pmatrix} \sum_{t=\alpha_{i-1}+1}^{\alpha_i} \mathbb{1}_u a(\kappa_i) a(\kappa_i) & \sum_{t=\alpha_{i-1}+1}^{\alpha_i} \mathbb{1}_u a(\kappa_i) b(\kappa_i) \\ \sum_{t=\alpha_{i-1}+1}^{\alpha_i} \mathbb{1}_u b(\kappa_i) a(\kappa_i) & \sum_{t=\alpha_{i-1}+1}^{\alpha_i} \mathbb{1}_u b(\kappa_i) b(\kappa_i) \end{pmatrix} \begin{pmatrix} A_i(u,v) \\ B_i(u,v) \end{pmatrix} =$$

$$\begin{pmatrix} \sum_{t=\alpha_{i-1}+1}^{\alpha_i} \mathbb{1}_{uv} a(\kappa_i) - \sum_{t=\alpha_{i-1}+1}^{\alpha_i} \mathbb{1}_u a(\kappa_i) a(\delta_i) A_{i+1}(u,v) - \nu_i \sum_{t=\alpha_{i-1}+1}^{\alpha_i} \mathbb{1}_u a(\kappa_i) b(\delta_i) B_{i+1}(u,v) \\ \sum_{t=\alpha_{i-1}+1}^{\alpha_i} \mathbb{1}_{uv} b(\kappa_i) - \sum_{t=\alpha_{i-1}+1}^{\alpha_i} \mathbb{1}_u b(\kappa_i) a(\delta_i) A_{i+1}(u,v) - \nu_i \sum_{t=\alpha_{i-1}+1}^{\alpha_i} \mathbb{1}_u b(\kappa_i) b(\delta_i) B_{i+1}(u,v) \end{pmatrix}.$$

Estimation de A_1 et B_1 :

$$\begin{pmatrix} \sum_{t=k}^{\alpha_1} \mathbb{1}_u a(\kappa_1) a(\kappa_1) & \sum_{t=k}^{\alpha_1} \mathbb{1}_u a(\kappa_1) b(\kappa_1) \\ \sum_{t=k}^{\alpha_1} \mathbb{1}_u b(\kappa_1) a(\kappa_1) & \sum_{t=k}^{\alpha_1} \mathbb{1}_u b(\kappa_1) b(\kappa_1) \end{pmatrix} \begin{pmatrix} A_1(u,v) \\ B_1(u,v) \end{pmatrix} =$$

$$\begin{pmatrix} \sum_{t=k}^{\alpha_1} \mathbb{1}_{uv} a(\kappa_1) - \sum_{t=k}^{\alpha_1} \mathbb{1}_u a(\kappa_1) a(\delta_1) A_2(u,v) - \nu_1 \sum_{t=k}^{\alpha_1} \mathbb{1}_u a(\kappa_1) b(\delta_1) B_2(u,v) \\ \sum_{t=k}^{\alpha_1} \mathbb{1}_{uv} b(\kappa_1) - \sum_{t=k}^{\alpha_1} \mathbb{1}_u b(\kappa_1) a(\delta_1) A_2(u,v) - \nu_1 \sum_{t=k}^{\alpha_1} \mathbb{1}_u b(\kappa_1) b(\delta_1) B_2(u,v) \end{pmatrix}.$$

Théorème 13 *Les matrices $\Pi^i_{\frac{t}{n}}$ sont stochastiques pour i allant de 1 à N.*

Preuve *Sachant que $a(\kappa_i)+a(\delta_i)=1$, nous avons pour tout i allant de 1 à $N+1$, $\sum_v A_i(u,v)=1$, $\sum_v B_i(u,v)=0$ solutions des systèmes dans lequel nous sommons sur tous les $v \in \mathcal{A}$. Ainsi la stochasticité des matrices $\Pi^i_{\frac{t}{n}}$ est vérifiée. La même remarque que précédemment (voir A.1.2) peut s'appliquer : il est théoriquement possible d'obtenir des termes négatifs.*

3.4 Perspectives

Nous verrons dans la partie suivante (Partie II) que les DMM par splines ne donnent pas les résultats escomptés. Afin d'améliorer ceci, plusieurs solutions sont envisageables. Il s'agit d'adapter à notre cas les différentes splines définies dans la littérature :

- Une fonction spline se présente au degré 3 sous la forme suivante :

$$s(x) = m_0 + m_1 x + m_2 x^2 + m_3 x^3 + \sum_{j=1}^{K} \theta_j (x - \alpha_j)_+^3$$

  où $m_0, m_1, m_2, m_3, \theta_j$ sont des constantes, α_j la position des noeuds et où le dernier terme (partie positive de $(x-\alpha_j)^3$) permet de satisfaire aux conditions de continuité. On parle alors de splines polynomiales. C'est celles que nous avons utilisées.
- Les splines naturelles (*natural splines*) en sont une variante assez proche. Une contrainte supplémentaire est imposée : au delà des noeuds extrêmes, la fonction doit être linéaire. Ainsi au voisinage de ces noeuds, la dérivée seconde et la dérivée d'ordre 3 doit être nulle. La fonction est ainsi moins libre, mais la variance des estimations aux extrémités est stabilisée (alors qu'elle est élevée dans le cas des splines polynomiales classiques).
- Les splines pénalisées (*penalized splines*) permettent de résoudre le problème de la localisation des noeuds qui apparaissent par construction au niveau des quantiles de la distribution des observations. Ces splines étant très sensibles à la position des noeuds, la solution est de disposer de nombreux noeuds puis de supprimer les noeuds inutiles au vu de la pénalité.
- Les splines de lissage (*cubic smoothing splines*) minimisent la somme des carrés des résidus (comme classiquement) mais cette somme est pénalisée par un terme représentant la courbure de la fonction :

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_a^b (f''(t))^2 dt$$

  f est la fonction spline recherchée, λ un paramètre de lissage, a et b sont tels que $a \leq x_1 \leq \ldots \leq x_n \leq b$. Plus λ est élevé, plus le lissage est important.

Deuxième partie

DRIMM et Applications

Chapitre 4

DRIMM

Nous présentons dans ce chapitre le logiciel que nous avons mis au point pour l'estimation des DMM : **DRIMM** pour **DRI**FTING MARKOV MODEL.

Écrit en `C++` et développé sur un système x86 GNU/Linux avec GCC 3.4, ce software a été testé avec succès pour les dernières versions de GCC sur les systèmes Sun et Apple Mac OSX. Il repose sur la librairie `seq++` ([MBR+05]) et y sera bientôt intégré, au moins en partie. Le reste du programme sera disponible à l'adresse http://stat.genopole.cnrs.fr/software/drimm. La compilation et l'installation sont en accord avec la procédure GNU standard. Ce software est sous licence : la GNU General Public License (http://www.gnu.org).

4.1 Estimations

DRIMM permet l'estimation de différents modèles de Markov régulés :

- DMM avec dérive polynomiale : `-Poly` pour le modèle des points d'appui, `-PolyM` pour le modèle des polynômes ;
- DMM par splines polynomiales : `-SpM` pour l'estimation globale (voir I.3.2), `-SpH` pour l'estimation sans fonction de base (voir I.3.3.1), `-Spl` pour l'estimation avec fonctions de base (voir I.3.3.2) .

L'estimation se fait sur une ou plusieurs séquences (`-fs`), un jeu de séquences de même taille (`-fsm`) ou de tailles éventuellement différentes (`-fss`).

Dans le cas d'une estimation sur une ou plusieurs séquences, un modèle est estimé pour chacune des séquences par une des méthodes explicitées en Partie I. Le ou les modèles sont renvoyés dans un ou plusieurs fichiers `config_<num>.out` où `<num>` désigne le numéro de la séquence dans le fichier (voir exemple 9).

Dans le cas d'une estimation sur un jeu de *nbs* séquences de même taille, nous estimons un modèle en moyennant les modèles de chaque séquence :

$$\Pi_{\frac{t}{n}} = \frac{\sum_{i=1}^{nbs} \Pi^i_{\frac{t}{n}}}{nbs},$$

où les $\Pi^i_{\frac{t}{n}}$ sont les matrices dérivantes de chacune des séquences, estimées comme à l'accoutumée en minimisant

$$\sum_{t=k}^{n} \left[\sum_{u \in \mathcal{A}^k} \mathbb{1}\{X_{t-k} \dots X_{t-1} = u\} \left[\sum_{v \in \mathcal{A}} \left(\Pi_{\frac{t}{n}}(u,v) - \mathbb{1}_{\{X_t = v\}} \right)^2 \right] \right].$$

Le modèle est renvoyé dans un fichier `config_sm.out` (voir exemple 9).

Dans le cas de séquences de tailles éventuellement différentes, nous estimons un modèle global en minimisant sur toutes les séquences en même temps :

$$\sum_{i=1}^{nbs} \left[\sum_{t=k}^{n} \left[\sum_{u \in \mathcal{A}^k} \mathbb{1}\{X_{t-k} \dots X_{t-1} = u\}(i) \left[\sum_{v \in \mathcal{A}} \left(\Pi_{\frac{t}{n}}(u,v) - \mathbb{1}_{\{X_t = v\}}(i) \right)^2 \right] \right] \right],$$

où $\mathbb{1}\{X_{t-k} \dots X_{t-1} = u\}(i)$ et $\mathbb{1}_{\{X_t = v\}}(i)$ représente l'apparition de u et v dans la séquence i. Le modèle est renvoyé dans un fichier `config_ss.out` (voir exemple 9).

Exemple 9 *Présentons le fichier de configuration pour une estimation polynomiale (avec points d'appui) de degré* 2 *et d'ordre* 1 *sur le phage lambda :*

```
0
4
48502
1
2

0
0.263741 0.275344 0.262339 0.198576
0.255933 0.288868 0.308969 0.146229
0.210733 0.37116 0.258175 0.159933
0.120091 0.487338 0.201123 0.191447

0.5
0.29305 0.215561 0.204394 0.286994
0.242379 0.244492 0.276049 0.23708
0.297588 0.253298 0.212724 0.23639
0.185585 0.289102 0.230062 0.295251

1
0.347986 0.201688 0.183321 0.267006
0.306111 0.203159 0.268609 0.222121
0.307255 0.245698 0.203744 0.243304
0.207328 0.299651 0.213196 0.279824
```

La première ligne correspond à la méthode utilisée :

- 0 *pour* `-Poly` ;
- 1 *pour* `-SpM` ;
- 2 *pour* `-SpH` ;
- 3 *pour* `-Spl` ;
- 4 *pour* `-PolyM`.

La seconde ligne correspond à la taille de l'alphabet, la troisième à la taille de la séquence, la quatrième à l'ordre du modèle et la cinquième au degré de la dérive. Viennent ensuite les trois matrices de transition points d'appui Π_0, $\Pi_{0.5}$ *et* Π_1.

Ajoutant les paramètres des modèles, les lignes de commandes pour l'estimation d'un DMM sont :

- `DRIMM -Poly+ -fs+ <s> -fa <a> -order <k>  -deg <d> -cf`
- `DRIMM -Spl -fs+ <s> -fa <a> -order <k> -deg <d> -N <N> -cf`
- `DRIMM -Sp+ -fs+ <s> -fa <a> -order <k> -N <N> -Nar <Nar> -cf`

où

- `-Poly+` signifie `-Poly` ou `-PolyM` ;
- `-fs+` signifie `-fs`, `fsm` ou `fss` ;
- `-Sp+` signifie `-SpM` ou `-SpH` ;
- `<s>` est le fichier contenant une ou plusieurs séquences au format FASTA (voir Exemple 10) ;
- `<a>` est le fichier contenant l'alphabet (voir Exemple 11) ;
- `<k>` est l'ordre du modèle de Markov ;
- `<d>` est le degré de la dérive polynomiale ;
- `<N>` est le nombre de segments dans le cas d'une estimation par splines polynomiales ;
- `<Nar>` est le nombre d'aller-retour dans le cas d'une estimation non-globale par splines polynomiales ;
- `-cf` permet de renvoyer le ou les fichiers de configuration correspondant au modèle estimé.

Exemple 10 *Un fichier fasta se compose d'une ligne commençant par > et décrivant la séquence (en général son nom), puis de la séquence elle-même. Il est possible de définir plusieurs séquences dans le même fichier FASTA. Nous en présentons un exemple pour deux séquences protéiques :*

```
>SEQUENCE_1
MTEITAAMVKELRESTGAGMMDCKNALSETNGDFDKAVQLLREKGLGKAAKKADRLAAEG
LVSVKVSDDFTIAAMRPSYLSYEDLDMTFVENEYKALVAELEKENEERRRLKDPNKPEHK
IPQFASRKQLSDAILKEAEEKIKEELKAQGKPEKIWDNIIPGKMNSFIADNSQLDSKLTL
MGQFYVMDDKKTVEQVIAEKEKEFGGKIKIVEFICFEVGEGLEKKTEDFAAEVAAQL
>SEQUENCE_2
```

```
SATVSEINSETDFVAKNDQFIALTKDTTAHIQSNSLQSVEELHSSTINGVKFEEYLKSQI
ATIGENLVVRRFATLKAGANGVVNGYIHTNGRVGVVIAAACDSAEVASKSRDLLRQICMH
```

Exemple 11 *Un alphabet est défini dans un simple fichier de configuration (voir http://stat.genopole.cnrs.fr/seqpp/classTranslator.html). Nous donnons un exemple pour l'alphabet des nucléotides :*

```
#dna
a A : t T
g G : c C
c C : g G
t T : a A
? n N
```

4.2 Possibilités

Le programme permet non seulement d'estimer un modèle mais aussi de l'utiliser pour certaines applications. Ainsi pour utiliser un DMM à l'aide de **DRIMM**, nous pouvons :

- estimer un modèle ;
- entrer un modèle à l'aide d'un fichier de configuration (voir Exemple 9) et du paramètre `-model`.

Nous allons lister les principales possibilités offertes par `DRIMM` :

- `-L` : calcule la log-vraisemblance des séquences estimées ;
- `-AIC` : calcule l'AIC des séquences estimées ;
- `-BIC` : calcule le BIC des séquences estimées ;
- `-l <sequence_file_l>` : calcule la log-vraisemblance, sous le modèle donné (estimé ou non), des séquences du fichier `<sequence_file_l>` ;
- `-aic <sequence_file_aic>` : calcule l'AIC, sous le modèle donné (estimé ou non), des séquences du fichier `<sequence_file_aic>` ;
- `-bic <sequence_file_bic>` : calcule le BIC, sous le modèle donné (estimé ou non), des séquences du fichier `<sequence_file_bic>` ;
- `-law` : donne dans le fichier `trace_stat_<num>.out` (`<num>` désignant le numéro de la séquence), la loi stationnaire de chaque matrice de transition du DMM ;
- `-slaw <in>0> <out>0>` donne dans le fichier `trace_stat_segment_<num>.out`, la loi stationnaire des matrices de transition du DMM entre la position `<in>` et `<out>` ;
- `-dist` : donne dans le fichier `trace_dist_<num>.out` (`<num>` désignant le numéro de la séquence), les distributions de probabilité des nucléotides sous le modèle donné, en chaque position ;
- `-sdist <in>0> <out>0>` donne dans le fichier `trace_dist_segment_<num>.out`, les distributions de probabilité des nucléotides sous le modèle donné, entre la position `<in>` et `<out>` ;
- `-pi` : donne dans les fichiers `Pit_<num>.out`, `Pit_sm.out`, `Pit_ss.out` ou `Pit.out` les matrices de transition du modèle en chaque position.
- `-simu` : donne dans les fichiers `simulation_<num>.out`, `simulation_sm.out`, `simulation_ss.out` ou `simulation.out` une séquence simulée à l'aide du modèle donné (estimé ou non).

DRIMM offre ainsi un panel d'outils pour l'analyse de séquences biologiques. Nous allons ratisser un éventail de ces possibilités dans les prochains chapitres.

Chapitre 5

Validation des modèles de Markov régulés

5.1 Les différents modèles

5.1.1 Dérive polynomiale

Dans le cadre d'une dérive polynomiale (voir I.2), il existe quelques différences entre la méthode de régression matricielle et la méthode point par point. La méthode de régression matricielle utilise des estimations préliminaires sur chaque segment et les estimateurs globaux sont calculés sur un point unique de chaque segment (le centre τ_ℓ). La méthode point par point propose une estimation directe sur tous les points de la séquence. Nous utilisons la log-vraisemblance pour comparer ces deux méthodes d'estimations des DMM. Nous estimons nos modèles sur le génome complet du *phage Lambda* (voir [WT71]) et considérons ces modèles comme étant les vrais modèles. Ensuite, nous simulons une séquence avec chacun de ces modèles et calculons la log-vraisemblance pour les deux méthodes d'estimation (voir Tableau 5.1). Notons que quelque soit l'ordre ou le degré, la méthode point par point

TAB. 5.1 – Log-vraisemblances de DMM avec dérive polynomiale, sur une séquence simulée par chacun de ces modèles.

Degré		0	1	2	3	4	5
Ordre 0	Régression	−67191	−66999	−66962	−66910	−66909	−66907
	Point par point	**−67191**	**−66999**	**−66962**	**−66910**	**−66909**	**−66907**
Ordre 1	Régression	−66718	−66504	−66448	−66382	−66376	−66368
	Point par point	**−66710**	**−66501**	**−66445**	**−66380**	**−66374**	**−66366**
Ordre 2	Régression	−66706	−66482	−66407	−66321	−66295	−66275
	Point par point	**−66693**	**−66477**	**−66402**	**−66317**	**−66290**	**−66270**
Ordre 3	Régression	−66630	−66331	−66186	−66038	−65938	−65883
	Point par point	**−66612**	**−66320**	**−66169**	**−66014**	**−65898**	**−65817**

donne de meilleures vraisemblances que la méthode de régression matricielle. Cet effet est encore plus apparent lorsque nous calculons la log-vraisemblance sur la séquence réelle du *phage Lambda* (voir Tableau 5.2).

TAB. 5.2 – Log-vraisemblances de DMM avec dérive polynomiale, sur le génome du *phage Lambda*.

Degré		0	1	2	3	4	5
Ordre 0	Régression	−67191	−66973	−66934	−66873	−66760	−66680
	Point par point	**−67191**	**−66973**	**−66934**	**−66873**	**−66760**	**−66680**
Ordre 1	Régression	−66743	−66500	−66439	−66362	−66234	−66146
	Point par point	**−66714**	**−66483**	**−66419**	**−66345**	**−66220**	**−66135**
Ordre 2	Régression	−66052	−65657	−65577	−65438	−65281	−65160
	Point par point	**−66005**	**−65631**	**−65544**	**−65410**	**−65255**	**−65139**
Ordre 3	Régression	−65661	−65168	−65033	−64809	−64597	−64432
	Point par point	**−65579**	**−65116**	**−64951**	**−64746**	**−64497**	**−64329**

Ces résultats nous amènent à ne considérer pour la suite que la méthode d'estimation point par point.

5.1.2 Dérive par splines polynomiales

Le Tableau 5.3 compare les log-vraisemblances des modèles par splines polynomiales, selon la méthode d'estimation utilisée.

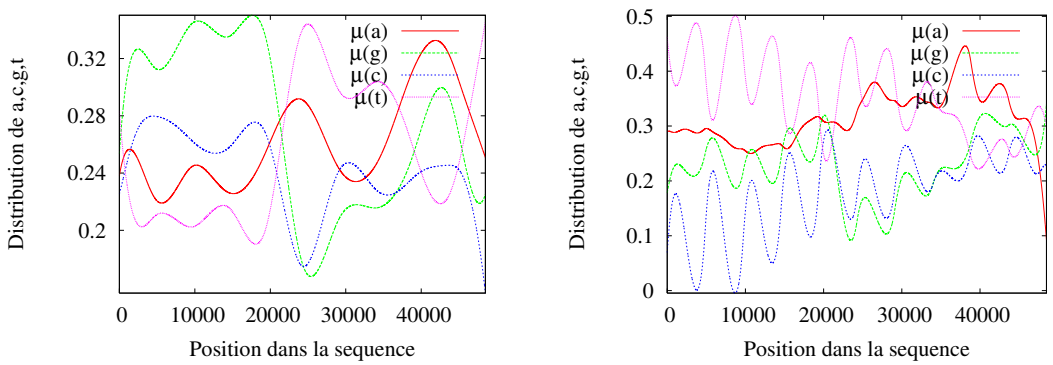
TAB. 5.3 – Log-vraisemblances de DMM par splines polynomiales, sur le génome du *phage Lambda* (DMM de degré 3 avec 1 aller-retour pour les méthodes non-globales).

Nombre de segments		2	3	4	5	10
Ordre 0	Globale	−66753	**−66681**	**−66658**	**−66661**	**−66543**
	Sans base	**−66665**	−67587	−72329	71979	−72680
	Bases	−67439	−68049	−70115	−68473	−83406
Ordre 1	Globale	−66213	**−66136**	**−66110**	**−66101**	**−65930**
	Sans base	**−66119**	−67429	−71213	−72184	−80817
	Bases	−66578	−68377	−76442	−70565	−780727
Ordre 2	Globale	−65244	**−65135**	**−65073**	**−65047**	**−64761**
	Sans base	**−65112**	−66843	−71439	−75769	−82707
	Bases	−67853	−69738	−75917	−72260	−79626
Ordre 3	Globale	−64486	**−64319**	**−64170**	**−64059**	**−63406**
	Sans base	**−64290**	−67234	−72830	−77288	−80974
	Bases	−70331	−70894	−76871	−75301	−80273

Remarque 9 *Les Tableaux 5.3 et 5.2 montrent que les modèles par splines polynomiales estimées avec la méthode globale ont de meilleures vraisemblances que les DMM avec dérive polynomiale.*

Nous constatons qu'excepté dans le cas d'un découpage en deux segments, la méthode d'estimation globale est la plus performante en terme de vraisemblance. Les méthodes par estimations successives, que ce soit avec ou sans fonction de base, s'éloignent de plus en plus de la séquence. Ce problème est directement lié à la récurrence et aux conditions imposées aux noeuds. En observant de plus près les formules d'estimation (voir I.3.3), nous constatons que nous nous réduisons à estimer entre deux noeuds un polynôme de degré 3 en connaissant sa valeur et la valeur de sa dérivée au premier noeud. Plus la récurrence progresse, plus la valeur de la dérivée augmente (afin de recoller aux données) et nous obtenons rapidement des fonctions très éloignées de la séquence. Observons ce phénomène sur la Figure 5.1 représentant les distributions μ des différents nucléotide : un seul aller-retour sur 10 segments suffit à constater la différence entre un DMM estimé globalement et un DMM estimé par récurrence.

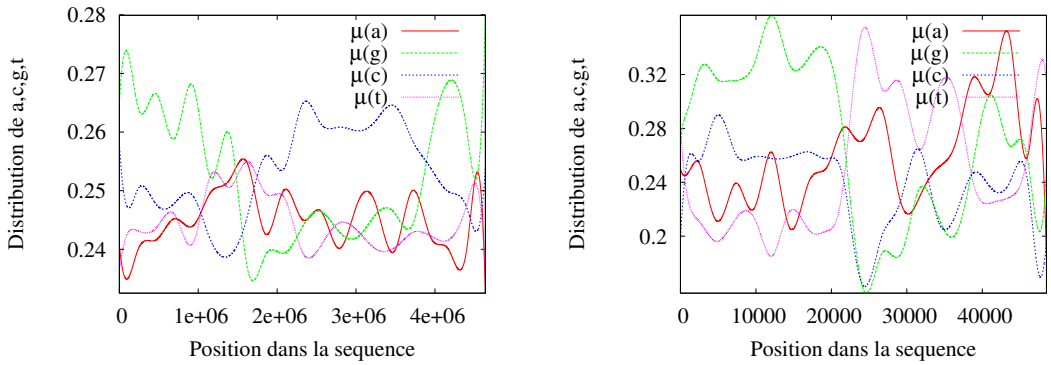
FIG. 5.1 – DMM par splines polynomiales (10 segments, ordre 1 et degré 3), sur le *phage Lambda* : estimation globale à gauche, estimation par récurrence à droite.



Remarque 10 *Le phénomène de divergence observé à la Figure 5.1 peut être contrôlé en imposant des bornes pour les valeurs aux noeuds de la dérivée, mais cela engendre une minimisation sous contraintes non élémentaire si nous voulons que les matrices soient stochastiques.*

La Figure 5.1 montre aussi le phénomène d'oscillations qui intervient lors de la modélisation par splines. Ce phénomène est lié au nombre de segments utilisés, quelque soit la méthode. La longueur des segments ou de la séquence ne modifie pas ce problème : une séquence de longueur 4639221 comme *Escherichia coli* ([BPB+97]) et une séquence de longueur 48502 comme le *phage Lambda*, divisées en 10 segments auront le même nombre d'oscillations (voir Figure 5.2).

FIG. 5.2 – Oscillations de DMM par splines polynomiales (20 segments, ordre 1 et degré 3), sur *Escherichia coli* (à gauche) et le *phage Lambda* (à droite).



Ces résultats nous amènent à ne considérer pour la suite que la méthode d'estimation globale.

5.1.3 Consistance des estimateurs

Nos estimateurs sont asymptotiquement sans biais et leurs variances tendent vers zéro. Ainsi nos estimateurs sont consistants. Afin de montrer cette consistance empiriquement, nous simulons quelques données où le vrai modèle est connu. Tout d'abord, nous estimons un modèle sur le génome complet du *phage Lambda* et considérons ce modèle comme le vrai modèle. Ensuite, nous simulons plusieurs séquences sous ce modèle et estimons un modèle moyen sur toutes ces séquences. Pour chaque paramètre du modèle, nous notons la valeur absolue de la différence entre le vrai paramètre et le paramètre estimé. Nous donnons dans les Tableaux 5.4 et 5.5 les moyennes de ces différences pour quelques DMM respectivement polynomiales et par splines, dans le but de montrer la consistance.

TAB. 5.4 – Comparaison entre les vrais modèles et les modèles estimés (dérive polynomiale). Nous donnons la moyenne des valeurs absolues des différences entre les vrais paramètres et les paramètres estimés. Le nombre de séquences simulées est donné par ns.

Degré		0	2	4	6
Ordre 1	$ns = 1$	0.0026691	0.00943604	0.00920864	0.00996372
	$ns = 10$	0.0018065	0.00323508	0.00381477	0.00338738
	$ns = 100$	0.0001906	0.00081519	0.00092941	0.00114088
Ordre 2	$ns = 1$	0.0356951	0.0499792	0.0484183	0.0549545
	$ns = 10$	0.0346129	0.0447842	0.0421838	0.0500320
	$ns = 100$	0.0333321	0.0442818	0.0406379	0.0482310

Pour chaque modèle (dérive polynomiale ou dérive par splines polynomiales), nous donnons un exemple contenant le vrai modèle et le modèle moyen estimé pour constater qu'ils sont très proches (voir Exemples 12 et 13).

Exemple 12 *Nous estimons un DMM d'ordre* 1 *et de degré* 2 *sur le génome complet du phage Lambda et nous considérons ce modèle comme le vrai modèle. Ensuite, nous simulons* 10 *séquence à l'aide de ce modèle puis estimons un modèle moyen sur ces* 10 *séquences. Les deux modèles (le vrai et l'estimé) sont donnés dans le Tableau 5.6. Nous observons une grande similarité entre les matrices.*

TAB. 5.5 – Comparaison entre les vrais modèles et les modèles estimés (dérive par splines polynomiales). Nous donnons la moyenne des valeurs absolues des différences entre les vrais paramètres et les paramètres estimés. Le nombre de séquences simulées est donné par ns. N est le nombre de segments.

Degré / N		1/2	2/2	3/2	1/3	2/3	3/3
Ordre 1	$ns = 1$	0.0211753	0.0844391	0.531507	0.0350229	0.262254	1.40359
	$ns = 10$	0.00887746	0.054083	0.216836	0.0180494	0.0749811	0.435706
	$ns = 100$	0.00281597	0.0122216	0.0578472	0.00447553	0.0186293	0.175117
Ordre 2	$ns = 1$	0.0996695	0.399781	2.04944	0.135387	0.778071	4.94055
	$ns = 10$	0.0951058	0.349734	1.81246	0.126934	0.723412	4.49569
	$ns = 100$	0.0841521	0.33828	1.69889	0.120671	0.659606	4.13114

TAB. 5.6 – Comparaison entre les matrices de transition des vrais modèles et des modèles estimés (dérive polynomiale)

Matrices	Vrai modèle				Modèle estimé			
Π_0	0.2637	0.2753	0.2623	0.1986	0.2634	0.2811	0.2596	0.1958
	0.2559	0.2889	0.3090	0.1462	0.2573	0.2839	0.3114	0.1472
	0.2107	0.3712	0.2582	0.1600	0.2077	0.3656	0.2662	0.1606
	0.1201	0.4873	0.2011	0.1914	0.1240	0.4884	0.1967	0.1910
$\Pi_{0.5}$	0.2931	0.2156	0.2044	0.2870	0.2913	0.2145	0.2065	0.2878
	0.2424	0.2445	0.2760	0.2371	0.2422	0.2469	0.2733	0.2376
	0.2976	0.2533	0.2127	0.2364	0.2954	0.2558	0.2127	0.2361
	0.1856	0.2891	0.2301	0.2953	0.1838	0.2883	0.2320	0.2959
Π_1	0.3480	0.2017	0.1833	0.2670	0.3559	0.1972	0.1781	0.2689
	0.3061	0.2032	0.2686	0.2221	0.3047	0.2034	0.2682	0.2238
	0.3076	0.2457	0.2037	0.2433	0.3047	0.2460	0.2017	0.2476
	0.2073	0.2997	0.2132	0.2798	0.2092	0.3044	0.2119	0.2744

Exemple 13 *Nous estimons un DMM d'ordre* 1 *et de degré* 3 *par splines polynomiales (avec* 4 *segments) sur le génome complet du phage Lambda et nous considérons ce modèle comme le vrai modèle. Ensuite, nous simulons* 10 *séquences à l'aide de ce modèle puis estimons un modèle moyen sur ces* 10 *séquences. Les deux modèles (le vrai et l'estimé) donnent les matrices de transition du Tableau 5.7. Nous observons une grande similarité entre les matrices.*

TAB. 5.7 – Comparaison entre les matrices de transition des vrais modèles et des modèles estimés (dérive par splines polynomiales)

Matrices	Vrai modèle				Modèle estimé			
Π_0	0.3418	0.2297	0.2093	0.2192	0.3532	0.2264	0.2101	0.2103
	0.2569	0.3030	0.2759	0.1642	0.2618	0.3030	0.2792	0.1560
	0.2094	0.3301	0.2503	0.2102	0.2042	0.3381	0.2486	0.2092
	0.1470	0.3687	0.1988	0.2856	0.1490	0.3612	0.2033	0.2865
$\Pi_{0.5}$	0.3127	0.1864	0.1937	0.3072	0.3165	0.1870	0.1893	0.3071
	0.2568	0.2070	0.2643	0.2719	0.2616	0.1959	0.2686	0.2739
	0.3145	0.2287	0.1934	0.2635	0.3235	0.2244	0.1916	0.2606
	0.2124	0.2461	0.2153	0.3262	0.2144	0.2404	0.2205	0.3246
Π_1	0.2563	0.1807	0.1491	0.4139	0.2585	0.1765	0.1542	0.4109
	0.2564	0.1478	0.2222	0.3736	0.2709	0.1467	0.2015	0.3809
	0.2803	0.1987	0.2193	0.3017	0.2877	0.1948	0.2208	0.2967
	0.1819	0.2519	0.1921	0.3741	0.1851	0.2576	0.2071	0.3502

5.2 Lois stationnaires et distributions de probabilité

5.2.1 Définitions

Les DMM offrent des modèles décrivant fidèlement les séquences réelles. Ce fait est particulièrement mis en évidence par l'étude des distributions de probabilités des nucléotides. En effet, analyser μ_t, la distribution de probabilité en position t associée à nos modèles, est le meilleur moyen d'évaluer leur qualité. À l'ordre k, la distribution μ_t est définie par récurrence par la formule suivante :

$$\mu_t(v) = \sum_{u \in \mathcal{A}^k} \mu_{t-k}(u_1) \dots \mu_{t-1}(u_k) \Pi_{\frac{t}{n}}(u, v) \quad \forall v \in \mathcal{A}$$

où $u = u_1 \dots u_k$. Nous rappelons qu'une chaîne de Markov ergodique Π sur un espace d'états fini possède une unique distribution de probabilité stationnaire ν telle que $\nu\Pi = \nu$. Les matrices de transition Π_i pour i allant de 0 à $k-1$ étant ergodiques, nous choisissons μ_i comme étant la loi stationnaire de Π_i pour i allant de 0 à $k-1$. Ainsi la récurrence est initialisée.

Exemple 14 *À l'ordre* 1, *la distribution* μ_t *est définie par récurrence par la formule suivante :*

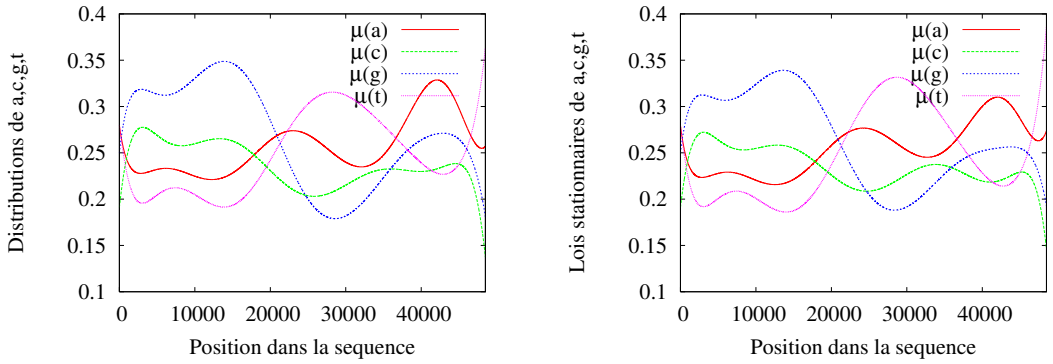
$$\mu_t(v) = \sum_{u \in \mathcal{A}} \mu_{t-1}(u) \Pi_{\frac{t}{n}}(u, v) \quad \forall v \in \mathcal{A}$$

avec μ_0 *loi stationnaire de* Π_0.

Nous calculons cette distribution de probabilité μ_t pour chaque position t pour analyser la composition des séquences.

Remarque 11 *L'étude de ces distributions de probabilités est très semblable à l'étude des lois stationnaires de chaque matrice* $\Pi_{\frac{t}{n}}$. *Les courbes obtenues sont similaires. Une première différence numérique de taille : la récurrence nous impose de calculer toutes les distributions pour tracer les courbes alors que les lois stationnaires nous permettent de ne prendre en compte qu'un certain nombre de points. Une seconde différence se situe au niveau de la définition : alors que les distributions de probabilité sont définies par récurrence (ce qui peut engendrer quelques problèmes si des matrices sont mal estimées, le mal pouvant s'amplifier), les lois stationnaires sont calculées ponctuellement. Une dernière différence se situe dans l'approche mathématique : alors qu'il semble raisonnable de calculer les distributions de probabilité récursivement à l'aide de nos matrices de transition, il peut paraître exagéré de calculer les lois stationnaires de chacune de matrice de transition, le régime n'étant pas en état stationnaire. Pourtant, ce n'est pas si déraisonnable : les DMM supposent une variation douce de la matrice de transition, une chaîne de Markov évoluant très vite vers son état stationnaire (une dizaine de points suffit souvent), utiliser les lois stationnaires n'est donc pas si insensé. L'expérience nous donne raison, les courbes étant très similaires voire identiques (Figure 5.3). Nous utiliserons donc indifféremment l'une ou l'autre des définitions.*

FIG. 5.3 – Distributions (à gauche) et lois stationnaires (à droite) pour des DMM d'ordre 1 et de degré 8, sur le *phage Lambda*.



La Figure 5.4 représente les variations des lois stationnaires pour des DMM d'ordre 1 et de degrés variant de 1 à 8 : dérive polynomiale sur le génome complet de *Chlamydia trachomatis* ([SKL+98]). La Figure 5.5 montre

ces mêmes lois stationnaires pour une dérive par splines polynomiales avec 4 segments. Nous observons une très nette symétrie entre les distributions de `a` et `t` ainsi qu'entre les distributions de `c` et `g`. Ces propriétés de symétrie de certaines séquences biologiques ont été introduites dans les années 60 : les règles de Chargaff ([LC67, CHA79]. Elles sont à l'origine de la découverte de la structure en double hélice de l'ADN.

Remarque 12 *La dérive par splines polynomiales de degré 4 avec 4 morceaux donne des résultats similaires à la dérive polynomiale de degré 8. Par ailleurs, la dérive par splines polynomiales de degré 8 provoque déjà les oscillations que nous avons évoquées en Figure 5.2. Il convient donc de choisir un degré de dérive et un nombre de morceaux plutôt petits (inférieurs à 5) pour l'estimation par splines.*

5.2.2 Comparaisons fréquences/lois

Afin d'établir la validité de nos modèles, nous comparons les évolutions des distributions de probabilités avec les fréquences des nucléotides. Pour calculer ces fréquences, nous utilisons une fenêtre glissante. La taille de la fenêtre a été choisie de manière à ne pas obtenir de courbes trop irrégulières (avec des pics) : nous avons opté pour une taille égale au vingtième de la taille de la séquence (de l'ordre de 100000 nucléotides pour *Chlamydia trachomatis*). Nous avons ainsi obtenu la fréquence de chaque lettre en chaque position de la séquence. Pour tracer ces courbes en un temps raisonnable, nous avons considéré 10000 points (donc 10000 positions t uniformément réparties), quelle que soit la séquence étudiée.

Les Figures 5.6 à 5.13 représentent ces comparaisons pour des DMM d'ordre 1 et de degrés respectifs variant de 1 à 8 dans le cas d'une dérive polynomiale sur *Chlamydia trachomatis*. Nous nous apercevons que plus le degré du modèle augmente, mieux il s'ajuste à la séquence.

Nous montrons seulement le degré 4 pour une dérive par splines polynomiales avec 4 segments (voir Figure 5.14) pour constater que cela s'ajuste tout aussi bien.

5.2.3 Comparaison MM / HMM / DMM

De la même manière que précédemment, nous traçons les lois stationnaires sur le génome du *phage Lambda* et retrouvons le long segment riche en `gc` déjà trouvé par un algorithme HMM développé par F. Muri [Mur97]. De plus, comparant cette segmentation HMM et l'évolution des lois stationnaires sous un DMM, nous observons des distributions de probabilité des lettres qui correspondent à la segmentation HMM (voir Figure 5.15). Cette comparaison est très intéressante car elle met en valeur les limites des HMM. Bien qu'un long segment riche en *gc* est connu et donné par les HMM, les autres parties de la segmentation HMM ne trouvent pas de significations réellement convaincantes au regard de l'évolution des probabilités de transition. Les DMM offrent une évolution douce contrairement à la segmentation parfois brutale des HMM. Il est important de noter que les DMM peuvent être vus aussi bien comme un outil concurrent des HMM que comme un outil complémentaire des HMM. En effet, chaque état d'un HMM pourrait cacher un DMM.

Afin de comparer nos DMM aux autres modèles de Markov, nous traçons en Figure 5.16 les évolutions des distributions de probabilité sous un DMM d'ordre 1 et de degré 0 (qui correspond au modèle de Markov d'ordre 1 classique). Il est évident que la distance entre les deux courbes est plus petite dans le cas d'un DMM de degré 8. Dans le cas des HMM, nous n'observons pas seulement une probabilité constante pour chaque lettre comme dans le modèle de Markov classique, mais plusieurs régions avec des probabilités constantes correspondant à la segmentation HMM. La Figure 5.17 compare les évolutions des distributions de probabilité pour un HMM d'ordre 1 à 3 états cachés et un DMM d'ordre 1 et de degré 3 sur le génome complet du *phage T4* (voir [MKM$^+$03]).

Nous avons aussi calculé une distance d_{df} entre les évolutions des distributions et les évolutions des fréquences des nucléotides :

$$d_{df} = \sum_{v \in \mathcal{A}} \sum_{t \in \mathcal{P}} (f_t(v) - \mu_t(v))^2,$$

où $\mathcal{P}$ est l'ensemble des 10000 points choisis pour l'évolution des lois stationnaires (voir 5.2.2), $f_t(v)$ est la fréquence du nucléotide v à la position t et $\mu_t(v)$ la probabilité stationnaire d'obtenir un v en position t. Un HMM d'ordre 1 à 3 états cachés a approximativement le même nombre de paramètres qu'un DMM d'ordre 1 et de degré 3 (en réalité, ce nombre est de 42 pour le HMM et de 48 pour le DMM). Pourtant, nous pouvons déjà noté que la distance d_{df} est légèrement plus faible pour le DMM : 5.865 contre 5.873. Évidemment, cette distance est encore plus petite pour un DMM de degré 8 ($d_{df} = 3.391$). En ce sens, nous montrons que les DMM représentent une nouvelle classe de modèles plus flexibles pour les séquences d'ADN qui proposent de meilleurs ajustements que les HMM dans de nombreux cas.

Dans le but d'illustrer à nouveau ce phénomène d'une autre manière, nous avons tracé à la Figure 5.18, la fréquence de `gc` dans le génome complet du *phage Lambda*. Comme nous l'avons dit en introduction, les biologistes

FIG. 5.4 – Lois stationnaires, DMM d'ordre 1 et de degré variant de 1 à 8 : dérive polynomiale sur *Chlamydia trachomatis*

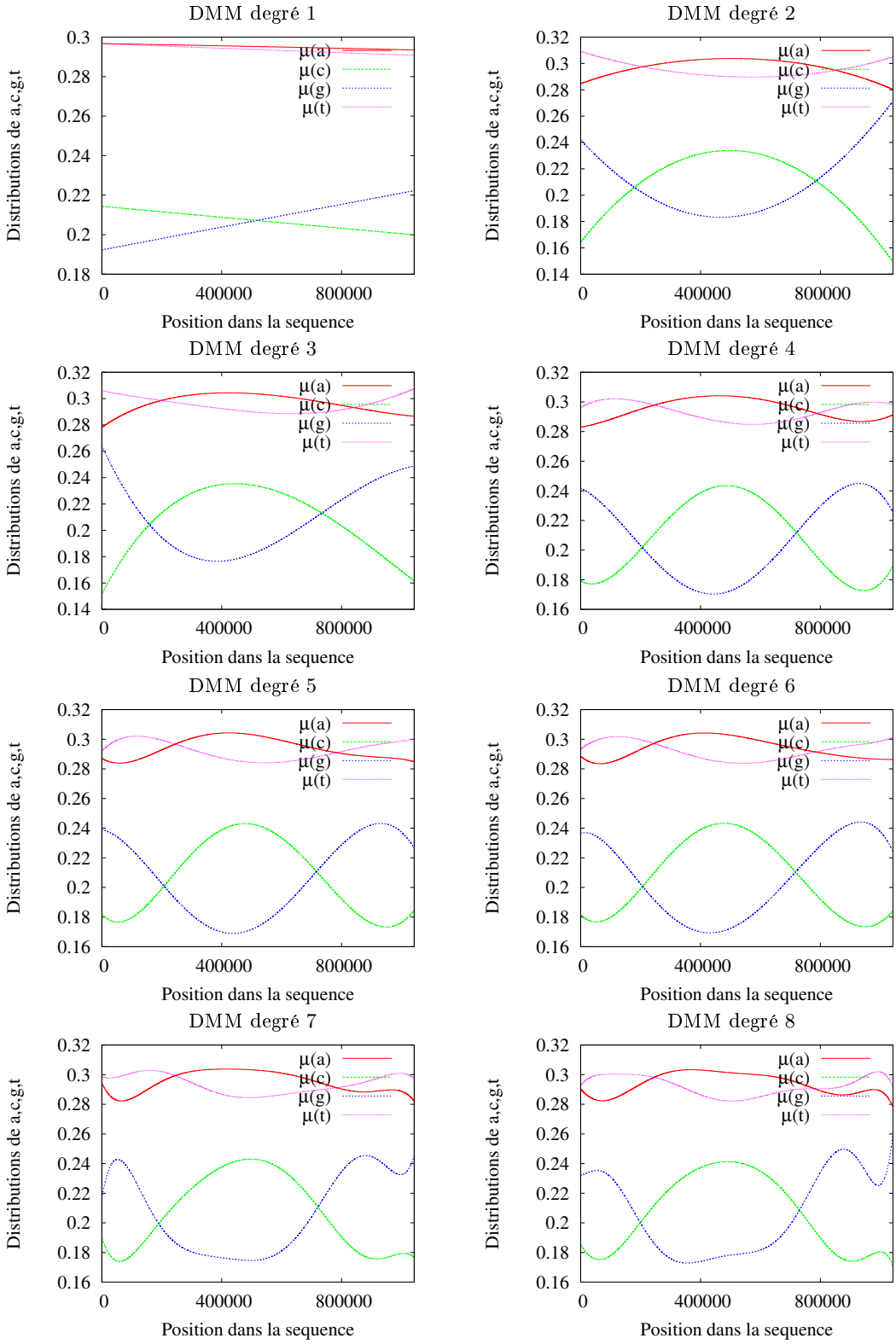


FIG. 5.5 – Lois stationnaires, DMM d'ordre 1 et de degré variant de 1 à 8 : dérive par splines polynomiales sur *Chlamydia trachomatis*

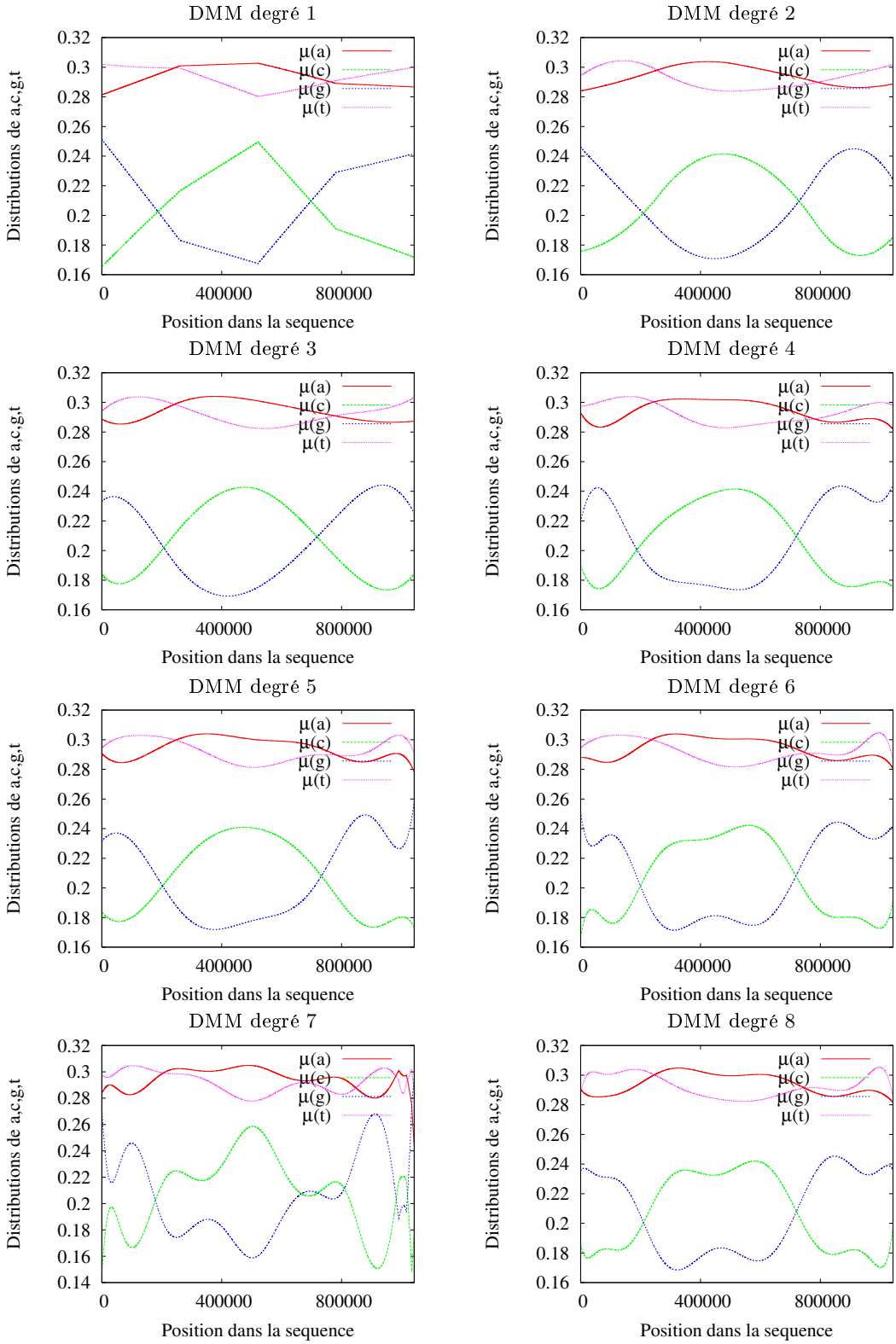


FIG. 5.6 – Comparaison lois stationnaires / fréquences, modèle de Markov régulé d'ordre 1 et de degré 1 : *Chlamydia trachomatis*

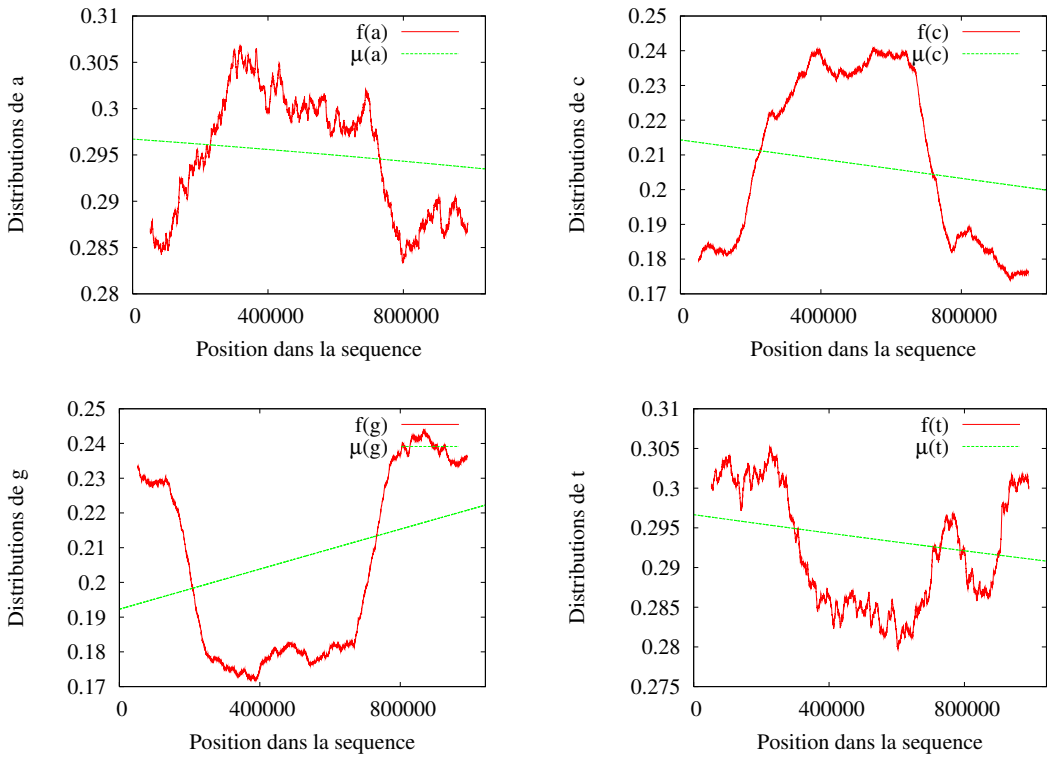


FIG. 5.7 – Comparaison lois stationnaires / fréquences, modèle de Markov régulé d'ordre 1 et de degré 2 : *Chlamydia trachomatis*

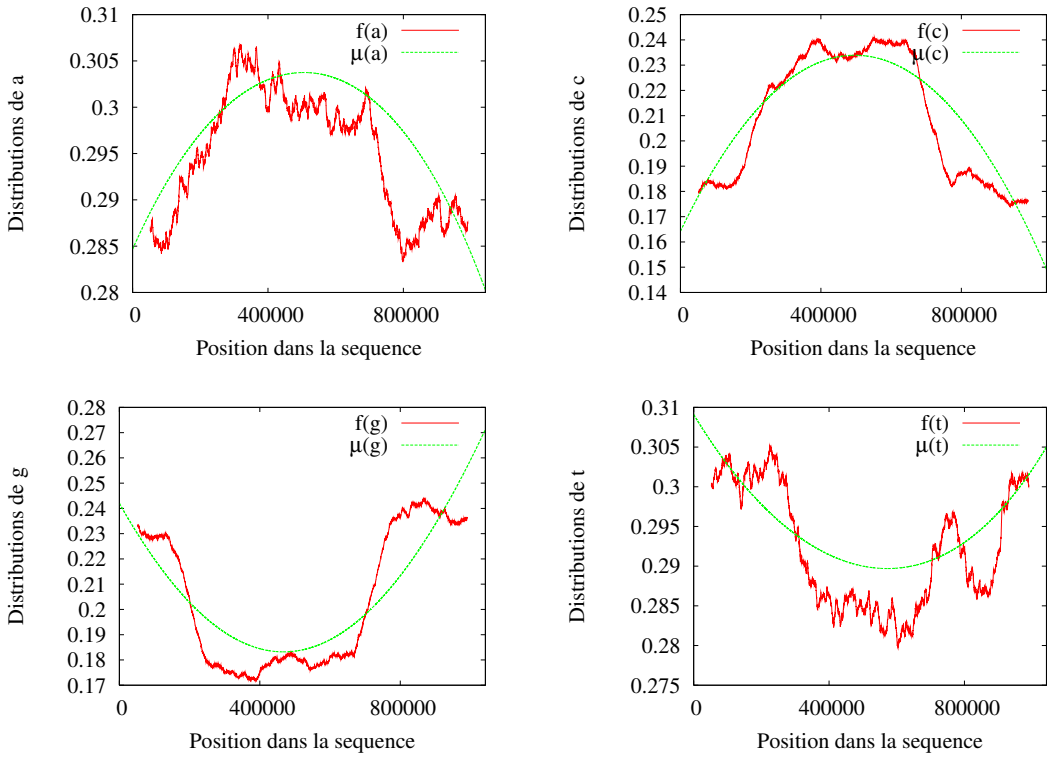


FIG. 5.8 – Comparaison lois stationnaires / fréquences, modèle de Markov régulé d'ordre 1 et de degré 3 : *Chlamydia trachomatis*

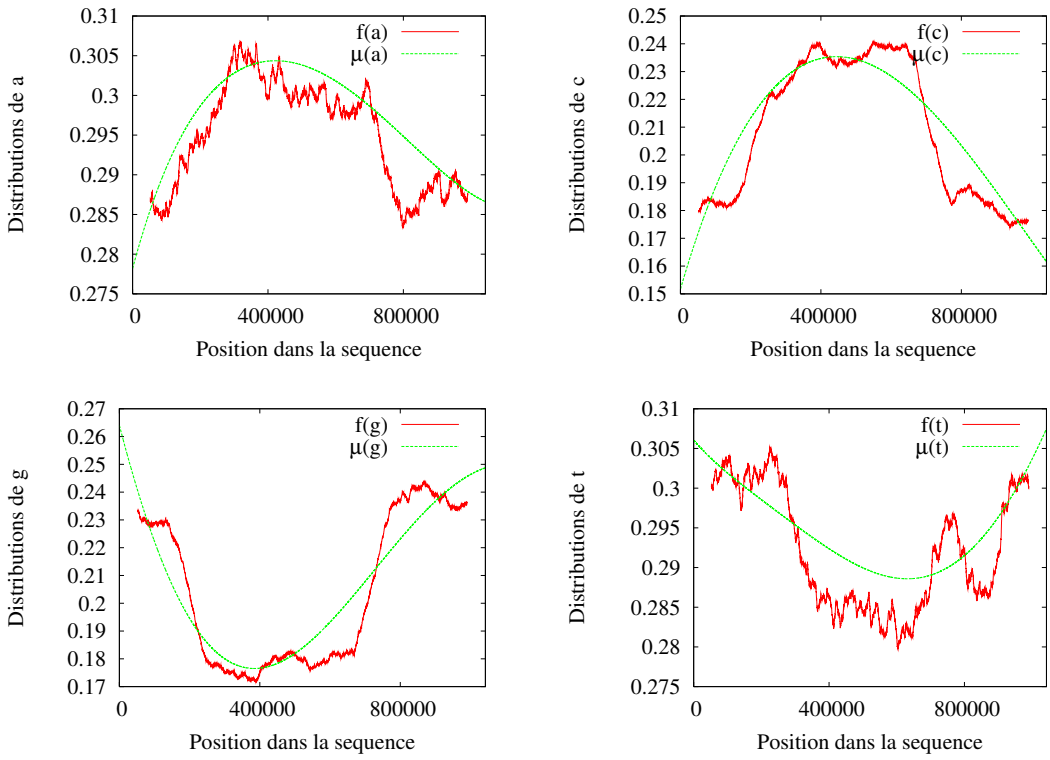


FIG. 5.9 – Comparaison lois stationnaires / fréquences, modèle de Markov régulé d'ordre 1 et de degré 4 : *Chlamydia trachomatis*

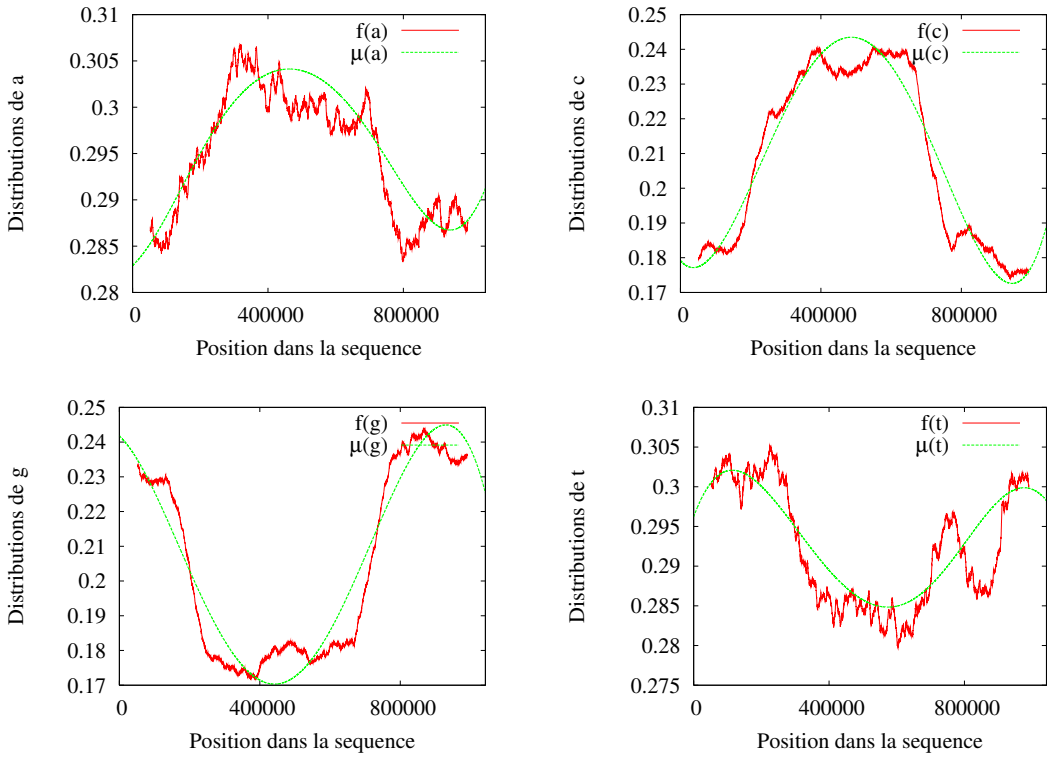


FIG. 5.10 – Comparaison lois stationnaires / fréquences, modèle de Markov régulé d'ordre 1 et de degré 5 : *Chlamydia trachomatis*

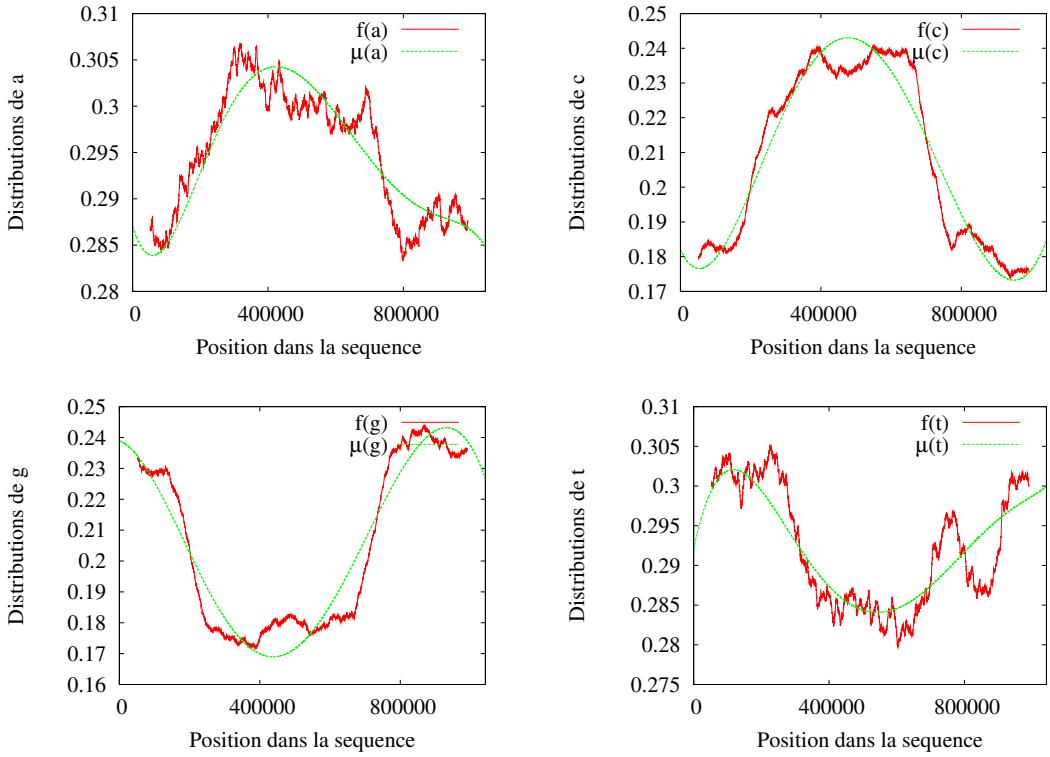


FIG. 5.11 – Comparaison lois stationnaires / fréquences, modèle de Markov régulé d'ordre 1 et de degré 6 : *Chlamydia trachomatis*

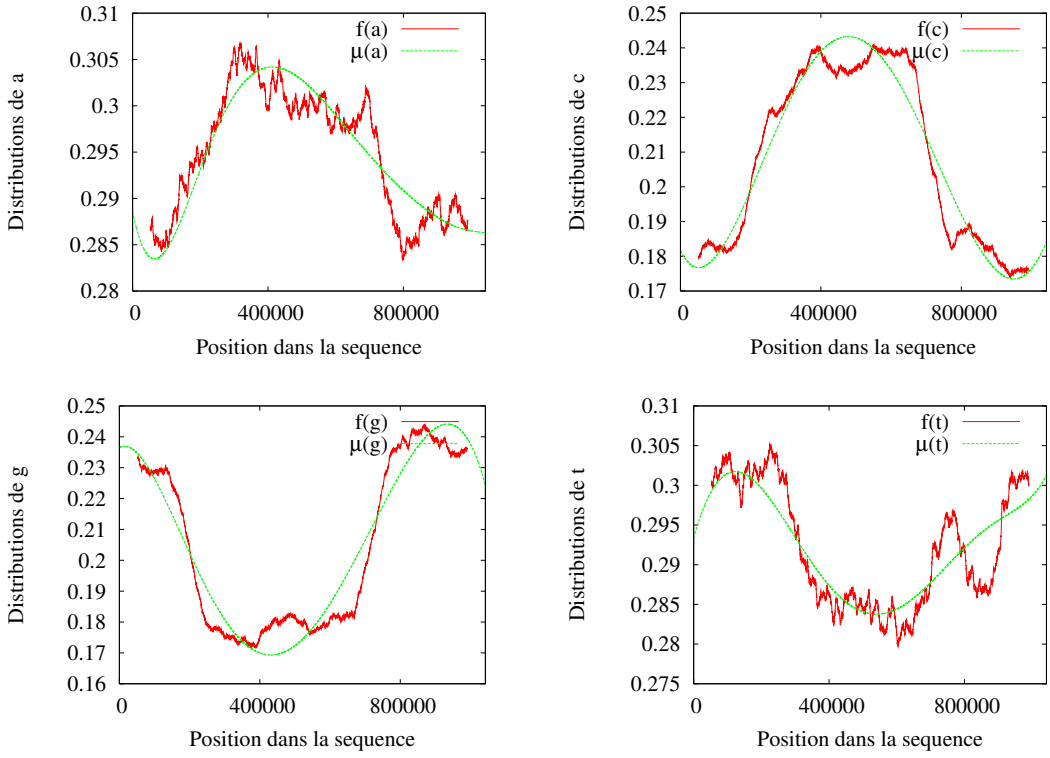


FIG. 5.12 – Comparaison lois stationnaires / fréquences, modèle de Markov régulé d'ordre 1 et de degré 7 : *Chlamydia trachomatis*

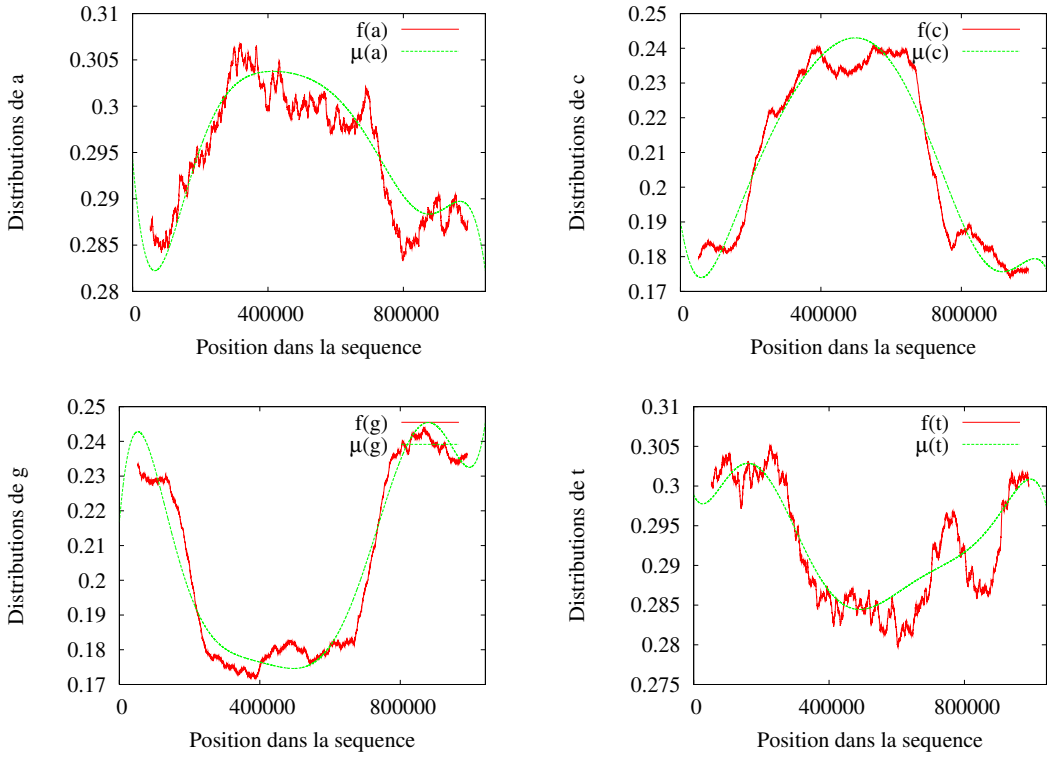


FIG. 5.13 – Comparaison lois stationnaires / fréquences, modèle de Markov régulé d'ordre 1 et de degré 8 : *Chlamydia trachomatis*

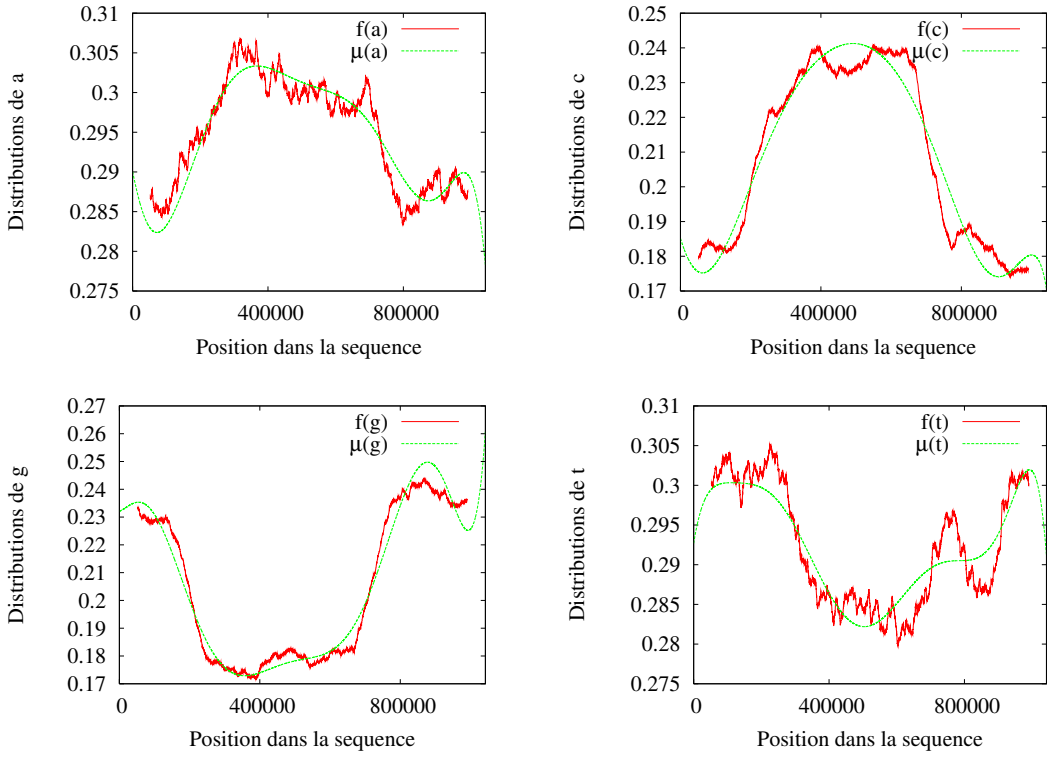


FIG. 5.14 – Comparaison lois stationnaires / fréquences, modèle de Markov régulé d'ordre 1 et de degré 4 par splines polynomiales (4 segments) : *Chlamydia trachomatis*

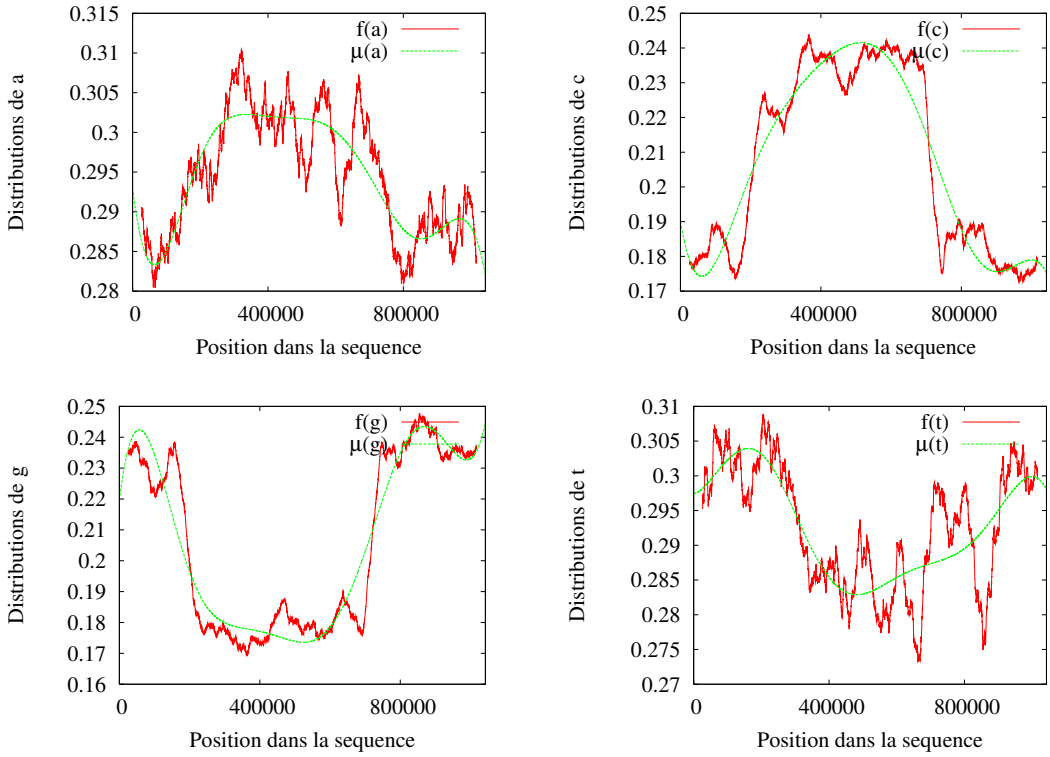


FIG. 5.15 – Lois stationnaires d'une DMM d'ordre 1 et de degré 8 sur le génome du *phage Lambda* et segmentation HMM avec trois états cachés (les états 0, 1 et 2 notés sur l'axe des ordonnées droit).

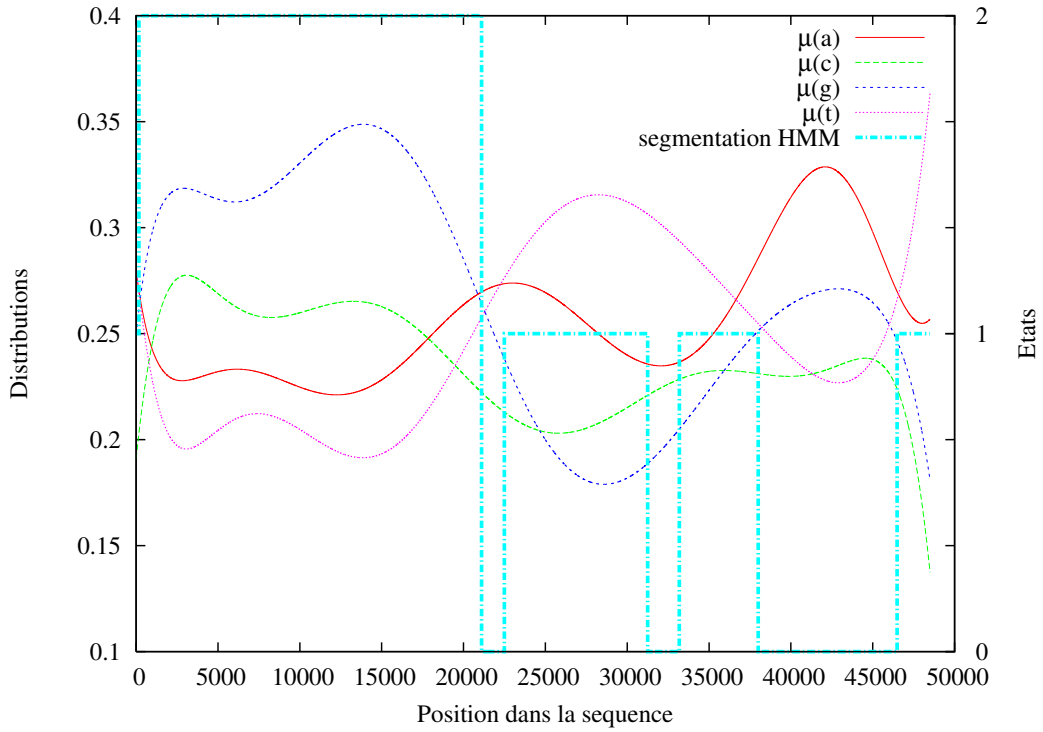
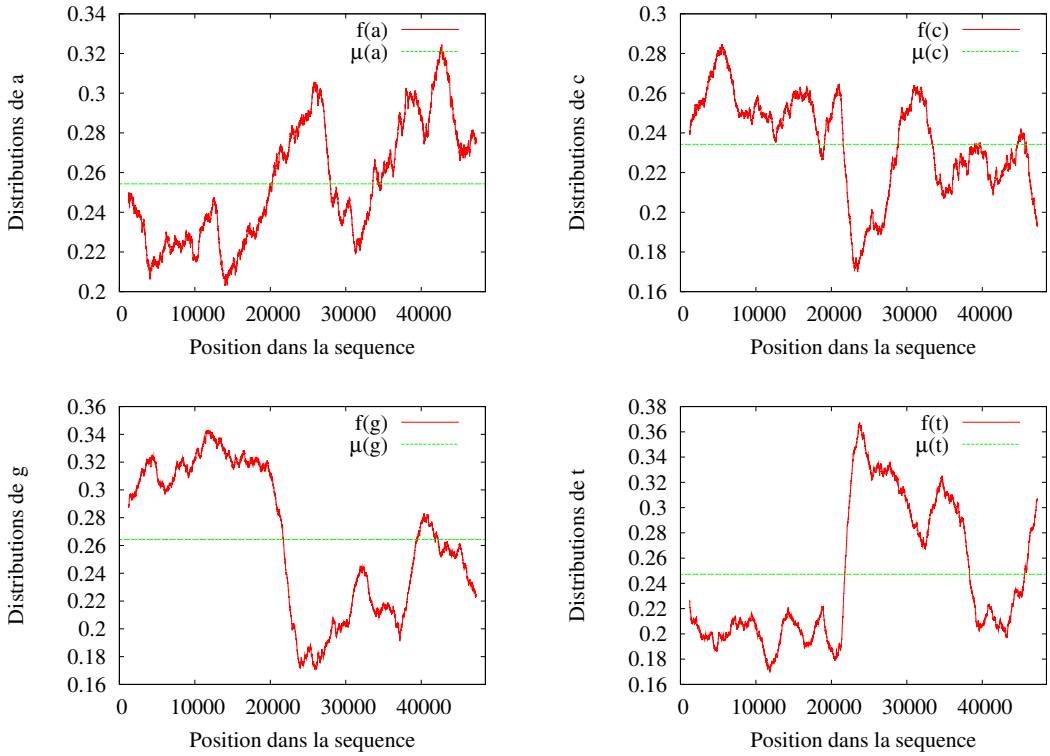


FIG. 5.16 – Fréquence f et distribution de probabilité μ des nucléotides pour un DMM de degré 0, à l'ordre 1 sur le *phage Lambda*.



sont très attentifs au pourcentage de gc parce qu'il suggère la présence de gènes. Ils considèrent cinq familles d'isochores : deux familles pauvres en gc (L1 et L2) et trois familles riches en gc (H1, H2, H3)([Ber93, OBGCRR01]). Mais les transitions entre deux familles sont souvent jugées trop soudaines lorsqu'elles sont modélisées par les HMM. Les DMM et leur évolution continue sont une bonne nouvelle manière de modéliser ces transitions. Par exemple, entre les positions 26000 et 32000 de la Figure 5.18, nous constatons une évolution linéaire du pourcentage en gc que nous modélisons à l'aide d'un DMM de degré 1. Ainsi les DMM forment un outil capable de modéliser les phénomènes hétérogènes, en particulier l'évolution linéaire du pourcentage en gc, alors que les HMM n'auraient pas prédit d'évolution ou bien auraient prédit un changement brutal.

Par ailleurs, il est important de constater que dans toutes les courbes représentant les fréquences (voir Figure 5.6 et Figure 5.16 par exemple), l'évolution de ces fréquences est continue. Même lorsque nous pourrions constater un changement brutal pour trois des nucléotides, le quatrième a une évolution continue (aux alentours de la position 22000 sur la Figure 5.16). Même lorsque ces courbes nous montrent un état relativement constant, la transition entre cet état et un autre état relativement constant est continue (sur la Figure 5.6 un état relativement constant est observé entre les positions 350000 et 700000 mais le reste de la séquence subit une évolution continue et non un changement brutal comme le suggèrent les HMM). Il en va de même pour la grande majorité des courbes biologiques, voire la totalité. À la vue de ces courbes, nous pouvons (ré)affirmer qu'une modélisation continue des séquences biologiques est très réaliste et les DMM offrent cette modélisation.

FIG. 5.17 – Fréquence f et distribution de probabilité μ des nucléotides pour un DMM de degré 8 et un HMM à 3 états, à l'ordre 1 sur le *phage T4*.

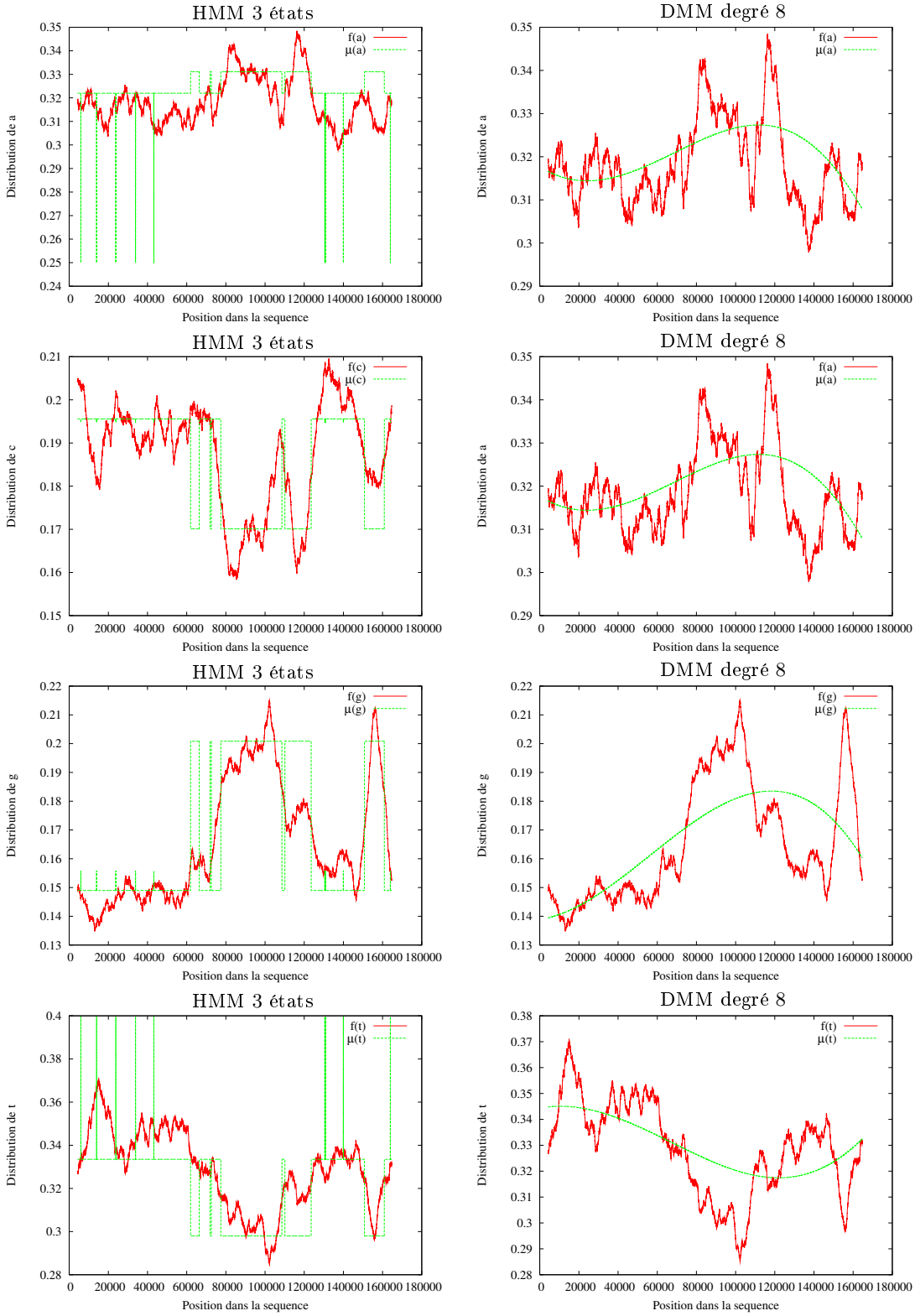
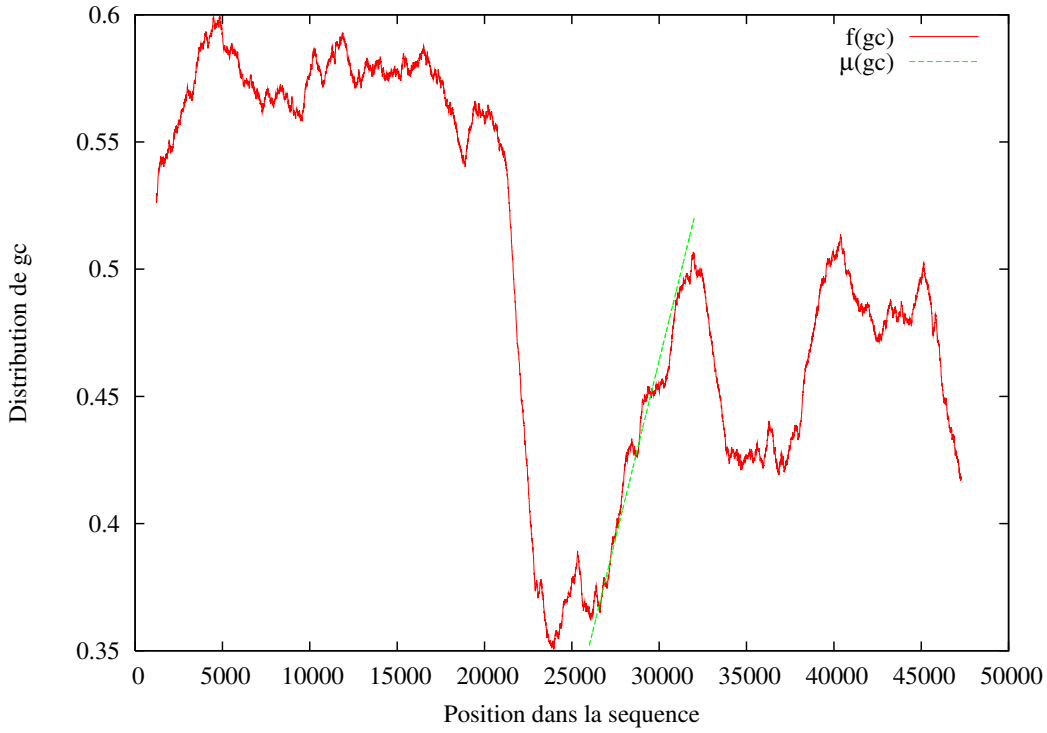


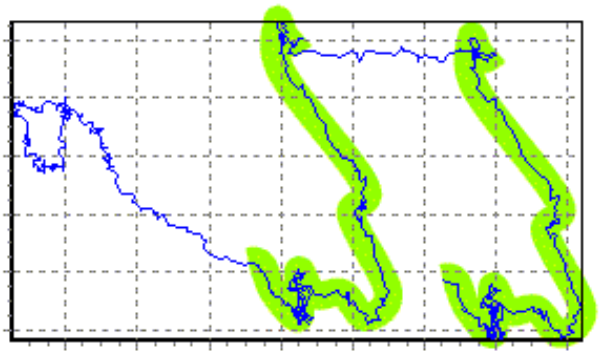
FIG. 5.18 – Fréquence f et probabilité μ de `gc` pour un DMM de degré 1 entre les positions 26000 et 32000 dans le génome du *phage Lambda*.



5.3 Origine de réplication

Nous présentons une application de nos modèles qui leur sert dans le même temps de validation : la recherche des origines de réplications chez les bactéries. Cette applications prend son inspiration dans le programme ORILOC (voir [Lob00] et http://pbil.univ-lyon1.fr/software/oriloc.html). La réplication de l'ADN est le procédé de recopie d'un double brin d'ADN en deux doubles brins identiques (aux erreurs de réplication près). Chacun des deux double brins d'ADN obtenus est composé d'un brin original et d'un brin synthétisé. L'origine de la réplication est la position à partir de laquelle la réplication est initialisée. La réplication de l'ADN peut commencer à partir de ce point bidirectionnellement ou bien unidirectionnellement. Fondé sur les asymétries de composition en base `a`, `c`, `g` et `t` entre le brin direct (*leading strand*) et le brin retardé (*lagging strand*) de réplication, le programme effectue une marche aléatoire ([Lob99]) pour obtenir la position de l'origine de réplication. Le principe de la *DNA Walk* est de tracer une courbe en avançant d'un pas vers la gauche lorsque l'on rencontre un `a`, d'un pas vers la droite lorsque l'on rencontre un `t`, d'un pas vers le haut lorsque l'on rencontre un `c` et d'un pas vers le bas lorsque l'on rencontre un `g`. Nous obtenons ainsi un graphe permettant par exemple de retrouver aisément les longues séquences répétées (voir Figure 5.19 et http://www.genometrician.com/).

FIG. 5.19 – DNA Walk de la plante parasite *Epifagus virginiananon*



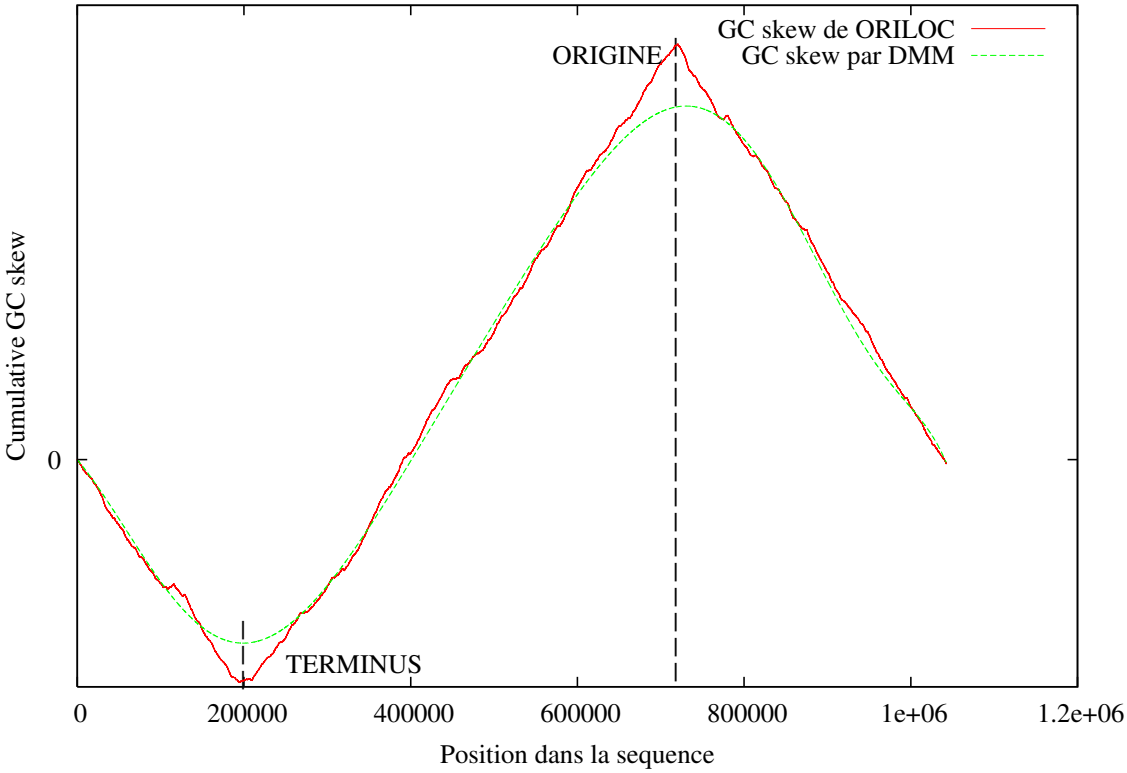
La méthode de détection de l'origine de réplication consiste en le tracé de l'évolution du *cumulated gc-skew* (voir [Lob99]) : la différence entre le nombre de `c` et le nombre de `g` présents jusqu'à une position t en fonction de cette position t. Une courbe est tracée par le programme ORILOC et le pic de cette courbe correspond à l'origine de réplication. Les valeurs permettant de tracer la courbe sont calculées comme suit : la première valeur est 0 et durant la marche le long de la séquence, ORILOC ajoute 1 chaque fois que la lettre `g` est rencontrée, et soustrait 1 chaque fois que la lettre `c` est rencontrée. Ainsi, ORILOC ne repose pas sur un modèle probabiliste, il trace une courbe en parcourant la séquence.

Nous utilisons les mêmes propriétés d'asymétries dans les génomes bactériens pour mettre en oeuvre une détection des origines de réplication fondée sur les DMM. En effet, grâce aux calculs des probabilités stationnaires d'apparition des nucléotides en chaque position t dans la séquence, nous traçons une courbe similaire à celle d'ORILOC en calculant une valeur équivalente au *cumulated gc-skew*. Les valeurs de notre courbe sont calculées comme suit : la première valeur est 0 et à chaque position dans la séquence, nous ajoutons la probabilité d'apparition de la lettre `g` et soustrayons la probabilité d'apparition de la lettre `c` :

$$gc(t) = \sum_{i=0}^{t} \mu_i(c) - \sum_{i=0}^{t} \mu_i(g)$$

où t est la position dans la séquence. Ce travail a été effectué sur le génome complet de *Chlamydia trachomatis* ([SKL+98]). Nous notons, sur la Figure 5.20, la grande similarité entre la courbe obtenue par ORILOC et celle obtenue par les DMM. Notons aussi que notre courbe est plus arrondie que celle d'ORILOC du fait que les DMM modélisent des transitions douce. Bien que la recherche d'origines de réplication soit un problème de détection de points de rupture, notre méthode fonctionne, dans le sens qu'elle offre aux biologistes une fenêtre permettant de trouver l'origine de réplication *in vivo*. Ainsi les transitions douces n'empêchent pas de trouver les origines de réplication, d'autant plus qu'il est possible d'affiner le résultat en se réduisant à la portion de génome qui nous intéresse. Bien entendu, il est toujours possible d'affiner les résultats à la portion de génome qui nous intéresse afin de déterminer plus précisément les origines de réplications recherchées.

FIG. 5.20 – Comparaison ORILOC / Chaînes de Markov régulées d'ordre 1 et de degré 8 : *Chlamydia trachomatis*



5.4 AIC, BIC

Il existe de nombreux critères de comparaison de modèles et deux sont assez unanimement reconnus : l'*AIC* et le *BIC* (pour *Akaïke Information Criterion* et *Bayesian Information Criterion* introduits respectivement par [Aka74] et [Sch78]). Rappelons les définitions de ces deux critères :

$$\begin{aligned} AIC &= -(2\ell(\theta) - 2K) \\ BIC &= -(2\ell(\theta) - K\log n) \end{aligned}$$

où $\ell(\theta)$ est la log-vraisemblance du modèle, K le nombre de paramètres et n la taille de l'échantillon. Le modèle qui a le plus petit *AIC* ou *BIC* est considéré comme le "meilleur" modèle selon ce critère.

AIC et *BIC* sont généralement construits en ajoutant une pénalisation à la log-vraisemblance évaluée par maximum de vraisemblance. Puisque l'estimateur des moindres carrés est le seul dont nous disposons, nous adaptons ces critères à l'aide d'une pénalisation de la log-vraisemblance calculé à l'aide de cet estimateur. Cela peut être justifié par le fait que pour des chaînes de Markov l'estimateur des moindres carrés (fondé sur une formule similaire à notre formule 2.1 p38) est asymptotiquement équivalent à l'estimateur du maximum de vraisemblance.

Dans les tableaux 5.8 et 5.9, nous avons calculé les *AIC* et *BIC* pour des DMM d'ordre 0, 1, 2 et 3 avec dérive polynomiale de degré variant de 0 à 5. Ces résultats ont été obtenus sur le génome complet d'*Haemophilus influenzae* (voir [FAW+95]).

TAB. 5.8 – *AIC* de DMM sur *Haemophilus influenzae*.

Degré	0	1	2	3	4	5
Ordre 0	4970436	4970420	4969422	4969322	4969286	**4969134**
Ordre 1	4907696	4907649	4906564	4906455	4906363	**4906223**
Ordre 2	4892311	4892250	4891118	4890996	4890826	**4890648**
Ordre 3	4865655	4865652	4864567	4864540	4864472	**4864342**

TAB. 5.9 – *BIC* de DMM sur *Haemophilus influenzae*.

Degré	0	1	2	3	4	5
Ordre 0	4970473	4970494	4969534	4969471	4969472	**4969358**
Ordre 1	4907845	4907947	**4907011**	4907051	4907108	4907117
Ordre 2	4892907	4893442	4892907	**4890996**	4893807	4894224
Ordre 3	**4868040**	4870422	4871721	4874079	4876395	4878650

Alors que l'*AIC* préfère les modèles avec de nombreux paramètres (l'*AIC* décroît généralement avec l'ordre et le degré), le *BIC* préfère les modèles avec un petit nombre de paramètres. C'est pourquoi les DMM avec de grands degrés sont partiellement ignorés par le *BIC*. En effet, pour un DMM d'ordre k et de degré d, le nombre de paramètres est égal à $(d+1)|\mathcal{A}|^k(|\mathcal{A}|-1)$. Il est possible de choisir ordre et degré d'un DMM avec le *BIC*, mais le grand nombre de paramètres de ces modèles apporte de meilleurs ajustement à la séquence comme nous avons pu le remarquer en 5.2 et le *BIC* ne prend pas en compte cet argument.

Dans le but de comparer les DMM avec les autres modèles, nous répétons que la meilleure manière est d'observer sur les figures 5.6 ou 5.16 que les variations des nucléotides sont continues.

Remarque 13 *Un DMM de degré* 0 *est un modèle de Markov classique. Nous avons constaté dans les Tableaux 5.8 et 5.9 que les AIC et BIC des DMM sont meilleurs que ces modèles de Markov classiques.*

Les comparaisons selon l'*AIC* et le *BIC*, effectuées dans les tableaux suivants, reflètent le comportement général de ces critères face à nos modèles. Selon l'*AIC* les DMM sont de meilleurs modèles que les HMM quel que soit l'ordre (voir Tableau 5.10). Le degré 1 est le degré pour lequel l'*AIC* est le plus grand, pourtant il est déjà plus petit que pour un HMM. Selon le *BIC*, les HMM sont de meilleurs modèles que les DMM (voir Tableau 5.11). Le degré 1 est cette fois le degré pour lequel le *BIC* est le plus petit (ce n'est pas toujours le cas : pour des séquences plus longues comme nous l'avons vu au Tableau 5.9), pourtant il est déjà plus grand que pour un HMM.

TAB. 5.10 – *AIC* de DMM et de HMM sur le virus du sida *HIV1*.

Ordre	0	1	2	3
DMM degré 1	25470	24959	24893	25123
HMM 2 états	25566	25234	26338	31145
HMM 3 états	25598	25482	27173	34392

TAB. 5.11 – *BIC* de DMM et de HMM sur le virus du sida *HIV1*

Ordre	0	1	2	3
DMM degré 1	25513	25130	25579	27869
HMM 2 états	25493	24996	25441	27613
HMM 3 états	25461	25097	25800	29066

Comme nous l'avons déjà dit, il ne faut pas seulement voir dans les DMM un modèle en compétition avec les HMM mais plutôt un outil complémentaire à la modélisation des séquences. Même si la différence entre les *BIC* n'est pas très grande, l'essentiel est de se rappeler que nous proposons les premiers modèles incluant la possibilité d'une variation continue de la matrice de transition. Combinés avec les qualités des HMM, les DMM apporteront des outils puissants pour l'analyse de séquences.

Chapitre 6

Les mots exceptionnels

6.1 La théorie

Un exemple fondamental d'application des DMM est la recherche de mots exceptionnels dans les séquences d'ADN. De nombreuses analyses sur les séquences d'ADN sont fondées sur la distribution des occurrences de motifs possédant une fonction biologique particulière. Un problème important est de déterminer la significativité statistique de la fréquence d'un mot dans une séquence d'ADN. [NDV02] relate cette importance de trouver des mots sur- ou sous- représentés. L'idée naïve est la suivante : un mot peut avoir une faible fréquence dans une séquence d'ADN parce qu'il empêche la réplication ou bien l'expression d'un gène, alors qu'un mot significativement fréquent peut avoir une activité fondamentale pour la stabilité du génome ou bien tout simplement la survie de l'organisme. Des exemples bien connus de mots de fréquences exceptionnelles dans les séquences d'ADN sont les palindromes correspondants aux sites de restriction évités par exemple dans *E. coli* [KBM92], ou bien les motifs CHI (Cross-over Hotspot Instigator) dans plusieurs bactéries telle *E. coli* par exemple ([SKS+81, EKBSG99]), ou bien encore les séquences uptake [SGS99] ou les signaux de polyadénylation [vHdOPO00]. L'approche la plus populaire consiste en l'ajustement d'un modèle de Markov sur la séquence et le calcul de la p-valeur qui est définie par $\mathbb{P}(N > N_{obs})$ pour un mot sur-représenté et par $\mathbb{P}(N < N_{obs})$ pour un mot sous-représenté, où N est la variable aléatoire du nombre d'occurrences du mots étudié et N_{obs} le nombre d'occurrences observées. Nous définissons la statistique de motif S associé à n'importe quel nombre N_{obs} par :

$$S = \begin{cases} -\log_{10} \mathbb{P}(N > N_{obs}) & \text{si } N \geq \mathbb{E}(N) \\ +\log_{10} \mathbb{P}(N < N_{obs}) & \text{si } N < \mathbb{E}(N) \end{cases}.$$

De cette manière, un motif a une statistique positive si il est vu plus de fois qu'il est attendu et une statistique négative si il est vu moins qu'il n'est attendu. Dans les deux cas, la p-valeur correspondante est donnée (en échelle logarithmique) par l'amplitude de la statistique. Voir [Nue06b] pour une étude des différentes méthodes disponibles pour calculer les statistiques de motifs sur des textes générés par des modèles de Markov.

Comme ces probabilités sont calculées sous un modèle, de petites p-valeurs peuvent apparaître pour des mots sans intérêt biologique si le modèle n'est pas fiable. C'est pourquoi il est préférable de proposer un modèle de référence le plus proche possible de la vraie séquence. Les DMM offrent un tel modèle. Il est toujours plus convaincant d'obtenir des p-valeurs pour les modèles les plus réalistes. Dans ce sens, considérer un DMM pour la recherche de mots exceptionnels dans les séquences biologiques semble être une meilleure approche qu'utiliser des modèles de Markov (comme nous l'avons vu dès l'introduction (voir Figure 7) et dans le chapitre précédent (voir 5.2.3), les DMM et leur variation continue offrent des modèles plus proches de la réalité que les modèles de Markov).

Des complexités numériques apparaissent lorsque nous voulons calculer les p-valeurs exactes de modèles de Markov inhomogènes mais une nouvelle approche proposée par [Nue06a], utilisant les *finite Markov chain imbedding* (FMCI, voir [Lou96]) apporte des solutions à ce problème. Une description détaillée de cette méthode est donnée dans [NP07].

6.2 Premières applications

Nous ne donnons qu'un exemple de recherche de mots exceptionnels. Nous choisissons le mot le plus populaire dans ce domaine, le CHI d'*Escherichia coli K12* (voir [BPB+97]). Nous considérons le génome complet de la bactérie où le motif CHI `gctggtgg` apparaît 499 fois. Comme nous pouvons le constater dans le Tableau 6.1, le motif CHI est attendu 70.10 par un DMM d'ordre 1 et de degré 0 et 175.31 fois par un DMM d'ordre 2 et de

TAB. 6.1 – Statistique S (log p-valeur) du mot `gctggtgg` sur-représenté pour des DMM de différents ordres et degrés : le CHI d'*E. coli* qui apparaît 499 fois dans la séquence. Notons qu'un DMM de degré 0 correspond à un modèle de Markov classique.

Ordre	Degré	Espérance	S
1	0	70.10	240.814
1	1	70.26	240.398
1	2	71.88	238.766
1	3	71.87	238.774
1	8	71.94	238.605
2	0	173.84	88.902
2	1	174.03	88.747
2	2	175.16	87.837
2	3	175.10	87.881
2	8	175.31	87.717

degré 8. Dans des modèles plus réalistes tels les DMM, les motifs CHI sont plus attendus que dans d'autres modèles (même si malheureusement, la différence n'est pas si flagrante).

Comme nous l'avons déjà dit, nous ne pouvons comparer les p-valeurs de différents modèles entre elles. Mais nous pouvons comparer les différentes classifications apportées pas ces différents modèles. Quelle classification devons-nous préférer ? Celle donnée par les HMM et leur segmentation ou celle donnée par les DMM et son évolution douce ? Évidemment, il semble plus raisonnable de considérer les p-valeurs dans le modèle qui apporte le meilleur ajustement aux données même si ce modèle contient plus de paramètres. Ainsi les DMM forment un outil très utile à la recherche de mots exceptionnels dans les séquences d'ADN. Le Tableau 6.2 donne la classification des mots de 5 lettres, pour des modèles de Markov classiques, un HMM à 3 états et un DMM de degré 1, à l'ordre 1.

TAB. 6.2 – Classification des mots de taille 5 dans le génome complet du *phage Lambda*, pour différents modèles, selon leur statistique de motifs S. N_{obs} désigne le nombre observé d'occurrences de ce mot. $\mathbb{E}(N)$ désigne l'espérance du nombre N de mots. Nous donnons seulement les cinq mots les plus sous-représentés et les cinq mots les plus sur-représentés pour chaque modèle.

MM				HMM 3 états				DMM degré 1			
Mots	N_{obs}	$\mathbb{E}(N)$	S	Mots	N_{obs}	$\mathbb{E}(N)$	S	Mots	N_{obs}	$\mathbb{E}(N)$	S
aattg	32	88.22	−11.41	aattg	32	83.38	−10.07	aattg	32	86.53	−10.94
ttggg	20	65.12	−10.33	acttg	13	47.59	−8.57	ttgga	21	64.94	−9.76
ttgga	21	66.70	−10.29	tctag	2	24.60	−8.19	ttggg	20	62.94	−9.66
acttg	13	50.74	−9.59	ttgga	21	59.47	−8.15	acttg	13	50.27	−9.44
taggg	3	29.60	−9.21	tcgag	9	39.01	−8.11	tcgag	9	40.69	−8.68
gccgg	114	53.97	12.13	gctgg	127	65.44	14.23	gctgg	127	64.80	11.77
ctgaa	124	61.02	12.16	ctgaa	124	61.34	14.90	ctgaa	124	60.85	12.21
tccgg	100	39.98	15.08	ccgga	112	44.00	20.58	tccgg	100	38.81	16.18
ccgga	112	43.11	17.93	tccgg	100	36.50	20.65	ccgga	112	43.57	18.10
gcaga	141	57.51	20.20	gcaga	141	58.35	22.66	gcaga	141	57.59	20.31

6.3 Perspectives : les facteurs de transcription

Conscient que les exemples d'application donnés plus haut ne sont qu'illustratifs et ne sont pas des "résultats", nous avons envisagé une application ambitieuse au sujet des sites de fixation de facteurs de transcription (SFFT). Nous avons parlé de la transcription en Introduction sans préciser que cette transcription s'initie à certains endroits de la séquence reconnus, au moins en partie, par la présence de sites de fixation de facteurs de transcription (des motifs composites correspondant le plus souvent à deux mots de 3 à 8 nucléotides). Ces facteurs sont donc très importants car ils sont indispensables à la première étape de l'expression d'un gène. Certains motifs sont communs à plusieurs bactéries ou diffèrent seulement pas 1 nucléotide. Certains autres sont propres à la bactérie. La recherche

de ces facteurs de transcription est un domaine d'actualité car la plupart de ces motifs restent encore inconnus. Une littérature abondante existe en ce domaine. [Her99] pose les bases de la structure des motifs, et détermine leurs caractéristiques. De nombreux algorithmes de recherche de sites de fixation des facteurs de transcription existent reflétant de multiples approches : [TLB+05] compare ces différentes approches. Des méthodes de Gibbs sampling ([BE94]) ou bien plus proche de notre sujet, des méthodes fondées sur la sur-représentation statistique de ces sites de fixation ([vHACV98, TSDR+08]).

La recherche de mots exceptionnels nécessite un modèle de référence, reflétant au mieux la séquence. Les méthodes de détection utilisent très souvent les modèles de Markov et il semble très intéressant de posséder de nouveaux modèles tels les DMM pour affiner les résultats de ces méthodes voire détecter de nouveaux motifs. Ce travail fait l'objet d'une collaboration avec Grégory Nuel[8]

[8]MAP5 - UMR CNRS 8145, Université Paris Descartes, `nuel@math-info.univ-paris5.fr`

Conclusion

DMM et perspectives

Nous avons introduit une nouvelle classe de modèles de Markov inhomogènes, les modèles de Markov régulés ou *drifting Markov models*. Ces nouveaux modèles permettent à la matrice de transition de varier le long de la séquence. Les modèles de Markov sont homogènes et les modèles de Markov cachés ne peuvent pas modéliser toutes les structures hétérogènes. L'hétérogénéité des séquences nous encourage à considérer des modèles plus flexibles tels que les DMM et la variation continue de leur matrice de transition. Une illustration important de ces modèles concerne le pourcentage en `gc` d'une séquence d'ADN. Il est communément admis qu'un fort pourcentage en `gc` peut induire la présence de gène ([ZCB96]). Du fait qu'ils offrent une évolution douce et une matrice de transition différente en chaque position de la séquence, les DMM proposent un meilleur ajustement au pourcentage en `gc` que les HMM et leur changements d'états brutaux. D'autres applications comme la recherche d'origines de réplications et particulièrement la recherche de mots exceptionnels sont des exemples importants des possibilités des DMM. Nous concluons que les DMM forment un outil très utile à l'analyse statistique des séquences biologiques. Ils offrent une description détaillée de la séquence et peuvent être utilisés pour l'analyse structurale ou bien pour des applications biologiques directes.

De plus, il serait intéressant de ne pas limiter notre étude à la dérive polynomiale et à la dérive par splines polynomiales. L'ajustement de nouveaux modèles avec des covariables telles le pourcentage en `gc`, le degré d'hydrophobicité ou bien un indicateur de la structure de la protéine (hélice α, feuillet β ...). Des dérives autres que polynomiales sont envisageables : dérives exponentielles, sinusoïdales.... Toute fonction est envisageable sans perdre de vue l'intérêt biologique de cette fonction.

L'utilisation des DMM en phylogénie est aussi une source de travail intéressant : au lieu de modéliser l'évolution d'une séquence, nous modéliserions l'évolution des séquences par un DMM.

Une autre perspective plus théorique est de réduire le nombre de paramètres des modèles de Markov dérivants. Pour cela, l'utilisation de modèles MTD (*mixture transition distribution*) introduit par Raftery ([BR02, Raf85]) et/ou de chaîne de Markov à longueurs variables [BW99] est envisageable.

La perspective la plus intéressante reste finalement celle évoquée de multiples fois dans ce manuscrit : combiner HMM et DMM. Nous estimerions une chaîne de Markov cachée pour laquelle chaque état serait un modèle de Markov dérivant.

Annexes

Annexe A

Estimation de Π_0 et Π_1

A.1 Estimation de Π_0 et Π_1 par régression matricielle

Nous présentons ici les calculs complémentaires relatifs au 1.1.2 de la partie I.

A.1.1 Stochasticité des matrices $\widehat{\Pi_0}$ et $\widehat{\Pi_1}$

Preuve (Théorème 3 p29 : Les matrices $\widehat{\Pi_0}$ et $\widehat{\Pi_1}$ sont stochastiques) *Il suffit de montrer que la somme des termes de chaque ligne est égale à 1 et que tous les termes sont positifs. Rappelons que les matrices $\widehat{\Pi_{S_\ell}}$ sont stochastiques. Nous avons donc*

- $\forall \ell \in [\![1, N]\!] \sum_{v \in \mathcal{A}} \widehat{\Pi_{S_\ell}}(u, v) = 1$
- $\forall \ell \in [\![1, N]\!] \forall (u, v) \in \mathcal{A}^k \times \mathcal{A}, \quad \widehat{\Pi_{S_\ell}}(u, v) > 0$

• *Montrons tout d'abord que la somme sur chaque ligne vaut* 1.

$$\begin{aligned}
\sum_{v \in \mathcal{A}} \widehat{\Pi_0}(u, v) &= \frac{\left(\sum_{\ell=1}^{N} \tau_\ell\right)\left(\sum_{\ell=1}^{N} \tau_\ell \sum_{v \in \mathcal{A}} \widehat{\Pi_{S_\ell}}(u, v)\right) - \left(\sum_{\ell=1}^{N} {\tau_\ell}^2\right)\left(\sum_{\ell=1}^{N} \sum_{v \in \mathcal{A}} \widehat{\Pi_{S_\ell}}(u, v)\right)}{\left(\sum_{\ell=1}^{N} \tau_\ell(1-\tau_\ell)\right)\left(\sum_{\ell=1}^{N} \tau_\ell\right) - \left(\sum_{\ell=1}^{N} 1-\tau_\ell\right)\left(\sum_{\ell=1}^{N} {\tau_\ell}^2\right)} \\
&= \frac{\sum_{\ell=1}^{N} \tau_\ell \sum_{\ell=1}^{N} \tau_\ell - \sum_{\ell=1}^{N} {\tau_\ell}^2 \sum_{\ell=1}^{N} 1}{\sum_{\ell=1}^{N} \tau_\ell \sum_{\ell=1}^{N} \tau_\ell - \sum_{\ell=1}^{N} {\tau_\ell}^2 \sum_{\ell=1}^{N} \tau_\ell - \sum_{\ell=1}^{N} 1 \sum_{\ell=1}^{N} {\tau_\ell}^2 + \sum_{\ell=1}^{N} \tau_\ell \sum_{\ell=1}^{N} {\tau_\ell}^2} \\
&= \frac{\sum_{\ell=1}^{N} \tau_\ell \sum_{\ell=1}^{N} \tau_\ell - \sum_{\ell=1}^{N} {\tau_\ell}^2 \sum_{\ell=1}^{N} 1}{\sum_{\ell=1}^{N} \tau_\ell \sum_{\ell=1}^{N} \tau_\ell - \sum_{\ell=1}^{N} {\tau_\ell}^2 \sum_{\ell=1}^{N} 1} = 1.
\end{aligned}$$

$$\begin{aligned}
\sum_{v \in \mathcal{A}} \widehat{\Pi_1}(u, v) &= \frac{\left(\sum_{\ell=1}^{N} \tau_\ell(1-\tau_\ell)\right)\left(\sum_{\ell=1}^{N} \sum_{v \in \mathcal{A}} \widehat{\Pi_{S_\ell}}(u, v)\right) - \left(\sum_{\ell=1}^{N} 1-\tau_\ell\right)\left(\sum_{\ell=1}^{N} \tau_\ell \sum_{v \in \mathcal{A}} \widehat{\Pi_{S_\ell}}(u, v)\right)}{\left(\sum_{\ell=1}^{N} \tau_\ell(1-\tau_\ell)\right)\left(\sum_{\ell=1}^{N} \tau_\ell\right) - \left(\sum_{\ell=1}^{N} 1-\tau_\ell\right)\left(\sum_{\ell=1}^{N} {\tau_\ell}^2\right)} \\
&= \frac{\sum_{\ell=1}^{N} \tau_\ell \sum_{\ell=1}^{N} 1 - \sum_{\ell=1}^{N} {\tau_\ell}^2 \sum_{\ell=1}^{N} 1 - \sum_{\ell=1}^{N} 1 \sum_{\ell=1}^{N} \tau_\ell + \sum_{\ell=1}^{N} \tau_\ell \sum_{\ell=1}^{N} \tau_\ell}{\sum_{\ell=1}^{N} \tau_\ell \sum_{\ell=1}^{N} \tau_\ell - \sum_{\ell=1}^{N} {\tau_\ell}^2 \sum_{\ell=1}^{N} \tau_\ell - \sum_{\ell=1}^{N} 1 \sum_{\ell=1}^{N} {\tau_\ell}^2 + \sum_{\ell=1}^{N} \tau_\ell \sum_{\ell=1}^{N} {\tau_\ell}^2}
\end{aligned}$$

$$= \frac{\sum_{\ell=1}^{N}\tau_\ell \sum_{\ell=1}^{N}\tau_\ell - \sum_{\ell=1}^{N}{\tau_\ell}^2 \sum_{\ell=1}^{N}1}{\sum_{\ell=1}^{N}\tau_\ell \sum_{\ell=1}^{N}\tau_\ell - \sum_{\ell=1}^{N}{\tau_\ell}^2 \sum_{\ell=1}^{N}1} = 1.$$

• *Montrons désormais que tous les termes sont positifs.*

$$\begin{aligned}
\widehat{\Pi_0}(u,v) &= \frac{\left(\sum_{\ell=1}^{N}\tau_\ell\right)\left(\sum_{\ell=1}^{N}\tau_\ell\widehat{\Pi_{S_\ell}}(u,v)\right) - \left(\sum_{\ell=1}^{N}{\tau_\ell}^2\right)\left(\sum_{\ell=1}^{N}\widehat{\Pi_{S_\ell}}(u,v)\right)}{\sum_{\ell=1}^{N}\tau_\ell \sum_{\ell=1}^{N}\tau_\ell - \sum_{\ell=1}^{N}{\tau_\ell}^2 \sum_{\ell=1}^{N}1} \\
&= \frac{\sum_{\ell=1}^{N}{\tau_\ell}^2\widehat{\Pi_{S_\ell}}^2(u,v) - \sum_{\ell=1}^{N}{\tau_\ell}^2\widehat{\Pi_{S_\ell}}^2(u,v)}{\sum_{\ell=1}^{N}{\tau_\ell}^2 + \left(\sum_{(\ell_1,\ell_2)\in[\![1,N]\!]^2,\ell_1\neq\ell_2}\tau_{\ell_1}\tau_{\ell_2}\right) - \sum_{\ell=1}^{N}{\tau_\ell}^2 - \left(\sum_{(\ell_1,\ell_2)\in[\![1,N]\!]^2,\ell_1\neq\ell_2}{\tau_{\ell_1}}^2\right)} \\
&+ \frac{\sum_{(\ell_1,\ell_2)\in[\![1,N]\!]^2,\ell_1\neq\ell_2}\tau_{\ell_1}\tau_{\ell_2}\widehat{\Pi_{S_{\ell_2}}}(u,v) - \sum_{(\ell_1,\ell_2)\in[\![1,N]\!]^2,\ell_1\neq\ell_2}{\tau_{\ell_1}}^2\widehat{\Pi_{S_{\ell_2}}}(u,v)}{\sum_{\ell=1}^{N}{\tau_\ell}^2 + \left(\sum_{(\ell_1,\ell_2)\in[\![1,N]\!]^2,\ell_1\neq\ell_2}\tau_{\ell_1}\tau_{\ell_2}\right) - \sum_{\ell=1}^{N}{\tau_\ell}^2 - \left(\sum_{(\ell_1,\ell_2)\in[\![1,N]\!]^2,\ell_1\neq\ell_2}{\tau_{\ell_1}}^2\right)} \\
&= \frac{\sum_{(\ell_1,\ell_2)\in[\![1,N]\!]^2,\ell_1\neq\ell_2}(\tau_{\ell_2}-\tau_{\ell_1})\tau_{\ell_1}\widehat{\Pi_{S_{\ell_2}}}(u,v)}{\sum_{(\ell_1,\ell_2)\in[\![1,N]\!]^2,\ell_1\neq\ell_2}(\tau_{\ell_2}-\tau_{\ell_1})\tau_{\ell_1}}.
\end{aligned}$$

$$\begin{aligned}
\widehat{\Pi_1}(u,v) &= \frac{\left(\sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell)\right)\left(\sum_{\ell=1}^{N}\widehat{\Pi_{S_\ell}}(u,v)\right) - \left(\sum_{\ell=1}^{N}1-\tau_\ell\right)\left(\sum_{\ell=1}^{N}\tau_\ell\widehat{\Pi_{S_\ell}}(u,v)\right)}{\sum_{\ell=1}^{N}\tau_\ell \sum_{\ell=1}^{N}\tau_\ell - \sum_{\ell=1}^{N}{\tau_\ell}^2 \sum_{\ell=1}^{N}1} \\
&= \frac{\sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell)\widehat{\Pi_{S_\ell}}(u,v) - \sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell)\widehat{\Pi_{S_\ell}}(u,v)}{\sum_{(\ell_1,\ell_2)\in[\![1,N]\!]^2,\ell_1\neq\ell_2}(\tau_{\ell_2}-\tau_{\ell_1})\tau_{\ell_1}} \\
&+ \frac{\sum_{(\ell_1,\ell_2)\in[\![1,N]\!]^2,\ell_1\neq\ell_2}\tau_{\ell_1}(1-\tau_{\ell_1})\widehat{\Pi_{S_{\ell_2}}}(u,v) - \sum_{(\ell_1,\ell_2)\in[\![1,N]\!]^2,\ell_1\neq\ell_2}(1-\tau_{\ell_1})\tau_{\ell_2}\widehat{\Pi_{S_{\ell_2}}}(u,v)}{\sum_{(\ell_1,\ell_2)\in[\![1,N]\!]^2,\ell_1\neq\ell_2}(\tau_{\ell_2}-\tau_{\ell_1})\tau_{\ell_1}} \\
&= \frac{\sum_{(\ell_1,\ell_2)\in[\![1,N]\!]^2,\ell_1\neq\ell_2}(\tau_{\ell_1}-\tau_{\ell_2})(1-\tau_{\ell_1})\widehat{\Pi_{S_{\ell_2}}}(u,v)}{\sum_{(\ell_1,\ell_2)\in[\![1,N]\!]^2,\ell_1\neq\ell_2}(\tau_{\ell_2}-\tau_{\ell_1})\tau_{\ell_1}}.
\end{aligned}$$

Nous savons que $\forall \ell \in [\![1,N]\!], 0 < \tau_\ell < 1$. *Ainsi,*

$$\sum_{(\ell_1,\ell_2)\in[\![1,N]\!]^2,\ell_1\neq\ell_2}(\tau_{\ell_2}-\tau_{\ell_1})\tau_{\ell_1} = \sum_{(\ell_1,\ell_2)\in[\![1,N]\!]^2,\ell_1<\ell_2}(\tau_{\ell_2}-\tau_{\ell_1})\tau_{\ell_1} + (\tau_{\ell_1}-\tau_{\ell_2})\tau_{\ell_2}$$

$$= - \sum_{(\ell_1,\ell_2)\in[\![1,N]\!]^2,\ell_1<\ell_2} (\tau_{\ell_1} - \tau_{\ell_2})^2 \leq 0.$$

Il resterait à montrer que

$$\sum_{(\ell_1,\ell_2)\in[\![1,N]\!]^2,\ell_1\neq\ell_2} (\tau_{\ell_2} - \tau_{\ell_1})\tau_{\ell_1}\widehat{\Pi_{S_{\ell_2}}}(u,v) < 0$$

et que

$$\sum_{(\ell_1,\ell_2)\in[\![1,N]\!]^2,\ell_1\neq\ell_2} (\tau_{\ell_1} - \tau_{\ell_2})(1-\tau_{\ell_1})\widehat{\Pi_{S_{\ell_2}}}(u,v) < 0.$$

Les termes ne sont donc pas toujours positifs. Lors de l'estimation nous proposons un réajustement proportionnel des valeurs (voir A.1.2).

A.1.2 Valeurs négatives dans les matrices estimées

Ce paragraphe a pour but de clarifier une fois pour toutes le réajustement d'une matrice dont les lignes somment à 1 mais comportant des valeurs négatives. Les valeurs négatives étant en pratique toujours très proches de 0, nous avons choisi la solution simple de remplacer la ligne $(\ a \quad b \quad c \quad d\ )$ d'une matrice par $\left(\ \frac{|a|}{T} \quad \frac{|b|}{T} \quad \frac{|c|}{T} \quad \frac{|d|}{T}\ \right)$ où $T = |a| + |b| + |c| + |d|$. Nous pouvons bien sûr envisager d'autres méthodes plus complexes mais cela semble inutile en pratique.

Exemple 15 *On estime un modèle d'ordre* 1 *et de degré* 8 *sur le virus du sida. Le Tableau A.1 montre la matrice de transition* Π *en position* 0 *avant et après réajustement pour constater qu'elles sont très ressemblantes.*

TAB. A.1 – Réajustement des valeurs négatives

Π avant				Π après			
0.253113	0.292272	0.322449	0.132166	0.253113	0.292272	0.322449	0.132166
0.300977	0.440376	0.150777	0.10787	0.300977	0.440376	0.150777	0.10787
0.394985	−0.007905	0.378852	0.234068	0.388837	0.007782	0.372956	0.230425
0.228737	0.294001	0.113683	0.363579	0.228737	0.294001	0.113683	0.363579

Remarque 14 *Le principe d'une dérive linéaire est de supposer que les données forment approximativement une droite dans l'espace des matrices considéré. Si nous obtenons des valeurs négatives pour les matrices de transition, le modèle de variation linéaire de la matrice n'est sans doute pas raisonnable. Il en va de même pour des dérives de degrés supérieurs. Dans ce cas, la recherche d'un autre modèle est nécessaire.*

A.1.3 Distances : deuxième méthode

Au lieu de choisir un seul point par segment, nous choisissons tous les points du segment pour ajuster notre modèle. Nous minimisons donc la somme suivante, pour une distance matricielle d choisie

$$\sum_{t=k}^{n} d\left(\widehat{\Pi_{S_{\ell_t}}}, \left(1-\frac{t}{n}\right)\Pi_0 + \frac{t}{n}\Pi_1\right),$$

où $\ell_t = \left[\frac{t-k}{m}\right]$ avec [.] désignant la partie entière. Nous devons donc minimiser la fonction suivante :

$$\sum_{t=k}^{n}\sum_{u\in\mathcal{A}^k}\sum_{v\in\mathcal{A}}\left(\widehat{\Pi_{S_{\ell_t}}}(u,v) - \left(1-\frac{t}{n}\right)\Pi_0(u,v) - \frac{t}{n}\Pi_1(u,v)\right)^2$$

De la même manière que pour un seul point, nous obtenons les estimateurs de Π_0 et Π_1. La stochasticité de ces estimateurs se démontre de la même manière.

Nous constatons dans le tableau A.2 que la différence d'estimation est minime alors que le gain de calculs du choix d'un seul point est grand. La log-vraisemblance est meme quasiment identique (à l'unité près). Nous avons donc décidé de ne pas approfondir cette méthode.

TAB. A.2 – Sur le phage Lambda, comparaison entre deux méthodes d'estimation : un seul point / tous les points. La moyenne des écarts est de $2.9.10^{-4}$ pour Π_0 et de $5.9.10^{-4}$ pour Π_1.

	Un seul point				Tous les points			
Π_0	0.246962	0.262711	0.253974	0.236353	0.246383	0.262735	0.254068	0.236814
	0.233381	0.281842	0.299379	0.185398	0.233398	0.281871	0.29892	0.185811
	0.235837	0.331655	0.244713	0.187795	0.235421	0.332072	0.244999	0.187508
	0.131972	0.43156	0.219167	0.217301	0.131662	0.431773	0.218864	0.217701
Π_1	0.331785	0.188344	0.171998	0.307873	0.332946	0.188298	0.171814	0.306942
	0.283362	0.193407	0.259235	0.263996	0.283326	0.193354	0.260158	0.263162
	0.336346	0.206012	0.186051	0.271591	0.337179	0.205179	0.185478	0.272164
	0.220492	0.242692	0.235669	0.301147	0.221114	0.242272	0.236275	0.300339

A.1.4 Espérances et variances de $\widehat{\Pi_{S_\ell}}$ pour le modèle des chaînes de Markov classiques

L'unique but de ce calcul est de le comparer au suivant. Nous calculons sur chaque segment, espérances et variances des estimateurs des matrices de transition sous un modèle de Markov classique (homogène le long du segment).

Théorème 14 *Les estimateurs $\widehat{\Pi_{S_\ell}}(u,v)$ sont convergents.*

Preuve *Nous posons $N_{S_\ell}(uv)$ le nombre de u suivis d'un v et $N_{S_\ell}(u+)$ le nombre de u du segment S_ℓ. Sur le segment S_ℓ, la probabilité de voir apparaître un v après un u ne change pas, la variable aléatoire du nombre de uv sachant le nombre de u suit donc une loi binomiale :*

$$N_{S_\ell}(uv)|N_{S_\ell}(u+) \sim \mathcal{B}(N_{S_\ell}(u+), \Pi_{S_\ell}(u,v)).$$

Nous en déduisons l'espérance de $\widehat{\Pi_{S_\ell}}(u,v)$ pour $\ell \in [\![1,N]\!]$:

$$\begin{aligned}
\mathbb{E}(N_{S_\ell}(uv)|N_{S_\ell}(u+)) &= N_{S_\ell}(u+)\Pi_{S_\ell}(u,v) \\
\mathbb{E}\left(\frac{N_{S_\ell}(uv)}{N_{S_\ell}(u+)}\middle| N_{S_\ell}(u+)\right) &= \Pi_{S_\ell}(u,v) \\
\mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right) &= \mathbb{E}\left(\frac{N_{S_\ell}(uv)}{N_{S_\ell}(u+)}\right) = \mathbb{E}\left(\mathbb{E}\left(\frac{N_{S_\ell}(uv)}{N_{S_\ell}(u+)}\middle| N_{S_\ell}(u+)\right)\right) = \Pi_{S_\ell}(u,v).
\end{aligned}$$

Nous avons

$$\mathbb{V}(N_{S_\ell}(uv)|N_{S_\ell}(u+)) = N_{S_\ell}(u+)\Pi_{S_\ell}(u,v)\left(1-\Pi_{S_\ell}(u,v)\right).$$

À l'aide du théorème de la variance totale, nous obtenons la variance de $\widehat{\Pi_{S_\ell}}(u,v)$ pour $\ell \in [\![1,N]\!]$:

$$\begin{aligned}
\mathbb{V}(\widehat{\Pi_{S_\ell}}(u,v)) &= \mathbb{V}\left(\frac{N_{S_\ell}(uv)}{N_{S_\ell}(u+)}\right) \\
&= \mathbb{E}\left(\mathbb{V}\left(\frac{N_{S_\ell}(uv)}{N_{S_\ell}(u+)}\middle| N_{S_\ell}(u+)\right)\right) + \mathbb{V}\left(\mathbb{E}\left(\frac{N_{S_\ell}(uv)}{N_{S_\ell}(u+)}\middle| N_{S_\ell}(u+)\right)\right) \\
&= \mathbb{E}\left(\frac{\Pi_{S_\ell}(u,v)(1-\Pi_{S_\ell}(u,v))}{N_{S_\ell}(u+)}\right) + \mathbb{V}\left(\Pi_{S_\ell}(u,v)\right) \\
&= \Pi_{S_\ell}(u,v)(1-\Pi_{S_\ell}(u,v)) \sum_{i=1}^{|S_\ell|-k} \frac{1}{i}\mathbb{P}(N_{S_\ell}(u+)=i).
\end{aligned}$$

Un modèle de Markov se retrouvant rapidement en régime stationnaire, il est raisonnable de considérer que

$$N_{S_\ell}(u+)+1 \sim \mathcal{B}(|S_\ell|-k-1, \mu_\ell(u)).$$

Ainsi

$$\begin{aligned}\mathbb{V}(\widehat{\Pi_{S_\ell}}(u,v)) &= \Pi_{S_\ell}(u,v)(1-\Pi_{S_\ell}(u,v))\sum_{i=1}^{|S_\ell|-k}\frac{1}{i}\mathrm{C}^{i-1}_{|S_\ell|-k-1}\mu_\ell(u)^{i-1}(1-\mu_\ell(u))^{|S_\ell|-k-i}\\ &= \frac{\Pi_{S_\ell}(u,v)(1-\Pi_{S_\ell}(u,v))}{\mu_\ell(u)(|S_\ell|-k)}\sum_{i=1}^{|S_\ell|-k}\mathrm{C}^{i}_{|S_\ell|-k}\mu_\ell(u)^{i}(1-\mu_\ell(u))^{|S_\ell|-k-i}\\ &= \Pi_{S_\ell}(u,v)(1-\Pi_{S_\ell}(u,v))\left(\frac{1-(1-\mu_\ell(u))^{|S_\ell|-k}}{\mu_\ell(u)(|S_\ell|-k)}\right).\end{aligned}$$

Les estimateurs $\widehat{\Pi_{S_\ell}}(u,v)$ sont donc sans biais et de variance qui tend vers zéro quand $|S_\ell|\to+\infty$, et par conséquent convergents.

A.1.5 Espérances et variances de $\widehat{\Pi_{S_\ell}}$ pour le modèle des chaînes de Markov régulées

Pour le modèle de Markov régulé, la loi de $N_{S_\ell}(uv)|N_{S_\ell}(u+)$ ne peut pas être approchée par une binomiale de paramètres $N_{S_\ell}(u+)$ et $\Pi_{S_\ell}(u,v)$. En effet, la probabilité d'apparition d'un v derrière un u dépend de la position dans le segment.

Calcul de $\mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)$

Preuve (Théorème 4 p29)

$$\begin{aligned}&\mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)\\ =\ &\sum_{t\in S^\star_\ell}\mathbb{E}\left(\frac{\mathbb{1}\{X_{t-k}\dots X_{t-1}=u,X_t=v\}}{\sum_{j\in S^\star_\ell}\mathbb{1}\{X_{j-k}\dots X_{j-1}=u\}}\right)\\ =\ &\sum_{t\in S^\star_\ell}\sum_{i\in[\![1,|S^\star_\ell|]\!]}\frac{1}{i}\mathbb{P}\left(X_{t-k}\dots X_{t-1}=u,X_t=v,\sum_{j\in S^\star_\ell}\mathbb{1}\{X_{j-k}\dots X_{j-1}=u\}=i\right)\\ =\ &\sum_{t\in S^\star_\ell}\sum_{i\in[\![1,|S^\star_\ell|]\!]}\frac{1}{i}\mathbb{P}\left(\sum_{j\in S^\star_\ell}\mathbb{1}\{X_{j-k}\dots X_{j-1}=u\}=i\,\middle|\,X_{t-k}\dots X_{t-1}=u,X_t=v\right)\\ &\mathbb{P}(X_t=v|X_{t-k}\dots X_{t-1}=u)\mathbb{P}(X_{t-k}\dots X_{t-1}=u)\\ =\ &\sum_{t\in S^\star_\ell}\Pi_{\frac{t}{n}}(u,v)\mu_\ell(u)\sum_{i\in[\![1,|S^\star_\ell|]\!]}\frac{1}{i}\mathbb{P}\left(\sum_{j\in S^\star_\ell}\mathbb{1}\{X_{j-k}\dots X_{j-1}=u\}=i\,\middle|\,X_{t-k}\dots X_{t-1}=u,X_t=v\right).\end{aligned}$$

La troisième égalité vient de la formule de multiplication. Pour approcher la loi du nombre de u dans le segment S_ℓ, nous définissons μ_ℓ loi stationnaire de Π_{τ_ℓ}. Les modèles de Markov régulés décrivant une variation douce de la matrice il est raisonnable de penser qu'en moyenne, la probabilité d'apparition d'un u sera définie par μ_ℓ et qu'ainsi le nombre de u d'un segment suit une loi binomiale :

$$\forall t\in S^\star_\ell\ \sum_{j\in S^\star_\ell,j\neq t}\mathbb{1}\{X_{j-k}\dots X_{j-1}=u\}\sim\mathcal{B}(m-k-1,\mu_\ell(u)).$$

Cette supposition est moins forte que celle qui associe directement au nombre de uv une loi binomiale. Nous obtenons ainsi

$$\begin{aligned}&\sum_{i\in[\![1,|S^\star_\ell|]\!]}\frac{1}{i}\mathbb{P}\left(\sum_{j\in S^\star_\ell}\mathbb{1}\{X_{j-k}\dots X_{j-1}=u\}=i\,\middle|\,X_{t-k}\dots X_{t-1}=u,X_t=v\right)\\ \simeq\ &\sum_{i\in[\![1,|S^\star_\ell|]\!]}\frac{1}{i}\mathbb{P}\left(\sum_{j\in S^\star_\ell,j\neq t}\mathbb{1}\{X_{j-k}\dots X_{j-1}=u\}=i-1\right)\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{|S_\ell^\star|} \frac{1}{i} \mathrm{C}_{|S_\ell^\star|-1}^{i-1} (1-\mu_\ell(u))^{|S_\ell^\star|-i} \mu_\ell(u)^{i-1} \\
&= \frac{1}{|S_\ell^\star|\mu_\ell(u)} \sum_{i=1}^{|S_\ell^\star|} \mathrm{C}_{|S_\ell^\star|}^{i} (1-\mu_\ell(u))^{|S_\ell^\star|-i} \mu_\ell(u)^{i} \\
&= \frac{1}{|S_\ell^\star|\mu_\ell(u)} \left(\sum_{i=1}^{|S_\ell^\star|} \mathrm{C}_{|S_\ell^\star|}^{i} (1-\mu_\ell(u))^{|S_\ell^\star|-i} \mu_\ell(u)^{i} \right) \\
&= \frac{1}{|S_\ell^\star|\mu_\ell(u)} \left(1-(1-\mu_\ell(u))^{|S_\ell^\star|} \right).
\end{aligned}
$$

L'approximation faite lors de la première égalité est négligeable. La connaissance de $X_t = v$ n'influe en effet que très faiblement sur le nombre de u du segment. Nous obtenons alors

$$
\begin{aligned}
\mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right) &= \left(1-(1-\mu_\ell(u))^{|S_\ell^\star|}\right) \left(\sum_{t\in S_\ell^\star} \frac{\left(1-\frac{t}{n}\right)\Pi_0(u,v) + \left(\frac{t}{n}\right)\Pi_1(u,v)}{|S_\ell^\star|} \right) \\
&= \left(1-(1-\mu_\ell(u))^{|S_\ell^\star|}\right) \left(\left(1 - \sum_{t\in S_\ell^\star} \frac{t}{|S_\ell^\star| n}\right)\Pi_0(u,v) + \sum_{t\in S_\ell^\star} \frac{t}{|S_\ell^\star| n}\Pi_1(u,v) \right).
\end{aligned}
$$

Sachant que $\displaystyle\sum_{t\in S_\ell^\star} \frac{t}{|S_\ell^\star|} = \frac{a+b}{2}$ *(avec a le premier élément de $S_\ell^\star$ et b le dernier), nous obtenons les approximations suivantes pour les estimateurs de Π_{S_ℓ} :*

$$
\begin{aligned}
\mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right) &\approx (1-(1-\mu_\ell(u))^{m-k}) \Pi_{\frac{2\ell m - m + k - 1}{2n}}(u,v), \\
\mathbb{E}\left(\widehat{\Pi_{S_N}}(u,v)\right) &\approx (1-(1-\mu_\ell(u))^{n-(N-1)m-k+1}) \Pi_{\frac{(N-1)m+k+n}{2n}}(u,v).
\end{aligned}
$$

Les estimateurs $\widehat{\Pi_{S_\ell}}(u,v)$ sont donc asymptotiquement sans biais.

Calcul de $\mathbb{V}\left(\widehat{\Pi_{S_\ell}}\right)$

Preuve (Théorème 5 p29) *Nous savons que :*

$$
\mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right) = \mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)^2\right) - \mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)^2.
$$

Nous avons l'égalité $(\star)$:

$$
\begin{aligned}
&\mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)^2\right) \\
&= \sum_{t\in S_\ell^\star} \mathbb{E}\left(\frac{\mathbb{1}\{X_{t-k}\dots X_{t-1} = u, X_t = v\}}{\left(\sum_{j\in S_\ell^\star} \mathbb{1}\{X_{j-k}\dots X_{j-1} = u\}\right)^2} \right) \\
&+ 2 \sum_{(t_1,t_2)\in S_\ell^\star \times S_\ell^\star, t_1<t_2} \mathbb{E}\left(\frac{\mathbb{1}\{X_{t_1-k}\dots X_{t_1-1} = u, X_{t_1} = v, X_{t_2-k}\dots X_{t_2-1} = u, X_{t_2} = v\}}{\left(\sum_{j\in S_\ell^\star} \mathbb{1}\{X_{j-k}\dots X_{j-1} = u\}\right)^2} \right)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{t\in S_\ell^\star} \Pi_{\frac{t}{n}}(u,v)\mu_\ell(u) \sum_{i\in[\![1,|S_\ell^\star|]\!]} \frac{1}{i^2}\mathbb{P}\left(\sum_{j\in S_\ell^\star} \mathbb{1}\{X_{j-k}\dots X_{j-1}=u\}=i \,\middle|\, X_{t-k}\dots X_{t-1}=u, X_t=v\right)\\
&+ 2\sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star, t_1<t_2} \Pi_{\frac{t_1}{n}}(u,v)\Pi_{\frac{t_2}{n}}(u,v)\mu_\ell(u)^2 \sum_{i\in[\![2,|S_\ell^\star|]\!]}\frac{1}{i^2}\\
&\quad \mathbb{P}\left(\sum_{j\in S_\ell^\star} \mathbb{1}\{X_{j-k}\dots X_{j-1}=u\}=i \,\middle|\, X_{t-k}\dots X_{t_1-1}=u, X_{t_1}=v, X_{t_2-k}\dots X_{t_2-1}=u, X_{t_2}=v\right).
\end{aligned}
$$

La formule de multiplication nous donne

$$
\begin{aligned}
&\mathbb{P}\left(\sum_{j\in S_\ell^\star} \mathbb{1}\{X_{j-k}\dots X_{j-1}=u\}=i, X_{t_1-k}\dots X_{t_1-1}=u, X_{t_1}=v, X_{t_2-k}\dots X_{t_2-1}=u, X_{t_2}=v\right)\\
=\ &\mathbb{P}\left(\sum_{j\in S_\ell^\star} \mathbb{1}\{X_{j-k}\dots X_{j-1}=u\}=i \,\middle|\, X_{t_1-k}\dots X_{t_1-1}=u, X_{t_1}=v, X_{t_2-k}\dots X_{t_2-1}=u, X_{t_2}=v\right)\\
&\mathbb{P}\left(X_{t_2}=v|X_{t_1-k}\dots X_{t_1-1}=u, X_{t_1}=v, X_{t_2-k}\dots X_{t_2-1}=u\right)\\
&\mathbb{P}\left(X_{t_2-k}\dots X_{t_2-1}=u|X_{t_1-k}\dots X_{t_1-1}=u, X_{t_1}=v\right)\\
&\mathbb{P}\left(X_{t_1}=v|X_{t_1-k}\dots X_{t_1-1}=u\right)\mathbb{P}\left(X_{t_1-k}\dots X_{t_1-1}=u\right)
\end{aligned}
$$

et nous avons utilisé les approximations suivantes (raisonnables du fait de la faible dépendance des événements) :

- $\mathbb{P}\left(X_{t_2}=v|X_{t_1-k}\dots X_{t_1-1}=u, X_{t_1}=v, X_{t_2-k}\dots X_{t_2-1}=u\right)$ *par* $\Pi_{\frac{t_2}{n}}(u,v)$
- $\mathbb{P}\left(X_{t_2-k}\dots X_{t_2-1}=u|X_{t_1-k}\dots X_{t_1-1}=u, X_{t_1}=v\right)$ *par* $\mu_\ell(u)$
- $\mathbb{P}\left(X_{t_1}=v|X_{t_1-k}\dots X_{t_1-1}=u\right)$ *par* $\Pi_{\frac{t_1}{n}}(u,v)$
- $\mathbb{P}\left(X_{t_1-k}\dots X_{t_1-1}=u\right)$ *par* $\mu_\ell(u)$.

• *Calcul du premier terme de la somme* $(\star)$

$$
\begin{aligned}
&\sum_{t\in S_\ell^\star} \Pi_{\frac{t}{n}}(u,v)\mu_\ell(u) \sum_{i\in[\![1,|S_\ell^\star|]\!]} \frac{1}{i^2}\mathbb{P}\left(\sum_{j\in S_\ell^\star} \mathbb{1}\{X_{j-k}\dots X_{j-1}=u\}=i \,\middle|\, X_{t-k}\dots X_{t-1}=u, X_t=v\right)\\
=\ &\sum_{t\in S_\ell^\star} \Pi_{\frac{t}{n}}(u,v)\mu_\ell(u) \sum_{i\in[\![1,|S_\ell^\star|]\!]} \frac{1}{i^2}\mathbb{P}\left(\sum_{j\in S_\ell^\star, j\neq t} \mathbb{1}\{X_{j-k}\dots X_{j-1}=u\}=i-1\right)\\
=\ &\sum_{t\in S_\ell^\star} \Pi_{\frac{t}{n}}(u,v)\mu_\ell(u) \sum_{i=1}^{|S_\ell^\star|} \frac{1}{i^2}\mathrm{C}^{i-1}_{|S_\ell^\star|-1}(1-\mu_\ell(u))^{|S_\ell^\star|-i}\mu_\ell(u)^{i-1}\\
=\ &\sum_{t\in S_\ell^\star} \frac{\Pi_{\frac{t}{n}}(u,v)}{|S_\ell^\star|} \sum_{i=1}^{|S_\ell^\star|} \frac{1}{i}\mathrm{C}^{i}_{|S_\ell^\star|}(1-\mu_\ell(u))^{|S_\ell^\star|-i}\mu_\ell(u)^{i}\\
=\ &A\sum_{t\in S_\ell^\star} \frac{\Pi_{\frac{t}{n}}(u,v)}{|S_\ell^\star|}.
\end{aligned}
$$

Nous avons

$$
\begin{aligned}
&\sum_{i=1}^{|S_\ell^\star|} \frac{1}{i+1}\mathrm{C}^{i}_{|S_\ell^\star|}(1-\mu_\ell(u))^{|S_\ell^\star|-i}\mu_\ell(u)^{i}\\
=\ &\frac{1}{|S_\ell^\star|\mu_\ell(u)} \sum_{i=1}^{|S_\ell^\star|} \mathrm{C}^{i+1}_{|S_\ell^\star|+1}(1-\mu_\ell(u))^{|S_\ell^\star|-i}\mu_\ell(u)^{i+1}\\
=\ &\frac{1}{|S_\ell^\star|\mu_\ell(u)} \sum_{i=2}^{|S_\ell^\star|+1} \mathrm{C}^{i}_{|S_\ell^\star|+1}(1-\mu_\ell(u))^{|S_\ell^\star|+1-i}\mu_\ell(u)^{i}
\end{aligned}
$$

$$= \frac{1-(1-\mu_\ell(u))^{|S_\ell^\star|+1}-(|S_\ell^\star|+1)\mu_\ell(u)(1-\mu_\ell(u))^{|S_\ell^\star|}}{|S_\ell^\star|\mu_\ell(u)}.$$

Sachant que $\frac{1}{i+1} \leq \frac{1}{i} \leq \frac{2}{i+1}$,

$$A_- \leq A \leq A_+$$

avec

$$A_- = \frac{1-(1-\mu_\ell(u))^{|S_\ell^\star|+1}-(|S_\ell^\star|+1)\mu_\ell(u)(1-\mu_\ell(u))^{|S_\ell^\star|}}{|S_\ell^\star|\mu_\ell(u)}$$

$$A_+ = \frac{2\left(1-(1-\mu_\ell(u))^{|S_\ell^\star|+1}-(|S_\ell^\star|+1)\mu_\ell(u)(1-\mu_\ell(u))^{|S_\ell^\star|}\right)}{|S_\ell^\star|\mu_\ell(u)}.$$

• *Calcul du second terme de la somme* $(\star)$

$$\begin{aligned}
& 2\sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star, t_1<t_2} \Pi_{\frac{t_1}{n}}(u,v)\Pi_{\frac{t_2}{n}}(u,v)\mu_\ell(u)^2 \sum_{i\in[\![2,|S_\ell^\star|]\!]}\frac{1}{i^2} \\
& \mathbb{P}\left(\sum_{j\in S_\ell^\star}\mathbb{1}\{X_{j-k}\dots X_{j-1}=u\}=i \,\middle|\, X_{t-k}\dots X_{t_1-1}=u, X_{t_1}=v, X_{t_2-k}\dots X_{t_2-1}=u, X_{t_2}=v\right) \\
= \; & 2\sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star, t_1<t_2} \Pi_{\frac{t_1}{n}}(u,v)\Pi_{\frac{t_2}{n}}(u,v)\mu_\ell(u)^2 \\
& \sum_{i\in[\![2,|S_\ell^\star|]\!]}\frac{1}{i^2}\mathbb{P}\left(\sum_{j\in S_\ell^\star, j\neq t_1,t_2}\mathbb{1}\{X_{j-k}\dots X_{j-1}=u\}=i-2\right) \\
= \; & 2\sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star, t_1<t_2} \Pi_{\frac{t_1}{n}}(u,v)\Pi_{\frac{t_2}{n}}(u,v)\mu_\ell(u)^2 \sum_{i\in[\![2,|S_\ell^\star|]\!]}\frac{1}{i^2}\mathrm{C}_{|S_\ell^\star|-2}^{i-2}(1-\mu_\ell(u))^{|S_\ell^\star|-i}\mu_\ell(u)^{i-2} \\
= \; & 2\sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star, t_1<t_2} \frac{\Pi_{\frac{t_1}{n}}(u,v)\Pi_{\frac{t_2}{n}}(u,v)}{|S_\ell^\star|(|S_\ell^\star|-1)} \sum_{i\in[\![2,|S_\ell^\star|]\!]}\frac{i-1}{i}\mathrm{C}_{|S_\ell^\star|}^{i}(1-\mu_\ell(u))^{|S_\ell^\star|-i}\mu_\ell(u)^{i} \\
= \; & 2B\sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star, t_1<t_2} \frac{\Pi_{\frac{t_1}{n}}(u,v)\Pi_{\frac{t_2}{n}}(u,v)}{|S_\ell^\star|(|S_\ell^\star|-1)}.
\end{aligned}$$

Nous avons

$$\begin{aligned}
B &= \sum_{i\in[\![2,|S_\ell^\star|]\!]}\frac{i-1}{i}\mathrm{C}_{|S_\ell^\star|}^{i}(1-\mu_\ell(u))^{|S_\ell^\star|-i}\mu_\ell(u)^{i} \\
&= \sum_{i\in[\![2,|S_\ell^\star|]\!]}\mathrm{C}_{|S_\ell^\star|}^{i}(1-\mu_\ell(u))^{|S_\ell^\star|-i}\mu_\ell(u)^{i} \\
&- \sum_{i\in[\![2,|S_\ell^\star|]\!]}\frac{1}{i}\mathrm{C}_{|S_\ell^\star|}^{i}(1-\mu_\ell(u))^{|S_\ell^\star|-i}\mu_\ell(u)^{i} \\
&= \left(1-(1-\mu_\ell(u))^{|S_\ell^\star|}-|S_\ell^\star|\mu_\ell(u)(1-\mu_\ell(u))^{|S_\ell^\star|-1}\right) \\
&- \sum_{i\in[\![2,|S_\ell^\star|]\!]}\frac{1}{i}\mathrm{C}_{|S_\ell^\star|}^{i}(1-\mu_\ell(u))^{|S_\ell^\star|-i}\mu_\ell(u)^{i}
\end{aligned}$$

et

$$\begin{aligned}
& \sum_{i\in[\![2,|S_\ell^\star|]\!]}\frac{1}{i+1}\mathrm{C}_{|S_\ell^\star|}^{i}(1-\mu_\ell(u))^{|S_\ell^\star|-i}\mu_\ell(u)^{i} \\
= \; & \frac{1}{\mu_\ell(u)\left(|S_\ell^\star|+1\right)}\sum_{i\in[\![2,|S_\ell^\star|]\!]}\mathrm{C}_{|S_\ell^\star|+1}^{i+1}(1-\mu_\ell(u))^{|S_\ell^\star|-i}\mu_\ell(u)^{i+1}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\mu_\ell(u)\left(|S_\ell^\star|+1\right)} \sum_{i\in[\![3,|S_\ell^\star|+1]\!]} \mathrm{C}^i_{|S_\ell^\star|+1}(1-\mu_\ell(u))^{|S_\ell^\star|+1-i}\mu_\ell(u)^i \\
&= \frac{1-(1-\mu_\ell(u))^{|S_\ell^\star|+1}-\left(|S_\ell^\star|+1\right)(1-\mu_\ell(u))^{|S_\ell^\star|}\mu_\ell(u)-\frac{|S_\ell^\star|(|S_\ell^\star|+1)}{2}(1-\mu_\ell(u))^{|S_\ell^\star|-1}\mu_\ell(u)^2}{\mu_\ell(u)\left(|S_\ell^\star|+1\right)}.
\end{aligned}
$$

Sachant que $\frac{1}{i+1} \leq \frac{1}{i} \leq \frac{2}{i+1}$,

$$B_- \leq B \leq B_+$$

avec

$$
\begin{aligned}
B_- &= I - II_+ \\
B_+ &= I - II_- \\
I &= \left(1-(1-\mu_\ell(u))^{|S_\ell^\star|}-|S_\ell^\star|\mu_\ell(u)(1-\mu_\ell(u))^{|S_\ell^\star|-1}\right) \\
II_- &= \frac{1-(1-\mu_\ell(u))^{|S_\ell^\star|+1}-\left(|S_\ell^\star|+1\right)(1-\mu_\ell(u))^{|S_\ell^\star|}\mu_\ell(u)-\frac{|S_\ell^\star|(|S_\ell^\star|+1)}{2}(1-\mu_\ell(u))^{|S_\ell^\star|-1}\mu_\ell(u)^2}{\mu_\ell(u)\left(|S_\ell^\star|+1\right)} \\
II_+ &= \frac{2\left(1-(1-\mu_\ell(u))^{|S_\ell^\star|+1}-\left(|S_\ell^\star|+1\right)(1-\mu_\ell(u))^{|S_\ell^\star|}\mu_\ell(u)-\frac{|S_\ell^\star|(|S_\ell^\star|+1)}{2}(1-\mu_\ell(u))^{|S_\ell^\star|-1}\mu_\ell(u)^2\right)}{\mu_\ell(u)\left(|S_\ell^\star|+1\right)}.
\end{aligned}
$$

• *Encadrement de la variance*
Nous obtenons

$$
\begin{aligned}
A_- \sum_{t\in S_\ell^\star} \frac{\Pi_{\frac{t}{n}}(u,v)}{|S_\ell^\star|} + 2B_- \sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star, t_1<t_2} \frac{\Pi_{\frac{t_1}{n}}(u,v)\Pi_{\frac{t_2}{n}}(u,v)}{|S_\ell^\star|(|S_\ell^\star|-1)} - \mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)^2 \\
\leq \mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right) \leq \\
A_+ \sum_{t\in S_\ell^\star} \frac{\Pi_{\frac{t}{n}}(u,v)}{|S_\ell^\star|} + 2B_+ \sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star, t_1<t_2} \frac{\Pi_{\frac{t_1}{n}}(u,v)\Pi_{\frac{t_2}{n}}(u,v)}{|S_\ell^\star|(|S_\ell^\star|-1)} - \mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)^2
\end{aligned}
$$

Nous avons

$$
\begin{aligned}
& 2\sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star, t_1<t_2} \frac{\Pi_{\frac{t_1}{n}}(u,v)\Pi_{\frac{t_2}{n}}(u,v)}{|S_\ell^\star|(|S_\ell^\star|-1)} \\
= \Pi_0(u,v)^2 & \frac{2\sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star, t_1<t_2} 1+\frac{t_1}{n}\frac{t_2}{n}-\frac{t_1}{n}-\frac{t_2}{n}}{|S_\ell^\star|(|S_\ell^\star|-1)} \\
+ \Pi_0(u,v)\Pi_1(u,v) & \frac{2\sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star, t_1<t_2} \frac{t_1}{n}+\frac{t_2}{n}-2\frac{t_1}{n}\frac{t_2}{n}}{|S_\ell^\star|(|S_\ell^\star|-1)} \\
+ \Pi_1(u,v)^2 & \frac{2\sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star, t_1<t_2} \left(\frac{t_1}{n}\frac{t_2}{n}\right)}{|S_\ell^\star|(|S_\ell^\star|-1)}.
\end{aligned}
$$

Sachant que

- $\dfrac{2\sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star, t_1<t_2} 1}{|S_\ell^\star|(|S_\ell^\star|-1)} = 1$
- $T_1 = \dfrac{2\sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star, t_1<t_2} \frac{t_1}{n}}{|S_\ell^\star|(|S_\ell^\star|-1)} = \dfrac{2a+b-1}{3n}$

- $T_2 = \dfrac{2\sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star, t_1<t_2} \frac{t_2}{n}}{|S_\ell^\star|(|S_\ell^\star|-1)} = \dfrac{a+2b+1}{3n}$
- $T_3 = \dfrac{2\sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star, t_1<t_2} \frac{t_1}{n}\frac{t_2}{n}}{|S_\ell^\star|(|S_\ell^\star|-1)} = \dfrac{3a^2+a+6ab+3b^2-b-2}{12n^2}$
- $K = T_3 - {\tau_\ell}^2 = \dfrac{a-b-2}{12n^2}$

avec a et b premier et dernier élément de S_ℓ, nous obtenons

$$\begin{aligned}
& \Pi_0(u,v)^2\left(1-T_1-T_2+T_3\right)+\Pi_0(u,v)\Pi_1(u,v)\left(T_1+T_2-2T_3\right)+\Pi_1(u,v)^2T_3 \\
= \;& \Pi_0(u,v)^2\left(1-2\tau_\ell+{\tau_\ell}^2+K\right)+2\Pi_0(u,v)\Pi_1(u,v)\left(\tau_\ell-{\tau_\ell}^2-K\right)+\Pi_1(u,v)^2\left({\tau_\ell}^2+K\right) \\
= \;& {\Pi_{\tau_\ell}}^2(u,v)+K(\Pi_1(u,v)-\Pi_0(u,v))^2.
\end{aligned}$$

Finalement, nous avons un encadrement de la variance :

$$\begin{aligned}
& A_-\Pi_{\tau_\ell}(u,v)+B_-\left(\Pi_{\tau_\ell}(u,v)^2+K(\Pi_1(u,v)-\Pi_0(u,v))^2\right)-\Pi_{\tau_\ell}(u,v)^2\left(1-(1-\mu_\ell(u))^{|S_\ell^\star|}\right)^2 \\
& \qquad \le \mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right) \le \\
& A_+\Pi_{\tau_\ell}(u,v)+B_+\left(\Pi_{\tau_\ell}(u,v)^2+K(\Pi_1(u,v)-\Pi_0(u,v))^2\right)-\Pi_{\tau_\ell}(u,v)^2\left(1-(1-\mu_\ell(u))^{|S_\ell^\star|}\right)^2
\end{aligned}$$

Nous avons A_-, A_+ et K qui tendent vers 0, B_- et B_+ qui tendent vers 1.
La variance des estimateurs $\widehat{\Pi_{S_\ell}}(u,v)$ tend donc vers zéro quand $|S_\ell| \to +\infty$.

A.1.6 Espérances et variances de $\widehat{\Pi_0}$ et $\widehat{\Pi_1}$

Espérances

Preuve (Théorème 7 p30) *Connaissant les valeurs des $\mathbb{E}\left(\widehat{\Pi_{S_\ell}}\right)$ (voir p29), nous posons*

$$E_m = \min_\ell\left\{(1-(1-\mu_\ell(u))^{m-k}),(1-(1-\mu_\ell(u))^{n-(N-1)m-k+1})\right\}$$

et

$$E_p = \max_\ell\left\{(1-(1-\mu_\ell(u))^{m-k}),(1-(1-\mu_\ell(u))^{n-(N-1)m-k+1})\right\}.$$

Ainsi, nous avons $E_m\Pi_{\tau_\ell} \le \mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right) \le E_p\Pi_{\tau_\ell}$. Sachant que $\Pi_{\tau_\ell} = (1-\tau_\ell)\Pi_0 + \tau_\ell\Pi_1$, nous obtenons :

$$\begin{aligned}
& \frac{E_m\Pi_0\sum_{\ell=1}^N\tau_\ell\sum_{\ell=1}^N\tau_\ell\left(1-\tau_\ell\right)+E_m\Pi_1\sum_{\ell=1}^N\tau_\ell\sum_{\ell=1}^N{\tau_\ell}^2-E_p\Pi_0\sum_{\ell=1}^N{\tau_\ell}^2\sum_{\ell=1}^N(1-\tau_\ell)-E_p\Pi_1\sum_{\ell=1}^N{\tau_\ell}^2\sum_{\ell=1}^N\tau_\ell}{\left(\sum_{\ell=1}^N\tau_\ell(1-\tau_\ell)\right)\left(\sum_{\ell=1}^N\tau_\ell\right)-\left(\sum_{\ell=1}^N1-\tau_\ell\right)\left(\sum_{\ell=1}^N{\tau_\ell}^2\right)} \\
& \qquad \le \mathbb{E}\left(\widehat{\Pi_0}(u,v)\right) \le \\
& \frac{E_p\Pi_0\sum_{\ell=1}^N\tau_\ell\sum_{\ell=1}^N\tau_\ell\left(1-\tau_\ell\right)+E_p\Pi_1\sum_{\ell=1}^N\tau_\ell\sum_{\ell=1}^N{\tau_\ell}^2-E_m\Pi_0\sum_{\ell=1}^N{\tau_\ell}^2\sum_{\ell=1}^N(1-\tau_\ell)-E_m\Pi_1\sum_{\ell=1}^N{\tau_\ell}^2\sum_{\ell=1}^N\tau_\ell}{\left(\sum_{\ell=1}^N\tau_\ell(1-\tau_\ell)\right)\left(\sum_{\ell=1}^N\tau_\ell\right)-\left(\sum_{\ell=1}^N1-\tau_\ell\right)\left(\sum_{\ell=1}^N{\tau_\ell}^2\right)}
\end{aligned}$$

$$\frac{E_m\Pi_0\sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell)\sum_{\ell=1}^{N}(1-\tau_\ell)+E_m\Pi_1\sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell)\sum_{\ell=1}^{N}\tau_\ell-E_p\Pi_0\sum_{\ell=1}^{N}(1-\tau_\ell)\sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell)-E_p\Pi_1\sum_{\ell=1}^{N}(1-\tau_\ell)\sum_{\ell=1}^{N}{\tau_\ell}^2}{\left(\sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell)\right)\left(\sum_{\ell=1}^{N}\tau_\ell\right)-\left(\sum_{\ell=1}^{N}1-\tau_\ell\right)\left(\sum_{\ell=1}^{N}{\tau_\ell}^2\right)}$$
$$\leq \mathbb{E}\left(\widehat{\Pi_1}(u,v)\right)\leq$$
$$\frac{E_p\Pi_0\sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell)\sum_{\ell=1}^{N}(1-\tau_\ell)+E_p\Pi_1\sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell)\sum_{\ell=1}^{N}\tau_\ell-E_m\Pi_0\sum_{\ell=1}^{N}(1-\tau_\ell)\sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell)-E_m\Pi_1\sum_{\ell=1}^{N}(1-\tau_\ell)\sum_{\ell=1}^{N}{\tau_\ell}^2}{\left(\sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell)\right)\left(\sum_{\ell=1}^{N}\tau_\ell\right)-\left(\sum_{\ell=1}^{N}1-\tau_\ell\right)\left(\sum_{\ell=1}^{N}{\tau_\ell}^2\right)}$$

Donc

$$\Pi_0+f_0^-(m)\leq\mathbb{E}\left(\widehat{\Pi_0}(u,v)\right)\leq\Pi_0+f_0^+(m)$$

$$\Pi_1+f_1^-(m)\leq\mathbb{E}\left(\widehat{\Pi_1}(u,v)\right)\leq\Pi_1+f_1^+(m)$$

avec

$$f_0^-(m) = \frac{(E_m-1)\Pi_0\sum_{\ell=1}^{N}\tau_\ell\sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell)-(E_p-1)\Pi_0\sum_{\ell=1}^{N}{\tau_\ell}^2\sum_{\ell=1}^{N}(1-\tau_\ell)+(E_m-E_p)\Pi_1\sum_{\ell=1}^{N}{\tau_\ell}^2\sum_{\ell=1}^{N}\tau_\ell}{\left(\sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell)\right)\left(\sum_{\ell=1}^{N}\tau_\ell\right)-\left(\sum_{\ell=1}^{N}1-\tau_\ell\right)\left(\sum_{\ell=1}^{N}{\tau_\ell}^2\right)}$$

$$f_0^+(m) = \frac{(E_p-1)\Pi_0\sum_{\ell=1}^{N}\tau_\ell\sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell)-(E_m-1)\Pi_0\sum_{\ell=1}^{N}{\tau_\ell}^2\sum_{\ell=1}^{N}(1-\tau_\ell)+(E_p-E_m)\Pi_1\sum_{\ell=1}^{N}{\tau_\ell}^2\sum_{\ell=1}^{N}\tau_\ell}{\left(\sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell)\right)\left(\sum_{\ell=1}^{N}\tau_\ell\right)-\left(\sum_{\ell=1}^{N}1-\tau_\ell\right)\left(\sum_{\ell=1}^{N}{\tau_\ell}^2\right)}$$

$$f_1^-(m) = \frac{(E_m-E_p)\Pi_0\sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell)\sum_{\ell=1}^{N}(1-\tau_\ell)+(E_m-1)\Pi_1\sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell)\sum_{\ell=1}^{N}\tau_\ell-(E_p-1)\Pi_1\sum_{\ell=1}^{N}(1-\tau_\ell)\sum_{\ell=1}^{N}{\tau_\ell}^2}{\left(\sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell)\right)\left(\sum_{\ell=1}^{N}\tau_\ell\right)-\left(\sum_{\ell=1}^{N}1-\tau_\ell\right)\left(\sum_{\ell=1}^{N}{\tau_\ell}^2\right)}$$

$$f_1^+(m) = \frac{(E_p-E_m)\Pi_0\sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell)\sum_{\ell=1}^{N}(1-\tau_\ell)+(E_p-1)\Pi_1\sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell)\sum_{\ell=1}^{N}\tau_\ell-(E_m-1)\Pi_1\sum_{\ell=1}^{N}(1-\tau_\ell)\sum_{\ell=1}^{N}{\tau_\ell}^2}{\left(\sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell)\right)\left(\sum_{\ell=1}^{N}\tau_\ell\right)-\left(\sum_{\ell=1}^{N}1-\tau_\ell\right)\left(\sum_{\ell=1}^{N}{\tau_\ell}^2\right)}$$

où $f_0^-(m)$, $f_0^+(m)$, $f_1^-(m)$ et $f_1^+(m)$ tendent vers 0 avec m. Les estimateurs $\widehat{\Pi_0}(u,v)$ et $\widehat{\Pi_1}(u,v)$ sont donc asymptotiquement sans biais.

Variances

Plusieurs simulations nous ont conduit à considérer des segments de taille $m=\sqrt{n}$. En effet, nous avons constaté que la variance atteint son minimum quand les segments sont de cette taille. Notre choix de privilégier la méthode point par point nous amène à ne pas nous étendre davantage sur ce choix.

A.2 Estimation de Π_0 et Π_1 point par point

Nous présentons ici les calculs complémentaires relatifs au 1.1.3 de la partie I.

A.2.1 Stochasticité des matrices $\widehat{\Pi_0}$ et $\widehat{\Pi_1}$

Preuve (Théorème 8 p33) • *Montrons tout d'abord que la somme sur chaque ligne vaut* 1.

$$\begin{aligned}
\sum_{v\in\mathcal{A}}\widehat{\Pi_0}(u,v) &= \frac{\left(\begin{array}{c}\left(\sum_{t=k}^{n}\mathbb{1}_{\{X_{t-k}\dots X_{t-1}=u\}}\left(\frac{t}{n}\right)^2\right)\left(\sum_{t=k}^{n}\mathbb{1}_{\{X_{t-k}\dots X_{t-1}=u\}}\left(1-\frac{t}{n}\right)\right)-\\ \left(\sum_{t=k}^{n}\mathbb{1}_{\{X_{t-k}\dots X_{t-1}=u\}}\left(1-\frac{t}{n}\right)\left(\frac{t}{n}\right)\right)\left(\sum_{t=k}^{n}\mathbb{1}_{\{X_{t-k}\dots X_{t-1}=u\}}\left(\frac{t}{n}\right)\right)\end{array}\right)}{\left(\begin{array}{c}\left(\sum_{t=k}^{n}\mathbb{1}_{\{X_{t-k}\dots X_{t-1}=u\}}\left(1-\frac{t}{n}\right)^2\right)\left(\sum_{t=k}^{n}\mathbb{1}_{\{X_{t-k}\dots X_{t-1}=u\}}\left(\frac{t}{n}\right)^2\right)-\\ \left(\sum_{t=k}^{n}\mathbb{1}_{\{X_{t-k}\dots X_{t-1}=u\}}\left(1-\frac{t}{n}\right)\left(\frac{t}{n}\right)\right)\left(\sum_{t=k}^{n}\mathbb{1}_{\{X_{t-k}\dots X_{t-1}=u\}}\left(1-\frac{t}{n}\right)\left(\frac{t}{n}\right)\right)\end{array}\right)}\\
&= \frac{\left(\begin{array}{c}\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2}\mathbb{1}_{\{X_{t_1-1}=u,X_{t_2-1}=u\}}\left(\frac{t_2}{n}\right)^2\left(1-\frac{t_1}{n}\right)-\\ \sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2}\mathbb{1}_{\{X_{t_1-1}=u,X_{t_2-1}=u\}}\left(1-\frac{t_1}{n}\right)\left(\frac{t_1}{n}\right)\left(\frac{t_2}{n}\right)\end{array}\right)}{\left(\begin{array}{c}\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2}\mathbb{1}_{\{X_{t_1-1}=u,X_{t_2-1}=u\}}\left(1-\frac{t_1}{n}\right)^2\left(\frac{t_2}{n}\right)^2-\\ \sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2}\mathbb{1}_{\{X_{t_1-1}=u,X_{t_2-1}=u\}}\left(1-\frac{t_1}{n}\right)\left(\frac{t_1}{n}\right)\left(1-\frac{t_2}{n}\right)\left(\frac{t_2}{n}\right)\end{array}\right)}\\
&= \frac{\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2}\mathbb{1}_{\{X_{t_1-1}=u,X_{t_2-1}=u\}}\left(\frac{t_2}{n}\right)\left(1-\frac{t_1}{n}\right)\left(\frac{t_2}{n}-\frac{t_1}{n}\right)}{\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2}\mathbb{1}_{\{X_{t_1-1}=u,X_{t_2-1}=u\}}\left(1-\frac{t_1}{n}\right)\left(\frac{t_2}{n}\right)\left(\left(1-\frac{t_1}{n}\right)\frac{t_2}{n}-\left(1-\frac{t_2}{n}\right)\frac{t_1}{n}\right)}\\
&= 1.
\end{aligned}$$

$$\begin{aligned}
\sum_{v\in\mathcal{A}}\widehat{\Pi_1}(u,v) &= \frac{\left(\begin{array}{c}\left(\sum_{t=k}^{n}\mathbb{1}_{\{X_{t-k}\dots X_{t-1}=u\}}\left(1-\frac{t}{n}\right)^2\right)\left(\sum_{t=k}^{n}\mathbb{1}_{\{X_{t-k}\dots X_{t-1}=u\}}\left(\frac{t}{n}\right)\right)-\\ \left(\sum_{t=k}^{n}\mathbb{1}_{\{X_{t-k}\dots X_{t-1}=u\}}\left(1-\frac{t}{n}\right)\left(\frac{t}{n}\right)\right)\left(\sum_{t=k}^{n}\mathbb{1}_{\{X_{t-k}\dots X_{t-1}=u\}}\left(1-\frac{t}{n}\right)\right)\end{array}\right)}{\left(\begin{array}{c}\left(\sum_{t=k}^{n}\mathbb{1}_{\{X_{t-k}\dots X_{t-1}=u\}}\left(1-\frac{t}{n}\right)^2\right)\left(\sum_{t=k}^{n}\mathbb{1}_{\{X_{t-k}\dots X_{t-1}=u\}}\left(\frac{t}{n}\right)^2\right)-\\ \left(\sum_{t=k}^{n}\mathbb{1}_{\{X_{t-k}\dots X_{t-1}=u\}}\left(1-\frac{t}{n}\right)\left(\frac{t}{n}\right)\right)\left(\sum_{t=k}^{n}\mathbb{1}_{\{X_{t-k}\dots X_{t-1}=u\}}\left(1-\frac{t}{n}\right)\left(\frac{t}{n}\right)\right)\end{array}\right)}\\
&= \frac{\left(\begin{array}{c}\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2}\mathbb{1}_{\{X_{t_1-1}=u,X_{t_2-1}=u\}}\left(1-\frac{t_1}{n}\right)^2\left(\frac{t_2}{n}\right)-\\ \sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2}\mathbb{1}_{\{X_{t_1-1}=u,X_{t_2-1}=u\}}\left(1-\frac{t_2}{n}\right)\left(\frac{t_2}{n}\right)\left(1-\frac{t_1}{n}\right)\end{array}\right)}{\left(\begin{array}{c}\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2}\mathbb{1}_{\{X_{t_1-1}=u,X_{t_2-1}=u\}}\left(1-\frac{t_1}{n}\right)^2\left(\frac{t_2}{n}\right)^2-\\ \sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2}\mathbb{1}_{\{X_{t_1-1}=u,X_{t_2-1}=u\}}\left(1-\frac{t_1}{n}\right)\left(\frac{t_1}{n}\right)\left(1-\frac{t_2}{n}\right)\left(\frac{t_2}{n}\right)\end{array}\right)}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\displaystyle\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2} \mathbb{1}_{\{X_{t_1-1}=u,X_{t_2-1}=u\}}\left(\frac{t_2}{n}\right)\left(1-\frac{t_1}{n}\right)\left(\left(1-\frac{t_1}{n}\right)-\left(1-\frac{t_2}{n}\right)\right)}{\displaystyle\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2} \mathbb{1}_{\{X_{t_1-1}=u,X_{t_2-1}=u\}}\left(1-\frac{t_1}{n}\right)\left(\frac{t_2}{n}\right)\left(\left(1-\frac{t_1}{n}\right)\frac{t_2}{n}-\left(1-\frac{t_2}{n}\right)\frac{t_1}{n}\right)}\\
&= 1.
\end{aligned}
$$

• *Montrons désormais que tous les termes sont positifs.*

$$
\begin{aligned}
\widehat{\Pi_0}(u,v) &= \frac{\left(\begin{array}{c}\displaystyle\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2} \mathbb{1}_{\{X_{t_1-1}=u,X_{t_2-1}=u,X_{t_2}=v\}}\left(\frac{t_1}{n}\right)^2\left(1-\frac{t_2}{n}\right)-\\ \displaystyle\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2} \mathbb{1}_{\{X_{t_1-1}=u,X_{t_2-1}=u,X_{t_2}=v\}}\left(1-\frac{t_1}{n}\right)\left(\frac{t_1}{n}\right)\left(\frac{t_2}{n}\right)\end{array}\right)}{\displaystyle\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2} \mathbb{1}_{\{X_{t_1-1}=u,X_{t_2-1}=u\}}\left(1-\frac{t_1}{n}\right)\left(\frac{t_2}{n}\right)\left(\left(1-\frac{t_1}{n}\right)\frac{t_2}{n}-\left(1-\frac{t_2}{n}\right)\frac{t_1}{n}\right)}\\
&= \frac{\displaystyle\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2} \mathbb{1}_{\{X_{t_1-1}=u,X_{t_2-1}=u,X_{t_2}=v\}}\left(\frac{t_1}{n}\right)\left(\frac{t_1}{n}\left(1-\frac{t_2}{n}\right)-\frac{t_2}{n}\left(1-\frac{t_1}{n}\right)\right)}{\displaystyle\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2} \mathbb{1}_{\{X_{t_1-1}=u,X_{t_2-1}=u\}}\left(1-\frac{t_1}{n}\right)\left(\frac{t_2}{n}\right)\left(\frac{t_2}{n}-\frac{t_1}{n}\right)},
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\Pi_1}(u,v) &= \frac{\left(\begin{array}{c}\displaystyle\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2} \mathbb{1}_{\{X_{t_1-1}=u,X_{t_2-1}=u,X_{t_2}=v\}}\left(1-\frac{t_1}{n}\right)^2\left(\frac{t_2}{n}\right)-\\ \displaystyle\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2} \mathbb{1}_{\{X_{t_1-1}=u,X_{t_2-1}=u,X_{t_2}=v\}}\left(1-\frac{t_1}{n}\right)\left(\frac{t_1}{n}\right)\left(1-\frac{t_2}{n}\right)\end{array}\right)}{\displaystyle\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2} \mathbb{1}_{\{X_{t_1-1}=u,X_{t_2-1}=u\}}\left(1-\frac{t_1}{n}\right)\left(\frac{t_2}{n}\right)\left(\frac{t_2}{n}-\frac{t_1}{n}\right)}\\
&= \frac{\displaystyle\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2} \mathbb{1}_{\{X_{t_1-1}=u,X_{t_2-1}=u,X_{t_2}=v\}}\left(1-\frac{t_1}{n}\right)\left(\left(1-\frac{t_1}{n}\right)\frac{t_2}{n}-\left(1-\frac{t_2}{n}\right)\frac{t_1}{n}\right)}{\displaystyle\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2} \mathbb{1}_{\{X_{t_1-1}=u,X_{t_2-1}=u\}}\left(1-\frac{t_1}{n}\right)\left(\frac{t_2}{n}\right)\left(\frac{t_2}{n}-\frac{t_1}{n}\right)}.
\end{aligned}
$$

Nous avons

$$
\begin{aligned}
&\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2} \mathbb{1}_{\{X_{t_1-1}=u,X_{t_2-1}=u\}}\left(1-\frac{t_1}{n}\right)\left(\frac{t_2}{n}\right)\left(\frac{t_2}{n}-\frac{t_1}{n}\right)\\
=&\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1<t_2} \mathbb{1}_{\{X_{t_1-1}=u,X_{t_2-1}=u\}}\left(1-\frac{t_1}{n}\right)\left(\frac{t_2}{n}\right)\left(\frac{t_2}{n}-\frac{t_1}{n}\right)\\
+&\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1<t_2} \mathbb{1}_{\{X_{t_1-1}=u,X_{t_2-1}=u\}}\left(1-\frac{t_2}{n}\right)\left(\frac{t_1}{n}\right)\left(\frac{t_1}{n}-\frac{t_2}{n}\right)\\
=&\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1<t_2} \mathbb{1}_{\{X_{t_1-1}=u,X_{t_2-1}=u\}}\left(\frac{t_2}{n}-\frac{t_1}{n}\right)\left(\left(1-\frac{t_1}{n}\right)\frac{t_2}{n}-\left(1-\frac{t_2}{n}\right)\frac{t_1}{n}\right)\\
=&\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1<t_2} \mathbb{1}_{\{X_{t_1-1}=u,X_{t_2-1}=u\}}\left(\frac{t_2}{n}-\frac{t_1}{n}\right)^2\\
>&\ 0
\end{aligned}
$$

Il resterait à montrer que

$$\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2} \mathbb{1}_{\{X_{t_1-1}=u,X_{t_2-1}=u,X_{t_2}=v\}} \left(\frac{t_1}{n}\right)\left(\frac{t_1}{n}-\frac{t_2}{n}\right) > 0$$

et que

$$\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2} \mathbb{1}_{\{X_{t_1-1}=u,X_{t_2-1}=u,X_{t_2}=v\}} \left(1-\frac{t_1}{n}\right)\left(\frac{t_2}{n}-\frac{t_1}{n}\right) > 0.$$

Les termes ne sont donc pas toujours positifs. Lors de l'estimation nous proposons un réajustement proportionnel des valeurs (voir A.1.2).

A.2.2 Espérances et variances de $\widehat{\Pi_0}$ et $\widehat{\Pi_1}$

Afin de faciliter la lecture nous utilisons les notations suivantes :
- $\mathbb{P}_t(u) = \mathbb{P}(X_{t-k}\dots X_{t-1} = u)$;
- $\mathbb{P}_t(uv) = \mathbb{P}(X_{t-k}\dots X_{t-1} = u, X_t = v)$;
- $\mathbb{P}_{t_1t_2}(uu) = \mathbb{P}_(X_{t_1-k}\dots X_{t_1-1} = u, X_{t_2-k}\dots X_{t_2-1} = u)$;
- $\mathbb{P}_{t_1}(u) = \mathbb{P}(X_{t_1-k}\dots X_{t_1-1} = u)$;
- $\mathbb{P}_{t_2}(u) = \mathbb{P}(X_{t_2-k}\dots X_{t_2-1} = u)$;
- $\mathbb{P}_{t_1t_2}(uvuv) = \mathbb{P}(X_{t_1-k}\dots X_{t_1-1} = u, X_{t_1} = v, X_{t_2-k}\dots X_{t_2-1} = u, X_{t_2} = v)$.

Calculons les espérances.

$$\begin{aligned}
\mathbb{E}(P_1(u)) &= 2\sum_{t=k}^{n} \mathbb{P}_t(u)\left(1-\frac{t}{n}\right)^2.\\
\mathbb{E}(P_2(u)) &= 2\sum_{t=k}^{n} \mathbb{P}_t(u)\left(1-\frac{t}{n}\right)\left(\frac{t}{n}\right).\\
\mathbb{E}(P_3(u,v)) &= 2\sum_{t=k}^{n} \mathbb{P}_t(uv)\left(1-\frac{t}{n}\right) = 2\sum_{t=k}^{n} \mathbb{P}_t(u)\Pi_{\frac{t}{n}}(u,v)\left(1-\frac{t}{n}\right).\\
\mathbb{E}(P_4(u)) &= 2\sum_{t=k}^{n} \mathbb{P}_t(u)\left(1-\frac{t}{n}\right)\left(\frac{t}{n}\right).\\
\mathbb{E}(P_5(u)) &= 2\sum_{t=k}^{n} \mathbb{P}_t(u)\left(\frac{t}{n}\right)^2.\\
\mathbb{E}(P_6(u,v)) &= 2\sum_{t=k}^{n} \mathbb{P}_t(uv)\left(\frac{t}{n}\right) = 2\sum_{t=k}^{n} \mathbb{P}_t(u)\Pi_{\frac{t}{n}}(u,v)\left(\frac{t}{n}\right).
\end{aligned}$$

Par le théorème de la limite centrale, la loi de chacun des P_i (pour i allant de 1 à 6) est approchée par une loi normale. Ainsi la delta méthode (voir [Oeh92] par exemple) nous permet d'approcher les lois de nos estimateurs par des lois normales d'espérances :

$$\begin{aligned}
\mathbb{E}(\widehat{\Pi_0}(u,v)) &= \frac{\mathbb{E}(P_5(u))\mathbb{E}(P_3(u,v)) - \mathbb{E}(P_2(u))\mathbb{E}(P_6(u,v))}{\mathbb{E}(P_1(u))\mathbb{E}(P_5(u)) - \mathbb{E}(P_4(u))\mathbb{E}(P_2(u))}\\
\mathbb{E}(\widehat{\Pi_1}(u,v)) &= \frac{\mathbb{E}(P_1(u))\mathbb{E}(P_6(u,v)) - \mathbb{E}(P_4(u))\mathbb{E}(P_3(u,v))}{\mathbb{E}(P_1(u))\mathbb{E}(P_5(u)) - \mathbb{E}(P_4(u))\mathbb{E}(P_2(u))}.
\end{aligned}$$

Nous obtenons les espérances suivantes :

$$\mathbb{E}(\widehat{\Pi_0}(u,v)) = \frac{\displaystyle\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2} \mathbb{P}_{t_1}(u)\mathbb{P}_{t_2}(u)\Pi_{\frac{t_2}{n}}(u,v)\left(\frac{t_1}{n}\right)\left(\frac{t_1}{n}\left(1-\frac{t_2}{n}\right)-\frac{t_2}{n}\left(1-\frac{t_1}{n}\right)\right)}{\displaystyle\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2} \mathbb{P}_{t_1}(u)\mathbb{P}_{t_2}(u)\left(1-\frac{t_1}{n}\right)\left(\frac{t_2}{n}\right)\left(\frac{t_2}{n}-\frac{t_1}{n}\right)}$$

$$= \frac{\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2} \mathbb{P}_{t_1}(u)\mathbb{P}_{t_2}(u)\Pi_0(u,v)\left(1-\frac{t_2}{n}\right)\left(\frac{t_1}{n}\right)\left(\frac{t_1}{n}-\frac{t_2}{n}\right)}{\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2} \mathbb{P}_{t_1}(u)\mathbb{P}_{t_2}(u)\left(1-\frac{t_1}{n}\right)\left(\frac{t_2}{n}\right)\left(\frac{t_2}{n}-\frac{t_1}{n}\right)}$$

$$+ \frac{\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2} \mathbb{P}_{t_1}(u)\mathbb{P}_{t_2}(u)\Pi_1(u,v)\left(\frac{t_2}{n}\right)\left(\frac{t_1}{n}\right)\left(\frac{t_1}{n}-\frac{t_2}{n}\right)}{\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2} \mathbb{P}_{t_1}(u)\mathbb{P}_{t_2}(u)\left(1-\frac{t_1}{n}\right)\left(\frac{t_2}{n}\right)\left(\frac{t_2}{n}-\frac{t_1}{n}\right)}.$$

Ainsi $\mathbb{E}(\widehat{\Pi_0}(u,v)) = \Pi_0$.

$$\mathbb{E}(\widehat{\Pi_1}(u,v)) = \frac{\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2} \mathbb{P}_{t_1}(u)\mathbb{P}_{t_2}(u)\Pi_{\frac{t_2}{n}}(u,v)\left(1-\frac{t_1}{n}\right)\left(\frac{t_2}{n}-\frac{t_1}{n}\right)}{\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2} \mathbb{P}_{t_1}(u)\mathbb{P}_{t_2}(u)\left(1-\frac{t_1}{n}\right)\left(\frac{t_2}{n}\right)\left(\frac{t_2}{n}-\frac{t_1}{n}\right)}$$

$$= \frac{\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2} \mathbb{P}_{t_1}(u)\mathbb{P}_{t_2}(u)\Pi_0(u,v)\left(1-\frac{t_2}{n}\right)\left(1-\frac{t_1}{n}\right)\left(\frac{t_2}{n}-\frac{t_1}{n}\right)}{\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2} \mathbb{P}_{t_1}(u)\mathbb{P}_{t_2}(u)\left(1-\frac{t_1}{n}\right)\left(\frac{t_2}{n}\right)\left(\frac{t_2}{n}-\frac{t_1}{n}\right)}$$

$$= \frac{\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2} \mathbb{P}_{t_1}(u)\mathbb{P}_{t_2}(u)\Pi_1(u,v)\left(\frac{t_2}{n}\right)\left(1-\frac{t_1}{n}\right)\left(\frac{t_2}{n}-\frac{t_1}{n}\right)}{\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2} \mathbb{P}_{t_1}(u)\mathbb{P}_{t_2}(u)\left(1-\frac{t_1}{n}\right)\left(\frac{t_2}{n}\right)\left(\frac{t_2}{n}-\frac{t_1}{n}\right)}.$$

Ainsi $\mathbb{E}(\widehat{\Pi_1}(u,v)) = \Pi_1$.
Nos estimateurs sont asymptotiquement sans biais.

Calculons maintenant les variances.
Nous rappelons que :

$$\begin{aligned}\widehat{\Pi_0}(u,v) &= \frac{P_5(u)P_3(u,v)-P_2(u)P_6(u,v)}{P_1(u)P_5(u)-P_4(u)P_2(u)}\\ \widehat{\Pi_1}(u,v) &= \frac{P_1(u)P_6(u,v)-P_4(u)P_3(u,v)}{P_1(u)P_5(u)-P_4(u)P_2(u)}\end{aligned}.$$

La delta méthode nous permet d'approcher les lois de nos estimateurs par des lois normales de variances :

$$\mathbb{V}(\widehat{\Pi_0}(u,v)) = [d\widehat{\Pi_0}^t][\mathbb{C}\mathrm{ov}_P][d\widehat{\Pi_0}]$$

$$\mathbb{V}(\widehat{\Pi_1}(u,v)) = [d\widehat{\Pi_1}^t][\mathbb{C}\mathrm{ov}_P][d\widehat{\Pi_1}]$$

où $[d\widehat{\Pi_0}]_i = \dfrac{\partial\widehat{\Pi_0}}{\partial P_i}$, $[d\widehat{\Pi_1}]_i = \dfrac{\partial\widehat{\Pi_1}}{\partial P_i}$ et $[\mathbb{C}\mathrm{ov}_P]_{ij} = \mathbb{C}\mathrm{ov}(P_i,P_j)$.

Nous utilisons les approximations suivantes pour le cas où les ensembles $[\![t_1-k,t_1]\!]$ et $[\![t_2-k,t_2]\!]$ ont une intersection vide :

- $\mathbb{P}_{t_1t_2}(uu) \simeq \mathbb{P}_{t_1}(u)\mathbb{P}_{t_2}(u)$;
- $\mathbb{P}_{t_1t_2}(uvuv) \simeq \mathbb{P}_{t_1}(u)\mathbb{P}_{t_2}(u)\Pi_{\frac{t_1}{n}}(u,v)\Pi_{\frac{t_2}{n}}(u,v)$.

Nous n'expliciterons ici que la diagonale de la matrice de covariance des P_i, les autres termes étant assez similaires :

$$\mathbb{V}(P_1(u)) = 4\sum_{t_1=k}^{n}\sum_{t_2=k}^{n}\mathbb{P}_{t_1t_2}(uu)\left(1-\frac{t_1}{n}\right)^2\left(1-\frac{t_2}{n}\right)^2-(\mathbb{E}(P_1(u)))^2.$$

$$\mathbb{V}(P_2(u)) = 4\sum_{t_1=k}^{n}\sum_{t_2=k}^{n}\mathbb{P}_{t_1t_2}(uu)\left(1-\frac{t_1}{n}\right)\left(\frac{t_1}{n}\right)\left(1-\frac{t_2}{n}\right)\left(\frac{t_2}{n}\right)-(\mathbb{E}(P_2(u)))^2.$$

$$
\begin{aligned}
\mathbb{V}(P_3(u,v)) &= 4\sum_{t_1=k}^{n}\sum_{t_2=k}^{n}\mathbb{P}_{t_1t_2}(uvuv)\left(1-\frac{t_1}{n}\right)\left(1-\frac{t_2}{n}\right)-\left(\mathbb{E}(P_3(u,v))\right)^2\\
&= 4\sum_{t_1=k}^{n}\sum_{t_2=k}^{n}\mathbb{P}_{t_1t_2}(uu)\Pi_{\frac{t_1}{n}}(u,v)\Pi_{\frac{t_2}{n}}(u,v)\left(1-\frac{t_1}{n}\right)\left(1-\frac{t_2}{n}\right)-4\left(\sum_{t=k}^{n}\mathbb{P}_t(u)\Pi_{\frac{t}{n}}(u,v)\left(1-\frac{t}{n}\right)\right)^2.\\
\mathbb{V}(P_4(u)) &= 4\sum_{t_1=k}^{n}\sum_{t_2=k}^{n}\mathbb{P}_{t_1t_2}(uu)\left(1-\frac{t_1}{n}\right)\left(\frac{t_1}{n}\right)\left(1-\frac{t_2}{n}\right)\left(\frac{t_2}{n}\right)-\left(\mathbb{E}(P_4(u))\right)^2.\\
\mathbb{V}(P_5(u)) &= 4\sum_{t_1=k}^{n}\sum_{t_2=k}^{n}\mathbb{P}_{t_1t_2}(uu)\left(\frac{t_1}{n}\right)^2\left(\frac{t_2}{n}\right)^2-\left(\mathbb{E}(P_5(u))\right)^2.\\
\mathbb{V}(P_6(u,v)) &= 4\sum_{t_1=k}^{n}\sum_{t_2=k}^{n}\mathbb{P}_{t_1t_2}(uvuv)\left(\frac{t_1}{n}\right)\left(\frac{t_2}{n}\right)-\left(\mathbb{E}(P_6(u,v))\right)^2\\
&= 4\sum_{t_1=k}^{n}\sum_{t_2=k}^{n}\mathbb{P}_{t_1t_2}(uu)\Pi_{\frac{t_1}{n}}(u,v)\Pi_{\frac{t_2}{n}}(u,v)\left(\frac{t_1}{n}\right)\left(\frac{t_2}{n}\right)-4\left(\sum_{t=k}^{n}\mathbb{P}_t(u)\Pi_{\frac{t}{n}}(u,v)\left(\frac{t}{n}\right)\right)^2.
\end{aligned}
$$

Lorsque l'intersection de $[\![t_1-k,t_1]\!]$ et $[\![t_2-k,t_2]\!]$ n'est pas vide (t_1 proche de t_2), les termes des différentes sommes ne sont pas nuls car nous n'utilisons pas les approximations précédentes. Nous pourrons nous inspirer de [SPdT95] pour terminer le calcul.

Annexe B

Estimation de M_0 et M_1

Rappelons le modèle (voir Partie I 1.2) :

$$\Pi_{\frac{t}{n}}(u,v) = M_0(u,v) + \frac{t}{n} M_1(u,v).$$

B.1 Estimation par maximum de vraisemblance

Nous ne donnons qu'un essai de maximisation de la vraisemblance pour des modèles de Markov régulés d'ordre 1 et de degré 1 sur l'alphabet des nucléotides $\mathcal{A} = \{a, c, g, t\}$.

Exemple 16 *Pour $\mathcal{A} = \{a, c, g, t\}$, nous devons résoudre 4 systèmes de 6 équations à 6 inconnues. Nous présentons un seul de ces systèmes.*

$$\begin{cases}
\bullet \sum_{t=1}^{n} \frac{\mathbb{1}_{\{X_{t-1}=a, X_t=c\}}}{M_0(a,c) + \frac{t}{n} M_1(a,c)} = \\
\quad \sum_{t=1}^{n} \frac{\mathbb{1}_{\{X_{t-1}=a, X_t=a\}}}{1 - M_0(a,c) - M_0(a,g) - M_0(a,t) - tM_1(a,c) - tM_1(a,g) - tM_1(a,t)} \\
\bullet \sum_{t=1}^{n} \frac{\mathbb{1}_{\{X_{t-1}=a, X_t=g\}}}{M_0(a,g) + \frac{t}{n} M_1(a,g)} = \\
\quad \sum_{t=1}^{n} \frac{\mathbb{1}_{\{X_{t-1}=a, X_t=a\}}}{1 - M_0(a,c) - M_0(a,g) - M_0(a,t) - tM_1(a,c) - tM_1(a,g) - tM_1(a,t)} \\
\bullet \sum_{t=1}^{n} \frac{\mathbb{1}_{\{X_{t-1}=a, X_t=t\}}}{M_0(a,t) + \frac{t}{n} M_1(a,t)} = \\
\quad \sum_{t=1}^{n} \frac{\mathbb{1}_{\{X_{t-1}=a, X_t=a\}}}{1 - M_0(a,c) - M_0(a,g) - M_0(a,t) - tM_1(a,c) - tM_1(a,g) - tM_1(a,t)} \\
\bullet \sum_{t=1}^{n} \frac{t}{n} \frac{\mathbb{1}_{\{X_{t-1}=a, X_t=c\}}\left(\frac{t}{n}\right)}{M_0(a,c) + \frac{t}{n} M_1(a,c)} = \\
\quad \sum_{t=1}^{n} \frac{t}{n} \frac{\mathbb{1}_{\{X_{t-1}=a, X_t=a\}}}{1 - M_0(a,c) - M_0(a,g) - M_0(a,t) - tM_1(a,c) - tM_1(a,g) - tM_1(a,t)} \\
\bullet \sum_{t=1}^{n} \frac{t}{n} \frac{\mathbb{1}_{\{X_{t-1}=a, X_t=g\}}}{M_0(a,g) + \frac{t}{n} M_1(a,g)} = \\
\quad \sum_{t=1}^{n} \frac{t}{n} \frac{\mathbb{1}_{\{X_{t-1}=a, X_t=a\}}}{1 - M_0(a,c) - M_0(a,g) - M_0(a,t) - tM_1(a,c) - tM_1(a,g) - tM_1(a,t)} \\
\bullet \sum_{t=1}^{n} \frac{t}{n} \frac{\mathbb{1}_{\{X_{t-1}=a, X_t=t\}}}{M_0(a,t) + \frac{t}{n} M_1(a,t)} = \\
\quad \sum_{t=1}^{n} \frac{t}{n} \frac{\mathbb{1}_{\{X_{t-1}=a, X_t=a\}}}{1 - M_0(a,c) - M_0(a,g) - M_0(a,t) - tM_1(a,c) - tM_1(a,g) - tM_1(a,t)}
\end{cases}$$

Même si M_0 et M_1 ne sont pas des matrices stochastiques, nous pouvons réduire notre estimation aux paramètres $M_0(u,v)$ et $M_1(u,v)$ tels que $u \neq v$. Pour $u = v$ nous avons

$$\Pi_{\frac{t}{n}}(u,u) = 1 - \sum_{v \in \mathcal{A}\setminus\{u\}} \Pi_{\frac{t}{n}}(u,v).$$

La maximisation de la vraisemblance se déroule de la même façon que pour le modèle des points d'appuis (voir 1.1.1), et nous donne ainsi à nouveau $|\mathcal{A}|$ systèmes de $2(|\mathcal{A}|-1)$ équations à $2(|\mathcal{A}|-1)$ inconnues. Ces systèmes ne sont à nouveau pas linéaires et impossible à résoudre analytiquement ou numériquement. Encore une fois, nous abandonnons cette méthode.

B.2 Estimation de M_0 et M_1 par régression matricielle

B.2.1 Régression matricielle

Les calculs sont identiques à ceux du 1.1.2.

Distances et choix de m

Nous devons minimiser la fonction PR_1 suivante :

$$PR_1(M_0,M_1) = \sum_{\ell \in [\![1,N]\!]} \sum_{u \in \mathcal{A}^k} \sum_{v \in \mathcal{A}} \left(\widehat{\Pi_{S_\ell}}(u,v) - M_0(u,v) - \tau_\ell M_1(u,v) \right)^2$$

La recherche des zéros de la dérivée nous donne le système suivant pour chaque couple $(u,v) \in \mathcal{A}^k \times \mathcal{A}$:

$$\begin{cases} \widehat{M_0}(u,v) &= \dfrac{R_2R_6(u,v) - R_5R_3(u,v)}{R_4R_2 - R_1R_5} \\ \\ \widehat{M_1}(u,v) &= \dfrac{R_4R_3(u,v) - R_1R_6(u,v)}{R_4R_2 - R_1R_5} \end{cases}$$

avec

$$\begin{array}{lllllllll} R_1 &=& \displaystyle\sum_{\ell=1}^{N} 1, & R_2 &=& \displaystyle\sum_{\ell=1}^{N} \tau_\ell, & R_3(u,v) &=& \displaystyle\sum_{\ell=1}^{N} \widehat{\Pi_{S_\ell}}(u,v), \\ R_4 &=& \displaystyle\sum_{\ell=1}^{N} \tau_\ell, & R_5 &=& \displaystyle\sum_{\ell=1}^{N} {\tau_\ell}^2, & R_6(u,v) &=& \displaystyle\sum_{\ell=1}^{N} \tau_\ell \widehat{\Pi_{S_\ell}}(u,v). \end{array}$$

Nous obtenons autant de systèmes que pour le modèle des points d'appuis Nous avons alors à m fixé, les estimateurs de M_0 et M_1.

B.2.2 Stochasticité des matrices $\widehat{\Pi_{\frac{t}{n}}}$

Théorème 15 *Les matrices $\widehat{\Pi_{\frac{t}{n}}}$ sont stochastiques.*

Preuve *Rappelons que les matrices $\widehat{\Pi_{S_\ell}}$ sont stochastiques. Nous avons donc*

- $\forall \ell \in [\![1,N]\!] \sum_{v \in \mathcal{A}} \widehat{\Pi_{S_\ell}}(u,v) = 1$
- $\forall \ell \in [\![1,N]\!] \forall (u,v) \in \mathcal{A}^k \times \mathcal{A}, \quad \widehat{\Pi_{S_\ell}}(u,v) > 0$

• *Montrons tout d'abord que la somme sur chaque ligne vaut 1. Pour cela, nous montrons que $\sum_{v \in \mathcal{A}} \widehat{M_1}(u,v) = 0$ et que $\sum_{v \in \mathcal{A}} \widehat{M_0}(u,v) = 1$. Ainsi, $\forall t \in [\![0,n]\!]$, $\sum_{v \in \mathcal{A}} \widehat{\Pi_{\frac{t}{n}}}(u,v) = 1$.*

$$\sum_{v \in \mathcal{A}} \widehat{M_0}(u,v) = \frac{\left(\sum_{\ell=1}^{N} \tau_\ell\right)\left(\sum_{\ell=1}^{N} \tau_\ell \sum_{v \in \mathcal{A}} \widehat{\Pi_{S_\ell}}(u,v)\right) - \left(\sum_{\ell=1}^{N} {\tau_\ell}^2\right)\left(\sum_{\ell=1}^{N} \sum_{v \in \mathcal{A}} \widehat{\Pi_{S_\ell}}(u,v)\right)}{\left(\sum_{\ell=1}^{N} \tau_\ell\right)\left(\sum_{\ell=1}^{N} \tau_\ell\right) - \left(\sum_{\ell=1}^{N} 1\right)\left(\sum_{\ell=1}^{N} {\tau_\ell}^2\right)}$$

$$= \frac{\sum_{\ell=1}^{N}\tau_\ell\sum_{\ell=1}^{N}\tau_\ell-\sum_{\ell=1}^{N}{\tau_\ell}^2\sum_{\ell=1}^{N}1}{\sum_{\ell=1}^{N}\tau_\ell\sum_{\ell=1}^{N}\tau_\ell-\sum_{\ell=1}^{N}{\tau_\ell}^2\sum_{\ell=1}^{N}1}=1$$

$$\sum_{v\in\mathcal{A}}\widehat{M_1}(u,v) = \frac{\left(\sum_{\ell=1}^{N}\tau_\ell\right)\left(\sum_{\ell=1}^{N}\sum_{v\in\mathcal{A}}\widehat{\Pi_{S_\ell}}(u,v)\right)-\left(\sum_{\ell=1}^{N}1\right)\left(\sum_{\ell=1}^{N}\tau_\ell\sum_{v\in\mathcal{A}}\widehat{\Pi_{S_\ell}}(u,v)\right)}{\left(\sum_{\ell=1}^{N}\tau_\ell\right)\left(\sum_{\ell=1}^{N}\tau_\ell\right)-\left(\sum_{\ell=1}^{N}1\right)\left(\sum_{\ell=1}^{N}{\tau_\ell}^2\right)}$$

$$= \frac{\sum_{\ell=1}^{N}\tau_\ell\sum_{\ell=1}^{N}1-\sum_{\ell=1}^{N}1\sum_{\ell=1}^{N}\tau_\ell}{\sum_{\ell=1}^{N}\tau_\ell\sum_{\ell=1}^{N}\tau_\ell-\sum_{\ell=1}^{N}1\sum_{\ell=1}^{N}{\tau_\ell}^2}=0$$

Les termes ne sont donc pas toujours positifs. Lors de l'estimation nous proposons un réajustement proportionnel des valeurs (voir A.1.2).

B.2.3 Distances : deuxième méthode

Comme pour le modèle des points d'appuis (voir A.1.3), nous choisissons tous les points du segment pour ajuster notre modèle. Nous minimisons donc la somme

$$\sum_{t=k}^{n}\sum_{u\in\mathcal{A}^k}\sum_{v\in\mathcal{A}}\left(\widehat{\Pi_{S_{\ell_t}}}(u,v)-M_0(u,v)-\frac{t}{n}M_1(u,v)\right)^2.$$

De la même manière que précédemment nous obtenons les estimateurs de M_0 et M_1 et la stochasticité de ces estimateurs.

B.2.4 Espérances et variances de $\widehat{\Pi_{S_\ell}}$ pour le modèle des chaînes de Markov régulées

Calcul de $\mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)$

Théorème 16 *Les estimateurs des $\widehat{\Pi_{S_\ell}}(u,v)$ pour le modèle des chaînes de Markov régulées sont asymptotiquement sans biais.*

Preuve *Le calcul est semblable à celui du modèle des points d'appuis (voir A.1.5). Nous devons seulement être attentifs à la somme $\sum_{t\in S_\ell^\star}\Pi_{\frac{t}{n}}(u,v)$.*

$$\mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right) = \left(1-(1-\mu_\ell(u))^{|S_\ell^\star|}\right)\left(\sum_{t\in S_\ell^\star}\frac{M_0(u,v)+\frac{t}{n}M_1(u,v)}{|S_\ell^\star|}\right)$$

$$= \left(1-(1-\mu_\ell(u))^{|S_\ell^\star|}\right)\left(M_0(u,v)+\sum_{t\in S_\ell^\star}\frac{t}{|S_\ell^\star|n}M_1(u,v)\right).$$

Nous obtenons comme précédemment

$$\mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right) \approx (1-(1-\mu_\ell(u))^{m-k})\Pi_{\frac{2\ell m-m+k-1}{2n}}(u,v)$$

$$\mathbb{E}\left(\widehat{\Pi_{S_{N-1}}}(u,v)\right) \approx (1-(1-\mu_\ell(u))^{n-(N-1)m-k+1})\Pi_{\frac{(N-1)m+k+n}{2n}}(u,v).$$

Calcul de $\mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)$

Théorème 17 *. Les estimateurs $\widehat{\Pi_{S_\ell}}(u,v)$ ont une variance qui tend vers zéro quand $|S_\ell| \to +\infty$.*

Preuve *Dans le calcul des variances, seules changent les sommes suivantes :*

$$\sum_{t\in S_\ell^\star} \Pi_{\frac{t}{n}}(u,v) \qquad et \qquad \sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star, t_1<t_2} \Pi_{\frac{t_1}{n}}(u,v)\Pi_{\frac{t_2}{n}}(u,v).$$

Nous obtenons comme précédemment

$$\begin{aligned} A_- \sum_{t\in S_\ell^\star} \frac{\Pi_{\frac{t}{n}}(u,v)}{|S_\ell^\star|} + 2B_- \sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star, t_1<t_2} \frac{\Pi_{\frac{t_1}{n}}(u,v)\Pi_{\frac{t_2}{n}}(u,v)}{|S_\ell^\star|(|S_\ell^\star|-1)} - \mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)^2 \\ \leq \mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right) \leq \\ A_+ \sum_{t\in S_\ell^\star} \frac{\Pi_{\frac{t}{n}}(u,v)}{|S_\ell^\star|} + 2B_+ \sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star, t_1<t_2} \frac{\Pi_{\frac{t_1}{n}}(u,v)\Pi_{\frac{t_2}{n}}(u,v)}{|S_\ell^\star|(|S_\ell^\star|-1)} - \mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)^2 \end{aligned}$$

Nous avons

$$\begin{aligned} & 2 \sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star, t_1<t_2} \frac{\Pi_{\frac{t_1}{n}}(u,v)\Pi_{\frac{t_2}{n}}(u,v)}{|S_\ell^\star|(|S_\ell^\star|-1)} \\ = & \frac{2 \sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star, t_1<t_2} \left[M_0(u,v)^2 + M_0(u,v)M_1(u,v)\left(\frac{t_1}{n}+\frac{t_2}{n}\right) + \frac{t_1}{n}\frac{t_2}{n}M_1(u,v)^2\right]}{|S_\ell^\star|(|S_\ell^\star|-1)} \\ = & M_0(u,v)^2 \frac{2\sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star, t_1<t_2} 1}{|S_\ell^\star|(|S_\ell^\star|-1)} + M_0(u,v)M_1(u,v)\frac{2\sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star, t_1<t_2} \frac{t_1}{n}+\frac{t_2}{n}}{|S_\ell^\star|(|S_\ell^\star|-1)} \\ + & M_1(u,v)^2 \frac{2\sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star, t_1<t_2} \left(\frac{t_1}{n}\frac{t_2}{n}\right)}{|S_\ell^\star|(|S_\ell^\star|-1)}. \end{aligned}$$

En reprenant les notations précédentes (voir 1.1.2), nous obtenons

$$\begin{aligned} & M_0(u,v)^2 + M_0(u,v)M_1(u,v)\left(T_1+T_2\right) + M_1(u,v)^2 T_3 \\ = & M_0(u,v)^2 + 2\tau_\ell M_0(u,v)M_1(u,v) + M_1(u,v)^2\left({\tau_\ell}^2 + K\right) \\ = & \Pi_{\tau_\ell}(u,v)^2 + KM_1(u,v)^2. \end{aligned}$$

Finalement, nous avons un encadrement de la variance :

$$\begin{aligned} A_-\Pi_{\tau_\ell}(u,v) + B_-\left(\Pi_{\tau_\ell}(u,v)^2 + KM_1(u,v)^2\right) - \Pi_{\tau_\ell}(u,v)^2\left(1-(1-\mu_\ell(u))^{|S_\ell^\star|}\right)^2 \\ \leq \mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right) \leq \\ A_+\Pi_{\tau_\ell}(u,v) + B_+\left(\Pi_{\tau_\ell}(u,v)^2 + KM_1(u,v)^2\right) - \Pi_{\tau_\ell}(u,v)^2\left(1-(1-\mu_\ell(u))^{|S_\ell^\star|}\right)^2. \end{aligned}$$

Nous avons A_-, A_+ et K qui tendent vers 0, B_- et B_+ qui tendent vers 1.
Les estimateurs $\widehat{\Pi_{S_\ell}}(u,v)$ sont donc sans biais et de variance qui tend vers zéro quand $|S_\ell| \to +\infty$, et par conséquent convergents.

Théorème 18 *Les estimateurs $\widehat{\Pi_{S_\ell}}(u,v)$ sont convergents.*

Preuve *La convergence découle directement des théorèmes 16 et 17.*

B.2.5 Espérances et variances de $\widehat{M_0}$ et $\widehat{M_1}$

Espérances

$$\left\{\begin{array}{rcl}\widehat{M_0}(u,v) & = & \dfrac{R_2R_6(u,v)-R_5R_3(u,v)}{R_4R_2-R_1R_5}\\ \\ \widehat{M_1}(u,v) & = & \dfrac{R_4R_3(u,v)-R_1R_6(u,v)}{R_4R_2-R_1R_5}\end{array}\right.$$

$$\Longleftrightarrow \left\{\begin{array}{rcl}\mathbb{E}\left(\widehat{M_0}(u,v)\right) & = & \dfrac{R_2\mathbb{E}\left(R_6(u,v)\right)-R_5\mathbb{E}\left(R_3(u,v)\right)}{R_4R_2-R_1R_5}\\ \\ \mathbb{E}\left(\widehat{M_1}(u,v)\right) & = & \dfrac{R_4\mathbb{E}\left(R_3(u,v)\right)-R_1\mathbb{E}\left(R_6(u,v)\right)}{R_4R_2-R_1R_5}\end{array}\right.$$

Ainsi

$$\mathbb{E}\left(\widehat{M_0}(u,v)\right) = \frac{\left(\sum_{\ell=1}^{N}\tau_\ell\right)\left(\sum_{\ell=1}^{N}\tau_\ell\mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)\right)-\left(\sum_{\ell=1}^{N}{\tau_\ell}^2\right)\left(\sum_{\ell=1}^{N}\mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)\right)}{\left(\sum_{\ell=1}^{N}\tau_\ell\right)\left(\sum_{\ell=1}^{N}\tau_\ell\right)-\left(\sum_{\ell=1}^{N}1\right)\left(\sum_{\ell=1}^{N}{\tau_\ell}^2\right)},$$

$$\mathbb{E}\left(\widehat{M_1}(u,v)\right) = \frac{\left(\sum_{\ell=1}^{N}\tau_\ell\right)\left(\sum_{\ell=1}^{N}\mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)\right)-\left(\sum_{\ell=1}^{N}1\right)\left(\sum_{\ell=1}^{N}\tau_\ell\mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)\right)}{\left(\sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell)\right)\left(\sum_{\ell=1}^{N}\tau_\ell\right)-\left(\sum_{\ell=1}^{N}1-\tau_\ell\right)\left(\sum_{\ell=1}^{N}{\tau_\ell}^2\right)}.$$

Théorème 19 *Les estimateurs $\widehat{M_0}(u,v)$ et $\widehat{M_1}(u,v)$ sont asymptotiquement sans biais.*

Preuve *Connaissant les valeurs des $\mathbb{E}\left(\widehat{\Pi_{S_\ell}}\right)$ (voir p111), nous posons*

$$E_m=\min_\ell\left\{(1-(1-\mu_\ell(u))^{m-k+1}),(1-(1-\mu_\ell(u))^{m-k}),(1-(1-\mu_\ell(u))^{n-(N-1)m-k})\right\}$$

et

$$E_p=\max_\ell\left\{(1-(1-\mu_\ell(u))^{m-k+1}),(1-(1-\mu_\ell(u))^{m-k}),(1-(1-\mu_\ell(u))^{n-(N-1)m-k})\right\}.$$

Ainsi, nous avons $E_m\Pi_{\tau_\ell}\leq\mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)\leq E_p\Pi_{\tau_\ell}$. Sachant que $\Pi_{\tau_\ell}=M_0+\tau_\ell M_1$, nous obtenons :

$$\begin{array}{c}\dfrac{E_mM_0\sum_{\ell=1}^{N}\tau_\ell\sum_{\ell=1}^{N}\tau_\ell+E_mM_1\sum_{\ell=1}^{N}\tau_\ell\sum_{\ell=1}^{N}{\tau_\ell}^2-E_pM_0\sum_{\ell=1}^{N}{\tau_\ell}^2\sum_{\ell=1}^{N}1-E_pM_1\sum_{\ell=1}^{N}{\tau_\ell}^2\sum_{\ell=1}^{N}\tau_\ell}{\left(\sum_{\ell=1}^{N}\tau_\ell\right)\left(\sum_{\ell=1}^{N}\tau_\ell\right)-\left(\sum_{\ell=1}^{N}1\right)\left(\sum_{\ell=1}^{N}{\tau_\ell}^2\right)}\\ \leq\mathbb{E}\left(\widehat{M_0}(u,v)\right)\leq\\ \dfrac{E_pM_0\sum_{\ell=1}^{N}\tau_\ell\sum_{\ell=1}^{N}\tau_\ell+E_pM_1\sum_{\ell=1}^{N}\tau_\ell\sum_{\ell=1}^{N}{\tau_\ell}^2-E_mM_0\sum_{\ell=1}^{N}{\tau_\ell}^2\sum_{\ell=1}^{N}1-E_mM_1\sum_{\ell=1}^{N}{\tau_\ell}^2\sum_{\ell=1}^{N}\tau_\ell}{\left(\sum_{\ell=1}^{N}\tau_\ell\right)\left(\sum_{\ell=1}^{N}\tau_\ell\right)-\left(\sum_{\ell=1}^{N}1\right)\left(\sum_{\ell=1}^{N}{\tau_\ell}^2\right)}\end{array}$$

$$\frac{E_m M_0 \sum_{\ell=1}^{N}\tau_\ell \sum_{\ell=1}^{N} 1 + E_m M_1 \sum_{\ell=1}^{N}\tau_\ell \sum_{\ell=1}^{N}\tau_\ell - E_p M_0 \sum_{\ell=1}^{N} 1 \sum_{\ell=1}^{N}\tau_\ell - E_p M_1 \sum_{\ell=1}^{N} 1 \sum_{\ell=1}^{N}\tau_\ell{}^2}{\left(\sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell)\right)\left(\sum_{\ell=1}^{N}\tau_\ell\right) - \left(\sum_{\ell=1}^{N} 1-\tau_\ell\right)\left(\sum_{\ell=1}^{N}\tau_\ell{}^2\right)}$$
$$\leq \mathbb{E}\left(\widehat{\Pi_1}(u,v)\right) \leq$$
$$\frac{E_p M_0 \sum_{\ell=1}^{N}\tau_\ell \sum_{\ell=1}^{N} 1 + E_p M_1 \sum_{\ell=1}^{N}\tau_\ell \sum_{\ell=1}^{N}\tau_\ell - E_m M_0 \sum_{\ell=1}^{N} 1 \sum_{\ell=1}^{N}\tau_\ell - E_m M_1 \sum_{\ell=1}^{N} 1 \sum_{\ell=1}^{N}\tau_\ell{}^2}{\left(\sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell)\right)\left(\sum_{\ell=1}^{N}\tau_\ell\right) - \left(\sum_{\ell=1}^{N} 1-\tau_\ell\right)\left(\sum_{\ell=1}^{N}\tau_\ell{}^2\right)}$$

Donc

$$M_0 + f_0^-(m) \leq \mathbb{E}\left(\widehat{M_0}(u,v)\right) \leq M_0 + f_0^+(m)$$
$$M_1 + f_1^-(m) \leq \mathbb{E}\left(\widehat{M_1}(u,v)\right) \leq M_1 + f_1^+(m)$$

avec

$$f_0^-(m) = \frac{(E_m-1)M_0 \sum_{\ell=1}^{N}\tau_\ell \sum_{\ell=1}^{N}\tau_\ell - (E_p-1)M_0 \sum_{\ell=1}^{N}\tau_\ell{}^2 \sum_{\ell=1}^{N} 1 + (E_m-E_p)M_1 \sum_{\ell=1}^{N}\tau_\ell{}^2 \sum_{\ell=1}^{N}\tau_\ell}{\left(\sum_{\ell=1}^{N}\tau_\ell\right)\left(\sum_{\ell=1}^{N}\tau_\ell\right) - \left(\sum_{\ell=1}^{N} 1\right)\left(\sum_{\ell=1}^{N}\tau_\ell{}^2\right)}$$

$$f_0^+(m) = \frac{(E_p-1)M_0 \sum_{\ell=1}^{N}\tau_\ell \sum_{\ell=1}^{N}\tau_\ell - (E_m-1)M_0 \sum_{\ell=1}^{N}\tau_\ell{}^2 \sum_{\ell=1}^{N} 1 + (E_p-E_m)M_1 \sum_{\ell=1}^{N}\tau_\ell{}^2 \sum_{\ell=1}^{N}\tau_\ell}{\left(\sum_{\ell=1}^{N}\tau_\ell\right)\left(\sum_{\ell=1}^{N}\tau_\ell\right) - \left(\sum_{\ell=1}^{N} 1\right)\left(\sum_{\ell=1}^{N}\tau_\ell{}^2\right)}$$

$$f_1^-(m) = \frac{(E_m-E_p)M_0 \sum_{\ell=1}^{N}\tau_\ell \sum_{\ell=1}^{N} 1 + (E_m-1)M_1 \sum_{\ell=1}^{N}\tau_\ell \sum_{\ell=1}^{N}\tau_\ell - (E_p-1)M_1 \sum_{\ell=1}^{N} 1 \sum_{\ell=1}^{N}\tau_\ell{}^2}{\left(\sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell)\right)\left(\sum_{\ell=1}^{N}\tau_\ell\right) - \left(\sum_{\ell=1}^{N} 1-\tau_\ell\right)\left(\sum_{\ell=1}^{N}\tau_\ell{}^2\right)}$$

$$f_1^+(m) = \frac{(E_p-E_m)M_0 \sum_{\ell=1}^{N}\tau_\ell \sum_{\ell=1}^{N} 1 + (E_p-1)M_1 \sum_{\ell=1}^{N}\tau_\ell \sum_{\ell=1}^{N}\tau_\ell - (E_m-1)M_1 \sum_{\ell=1}^{N} 1 \sum_{\ell=1}^{N}\tau_\ell{}^2}{\left(\sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell)\right)\left(\sum_{\ell=1}^{N}\tau_\ell\right) - \left(\sum_{\ell=1}^{N} 1-\tau_\ell\right)\left(\sum_{\ell=1}^{N}\tau_\ell{}^2\right)}$$

où $f_0^-(m)$, $f_0^+(m)$, $f_1^-(m)$ et $f_1^+(m)$ tendent vers 0 avec m. Ces estimateurs sont donc asymptotiquement sans biais.

Variances

$$\begin{cases} \widehat{M_0}(u,v) = \dfrac{R_2 R_6(u,v) - R_5 R_3(u,v)}{R_4 R_2 - R_1 R_5} \\ \widehat{M_1}(u,v) = \dfrac{R_4 R_3(u,v) - R_1 R_6(u,v)}{R_4 R_2 - R_1 R_5} \end{cases}$$

$$\iff \begin{cases} \mathbb{V}\left(\widehat{M_0}(u,v)\right) = \dfrac{{R_2}^2 \mathbb{V}\left(R_6(u,v)\right) + {R_5}^2 \mathbb{V}\left(R_3(u,v)\right) - 2R_2R_5 \mathbb{C}\mathrm{ov}\left(R_3(u,v), R_6(u,v)\right)}{\left(R_4R_2 - R_1R_5\right)^2} \\ \mathbb{V}\left(\widehat{M_1}(u,v)\right) = \dfrac{{R_4}^2 \mathbb{V}\left(R_3(u,v)\right) - {R_1}^2 \mathbb{V}\left(R_6(u,v)\right) - 2R_1R_4 \mathbb{C}\mathrm{ov}\left(R_3(u,v), R_6(u,v)\right)}{\left(R_4R_2 - R_1R_5\right)^2} \end{cases}.$$

Nous avons comme auparavant l'indépendance des estimateurs $\widehat{\Pi_{S_\ell}}(u,v)$ des segments. Ainsi :

$$\begin{aligned}
\mathbb{V}\left(R_3(u,v)\right) &= \sum_{\ell=1}^{N} \mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right) \\
\mathbb{V}\left(R_6(u,v)\right) &= \sum_{\ell=1}^{N} {\tau_\ell}^2 \mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right) \\
\mathbb{C}\mathrm{ov}\left(R_3(u,v), R_6(u,v)\right) &= \mathbb{E}\left(R_3(u,v)R_6(u,v)\right) - \mathbb{E}\left(R_3(u,v)\right)\mathbb{E}\left(R_6(u,v)\right) \\
&= \sum_{(\ell 1,\ell 2)\in[\![1,N]\!]^2} \tau_{\ell 2}\mathbb{E}\left(\widehat{\Pi_{S_{\ell 1}}}(u,v)\widehat{\Pi_{S_{\ell 2}}}(u,v)\right) \\
&- \left(\sum_{\ell_1=0}^{N-1} \mathbb{E}\left(\widehat{\Pi_{S_{\ell_1}}}(u,v)\right)\right)\left(\sum_{\ell_2=0}^{N-1} \tau_{\ell_2}\mathbb{E}\left(\widehat{\Pi_{S_{\ell_2}}}(u,v)\right)\right) \\
&= \sum_{\ell=1}^{N} \tau_\ell \mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)
\end{aligned}$$

En effet les termes pour lesquels $\ell_1 \neq \ell_2$ s'annulent du fait de l'indépendance des estimateurs des segments. Ainsi :

$$\begin{aligned}
\mathbb{V}\left(\widehat{M_0}(u,v)\right) &= \frac{\left(\sum_{\ell=1}^{N}\tau_\ell\right)^2\left(\sum_{\ell=1}^{N}{\tau_\ell}^2\mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)\right) + \left(\sum_{\ell=1}^{N}{\tau_\ell}^2\right)^2\left(\sum_{\ell=1}^{N}\mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)\right)}{\left[\left(\sum_{\ell=1}^{N}\tau_\ell\right)\left(\sum_{\ell=1}^{N}\tau_\ell\right) - \left(\sum_{\ell=1}^{N}1\right)\left(\sum_{\ell=1}^{N}{\tau_\ell}^2\right)\right]^2} \\
&- \frac{2\left(\sum_{\ell=1}^{N}\tau_\ell\right)\left(\sum_{\ell=1}^{N}{\tau_\ell}^2\right)\left(\sum_{\ell=1}^{N}\tau_\ell\mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)\right)}{\left[\left(\sum_{\ell=1}^{N}\tau_\ell\right)\left(\sum_{\ell=1}^{N}\tau_\ell\right) - \left(\sum_{\ell=1}^{N}1\right)\left(\sum_{\ell=1}^{N}{\tau_\ell}^2\right)\right]^2} \\
\mathbb{V}\left(\widehat{M_1}(u,v)\right) &= \frac{\left(\sum_{\ell=1}^{N}\tau_\ell\right)^2\left(\sum_{\ell=1}^{N}\mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)\right) + \left(\sum_{\ell=1}^{N}1\right)^2\left(\sum_{\ell=1}^{N}{\tau_\ell}^2\mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)\right)}{\left[\left(\sum_{\ell=1}^{N}\tau_\ell\right)\left(\sum_{\ell=1}^{N}\tau_\ell\right) - \left(\sum_{\ell=1}^{N}1\right)\left(\sum_{\ell=1}^{N}{\tau_\ell}^2\right)\right]^2} \\
&- \frac{2\left(\sum_{\ell=1}^{N}\tau_\ell\right)\left(\sum_{\ell=1}^{N}1\right)\left(\sum_{\ell=1}^{N}\tau_\ell\mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)\right)}{\left[\left(\sum_{\ell=1}^{N}\tau_\ell\right)\left(\sum_{\ell=1}^{N}\tau_\ell\right) - \left(\sum_{\ell=1}^{N}1\right)\left(\sum_{\ell=1}^{N}{\tau_\ell}^2\right)\right]^2}
\end{aligned}$$

Comme précédemment, plusieurs simulations nous ont conduit à considérer des segments de taille $m = \sqrt{n}$. En effet, nous avons constaté que la variance atteint son minimum quand les segments sont de cette taille. Notre choix de privilégier la méthode point par point nous amène à ne pas nous étendre davantage sur ce choix.

B.3 Estimation de M_0 et M_1 point par point

Nous nous intéressons désormais à la troisième méthode. Notre modèle est toujours celui des polynômes :

$$\Pi_{\frac{t}{n}}(u,v) = M_0(u,v) + M_1(u,v)\frac{t}{n}$$

Il s'agit de minimiser la fonction PP_1 suivante :

$$PP_1\left(M_0, M_1\right) = \sum_{t=k}^{n}\left[\sum_{u\in\mathcal{A}^k}\mathbb{1}\{X_{t-k}\dots X_{t-1} = u\}\left[\sum_{v\in\mathcal{A}}\left(\Pi_{\frac{t}{n}}(u,v) - \mathbb{1}_{\{X_t=v\}}\right)^2\right]\right]$$

C'est la même fonction qu'auparavant excepté le fait que la matrice de transition s'exprime différemment (en fonction de M_0 et M_1 et non plus en fonction de Π_0 et Π_1). Nous notons toujours $\mathbb{1}_u = \mathbb{1}_{\{X_{t-k}\dots X_{t-1}=u\}}$, $\mathbb{1}_v = \mathbb{1}_{\{X_t=v\}}$ et $\mathbb{1}_{uv} = \mathbb{1}_{\{X_{t-k}\dots X_t=uv\}}$.

Déterminons la dérivée de PP_1.

$$\begin{aligned}\frac{\partial PP_1}{\partial M_0(u,v)}(M_0,M_1) &= \sum_{t=k}^{n} \mathbb{1}_u 2\left(M_0(u,v) + M_1(u,v)\frac{t}{n} - \mathbb{1}_v\right)\\ &= P_1(u)M_0(u,v) + P_2(u)M_1(u,v) - P_3\end{aligned}$$

avec $P_1(u) = 2\sum_{t=k}^{n}\mathbb{1}_u, P_2(u) = 2\sum_{t=k}^{n}\mathbb{1}_u\left(\frac{t}{n}\right)$ et $P_3(u,v) = 2\sum_{t=k}^{n}\mathbb{1}_{uv}$.

$$\begin{aligned}\frac{\partial PP_1}{\partial M_1(u,v)}(M_0,M_1) &= \sum_{t=k}^{n}\mathbb{1}_{\{X_{t-k}\dots X_{t-1}=u\}}2\frac{t}{n}\left(M_0(u,v) + M_1(u,v)\frac{t}{n} - \mathbb{1}_{\{X_t=v\}}\right)\\ &= P_4(u)M_0(u,v) + P_5(u)M_1(u,v) - P_6(u,v)\end{aligned}$$

avec $P_4(u) = 2\sum_{t=k}^{n}\mathbb{1}_u\left(\frac{t}{n}\right), P_5(u) = 2\sum_{t=k}^{n}\mathbb{1}_u\left(\frac{t}{n}\right)^2$ et $P_6(u,v) = 2\sum_{t=k}^{n}\mathbb{1}_{uv}\left(\frac{t}{n}\right)$.

Remarque 15 *Nous avons les affirmations suivantes :*
- $P_4(u) = P_2(u)$ *;*
- $P_1(u), P_4(u), P_2(u), P_5(u) \neq 0$ *si u apparaît dans la séquence ;*
- $P_3(u,v), P_6(u,v) \neq 0$ *si uv apparaît dans la séquence.*

Ces coefficients sont presque toujours non nuls pour une séquence assez longue (un exemple : à l'ordre 1*, il suffit que tous les mots de deux lettres apparaissent dans la séquence).*

Déterminons le minimum.

$$\begin{aligned}&\begin{cases} P_1(u)\widehat{M_0}(u,v) + P_2(u)\widehat{M_1}(u,v) - P_3(u,v) &= 0\\ P_4(u)\widehat{M_0}(u,v) + P_5(u)\widehat{M_1}(u,v) - P_6(u,v) &= 0\end{cases}\\ \iff &\begin{cases}\widehat{M_0}(u,v) &= \dfrac{P_5(u)P_3(u,v) - P_2(u)P_6(u,v)}{P_1(u)P_5(u) - P_4(u)P_2(u)}\\ \widehat{M_1}(u,v) &= \dfrac{P_1(u)P_6(u,v) - P_4(u)P_3(u,v)}{P_1(u)P_5(u) - P_4(u)P_2(u)}\end{cases}.\end{aligned}$$

Donc

$$\widehat{M_0}(u,v) = \frac{\begin{pmatrix}\left(\sum_{t=k}^{n}\mathbb{1}_u\left(\frac{t}{n}\right)^2\right)\left(\sum_{t=k}^{n}\mathbb{1}_{uv}\right) -\\ \left(\sum_{t=k}^{n}\mathbb{1}_u\left(\frac{t}{n}\right)\right)\left(\sum_{t=k}^{n}\mathbb{1}_{uv}\left(\frac{t}{n}\right)\right)\end{pmatrix}}{\begin{pmatrix}\left(\sum_{t=k}^{n}\mathbb{1}_u\right)\left(\sum_{t=k}^{n}\mathbb{1}_u\left(\frac{t}{n}\right)^2\right) -\\ \left(\sum_{t=k}^{n}\mathbb{1}_u\left(\frac{t}{n}\right)\right)\left(\sum_{t=k}^{n}\mathbb{1}_u\left(\frac{t}{n}\right)\right)\end{pmatrix}}, \quad \widehat{M_1}(u,v) = \frac{\begin{pmatrix}\left(\sum_{t=k}^{n}\mathbb{1}_u\right)\left(\sum_{t=k}^{n}\mathbb{1}_{uv}\left(\frac{t}{n}\right)\right) -\\ \left(\sum_{t=k}^{n}\mathbb{1}_u\left(\frac{t}{n}\right)\right)\left(\sum_{t=k}^{n}\mathbb{1}_{uv}\right)\end{pmatrix}}{\begin{pmatrix}\left(\sum_{t=k}^{n}\mathbb{1}_u\right)\left(\sum_{t=k}^{n}\mathbb{1}_u\left(\frac{t}{n}\right)^2\right) -\\ \left(\sum_{t=k}^{n}\mathbb{1}_u\left(\frac{t}{n}\right)\right)\left(\sum_{t=k}^{n}\mathbb{1}_u\left(\frac{t}{n}\right)\right)\end{pmatrix}}.$$

B.3.1 Stochasticité des matrices $\widehat{\Pi_{\frac{t}{n}}}$

Théorème 20 *Les matrices $\widehat{\Pi_{\frac{t}{n}}}$ sont stochastiques.*

Preuve *Il suffit de montrer que la somme des termes de chaque ligne est égale à* 1 *et que tous les termes sont positifs.*

• *Montrons tout d'abord que la somme sur chaque ligne vaut* 1. *Pour cela, nous montrons que* $\sum_{v\in\mathcal{A}} \widehat{M_1}(u,v) = 0$ *et que* $\sum_{v\in\mathcal{A}} \widehat{M_0}(u,v) = 1$. *Ainsi* $\forall t \in [\![0,n]\!]$, $\sum_{v\in\mathcal{A}} \widehat{\Pi_{\frac{t}{n}}}(u,v) = 1$.

$$\begin{aligned}
\sum_{v\in\mathcal{A}} \widehat{M_0}(u,v) &= \frac{\left(\begin{array}{c}\left(\sum_{t=k}^{n}\mathbb{1}_u\left(\frac{t}{n}\right)^2\right)\left(\sum_{t=k}^{n}\mathbb{1}_u\right) - \\ \left(\sum_{t=k}^{n}\mathbb{1}_u\left(\frac{t}{n}\right)\right)\left(\sum_{t=k}^{n}\mathbb{1}_u\left(\frac{t}{n}\right)\right)\end{array}\right)}{\left(\begin{array}{c}\left(\sum_{t=k}^{n}\mathbb{1}_u\right)\left(\sum_{t=k}^{n}\mathbb{1}_u\left(\frac{t}{n}\right)^2\right) - \\ \left(\sum_{t=k}^{n}\mathbb{1}_u\left(\frac{t}{n}\right)\right)\left(\sum_{t=k}^{n}\mathbb{1}_u\left(\frac{t}{n}\right)\right)\end{array}\right)} \\
&= \frac{\left(\begin{array}{c}\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2}\mathbb{1}_{\{X_{t_1-1}=u,X_{t_2-1}=u\}}\left(\frac{t_2}{n}\right)^2 - \\ \sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2}\mathbb{1}_{\{X_{t_1-1}=u,X_{t_2-1}=u\}}\left(\frac{t_1}{n}\right)\left(\frac{t_2}{n}\right)\end{array}\right)}{\left(\begin{array}{c}\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2}\mathbb{1}_{\{X_{t_1-1}=u,X_{t_2-1}=u\}}\left(\frac{t_2}{n}\right)^2 - \\ \sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2}\mathbb{1}_{\{X_{t_1-1}=u,X_{t_2-1}=u\}}\left(\frac{t_1}{n}\right)\left(\frac{t_2}{n}\right)\end{array}\right)} = 1.
\end{aligned}$$

$$\sum_{v\in\mathcal{A}} \widehat{M_1}(u,v) = \frac{\left(\begin{array}{c}\left(\sum_{t=k}^{n}\mathbb{1}_u\right)\left(\sum_{t=k}^{n}\mathbb{1}_u\left(\frac{t}{n}\right)\right) - \\ \left(\sum_{t=k}^{n}\mathbb{1}_u\left(\frac{t}{n}\right)\right)\left(\sum_{t=k}^{n}\mathbb{1}_u\right)\end{array}\right)}{\left(\begin{array}{c}\left(\sum_{t=k}^{n}\mathbb{1}_u\right)\left(\sum_{t=k}^{n}\mathbb{1}_u\left(\frac{t}{n}\right)^2\right) - \\ \left(\sum_{t=k}^{n}\mathbb{1}_u\left(\frac{t}{n}\right)\right)\left(\sum_{t=k}^{n}\mathbb{1}_u\left(\frac{t}{n}\right)\right)\end{array}\right)} = 0.$$

Les termes ne sont donc pas toujours positifs. Lors de l'estimation nous proposons un réajustement proportionnel des valeurs (voir A.1.2).

B.3.2 Espérances et variances de $\widehat{M_0}$ et $\widehat{M_1}$

Les estimateurs sont convergents. Nous ne traitons pas ce cas particulier. Voir le cas général en D.2.2.
Afin de faciliter la lecture nous gardons les notations suivantes :
- $\mathbb{P}_t(u) = \mathbb{P}(X_{t-k}\dots X_{t-1} = u)$;
- $\mathbb{P}_t(uv) = \mathbb{P}(X_{t-k}\dots X_{t-1} = u, X_t = v)$;
- $\mathbb{P}_{t_1t_2}(uu) = \mathbb{P}_(X_{t_1-k}\dots X_{t_1-1} = u, X_{t_2-k}\dots X_{t_2-1} = u)$;
- $\mathbb{P}_{t_1}(u) = \mathbb{P}(X_{t_1-k}\dots X_{t_1-1} = u)$;
- $\mathbb{P}_{t_2}(u) = \mathbb{P}(X_{t_2-k}\dots X_{t_2-1} = u)$;
- $\mathbb{P}_{t_1t_2}(uvuv) = \mathbb{P}(X_{t_1-k}\dots X_{t_1-1} = u, X_{t_1} = v, X_{t_2-k}\dots X_{t_2-1} = u, X_{t_2} = v)$.

Calculons les espérances.

$$\begin{aligned}
\mathbb{E}(P_1(u)) &= 2\sum_{t=k}^{n}\mathbb{P}_t(u). \\
\mathbb{E}(P_2(u)) &= 2\sum_{t=k}^{n}\mathbb{P}_t(u)\left(\frac{t}{n}\right).
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}(P_3(u,v)) &= 2\sum_{t=k}^{n}\mathbb{P}_t(uv) = 2\sum_{t=k}^{n}\mathbb{P}_t(u)\Pi_{\frac{t}{n}}(u,v).\\
\mathbb{E}(P_4(u)) &= 2\sum_{t=k}^{n}\mathbb{P}_t(u)\left(\frac{t}{n}\right).\\
\mathbb{E}(P_5(u)) &= 2\sum_{t=k}^{n}\mathbb{P}_t(u)\left(\frac{t}{n}\right)^2.\\
\mathbb{E}(P_6(u,v)) &= 2\sum_{t=k}^{n}\mathbb{P}_t(uv)\left(\frac{t}{n}\right) = 2\sum_{t=k}^{n}\mathbb{P}_t(u)\Pi_{\frac{t}{n}}(u,v)\left(\frac{t}{n}\right).
\end{aligned}$$

Par le théorème de la limite centrale, la loi de chacun des P_i (pour i allant de 1 à 6) est approchée par une loi normale. Ainsi la delta méthode nous permet d'approcher les lois de nos estimateurs par des lois normales d'espérances :

$$\begin{aligned}
\mathbb{E}(\widehat{M_0}(u,v)) &= \frac{\mathbb{E}(P_5(u))\mathbb{E}(P_3(u,v)) - \mathbb{E}(P_2(u))\mathbb{E}(P_6(u,v))}{\mathbb{E}(P_1(u))\mathbb{E}(P_5(u)) - \mathbb{E}(P_4(u))\mathbb{E}(P_2(u))}\\
\mathbb{E}(\widehat{M_1}(u,v)) &= \frac{\mathbb{E}(P_1(u))\mathbb{E}(P_6(u,v)) - \mathbb{E}(P_4(u))\mathbb{E}(P_3(u,v))}{\mathbb{E}(P_1(u))\mathbb{E}(P_5(u)) - \mathbb{E}(P_4(u))\mathbb{E}(P_2(u))}.
\end{aligned}$$

Nous obtenons les espérances suivantes :

$$\begin{aligned}
\mathbb{E}(\widehat{M_0}(u,v)) &= \frac{\displaystyle\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2}\mathbb{P}_{t_1}(u)\mathbb{P}_{t_2}(u)\Pi_{\frac{t_2}{n}}(u,v)\left(\frac{t_1}{n}\right)\left(\frac{t_1}{n}-\frac{t_2}{n}\right)}{\displaystyle\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2}\mathbb{P}_{t_1}(u)\mathbb{P}_{t_2}(u)\left(\frac{t_2}{n}\right)\left(\frac{t_2}{n}-\frac{t_1}{n}\right)}\\
&= \frac{\displaystyle\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2}\mathbb{P}_{t_1}(u)\mathbb{P}_{t_2}(u)M_0(u,v)\left(\frac{t_1}{n}\right)\left(\frac{t_1}{n}-\frac{t_2}{n}\right)}{\displaystyle\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2}\mathbb{P}_{t_1}(u)\mathbb{P}_{t_2}(u)\left(\frac{t_2}{n}\right)\left(\frac{t_2}{n}-\frac{t_1}{n}\right)}\\
&+ \frac{\displaystyle\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2}\mathbb{P}_{t_1}(u)\mathbb{P}_{t_2}(u)M_1(u,v)\left(\frac{t_2}{n}\right)\left(\frac{t_1}{n}\right)\left(\frac{t_1}{n}-\frac{t_2}{n}\right)}{\displaystyle\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2}\mathbb{P}_{t_1}(u)\mathbb{P}_{t_2}(u)\left(\frac{t_2}{n}\right)\left(\frac{t_2}{n}-\frac{t_1}{n}\right)}.
\end{aligned}$$

Ainsi $\mathbb{E}(\widehat{M_0}(u,v)) = M_0$.

$$\begin{aligned}
\mathbb{E}(\widehat{M_1}(u,v)) &= \frac{\displaystyle\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2}\mathbb{P}_{t_1}(u)\mathbb{P}_{t_2}(u)\Pi_{\frac{t_2}{n}}(u,v)\left(\frac{t_2}{n}-\frac{t_1}{n}\right)}{\displaystyle\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2}\mathbb{P}_{t_1}(u)\mathbb{P}_{t_2}(u)\left(1-\frac{t_1}{n}\right)\left(\frac{t_2}{n}\right)\left(\frac{t_2}{n}-\frac{t_1}{n}\right)}\\
&= \frac{\displaystyle\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2}\mathbb{P}_{t_1}(u)\mathbb{P}_{t_2}(u)M_0(u,v)\left(\frac{t_2}{n}-\frac{t_1}{n}\right)}{\displaystyle\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2}\mathbb{P}_{t_1}(u)\mathbb{P}_{t_2}(u)\left(\frac{t_2}{n}\right)\left(\frac{t_2}{n}-\frac{t_1}{n}\right)}\\
&+ \frac{\displaystyle\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2}\mathbb{P}_{t_1}(u)\mathbb{P}_{t_2}(u)M_1(u,v)\left(\frac{t_2}{n}\right)\left(\frac{t_2}{n}-\frac{t_1}{n}\right)}{\displaystyle\sum_{(t_1,t_2)\in[\![k,n]\!]^2,t_1\neq t_2}\mathbb{P}_{t_1}(u)\mathbb{P}_{t_2}(u)\left(\frac{t_2}{n}\right)\left(\frac{t_2}{n}-\frac{t_1}{n}\right)}.
\end{aligned}$$

Ainsi $\mathbb{E}(\widehat{M_1}(u,v)) = M_1$.
Nos estimateurs sont asymptotiquement sans biais.

Calculons maintenant les variances.

Nous rappelons que :
$$\begin{array}{rcl} \widehat{M_0}(u,v) & = & \dfrac{P_5(u)P_3(u,v) - P_2(u)P_6(u,v)}{P_1(u)P_5(u) - P_4(u)P_2(u)} \\ \widehat{M_1}(u,v) & = & \dfrac{P_1(u)P_6(u,v) - P_4(u)P_3(u,v)}{P_1(u)P_5(u) - P_4(u)P_2(u)} \end{array} .$$

La delta méthode nous permet d'approcher les lois de nos estimateurs par des lois normales de variances :

$$\mathbb{V}(\widehat{M_0}(u,v)) = [d\widehat{M_0}^t][\mathbb{C}\mathrm{ov}_P][d\widehat{M_0}]$$

$$\mathbb{V}(\widehat{M_1}(u,v)) = [d\widehat{M_1}^t][\mathbb{C}\mathrm{ov}_P][d\widehat{M_1}]$$

où $[d\widehat{M_0}]_i = \dfrac{\partial \widehat{M_0}}{\partial P_i}$, $[d\widehat{M_1}]_i = \dfrac{\partial \widehat{M_1}}{\partial P_i}$ et $[\mathbb{C}\mathrm{ov}_P]_{ij} = \mathbb{C}\mathrm{ov}(P_i, P_j)$.

Nous utilisons les approximations suivantes pour le cas où les ensembles $[\![t_1 - k, t_1]\!]$ et $[\![t_2 - k, t_2]\!]$ ont une intersection vide :

- $\mathbb{P}_{t_1t_2}(uu) \simeq \mathbb{P}_{t_1}(u)\mathbb{P}_{t_2}(u)$;
- $\mathbb{P}_{t_1t_2}(uvuv) \simeq \mathbb{P}_{t_1}(u)\mathbb{P}_{t_2}(u)\Pi_{\frac{t_1}{n}}(u,v)\Pi_{\frac{t_2}{n}}(u,v)$.

Nous n'expliciterons ici que la diagonale de la matrice de covariance des P_i, les autres termes étant assez similaires :

$$\begin{array}{rcl}
\mathbb{V}(P_1(u)) & = & 4\displaystyle\sum_{t_1=k}^{n}\sum_{t_2=k}^{n}\mathbb{P}_{t_1t_2}(uu) - (\mathbb{E}(P_1(u)))^2. \\
\mathbb{V}(P_2(u)) & = & 4\displaystyle\sum_{t_1=k}^{n}\sum_{t_2=k}^{n}\mathbb{P}_{t_1t_2}(uu)\left(\frac{t_1}{n}\right)\left(\frac{t_2}{n}\right) - (\mathbb{E}(P_2(u)))^2. \\
\mathbb{V}(P_3(u,v)) & = & 4\displaystyle\sum_{t_1=k}^{n}\sum_{t_2=k}^{n}\mathbb{P}_{t_1t_2}(uvuv) - (\mathbb{E}(P_3(u,v)))^2 \\
& = & 4\displaystyle\sum_{t_1=k}^{n}\sum_{t_2=k}^{n}\mathbb{P}_{t_1t_2}(uu)\Pi_{\frac{t_1}{n}}(u,v)\Pi_{\frac{t_2}{n}}(u,v) - 4\left(\sum_{t=k}^{n}\mathbb{P}_t(u)\Pi_{\frac{t}{n}}(u,v)\right)^2. \\
\mathbb{V}(P_4(u)) & = & 4\displaystyle\sum_{t_1=k}^{n}\sum_{t_2=k}^{n}\mathbb{P}_{t_1t_2}(uu)\left(\frac{t_1}{n}\right)\left(\frac{t_2}{n}\right) - (\mathbb{E}(P_4(u)))^2. \\
\mathbb{V}(P_5(u)) & = & 4\displaystyle\sum_{t_1=k}^{n}\sum_{t_2=k}^{n}\mathbb{P}_{t_1t_2}(uu)\left(\frac{t_1}{n}\right)^2\left(\frac{t_2}{n}\right)^2 - (\mathbb{E}(P_5(u)))^2. \\
\mathbb{V}(P_6(u,v)) & = & 4\displaystyle\sum_{t_1=k}^{n}\sum_{t_2=k}^{n}\mathbb{P}_{t_1t_2}(uvuv)\left(\frac{t_1}{n}\right)\left(\frac{t_2}{n}\right) - (\mathbb{E}(P_6(u,v)))^2 \\
& = & 4\displaystyle\sum_{t_1=k}^{n}\sum_{t_2=k}^{n}\mathbb{P}_{t_1t_2}(uu)\Pi_{\frac{t_1}{n}}(u,v)\Pi_{\frac{t_2}{n}}(u,v)\left(\frac{t_1}{n}\right)\left(\frac{t_2}{n}\right) - 4\left(\sum_{t=k}^{n}\mathbb{P}_t(u)\Pi_{\frac{t}{n}}(u,v)\left(\frac{t}{n}\right)\right)^2.
\end{array}$$

Lorsque l'intersection de $[\![t_1 - k, t_1]\!]$ et $[\![t_2 - k, t_2]\!]$ n'est pas vide (t_1 proche de t_2), les termes des différentes sommes ne sont pas nuls car nous n'utilisons pas les approximations précédentes. Nous pourrons nous inspirer de [SPdT95] pour terminer le calcul.

Annexe C

Estimation des $\Pi_{\frac{i}{d}}$

Nous présentons ici les calculs complémentaires relatifs au chapitre 2 de la Partie I.

C.1 Estimation des $\Pi_{\frac{i}{d}}$ par régression matricielle

C.1.1 Stochasticité des matrices $\widehat{\Pi_{\frac{t}{n}}}$

Preuve (Théorème 9 p38) *Montrons que* $\sum_{v\in\mathcal{A}} \widehat{\Pi_{\frac{i}{d}}}(u,v) = 1$. *Reprenant le système obtenu pour la minimisation (p37), et sachant que* $\sum_{v\in\mathcal{A}} \widehat{\Pi_{S_\ell}}(u,v) = 1$, *nous avons*

$$\left\{\begin{array}{rcl}
\left(\sum_{\ell=1}^{N} A_0(n\tau_\ell)A_0(n\tau_\ell)\right)\sum_{v\in\mathcal{A}}\widehat{\Pi_0}(u,v) + \ldots + \left(\sum_{\ell=1}^{N} A_0(n\tau_\ell)A_d(n\tau_\ell)\right)\sum_{v\in\mathcal{A}}\widehat{\Pi_1}(u,v) - \sum_{\ell=1}^{N} A_0(n\tau_\ell) &=& 0 \\
\vdots & & \\
\left(\sum_{\ell=1}^{N} A_j(n\tau_\ell)A_0(n\tau_\ell)\right)\sum_{v\in\mathcal{A}}\widehat{\Pi_0}(u,v) + \ldots + \left(\sum_{\ell=1}^{N} A_j(n\tau_\ell)A_d(n\tau_\ell)\right)\sum_{v\in\mathcal{A}}\widehat{\Pi_1}(u,v) - \sum_{\ell=1}^{N} A_j(n\tau_\ell) &=& 0 \\
\vdots & & \\
\left(\sum_{\ell=1}^{N} A_d(n\tau_\ell)A_0(n\tau_\ell)\right)\sum_{v\in\mathcal{A}}\widehat{\Pi_0}(u,v) + \ldots + \left(\sum_{\ell=1}^{N} A_d(n\tau_\ell)A_d(n\tau_\ell)\right)\sum_{v\in\mathcal{A}}\widehat{\Pi_1}(u,v) - \sum_{\ell=1}^{N} A_d(n\tau_\ell) &=& 0
\end{array}\right.$$

Ainsi, nous avons

$$\begin{pmatrix}
\sum_{\ell=1}^{N} A_0(n\tau_\ell)A_0(n\tau_\ell) & \cdots & \sum_{\ell=1}^{N} A_0(n\tau_\ell)A_d(n\tau_\ell) \\
\vdots & & \vdots \\
\sum_{\ell=1}^{N} A_d(n\tau_\ell)A_0(n\tau_\ell) & \cdots & \sum_{\ell=1}^{N} A_d(n\tau_\ell)A_d(n\tau_\ell)
\end{pmatrix}
\begin{pmatrix}
\sum_{v\in\mathcal{A}}\widehat{\Pi_0}(u,v) \\
\vdots \\
\sum_{v\in\mathcal{A}}\widehat{\Pi_1}(u,v)
\end{pmatrix} =
\begin{pmatrix}
\sum_{\ell=1}^{N} A_0(n\tau_\ell) \\
\vdots \\
\sum_{\ell=1}^{N} A_d(n\tau_\ell)
\end{pmatrix}$$

$X = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ *est la solution du système. Ainsi, nous aurons bien des matrices stochastiques si tous les termes des matrices sont positifs (ce qui n'est à nouveau pas forcément le cas (voir 1)).*

C.1.2 Distances : deuxième méthode

De la même manière qu'en A.1.3 au lieu de choisir un seul point par segment, nous choisissons tous les points du segment pour ajuster notre modèle. Nous minimisons la somme suivante :

$$\sum_{t=k}^{n}\sum_{u\in\mathcal{A}^k}\sum_{v\in\mathcal{A}}\left(\widehat{\Pi_{S_{\ell_t}}}(u,v)-\left(1-\frac{t}{n}\right)\Pi_0(u,v)-\frac{t}{n}\Pi_1(u,v)\right)^2.$$

De la même manière que pour un seul point, nous obtenons les estimateurs des $\Pi_{\frac{i}{d}}$ stochastiques.

C.1.3 Espérances et variances de $\widehat{\Pi_{S_\ell}}$ pour le modèle des chaînes de Markov régulées

Calcul de $\mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)$

Comme auparavant, il faut calculer espérances et variances par rapport au nouveau modèle considéré. Nous avons

$$\mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)=A\sum_{t\in S_\ell^\star}\frac{\Pi_{\frac{t}{n}}(u,v)}{|S_\ell^\star|}$$

Ce qu'il faut calculer, c'est

$$\begin{aligned}\sum_{t\in S_\ell^\star}\Pi_{\frac{t}{n}}(u,v) &= \Pi_0(u,v)\sum_{t\in S_\ell^\star}A_0(t)+\ldots+\Pi_1(u,v)\sum_{t\in S_\ell^\star}A_d\\ &= \sum_{i=0}^{d}\left(\Pi_{\frac{i}{d}}(u,v)\sum_{t\in S_\ell^\star}A_i(t)\right).\end{aligned}$$

Sachant que

$$\begin{aligned}\sum_{t\in S_\ell^\star}A_i(t) &= a_i^0\sum_{t\in S_\ell^\star}1+a_i^1\sum_{t\in S_\ell^\star}\frac{t}{n}+\ldots+a_i^d\sum_{t\in S_\ell^\star}\frac{t^d}{n^d}\\ &= \sum_{j=0}^{d}\left(a_i^j\sum_{t\in S_\ell^\star}\frac{t^j}{n^j}\right),\end{aligned}$$

nous obtenons

$$\sum_{t\in S_\ell^\star}\Pi_{\frac{t}{n}}(u,v)=\sum_{i=0}^{d}\left(\Pi_{\frac{i}{d}}(u,v)\sum_{j=0}^{d}\left(a_i^j\sum_{t\in S_\ell^\star}\frac{t^j}{n^j}\right)\right).$$

Nous aimerions bien

$$\sum_{t\in S_\ell^\star}\frac{\Pi_{\frac{t}{n}}(u,v)}{|S_\ell^\star|}=\Pi_{\tau_\ell}+\varepsilon_{\mathbb{E}}$$

avec, nous l'espèrons $\varepsilon_{\mathbb{E}}$ petit devant Π_{τ_ℓ}.

Calcul de $\mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)$

Pour la variance, nous avons comme auparavant

$$\mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)=A\sum_{t\in S_\ell^\star}\frac{\Pi_{\frac{t}{n}}(u,v)}{|S_\ell^\star|}+2B\sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star,t_1<t_2}\frac{\Pi_{\frac{t_1}{n}}(u,v)\Pi_{\frac{t_2}{n}}(u,v)}{|S_\ell^\star|(|S_\ell^\star|-1)}-\left(\mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)\right)^2.$$

Ce qu'il faut calculer c'est

$$\begin{aligned}&\sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star,t_1<t_2}\Pi_{\frac{t_1}{n}}(u,v)\Pi_{\frac{t_2}{n}}(u,v)\\ =&\sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star,t_1<t_2}\left(\sum_{i=0}^{d}A_i(t)\Pi_{\frac{i}{d}}(u,v)\right)\left(\sum_{j=0}^{d}A_j(t)\Pi_{\frac{j}{d}}(u,v)\right)\\ =&\sum_{(i,j)\in[\![0,d]\!]^2}\Pi_{\frac{i}{d}}(u,v)\Pi_{\frac{j}{d}}(u,v)\sum_{(x,y)\in[\![0,d]\!]^2}a_i^xa_j^y\left(\sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star,t_1<t_2}\frac{t_1^xt_2^y}{n^{x+y}}\right).\end{aligned}$$

Reste $\sum_{t\in S_\ell^\star} t^j$ et $\sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star, t_1<t_2} t_1^i t_2^j$ pour tous les i,j.

Nous aimerions bien

$$\frac{2\sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star,t_1<t_2}\Pi_{\frac{t_1}{n}}(u,v)\Pi_{\frac{t_2}{n}}(u,v)}{|S_\ell^\star|(|S_\ell^\star|-1)} = {\Pi_{\tau_\ell}}^2 + \varepsilon_{\mathbb{V}}$$

avec, nous l'espèrons $\varepsilon_{\mathbb{V}}$ petit devant ${\Pi_{\tau_\ell}}^2$.

Remarque 16 *Alors que pour une dérive linéaire nous pouvions conclure à la convergence des estimateurs (voir Théorème 6 p30 et ce qui suit), cela n'est plus possible ici. En effet, les fonctions $\varepsilon_{\mathbb{E}}$ et $\varepsilon_{\mathbb{V}}$ définies plus bas ne tendent pas vers zéro.*

$$\varepsilon_{\mathbb{E}} = \sum_{i=0}^{d}\left(\Pi_{\frac{i}{d}}(u,v)\sum_{j=0}^{d} a_i^j\left[\frac{\sum_{t\in S_\ell^\star}\frac{t^j}{n^j}}{|S_\ell^\star|} - {\tau_\ell}^j\right]\right)$$

$$\varepsilon_{\mathbb{V}} = \sum_{(i,j)\in[\![0,d]\!]^2}\Pi_{\frac{i}{d}}(u,v)\Pi_{\frac{j}{d}}(u,v)\sum_{(x,y)\in[\![0,d]\!]^2} a_i^x a_j^y\left[\frac{2\sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star,t_1<t_2}\frac{t_1^x t_2^y}{n^{x+y}}}{|S_\ell^\star|(|S_\ell^\star|-1)} - {\tau_\ell}^{x+y}\right]$$

Ceci explique en partie le fait que cette méthode d'estimation donne de moins bonnes vraisemblances que la méthode point par point.

C.2 Estimation des $\Pi_{\frac{i}{d}}$ point par point

C.2.1 Stochasticité des matrices $\widehat{\Pi_{\frac{i}{d}}}$

Preuve (Théorème 10 p39) *Montrons que* $\sum_{v\in\mathcal{A}}\widehat{\Pi_{\frac{i}{d}}}(u,v) = 1$.

Reprenant le système obtenu pour la minimisation (p39), et sachant que $\sum_{v\in\mathcal{A}}\mathbb{1}_{uv} = \mathbb{1}_u$, *nous avons*

$$\begin{cases}
\sum_{i=0}^{d}\sum_{v\in\mathcal{A}}\widehat{\Pi_{\frac{i}{d}}}(u,v)\left(\sum_{t=k}^{n}\mathbb{1}_u A_0(t)A_i(t)\right) - \sum_{t=k}^{n}A_0(t)\mathbb{1}_u &=& 0\\
\vdots & & \\
\sum_{i=0}^{d}\sum_{v\in\mathcal{A}}\widehat{\Pi_{\frac{i}{d}}}(u,v)\left(\sum_{t=k}^{n}\mathbb{1}_u A_j(t)A_i(t)\right) - \sum_{t=k}^{n}A_j(t)\mathbb{1}_u &=& 0\\
\vdots & & \\
\sum_{i=0}^{d}\sum_{v\in\mathcal{A}}\widehat{\Pi_{\frac{i}{d}}}(u,v)\left(\sum_{t=k}^{n}\mathbb{1}_u A_d(t)A_i(t)\right) - \sum_{t=k}^{n}A_d(t)\mathbb{1}_u &=& 0
\end{cases}$$

Nous avons ainsi

$$\begin{pmatrix}
\sum_{t=k}^{n}\mathbb{1}_u A_0(t)A_0(t) & \cdots & \sum_{t=k}^{n}\mathbb{1}_u A_0(t)A_d(t)\\
\vdots & & \vdots\\
\sum_{t=k}^{n}\mathbb{1}_u A_d(t)A_0(t) & \cdots & \sum_{t=k}^{n}\mathbb{1}_u A_d(t)A_d(t)
\end{pmatrix}\begin{pmatrix}
\sum_{v\in\mathcal{A}}\widehat{\Pi_0}(u,v)\\
\vdots\\
\sum_{v\in\mathcal{A}}\widehat{\Pi_1}(u,v)
\end{pmatrix} = \begin{pmatrix}
\sum_{t=k}^{n}A_0(t)\mathbb{1}_u\\
\vdots\\
\sum_{t=k}^{n}A_d(t)\mathbb{1}_u
\end{pmatrix}.$$

$X = \begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}$ *est la solution du système. Ainsi, nous aurons bien des matrices stochastiques si tous les termes des matrices sont positifs (ce qui n'est à nouveau pas forcément le cas (voir 1)).*

C.2.2 Espérances et variances de $\widehat{\Pi_{\frac{i}{d}}}$

De la même manière que pour le cas linéaire (voir A.2.2), nous montrons la convergence de nos estimateurs à l'aide de la delta méthode. Notre système 2.2 (page 39) devient

$$\begin{pmatrix} \mathbb{E}\left(\sum_{t=k}^{n} \mathbb{1}_u A_0(t)A_0(t)\right) & \cdots & \mathbb{E}\left(\sum_{t=k}^{n} \mathbb{1}_u A_0(t)A_d(t)\right) \\ \vdots & & \vdots \\ \mathbb{E}\left(\sum_{t=k}^{n} \mathbb{1}_u A_d(t)A_0(t)\right) & \cdots & \mathbb{E}\left(\sum_{t=k}^{n} \mathbb{1}_u A_d(t)A_d(t)\right) \end{pmatrix} \begin{pmatrix} \mathbb{E}\left(\widehat{\Pi_0}(u,v)\right) \\ \vdots \\ \mathbb{E}\left(\widehat{\Pi_1}(u,v)\right) \end{pmatrix} = \begin{pmatrix} \mathbb{E}\left(\sum_{t=k}^{n} A_0(t)\mathbb{1}_{uv}\right) \\ \vdots \\ \mathbb{E}\left(\sum_{t=k}^{n} A_d(t)\mathbb{1}_{uv}\right) \end{pmatrix}$$

Le vecteur $\begin{pmatrix} \Pi_0(u,v) \\ \vdots \\ \Pi_1(u,v) \end{pmatrix}$ est a solution de ce système. Ainsi nos estimateurs sont asymptotiquement sans biais.

Le calcul de la variance se fait de la même manière que dans le cas linéaire (voir Annexe A.2.2).

Annexe D

Estimation des M_d

Nous présentons ici les calculs complémentaires relatifs au chapitre 2 de la Partie I. Rappelons le modèle :

$$\Pi_{\frac{t}{n}}(u,v) = \sum_{i=0}^{d} M_i(u,v)\frac{t^i}{n^i}.$$

D.1 Estimation des M_i par régression matricielle

Nous devons minimiser la fonction PR_d suivante :

$$PR_d(M_0,\dots,M_1) = \sum_{\ell\in[\![1,N]\!]}\sum_{u\in\mathcal{A}^k}\sum_{v\in\mathcal{A}}\left(\widehat{\Pi_{S_\ell}}(u,v) - \sum_{i=0}^{d} M_i(u,v)\frac{t^i}{n^i}\right)^2$$

La recherche des zéros de la dérivée nous donne le système suivant pour chaque couple $(u,v)\in\mathcal{A}^k\times\mathcal{A}$:

$$\left\{\begin{array}{rcl} \dfrac{\partial PR_d(M_0,\dots,M_1)}{\partial M_0(u,v)} &=& 0 \\ &\vdots& \\ \dfrac{\partial PR_d(M_0,\dots,M_1)}{\partial M_j(u,v)} &=& 0 \\ &\vdots& \\ \dfrac{\partial PR_d(M_0,\dots,M_1)}{\partial M_d(u,v)} &=& 0 \end{array}\right.$$

$$\iff \left\{\begin{array}{rcl} \displaystyle\sum_{\ell=1}^{N}\left(\widehat{\Pi_{S_\ell}}(u,v) - \sum_{i=0}^{d}{\tau_\ell}^i\widehat{M_i}(u,v)\right) &=& 0 \\ &\vdots& \\ \displaystyle\sum_{\ell=1}^{N}{\tau_\ell}^j\left(\widehat{\Pi_{S_\ell}}(u,v) - \sum_{i=0}^{d}{\tau_\ell}^i\widehat{M_i}(u,v)\right) &=& 0 \\ &\vdots& \\ \displaystyle\sum_{\ell=1}^{N}{\tau_\ell}^d\left(\widehat{\Pi_{S_\ell}}(u,v) - \sum_{i=0}^{d}{\tau_\ell}^i\widehat{M_i}(u,v)\right) &=& 0 \end{array}\right.$$

$$\iff \left\{ \begin{array}{rcl} \sum_{i=0}^{d} \widehat{M_i}(u,v) \left(\sum_{\ell=1}^{N} {\tau_\ell}^i \right) - \sum_{\ell=1}^{N} \widehat{\Pi_{S_\ell}}(u,v) & = & 0 \\ \vdots & & \\ \sum_{i=0}^{d} \widehat{M_i}(u,v) \left(\sum_{\ell=1}^{N} {\tau_\ell}^j {\tau_\ell}^i \right) - \sum_{\ell=1}^{N} {\tau_\ell}^j \widehat{\Pi_{S_\ell}}(u,v) & = & 0 \\ \vdots & & \\ \sum_{i=0}^{d} \widehat{M_i}(u,v) \left(\sum_{\ell=1}^{N} {\tau_\ell}^d {\tau_\ell}^i \right) - \sum_{\ell=1}^{N} {\tau_\ell}^d \widehat{\Pi_{S_\ell}}(u,v) & = & 0 \end{array} \right.$$

$$\iff \left\{ \begin{array}{rcl} \left(\sum_{\ell=1}^{N} 1 \right) \widehat{M_0}(u,v) + \ldots + \left(\sum_{\ell=1}^{N} {\tau_\ell}^d \right) \widehat{M_d}(u,v) - \sum_{\ell=1}^{N} \widehat{\Pi_{S_\ell}}(u,v) & = & 0 \\ & \vdots & \\ \left(\sum_{\ell=1}^{N} {\tau_\ell}^j \right) \widehat{M_0}(u,v) + \ldots + \left(\sum_{\ell=1}^{N} {\tau_\ell}^j {\tau_\ell}^d \right) \widehat{M_d}(u,v) - \sum_{\ell=1}^{N} {\tau_\ell}^j \widehat{\Pi_{S_\ell}}(u,v) & = & 0 \\ & \vdots & \\ \left(\sum_{\ell=1}^{N} {\tau_\ell}^d \right) \widehat{M_0}(u,v) + \ldots + \left(\sum_{\ell=1}^{N} {\tau_\ell}^d {\tau_\ell}^d \right) \widehat{M_d}(u,v) - \sum_{\ell=1}^{N} {\tau_\ell}^d \widehat{\Pi_{S_\ell}}(u,v) & = & 0 \end{array} \right.$$

Nous obtenons autant de systèmes que pour le modèle des points d'appuis. Nous avons alors à m fixé, les estimateurs des $\widehat{M_i}$.

Nous réécrivons le système sous la forme matricielle suivante :

$$\begin{pmatrix} \sum_{\ell=1}^{N} 1 & \cdots & \sum_{\ell=1}^{N} {\tau_\ell}^d \\ \vdots & & \vdots \\ \sum_{\ell=1}^{N} {\tau_\ell}^d & \cdots & \sum_{\ell=1}^{N} {\tau_\ell}^{2d} \end{pmatrix} \begin{pmatrix} \widehat{M_0}(u,v) \\ \vdots \\ \widehat{M_d}(u,v) \end{pmatrix} = \begin{pmatrix} \sum_{\ell=1}^{N} \widehat{\Pi_{S_\ell}}(u,v) \\ \vdots \\ \sum_{\ell=1}^{N} {\tau_\ell}^d \widehat{\Pi_{S_\ell}}(u,v) \end{pmatrix}$$

Nous notons ce système $MX = P$ avec

$$M_{ij} = \sum_{\ell=1}^{N} {\tau_\ell}^{i+j}, X_i = \widehat{M_i}(u,v) \text{ et } P_i = \sum_{\ell=1}^{N} {\tau_\ell}^i \widehat{\Pi_{S_\ell}}(u,v).$$

D.1.1 Stochasticité des matrices $\widehat{\Pi_{\frac{t}{n}}}$

Théorème 21 *Les matrices $\widehat{\Pi_{\frac{t}{n}}}$ sont stochastiques.*

Preuve *Montrons que $\sum_{v \in \mathcal{A}} \widehat{M_0}(u,v) = 1$ et que $\sum_{v \in \mathcal{A}} \widehat{M_i}(u,v) = 0$ pour $i \neq 0$. Reprenant le système précédent, et sachant que $\sum_{v \in \mathcal{A}} \widehat{\Pi_{S_\ell}}(u,v) = 1$, nous avons*

$$\left\{ \begin{array}{rcl} \left(\sum_{\ell=1}^{N} 1 \right) \sum_{v \in \mathcal{A}} \widehat{M_0}(u,v) + \ldots + \left(\sum_{\ell=1}^{N} {\tau_\ell}^d \right) \sum_{v \in \mathcal{A}} \widehat{M_d}(u,v) - \sum_{\ell=1}^{N} 1 & = & 0 \\ & \vdots & \\ \left(\sum_{\ell=1}^{N} {\tau_\ell}^j \right) \sum_{v \in \mathcal{A}} \widehat{M_0}(u,v) + \ldots + \left(\sum_{\ell=1}^{N} {\tau_\ell}^j {\tau_\ell}^d \right) \sum_{v \in \mathcal{A}} \widehat{M_d}(u,v) - \sum_{\ell=1}^{N} {\tau_\ell}^j & = & 0 \\ & \vdots & \\ \left(\sum_{\ell=1}^{N} {\tau_\ell}^d \right) \sum_{v \in \mathcal{A}} \widehat{M_0}(u,v) + \ldots + \left(\sum_{\ell=1}^{N} {\tau_\ell}^d {\tau_\ell}^d \right) \sum_{v \in \mathcal{A}} \widehat{M_d}(u,v) - \sum_{\ell=1}^{N} {\tau_\ell}^d & = & 0 \end{array} \right.$$

Nous avons

$$\begin{pmatrix} \sum_{\ell=1}^{N} 1 & \cdots & \sum_{\ell=1}^{N} {\tau_\ell}^d \\ \vdots & & \vdots \\ \sum_{\ell=1}^{N} {\tau_\ell}^d & \cdots & \sum_{\ell=1}^{N} {\tau_\ell}^{2d} \end{pmatrix} \begin{pmatrix} \sum_{v\in\mathcal{A}} \widehat{M_0}(u,v) \\ \vdots \\ \sum_{v\in\mathcal{A}} \widehat{M_d}(u,v) \end{pmatrix} = \begin{pmatrix} \sum_{\ell=1}^{N} 1 \\ \vdots \\ \sum_{\ell=1}^{N} {\tau_\ell}^d \end{pmatrix}$$

$X = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ *est la solution du système. Ainsi, nous aurons bien* $\Pi_{\frac{t}{n}}$ *stochastiques si tous les termes de ces matrices sont positifs (ce qui n'est à nouveau pas forcément le cas (voir 1)).*

D.1.2 Distances : deuxième méthode

De la même manière qu'en A.1.3 au lieu de choisir un seul point par segment, nous choisissons tous les points du segment pour ajuster notre modèle. Nous minimisons la somme suivante :

$$\sum_{t=k}^{n} \sum_{u\in\mathcal{A}^k} \sum_{v\in\mathcal{A}} \left(\widehat{\Pi_{S_{\ell_t}}}(u,v) - \sum_{i=0}^{d} M_i(u,v) \frac{t^i}{n^i} \right)^2$$

De la même manière que pour un seul point, nous obtenons les estimateurs des M_i avec $\Pi_{\frac{t}{n}}$ stochastique.

D.1.3 Espérances et variances de $\widehat{\Pi_{S_\ell}}$ pour le modèle des chaînes de Markov régulées

Calcul de $\mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)$

Comme auparavant, il faut calculer espérances et variances par rapport au nouveau modèle considéré. Nous avons

$$\mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right) = A \sum_{t\in S_\ell^\star} \frac{\Pi_{\frac{t}{n}}(u,v)}{|S_\ell^\star|}.$$

Ce qu'il faut calculer, c'est

$$\begin{aligned} \sum_{t\in S_\ell^\star} \frac{\Pi_{\frac{t}{n}}(u,v)}{|S_\ell^\star|} &= M_0(u,v) \sum_{t\in S_\ell^\star} + \ldots + M_d(u,v) \sum_{t\in S_\ell^\star} \frac{t^d}{n^d} \\ &= \sum_{i=0}^{d} \left(M_i(u,v) \sum_{t\in S_\ell^\star} \frac{t^i}{n^i} \right) \\ &= \sum_{i=0}^{d} \left(M_i(u,v) \left(\sum_{t\in S_\ell^\star} \frac{t^i}{n^i} \right) \right). \end{aligned}$$

Nous aimerions bien

$$\sum_{t\in S_\ell^\star} \frac{\Pi_{\frac{t}{n}}(u,v)}{|S_\ell^\star|} = \Pi_{\tau_\ell} + \varepsilon_{\mathbb{E}}$$

avec, nous l'espèrons $\varepsilon_{\mathbb{E}}$ petit devant Π_{τ_ℓ}.

Calcul de $\mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)$

Pour la variance, nous avons comme auparavant

$$\mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right) = A \sum_{t\in S_\ell^\star} \frac{\Pi_{\frac{t}{n}}(u,v)}{|S_\ell^\star|} + 2B \sum_{(t_1,t_2)\in S_\ell^\star \times S_\ell^\star, t_1<t_2} \frac{\Pi_{\frac{t_1}{n}}(u,v)\Pi_{\frac{t_2}{n}}(u,v)}{|S_\ell^\star|(|S_\ell^\star|-1)} - \left(\mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)\right)^2.$$

Ce qu'il faut calculer c'est

$$\begin{aligned}
& \sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star, t_1<t_2} \Pi_{\frac{t_1}{n}}(u,v)\Pi_{\frac{t_2}{n}}(u,v) \\
= & \sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star, t_1<t_2} \left(\sum_{i=0}^{d} M_i \frac{t_1^i}{n^i}\right)\left(\sum_{j=0}^{d} M_j \frac{t_2^j}{n^j}\right) \\
= & \sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star, t_1<t_2} \sum_{(i,j)\in[\![0,d]\!]^2} M_i M_j \frac{t_1^i t_2^j}{n^{i+j}} \\
= & \sum_{(i,j)\in[\![0,d]\!]^2} M_i M_j \left(\sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star, t_1<t_2} \frac{t_1^i t_2^j}{n^{i+j}}\right)
\end{aligned}.$$

Reste $\sum_{t\in S_\ell^\star} t^j$ et $\sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star, t_1<t_2} t_1^i t_2^j$ pour tous les i,j.

Nous aimerions bien

$$\frac{2\sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star, t_1<t_2} \Pi_{\frac{t_1}{n}}(u,v)\Pi_{\frac{t_2}{n}}(u,v)}{|S_\ell^\star|(|S_\ell^\star|-1)} = {\Pi_{\tau_\ell}}^2 + \varepsilon_{\mathbb{V}}$$

avec, nous l'espèrons $\varepsilon_{\mathbb{V}}$ petit devant ${\Pi_{\tau_\ell}}^2$.

Remarque 17 *Alors que pour une dérive linéaire nous pouvions conclure à la convergence des estimateurs (voir Théorème 18 p112 et ce qui suit), cela n'est plus possible ici. En effet, les fonctions $\varepsilon_{\mathbb{E}}$ et $\varepsilon_{\mathbb{V}}$ définies plus bas ne tendent pas vers zéro.*

$$\varepsilon_{\mathbb{E}} = \sum_{i=0}^{d} M_i \left[\frac{\sum_{t\in S_\ell^\star} \frac{t^i}{n^i}}{|S_\ell^\star|} - {\tau_\ell}^i\right]$$

$$\varepsilon_{\mathbb{V}} = \sum_{(i,j)\in[\![0,d]\!]^2} M_i M_j \left[\frac{2\sum_{(t_1,t_2)\in S_\ell^\star\times S_\ell^\star, t_1<t_2} \frac{t_1^i t_2^j}{n^{i+j}}}{|S_\ell^\star|(|S_\ell^\star|-1)} - {\tau_\ell}^{i+j}\right]$$

Ceci explique en partie le fait que cette méthode d'estimation donne de moins bonnes vraisemblances que la méthode point par point.

D.1.4 Espérances et variances des $\widehat{M_i}$

Espérances

Nous avons $\mathbb{E}(X) = M^{-1}\mathbb{E}(P)$. Ainsi

$$\mathbb{E}(M) = \begin{pmatrix} \sum_{\ell=1}^{N} \mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right) \\ \vdots \\ \sum_{\ell=1}^{N} {\tau_\ell}^d \mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right) \end{pmatrix}$$

Ayant déjà calculé $\mathbb{E}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)$ (voir D.1.3), nous déduisons $\mathbb{E}(X)$.

Variances

Nous avons $\mathbb{V}(X) = M^{-1}\mathbb{V}(P)(M^t)^{-1}$. Le calcul de $\mathbb{V}(P)$ nécessite celui des deux termes suivants :

- $\mathbb{V}\left(\sum_{\ell=1}^{N} {\tau_\ell}^i \widehat{\Pi_{S_\ell}}(u,v)\right) = \sum_{\ell=1}^{N} {\tau_\ell}^{2i}\mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)$
- $\mathbb{C}\text{ov}\left(\sum_{\ell=1}^{N} {\tau_\ell}^i \widehat{\Pi_{S_\ell}}(u,v), \sum_{\ell=1}^{N} {\tau_\ell}^j \widehat{\Pi_{S_\ell}}(u,v)\right) = \sum_{\ell=1}^{N} {\tau_\ell}^{i+j}\mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)$

Nous nous servons de l'indépendance des segments pour la seconde égalité. Ayant déjà calculé $\mathbb{V}\left(\widehat{\Pi_{S_\ell}}(u,v)\right)$, nous déduisons $\mathbb{V}(X)$.

D.2 Estimation de M_i point par point

Rappelons le modèle des polynômes :

$$\Pi_{\frac{t}{n}}(u,v) = \sum_{i=0}^{d} M_i(u,v)\frac{t^i}{n^i}.$$

De la même façon que précédemment (voir 1.1.3), nous devons minimiser la fonction PP_d suivante :

$$PP_d\left(M_0,\ldots,M_d\right) = \sum_{t=k}^{n}\left[\sum_{u\in\mathcal{A}^k}\mathbb{1}\{X_{t-k}\ldots X_{t-1}=u\}\left[\sum_{v\in\mathcal{A}}\left(\Pi_{\frac{t}{n}}(u,v) - \mathbb{1}_{\{X_t=v\}}\right)^2\right]\right].$$

La recherche des zéros de la dérivée nous donne le système suivant pour chaque couple $(u,v)\in\mathcal{A}^k\times\mathcal{A}$:

$$\left\{\begin{array}{rcl}\frac{\partial PP_d(M_0,\ldots,M_d)}{\partial M_0(u,v)} &=& 0\\ &\vdots&\\ \frac{\partial PP_d(M_0,\ldots,M_d)}{\partial M_j(u,v)} &=& 0\\ &\vdots&\\ \frac{\partial PP_d(M_0,\ldots,M_d)}{\partial M_d(u,v)} &=& 0\end{array}\right.$$

$$\Longleftrightarrow \left\{\begin{array}{rcl}\sum_{t=k}^{n}\mathbb{1}_{\{X_{t-k}\ldots X_{t-1}=u\}}\left(\sum_{i=0}^{d}\frac{t^i}{n^i}\widehat{M_i}(u,v) - \mathbb{1}_{\{X_t=v\}}\right) &=& 0\\ &\vdots&\\ \sum_{t=k}^{n}\mathbb{1}_{\{X_{t-k}\ldots X_{t-1}=u\}}\frac{t^j}{n^j}\left(\sum_{i=0}^{d}\frac{t^i}{n^i}\widehat{M_i}(u,v) - \mathbb{1}_{\{X_t=v\}}\right) &=& 0\\ &\vdots&\\ \sum_{t=k}^{n}\mathbb{1}_{\{X_{t-k}\ldots X_{t-1}=u\}}\frac{t^d}{n^d}\left(\sum_{i=0}^{d}\frac{t^i}{n^i}\widehat{M_i}(u,v) - \mathbb{1}_{\{X_t=v\}}\right) &=& 0\end{array}\right.$$

$$\Longleftrightarrow \left\{\begin{array}{rcl}\sum_{i=0}^{d}\widehat{M_i}(u,v)\left(\sum_{t=k}^{n}\mathbb{1}_{\{X_{t-k}\ldots X_{t-1}=u\}}\frac{t^i}{n^i}\right) - \sum_{t=k}^{n}\mathbb{1}_{\{X_{t-k}\ldots X_{t-1}=u,X_t=v\}} &=& 0\\ &\vdots&\\ \sum_{i=0}^{d}\widehat{M_i}(u,v)\left(\sum_{t=k}^{n}\mathbb{1}_{\{X_{t-k}\ldots X_{t-1}=u\}}\frac{t^j}{n^j}\frac{t^i}{n^i}\right) - \sum_{t=k}^{n}\frac{t^j}{n^j}\mathbb{1}_{\{X_{t-k}\ldots X_{t-1}=u,X_t=v\}} &=& 0\\ &\vdots&\\ \sum_{i=0}^{d}\widehat{M_i}(u,v)\left(\sum_{t=k}^{n}\mathbb{1}_{\{X_{t-k}\ldots X_{t-1}=u\}}\frac{t^d}{n^d}\frac{t^i}{n^i}\right) - \sum_{t=k}^{n}\frac{t^d}{n^d}\mathbb{1}_{\{X_{t-k}\ldots X_{t-1}=u,X_t=v\}} &=& 0\end{array}\right.$$

Pour un alphabet $\mathcal{A}$ donné, nous obtenons ainsi $(|\mathcal{A}|)^{k+1}$ systèmes de $d+1$ équations à $d+1$ inconnues. Nous avons alors les estimateurs $\widehat{M_i}$.

En notant $\mathbb{1}_u = \mathbb{1}_{\{X_{t-k}\dots X_{t-1}=u\}}$ et $\mathbb{1}_{uv} = \mathbb{1}_{\{X_{t-k}\dots X_{t-1}=u, X_t=v\}}$, nous réécrivons le système sous la forme matricielle suivante :

$$\begin{pmatrix} \sum_{t=k}^{n} \mathbb{1}_u & \cdots & \sum_{t=k}^{n} \mathbb{1}_u \frac{t^d}{n^d} \\ \vdots & & \vdots \\ \sum_{t=k}^{n} \mathbb{1}_u \frac{t^d}{n^d} & \cdots & \sum_{t=k}^{n} \mathbb{1}_u \frac{t^{2d}}{n^{2d}} \end{pmatrix} \begin{pmatrix} \widehat{M_0}(u,v) \\ \vdots \\ \widehat{M_d}(u,v) \end{pmatrix} = \begin{pmatrix} \sum_{t=k}^{n} \mathbb{1}_{uv} \\ \vdots \\ \sum_{t=k}^{n} \frac{t^d}{n^d} \mathbb{1}_{uv} \end{pmatrix} \tag{D.1}$$

Nous notons ce système $MX = P$ avec

$$M_{ij} = \sum_{t=k}^{n} \mathbb{1}_u \frac{t^{i+j}}{n^{i+j}}, X_i = \widehat{M_i}(u,v) \text{ et } P_i = \sum_{t=k}^{n} \frac{t^i}{n^i} \mathbb{1}_{uv}.$$

D.2.1 Stochasticité des matrices $\widehat{\Pi_{\frac{t}{n}}}$

Théorème 22 *Les matrices $\widehat{\Pi_{\frac{t}{n}}}$ sont stochastiques.*

Preuve *Montrons que $\sum_{v\in\mathcal{A}} \widehat{M_0}(u,v) = 1$ et que $\sum_{v\in\mathcal{A}} \widehat{M_i}(u,v) = 0$ pour $i \neq 0$. Reprenant le système précédent, et sachant que $\sum_{v\in\mathcal{A}} \mathbb{1}_{uv} = \mathbb{1}_u$, nous avons*

$$\begin{cases} \sum_{i=0}^{d} \sum_{v\in\mathcal{A}} \widehat{M_i}(u,v) \left(\sum_{t=k}^{n} \mathbb{1}_u \frac{t^i}{n^i} \right) - \sum_{t=k}^{n} \mathbb{1}_u & = & 0 \\ & \vdots & \\ \sum_{i=0}^{d} \sum_{v\in\mathcal{A}} \widehat{M_i}(u,v) \left(\sum_{t=k}^{n} \mathbb{1}_u \frac{t^j}{n^j} \frac{t^i}{n^i} \right) - \sum_{t=k}^{n} \frac{t^j}{n^j} \mathbb{1}_u & = & 0 \\ & \vdots & \\ \sum_{i=0}^{d} \sum_{v\in\mathcal{A}} \widehat{M_i}(u,v) \left(\sum_{t=k}^{n} \mathbb{1}_u \frac{t^d}{n^d} \frac{t^i}{n^i} \right) - \sum_{t=k}^{n} \frac{t^d}{n^d} \mathbb{1}_u & = & 0 \end{cases}$$

Nous avons

$$\begin{pmatrix} \sum_{t=k}^{n} \mathbb{1}_u & \cdots & \sum_{t=k}^{n} \mathbb{1}_u \frac{t^d}{n^d} \\ \vdots & & \vdots \\ \sum_{t=k}^{n} \mathbb{1}_u \frac{t^d}{n^d} & \cdots & \sum_{t=k}^{n} \mathbb{1}_u \frac{t^{2d}}{n^{2d}} \end{pmatrix} \begin{pmatrix} \sum_{v\in\mathcal{A}} \widehat{M_0}(u,v) \\ \vdots \\ \sum_{v\in\mathcal{A}} \widehat{M_d}(u,v) \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{t=k}^{n} \mathbb{1}_u \\ \vdots \\ \sum_{t=k}^{n} \frac{t^d}{n^d} \mathbb{1}_u \end{pmatrix}$$

$X = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ *est la solution du système. Ainsi, nous aurons bien $\Pi_{\frac{t}{n}}$ stochastiques si tous les termes de ces matrices sont positifs (ce qui n'est à nouveau pas forcément le cas (voir 1)).*

D.2.2 Espérances et variances des $\widehat{M_i}$

De la même manière que pour le cas linéaire (voir B.3.2), nous montrons la convergence de nos estimateurs à l'aide de la delta méthode. Notre système D.1 devient

$$\begin{pmatrix} \mathbb{E}\left(\sum_{t=k}^{n} \mathbb{1}_u\right) & \cdots & \mathbb{E}\left(\sum_{t=k}^{n} \mathbb{1}_u \frac{t^d}{n^d}\right) \\ \vdots & & \vdots \\ \mathbb{E}\left(\sum_{t=k}^{n} \mathbb{1}_u \frac{t^d}{n^d}\right) & \cdots & \mathbb{E}\left(\sum_{t=k}^{n} \mathbb{1}_u \frac{t^{2d}}{n^{2d}}\right) \end{pmatrix} \begin{pmatrix} \mathbb{E}\left(\widehat{M_0}(u,v)\right) \\ \vdots \\ \mathbb{E}\left(\widehat{M_d}(u,v)\right) \end{pmatrix} = \begin{pmatrix} \mathbb{E}\left(\sum_{t=k}^{n} \mathbb{1}_{uv}\right) \\ \vdots \\ \mathbb{E}\left(\sum_{t=k}^{n} \frac{t^d}{n^d} \mathbb{1}_{uv}\right) \end{pmatrix} \tag{D.2}$$

(D.3)

Le vecteur $\begin{pmatrix} M_0(u,v) \\ \vdots \\ M_d(u,v) \end{pmatrix}$ est a solution de ce système. Ainsi nos estimateurs sont asymptotiquement sans biais.

Le calcul de la variance se fait de la même manière que dans le cas linéaire (voir Annexe B.3.2).

Annexe E

Splines

E.1 Estimation globale

Nous présentons ici quelques cas particuliers relatifs au 3.2 de la partie I.

E.1.1 Degré 1

Nous détaillons ici la méthode pour un découpage en 2 puis 3 morceaux.

2 morceaux

Sur chacun des morceaux i, nous avons une matrice de transition $\Pi^i_{\frac{t}{n}}$ calculée à partir de deux matrices "point d'appuis" estimées :

$$\begin{aligned}\Pi^1_{\frac{t}{n}}(u,v) &= \frac{\alpha-t}{\alpha}\Pi_0(u,v)+\frac{t}{\alpha}\Pi_{\frac{\alpha}{n}}(u,v)\\ \Pi^2_{\frac{t}{n}}(u,v) &= \frac{n-t}{n-\alpha}\Pi_{\frac{\alpha}{n}}(u,v)+\frac{t-\alpha}{n-\alpha}\Pi_1(u,v).\end{aligned}$$

Comme précédemment (voir 1.1.3), nous utilisons la méthode point par point pour estimer les paramètres du modèle. Nous minimisons la fonction f_1 suivante :

$$\begin{aligned}f_1(\Pi_0,\Pi_{\frac{\alpha}{n}},\Pi_1) &= \sum_{t=k}^{n}\left(\sum_{u\in\mathcal{A}^k}\mathbb{1}_u\left(\sum_{v\in\mathcal{A}}\left(\Pi_{\frac{t}{n}}(u,v)-\mathbb{1}_v\right)^2\right)\right)\\ &= \sum_{t=k}^{\alpha}\left(\sum_{u\in\mathcal{A}^k}\mathbb{1}_u\left(\sum_{v\in\mathcal{A}}\left(\frac{\alpha-t}{\alpha}\Pi_0(u,v)+\frac{t}{\alpha}\Pi_{\frac{\alpha}{n}}(u,v)-\mathbb{1}_v\right)^2\right)\right)\\ &+ \sum_{t=\alpha+1}^{n}\left(\sum_{u\in\mathcal{A}^k}\mathbb{1}_u\left(\sum_{v\in\mathcal{A}}\left(\frac{n-t}{n-\alpha}\Pi_{\frac{\alpha}{n}}(u,v)+\frac{t-\alpha}{n-\alpha}\Pi_1(u,v)-\mathbb{1}_v\right)^2\right)\right).\end{aligned}$$

Explicitons les dérivées :

$$\begin{aligned}\frac{\partial f_1}{\partial\Pi_0(u,v)} &= \sum_{t=k}^{\alpha}\left(\mathbb{1}_u 2\left(\frac{\alpha-t}{\alpha}\right)\left(\frac{\alpha-t}{\alpha}\Pi_0(u,v)+\frac{t}{\alpha}\Pi_{\frac{\alpha}{n}}(u,v)-\mathbb{1}_v\right)\right);\\ \frac{\partial f_1}{\partial\Pi_{\frac{\alpha}{n}}(u,v)} &= \sum_{t=k}^{\alpha}\left(\mathbb{1}_u 2\left(\frac{t}{\alpha}\right)\left(\frac{\alpha-t}{\alpha}\Pi_0(u,v)+\frac{t}{\alpha}\Pi_{\frac{\alpha}{n}}(u,v)-\mathbb{1}_v\right)\right);\\ &+ \sum_{t=\alpha+1}^{n}\left(\mathbb{1}_u 2\left(\frac{n-t}{n-\alpha}\right)\left(\frac{n-t}{n-\alpha}\Pi_{\frac{\alpha}{n}}(u,v)+\frac{t-\alpha}{n-\alpha}\Pi_1(u,v)-\mathbb{1}_v\right)\right);\\ \frac{\partial f_1}{\partial\Pi_1(u,v)} &= \sum_{t=\alpha+1}^{n}\left(\mathbb{1}_u 2\left(\frac{t-\alpha}{n-\alpha}\right)\left(\frac{n-t}{n-\alpha}\Pi_{\frac{\alpha}{n}}(u,v)+\frac{t-\alpha}{n-\alpha}\Pi_1(u,v)-\mathbb{1}_v\right)\right).\end{aligned}$$

La minimisation nous donne le système suivant :

$$\begin{pmatrix} A_1(u) & B_1(u) & C_1(u) \\ A_2(u) & B_2(u) & C_2(u) \\ A_3(u) & B_3(u) & C_3(u) \end{pmatrix} \begin{pmatrix} \Pi_0(u,v) \\ \Pi_{\frac{\alpha}{n}}(u,v) \\ \Pi_1(u,v) \end{pmatrix} = \begin{pmatrix} D_1(u,v) \\ D_2(u,v) \\ D_3(u,v) \end{pmatrix}$$

avec

$$\begin{aligned}
A_1(u) &= \sum_{t=k}^{\alpha} \mathbb{1}_u 2 \left(\frac{\alpha-t}{\alpha}\right)^2, \quad B_1(u) = \sum_{t=k}^{\alpha} \mathbb{1}_u 2 \left(\frac{\alpha-t}{\alpha}\right)\frac{t}{\alpha}, \quad C_1(u) = 0, \quad D_1(u,v) = \sum_{t=k}^{\alpha} \mathbb{1}_{uv} 2 \left(\frac{\alpha-t}{\alpha}\right) \\
A_2(u) &= \sum_{t=k}^{\alpha} \mathbb{1}_u 2 \left(\frac{\alpha-t}{\alpha}\right)\frac{t}{\alpha}, \quad B_2(u) = \sum_{t=k}^{\alpha} \mathbb{1}_u 2 \left(\frac{t}{\alpha}\right)^2 + \sum_{t=\alpha+1}^{n} \mathbb{1}_u 2 \left(\frac{n-t}{n-\alpha}\right)^2 \\
C_2(u) &= \sum_{t=\alpha+1}^{n} \mathbb{1}_u 2 \left(\frac{n-t}{n-\alpha}\right)\left(\frac{t-\alpha}{n-\alpha}\right), \quad D_2(u,v) = \sum_{t=k}^{\alpha} \mathbb{1}_{uv} 2 \left(\frac{t}{\alpha}\right) + \sum_{t=\alpha+1}^{n} \mathbb{1}_u 2 \left(\frac{n-t}{n-\alpha}\right), \quad A_3(u) = 0 \\
B_3(u) &= \sum_{t=\alpha+1}^{n} \mathbb{1}_u 2 \left(\frac{n-t}{n-\alpha}\right)\left(\frac{t-\alpha}{n-\alpha}\right), \quad C_3(u) = \sum_{t=\alpha+1}^{n} \mathbb{1}_u 2 \left(\frac{t-\alpha}{n-\alpha}\right)^2, \quad D_3(u,v) = \sum_{t=\alpha+1}^{n} \mathbb{1}_{uv} 2 \left(\frac{t-\alpha}{n-\alpha}\right).
\end{aligned}$$

Théorème 23 *Les matrices $\Pi^1_{\frac{t}{n}}$ et $\Pi^2_{\frac{t}{n}}$ sont stochastiques.*

Preuve *Nous sommons sur $v \in \mathcal{A}$ les équations. Nous obtenons le système suivant :*

$$\begin{pmatrix} A_1(u) & B_1(u) & C_1(u) \\ A_2(u) & B_2(u) & C_2(u) \\ A_3(u) & B_3(u) & C_3(u) \end{pmatrix} \begin{pmatrix} \sum_{v\in\mathcal{A}} \Pi_0(u,v) \\ \sum_{v\in\mathcal{A}} \Pi_{\frac{\alpha}{n}}(u,v) \\ \sum_{v\in\mathcal{A}} \Pi_1(u,v) \end{pmatrix} = \begin{pmatrix} \sum_{v\in\mathcal{A}} D_1(u,v) \\ \sum_{v\in\mathcal{A}} D_2(u,v) \\ \sum_{v\in\mathcal{A}} D_3(u,v) \end{pmatrix}$$

Le vecteur $(1 \cdots 1)^t$ est solution. Ainsi on montre que les lignes de ces matrices somment toutes à 1. Par conséquent, la stochasticité de $\Pi^1_{\frac{t}{n}}$ et $\Pi^2_{\frac{t}{n}}$ est vérifiée. La même remarque que précédemment (voir A.1.2) peut s'appliquer : il est théoriquement possible d'obtenir des termes négatifs.

3 morceaux

Sur chacun des morceaux i, nous avons une matrice de transition $\Pi^i_{\frac{t}{n}}$ calculée à partir de deux matrices "point d'appuis" estimées :

$$\begin{aligned}
\Pi^1_{\frac{t}{n}}(u,v) &= \frac{\alpha_1 - t}{\alpha_1}\Pi_0(u,v) + \frac{t}{\alpha_1}\Pi_{\frac{\alpha_1}{n}}(u,v) \\
\Pi^2_{\frac{t}{n}}(u,v) &= \frac{\alpha_2 - t}{\alpha_2 - \alpha_1}\Pi_{\frac{\alpha_1}{n}}(u,v) + \frac{t - \alpha_1}{\alpha_2 - \alpha_1}\Pi_{\frac{\alpha_2}{n}}(u,v) \\
\Pi^3_{\frac{t}{n}}(u,v) &= \frac{n - t}{n - \alpha_2}\Pi_{\frac{\alpha_2}{n}}(u,v) + \frac{t - \alpha_2}{n - \alpha_2}\Pi_1(u,v)
\end{aligned}$$

Comme précédemment (voir 1.1.3), nous utilisons la méthode point par point pour estimer les paramètres du modèle. Nous minimisons la fonction f_1 suivante :

$$\begin{aligned}
f_1(\Pi_0, \Pi_{\frac{\alpha_1}{n}}, \Pi_{\frac{\alpha_2}{n}}, \Pi_1) &= \sum_{t=k}^{n} \left(\sum_{u\in\mathcal{A}^k} \mathbb{1}_u \left(\sum_{v\in\mathcal{A}} \left(\Pi_{\frac{t}{n}}(u,v) - \mathbb{1}_v \right)^2 \right) \right) \\
&= \sum_{t=k}^{\alpha_1} \left(\sum_{u\in\mathcal{A}^k} \mathbb{1}_u \left(\sum_{v\in\mathcal{A}} \left(\frac{\alpha_1 - t}{\alpha_1}\Pi_0(u,v) + \frac{t}{\alpha_1}\Pi_{\frac{\alpha_1}{n}}(u,v) - \mathbb{1}_v \right)^2 \right) \right) \\
&+ \sum_{t=\alpha+1}^{\alpha_2} \left(\sum_{u\in\mathcal{A}^k} \mathbb{1}_u \left(\sum_{v\in\mathcal{A}} \left(\frac{\alpha_2 - t}{\alpha_2 - \alpha_1}\Pi_{\frac{\alpha_1}{n}}(u,v) + \frac{t - \alpha_1}{\alpha_2 - \alpha_1}\Pi_{\frac{\alpha_2}{n}}(u,v) - \mathbb{1}_v \right)^2 \right) \right) \\
&+ \sum_{t=\alpha_2+1}^{n} \left(\sum_{u\in\mathcal{A}^k} \mathbb{1}_u \left(\sum_{v\in\mathcal{A}} \left(\frac{n - t}{n - \alpha_2}\Pi_{\frac{\alpha_2}{n}}(u,v) + \frac{t - \alpha_2}{n - \alpha_2}\Pi_1(u,v) - \mathbb{1}_v \right)^2 \right) \right)
\end{aligned}$$

Explicitons les dérivées :

$$\begin{aligned}
\frac{\partial f_1}{\partial \Pi_0(u,v)} &= \sum_{t=k}^{\alpha_1}\left(\mathbb{1}_u 2\left(\frac{\alpha_1-t}{\alpha_1}\right)\left(\frac{\alpha_1-t}{\alpha_1}\Pi_0(u,v)+\frac{t}{\alpha_1}\Pi_{\frac{\alpha_1}{n}}(u,v)-\mathbb{1}_v\right)\right);\\
\frac{\partial f_1}{\partial \Pi_{\frac{\alpha_1}{n}}(u,v)} &= \sum_{t=k}^{\alpha_1}\left(\mathbb{1}_u 2\left(\frac{t}{\alpha_1}\right)\left(\frac{\alpha_1-t}{\alpha_1}\Pi_0(u,v)+\frac{t}{\alpha_1}\Pi_{\frac{\alpha_1}{n}}(u,v)-\mathbb{1}_v\right)\right)\\
&+ \sum_{t=\alpha_1+1}^{\alpha_2}\left(\mathbb{1}_u 2\left(\frac{\alpha_2-t}{\alpha_2-\alpha_1}\right)\left(\frac{\alpha_2-t}{\alpha_2-\alpha_1}\Pi_{\frac{\alpha_1}{n}}(u,v)+\frac{t-\alpha_2}{n-\alpha_2}\Pi_{\frac{\alpha_2}{n}}(u,v)-\mathbb{1}_v\right)\right);\\
\frac{\partial f_1}{\partial \Pi_{\frac{\alpha_2}{n}}(u,v)} &= \sum_{t=\alpha_1+1}^{\alpha_2}\left(\mathbb{1}_u 2\left(\frac{t-\alpha_1}{\alpha_2-\alpha_1}\right)\left(\frac{\alpha_2-t}{\alpha_2-\alpha_1}\Pi_{\frac{\alpha_1}{n}}(u,v)+\frac{t-\alpha_1}{\alpha_2-\alpha_1}\Pi_{\frac{\alpha_2}{n}}(u,v)-\mathbb{1}_v\right)\right)\\
&+ \sum_{t=\alpha_2+1}^{n}\left(\mathbb{1}_u 2\left(\frac{n-t}{n-\alpha_2}\right)\left(\frac{n-t}{n-\alpha_2}\Pi_{\frac{\alpha_2}{n}}(u,v)+\frac{t-\alpha_2}{n-\alpha_2}\Pi_1(u,v)-\mathbb{1}_v\right)\right);\\
\frac{\partial f_1}{\partial \Pi_1(u,v)} &= \sum_{t=\alpha_2+1}^{n}\left(\mathbb{1}_u 2\left(\frac{t-\alpha_2}{n-\alpha_2}\right)\left(\frac{n-t}{n-\alpha_2}\Pi_{\frac{\alpha_2}{n}}(u,v)+\frac{t-\alpha_2}{n-\alpha_2}\Pi_1(u,v)-\mathbb{1}_v\right)\right).
\end{aligned}$$

La minimisation nous donne le système suivant :

$$\begin{pmatrix} A_1(u) & B_1(u) & C_1(u) & D_1(u)\\ A_2(u) & B_2(u) & C_2(u) & D_2(u)\\ A_3(u) & B_3(u) & C_3(u) & D_3(u)\\ A_4(u) & B_4(u) & C_4(u) & D_4(u)\end{pmatrix}\begin{pmatrix}\Pi_0(u,v)\\ \Pi_{\frac{\alpha_1}{n}}(u,v)\\ \Pi_{\frac{\alpha_2}{n}}(u,v)\\ \Pi_1(u,v)\end{pmatrix}=\begin{pmatrix}E_1(u,v)\\ E_2(u,v)\\ E_3(u,v)\\ E_4(u,v)\end{pmatrix}$$

avec

$$\begin{aligned}
A_1(u) &= \sum_{t=k}^{\alpha_1}\mathbb{1}_u 2\left(\frac{\alpha_1-t}{\alpha_1}\right)^2,\quad B_1(u)=\sum_{t=k}^{\alpha_1}\mathbb{1}_u 2\left(\frac{\alpha_1-t}{\alpha_1}\right)\frac{t}{\alpha_1},\quad C_1(u)=0,\quad D_1(u)=0,\\
E_1(u,v) &= \sum_{t=k}^{\alpha_1}\mathbb{1}_{uv} 2\left(\frac{\alpha_1-t}{\alpha_1}\right),\quad A_2(u)=\sum_{t=k}^{\alpha_1}\mathbb{1}_u 2\left(\frac{\alpha_1-t}{\alpha_1}\right)\frac{t}{\alpha_1},\\
B_2(u) &= \sum_{t=k}^{\alpha_1}\mathbb{1}_u 2\left(\frac{t}{\alpha_1}\right)^2+\sum_{t=\alpha_1+1}^{\alpha_2}\mathbb{1}_u 2\left(\frac{\alpha_2-t}{\alpha_2-\alpha_1}\right)^2,\quad C_2(u)=\sum_{t=\alpha_1+1}^{\alpha_2}\mathbb{1}_u 2\left(\frac{\alpha_2-t}{\alpha_2-\alpha_1}\right)\left(\frac{t-\alpha_1}{\alpha_2-\alpha_1}\right),\\
D_2(u) &= 0,\quad E_2(u,v)=\sum_{t=k}^{\alpha_1}\mathbb{1}_{uv} 2\left(\frac{t}{\alpha_1}\right)+\sum_{t=\alpha_1+1}^{\alpha_2}\mathbb{1}_u 2\left(\frac{\alpha_2-t}{\alpha_2-\alpha_1}\right),\quad A_3(u)=0,\\
B_3(u) &= \sum_{t=\alpha_1+1}^{\alpha_2}\mathbb{1}_u 2\left(\frac{\alpha_2-t}{\alpha_2-\alpha_1}\right)\left(\frac{t-\alpha_1}{\alpha_2-\alpha_1}\right),\quad C_3(u)=\sum_{t=\alpha_1+1}^{\alpha_2}\mathbb{1}_u 2\left(\frac{t-\alpha_1}{\alpha_2-\alpha_1}\right)^2+\sum_{t=\alpha_2+1}^{n}\mathbb{1}_u 2\left(\frac{n-t}{n-\alpha_2}\right)^2,\\
D_3(u) &= \sum_{t=\alpha_2+1}^{n}\mathbb{1}_u 2\left(\frac{n-t}{n-\alpha_2}\right)\left(\frac{t-\alpha_2}{n-\alpha_2}\right),\quad E_3(u,v)=\sum_{t=\alpha_1+1}^{\alpha_2}\mathbb{1}_{uv} 2\left(\frac{t-\alpha_1}{\alpha_2-\alpha_1}\right)+\sum_{t=\alpha_2+1}^{n}\mathbb{1}_{uv} 2\left(\frac{n-t}{n-\alpha_2}\right),\\
A_4(u) &= 0,\quad B_4(u)=0,\quad C_4(u)=\sum_{t=\alpha_2+1}^{n}\mathbb{1}_u 2\left(\frac{n-t}{n-\alpha_2}\right)\left(\frac{t-\alpha_2}{n-\alpha_2}\right),\\
D_4(u) &= \sum_{t=\alpha_2+1}^{n}\mathbb{1}_u 2\left(\frac{t-\alpha_2}{n-\alpha_2}\right)^2,\quad E_4(u,v)=\sum_{t=\alpha_2+1}^{n}\mathbb{1}_{uv} 2\left(\frac{t-\alpha_2}{n-\alpha_2}\right).
\end{aligned}$$

Théorème 24 *Les matrices $\Pi^i_{\frac{t}{n}}$ sont stochastiques pour $i=1,2,3$.*

Preuve *Nous sommons sur $v \in \mathcal{A}$ les équations. Nous obtenons le système suivant :*

$$\begin{pmatrix} A_1(u) & B_1(u) & C_1(u) & D_1(u) \\ A_2(u) & B_2(u) & C_2(u) & D_2(u) \\ A_3(u) & B_3(u) & C_3(u) & D_3(u) \\ A_4(u) & B_4(u) & C_4(u) & D_4(u) \end{pmatrix} \begin{pmatrix} \sum_{v\in\mathcal{A}} \Pi_0(u,v) \\ \sum_{v\in\mathcal{A}} \Pi_{\frac{\alpha_1}{n}}(u,v) \\ \sum_{v\in\mathcal{A}} \Pi_{\frac{\alpha_2}{n}}(u,v) \\ \sum_{v\in\mathcal{A}} \Pi_1(u,v) \end{pmatrix} = \begin{pmatrix} \sum_{v\in\mathcal{A}} E_1(u,v) \\ \sum_{v\in\mathcal{A}} E_2(u,v) \\ \sum_{v\in\mathcal{A}} E_3(u,v) \\ \sum_{v\in\mathcal{A}} E_4(u,v) \end{pmatrix}.$$

Le vecteur $(1 \cdots 1)^t$ est solution. Ainsi on montre que les lignes de ces matrices somment toutes à 1. Par conséquent, la stochasticité des $\Pi^i_{\frac{t}{n}}$ est vérifiée pour $i = 1, 2, 3$. La même remarque que précédemment (voir A.1.2) peut s'appliquer : il est théoriquement possible d'obtenir des termes négatifs.

E.1.2 Degré 2, 2 morceaux

La modélisation par splines polynomiales de degré 2 revient à estimer un DMM de degré 2 sur chaque morceau en imposant la continuité aux noeuds ainsi que la continuité des dérivées premières. L'utilisation du modèle des points d'appuis va rendre les calculs complexes. Aussi, nous utilisons désormais le modèle des polynômes (voir Annexe B). Rappelons que ces deux modèles sont équivalents. Sur chaque segment i, nous utilisons trois matrices de paramètres (voir B). pour définir notre matrice de transition dérivante $\Pi^i_{\frac{t}{n}}$:

$$\begin{aligned} \Pi^1_{\frac{t}{n}}(u,v) &= M^1_0(u,v) + \frac{t}{n} M^1_1(u,v) + \frac{t^2}{n^2} M^1_2(u,v) \\ \Pi^2_{\frac{t}{n}}(u,v) &= M^2_0(u,v) + \frac{t}{n} M^2_1(u,v) + \frac{t^2}{n^2} M^2_2(u,v). \end{aligned}$$

Nous souhaitons minimiser la fonction f_2 suivante sous les contraintes de continuité précitées :

$$\begin{aligned} f_2(M^i_j) &= \sum_{t=k}^{n} \left(\sum_{u\in\mathcal{A}^k} \mathbb{1}_u \left(\sum_{v\in\mathcal{A}} \left(\Pi_{\frac{t}{n}}(u,v) - \mathbb{1}_v \right)^2 \right) \right) \\ &= \sum_{t=k}^{\alpha} \left(\sum_{u\in\mathcal{A}^k} \mathbb{1}_u \left(\sum_{v\in\mathcal{A}} \left(M^1_0 + \frac{t}{n} M^1_1 + \frac{t^2}{n^2} M^1_2 - \mathbb{1}_v \right)^2 \right) \right) \\ &+ \sum_{t=\alpha+1}^{n} \left(\sum_{u\in\mathcal{A}^k} \mathbb{1}_u \left(\sum_{v\in\mathcal{A}} \left(M^2_0 + \frac{t}{n} M^2_1 + \frac{t^2}{n^2} M^2_2 - \mathbb{1}_v \right)^2 \right) \right) \end{aligned}$$

Notons nos deux contraintes :
- $\Pi^1_{\frac{\alpha}{n}}(u,v) = \Pi^2_{\frac{\alpha}{n}}(u,v)$
- $\left(\Pi^1_{\frac{\alpha}{n}}\right)'(u,v) = \left(\Pi^2_{\frac{\alpha}{n}}\right)'(u,v)$

$\Longleftrightarrow$
- $M^1_0(u,v) + \frac{\alpha}{n} M^1_1(u,v) + \frac{\alpha^2}{n^2} M^1_2(u,v) = M^2_0(u,v) + \frac{\alpha}{n} M^2_1(u,v) + \frac{\alpha^2}{n^2} M^2_2(u,v)$
- $M^1_1(u,v) + 2\frac{\alpha}{n} M^1_2(u,v) = M^2_1(u,v) + 2\frac{\alpha}{n} M^2_2(u,v)$

Utilisons la méthode lagrangienne pour la minimisation sous contraintes. Ajoutons à la fonctions à minimiser nos contraintes avec des coefficients λ et γ. La fonction à minimiser devient donc la fonction F_2 suivante :

$$\begin{aligned} F_2(M^i_j, \lambda, \gamma) &= \sum_{t=k}^{\alpha} \left(\sum_{u\in\mathcal{A}^k} \mathbb{1}_u \left(\sum_{v\in\mathcal{A}} \left(M^1_0 + \frac{t}{n} M^1_1 + \frac{t^2}{n^2} M^1_2 - \mathbb{1}_v \right)^2 \right) \right) \\ &+ \sum_{t=\alpha+1}^{n} \left(\sum_{u\in\mathcal{A}^k} \mathbb{1}_u \left(\sum_{v\in\mathcal{A}} \left(M^2_0 + \frac{t}{n} M^2_1 + \frac{t^2}{n^2} M^2_2 - \mathbb{1}_v \right)^2 \right) \right) \\ &+ \lambda \left(M^1_0(u,v) + \frac{\alpha}{n} M^1_1(u,v) + \frac{\alpha^2}{n^2} M^1_2(u,v) - M^2_0(u,v) - \frac{\alpha}{n} M^2_1(u,v) - \frac{\alpha^2}{n^2} M^2_2(u,v) \right) \end{aligned}$$

$$+ \quad \gamma\left(M_1^1(u,v) + 2\frac{\alpha}{n}M_2^1(u,v) - M_1^2(u,v) - 2\frac{\alpha}{n}M_2^2(u,v)\right)$$

Explicitons les dérivées :

$$\begin{aligned}
\frac{\partial F_2(M_j^i,\lambda,\gamma)}{\partial M_0^1(u,v)} &= \sum_{t=k}^{\alpha} \mathbb{1}_u 2\left(M_0^1(u,v) + \frac{t}{n}M_1^1(u,v) + \frac{t^2}{n^2}M_2^1(u,v) - \mathbb{1}_v\right) + \lambda;\\
\frac{\partial F_2(M_j^i,\lambda,\gamma)}{\partial M_1^1(u,v)} &= \sum_{t=k}^{\alpha} \mathbb{1}_u 2\frac{t}{n}\left(M_0^1(u,v) + \frac{t}{n}M_1^1(u,v) + \frac{t^2}{n^2}M_2^1(u,v) - \mathbb{1}_v\right) + \lambda\frac{\alpha}{n} + \gamma;\\
\frac{\partial F_2(M_j^i,\lambda,\gamma)}{\partial M_2^1(u,v)} &= \sum_{t=k}^{\alpha} \mathbb{1}_u 2\frac{t^2}{n^2}\left(M_0^1(u,v) + \frac{t}{n}M_1^1(u,v) + \frac{t^2}{n^2}M_2^1(u,v) - \mathbb{1}_v\right) + \lambda\frac{\alpha^2}{n^2} + 2\gamma\frac{\alpha}{n};\\
\frac{\partial F_2(M_j^i,\lambda,\gamma)}{\partial M_0^2(u,v)} &= \sum_{t=\alpha+1}^{n} \mathbb{1}_u 2\left(M_0^2(u,v) + \frac{t}{n}M_1^2(u,v) + \frac{t^2}{n^2}M_2^2(u,v) - \mathbb{1}_v\right) - \lambda;\\
\frac{\partial F_2(M_j^i,\lambda,\gamma)}{\partial M_1^2(u,v)} &= \sum_{t=\alpha+1}^{n} \mathbb{1}_u 2\frac{t}{n}\left(M_0^2(u,v) + \frac{t}{n}M_1^2(u,v) + \frac{t^2}{n^2}M_2^2(u,v) - \mathbb{1}_v\right) - \lambda\frac{\alpha}{n} - \gamma;\\
\frac{\partial F_2(M_j^i,\lambda,\gamma)}{\partial M_2^2(u,v)} &= \sum_{t=\alpha+1}^{n} \mathbb{1}_u 2\frac{t^2}{n^2}\left(M_0^2(u,v) + \frac{t}{n}M_1^2(u,v) + \frac{t^2}{n^2}M_2^2(u,v) - \mathbb{1}_v\right) - \lambda\frac{\alpha^2}{n^2} - 2\gamma\frac{\alpha}{n};\\
\frac{\partial F_2(M_j^i,\lambda,\gamma)}{\partial \lambda} &= M_0^1(u,v) + \frac{\alpha}{n}M_1^1(u,v) + \frac{\alpha^2}{n^2}M_2^1(u,v) - M_0^2(u,v) - \frac{\alpha}{n}M_1^2(u,v) - \frac{\alpha^2}{n^2}M_2^2(u,v);\\
\frac{\partial F_2(M_j^i,\lambda,\gamma)}{\partial \gamma} &= M_1^1(u,v) + 2\frac{\alpha}{n}M_2^1(u,v) - M_1^2(u,v) - 2\frac{\alpha}{n}M_2^2(u,v).
\end{aligned}$$

Nous obtenons ainsi un système de 8 équations à 8 inconnues :

$$M \times \begin{pmatrix} M_0^1 \\ M_1^1 \\ M_2^1 \\ M_0^2 \\ M_1^2 \\ M_2^2 \\ \lambda \\ \gamma \end{pmatrix} = \begin{pmatrix} 2\sum_{t=k}^{\alpha} \mathbb{1}_{uv} \\ 2\sum_{t=k}^{\alpha} \mathbb{1}_{uv}\frac{t}{n} \\ 2\sum_{t=k}^{\alpha} \mathbb{1}_{uv}\frac{t^2}{n^2} \\ 2\sum_{t=\alpha+1}^{n} \mathbb{1}_{uv} \\ 2\sum_{t=\alpha+1}^{n} \mathbb{1}_{uv}\frac{t}{n} \\ 2\sum_{t=\alpha+1}^{n} \mathbb{1}_{uv}\frac{t^2}{n^2} \\ 0 \\ 0 \end{pmatrix}$$

avec

$$M = \begin{pmatrix}
2\sum_{t=k}^{\alpha}\mathbb{1}_u & 2\sum_{t=k}^{\alpha}\mathbb{1}_u\frac{t}{n} & 2\sum_{t=k}^{\alpha}\mathbb{1}_u\frac{t^2}{n^2} & 0 & 0 & 0 & 1 & 0\\
2\sum_{t=k}^{\alpha}\mathbb{1}_u\frac{t}{n} & 2\sum_{t=k}^{\alpha}\mathbb{1}_u\frac{t^2}{n^2} & 2\sum_{t=k}^{\alpha}\mathbb{1}_u\frac{t^3}{n^3} & 0 & 0 & 0 & \frac{\alpha}{n} & 1\\
2\sum_{t=k}^{\alpha}\mathbb{1}_u\frac{t^2}{n^2} & 2\sum_{t=k}^{\alpha}\mathbb{1}_u\frac{t^3}{n^3} & 2\sum_{t=k}^{\alpha}\mathbb{1}_u\frac{t^4}{n^4} & 0 & 0 & 0 & \frac{\alpha^2}{n^2} & 2\frac{\alpha}{n}\\
0 & 0 & 0 & 2\sum_{t=\alpha+1}^{n}\mathbb{1}_u\frac{t^2}{n^2} & 2\sum_{t=\alpha+1}^{n}\mathbb{1}_u\frac{t^3}{n^3} & 2\sum_{t=\alpha+1}^{n}\mathbb{1}_u\frac{t^4}{n^4} & -1 & 0\\
0 & 0 & 0 & 2\sum_{t=\alpha+1}^{n}\mathbb{1}_u\frac{t^2}{n^2} & 2\sum_{t=\alpha+1}^{n}\mathbb{1}_u\frac{t^3}{n^3} & 2\sum_{t=\alpha+1}^{n}\mathbb{1}_u\frac{t^4}{n^4} & -\frac{\alpha}{n} & -1\\
0 & 0 & 0 & 2\sum_{t=\alpha+1}^{n}\mathbb{1}_u\frac{t^2}{n^2} & 2\sum_{t=\alpha+1}^{n}\mathbb{1}_u\frac{t^3}{n^3} & 2\sum_{t=\alpha+1}^{n}\mathbb{1}_u\frac{t^4}{n^4} & -\frac{\alpha^2}{n^2} & -2\frac{\alpha}{n}\\
1 & \frac{\alpha}{n} & \frac{\alpha^2}{n^2} & -1 & -\frac{\alpha}{n} & -\frac{\alpha^2}{n^2} & 0 & 0\\
0 & 1 & 2\frac{\alpha}{n} & 0 & -1 & -2\frac{\alpha}{n} & 0 & 0
\end{pmatrix}$$

Théorème 25 *Les matrices $\Pi^i_{\frac{t}{n}}$ sont stochastiques pour $i = 1, 2$.*

Preuve *Nous sommons sur $v \in \mathcal{A}$ les équations de notre système. Nous obtenons le système suivant :*

$$M \times \begin{pmatrix}
\sum_{v\in\mathcal{A}} M_0^1(u,v)\\
\sum_{v\in\mathcal{A}} M_1^1(u,v)\\
\sum_{v\in\mathcal{A}} M_2^1(u,v)\\
\sum_{v\in\mathcal{A}} M_0^2(u,v)\\
\sum_{v\in\mathcal{A}} M_1^2(u,v)\\
\sum_{v\in\mathcal{A}} M_2^2(u,v)\\
\lambda\\
\gamma
\end{pmatrix} = \begin{pmatrix}
\sum_{v\in\mathcal{A}} 2\sum_{t=k}^{\alpha}\mathbb{1}_{uv}\\
\sum_{v\in\mathcal{A}} 2\sum_{t=k}^{\alpha}\mathbb{1}_{uv}\frac{t}{n}\\
\sum_{v\in\mathcal{A}} 2\sum_{t=k}^{\alpha}\mathbb{1}_{uv}\frac{t^2}{n^2}\\
\sum_{v\in\mathcal{A}} 2\sum_{t=\alpha+1}^{n}\mathbb{1}_{uv}\\
\sum_{v\in\mathcal{A}} 2\sum_{t=\alpha+1}^{n}\mathbb{1}_{uv}\frac{t}{n}\\
\sum_{v\in\mathcal{A}} 2\sum_{t=\alpha+1}^{n}\mathbb{1}_{uv}\frac{t^2}{n^2}\\
0\\
0
\end{pmatrix}$$

Le vecteur $(\ 1\quad 0\quad 0\quad 1\quad 0\quad 0\quad 0\quad 0\ )^t$ est solution. Ainsi on montre que les lignes de ces matrices somment toutes à 0 sauf pour M_0^i, avec $i = 1, 2$ qui somment à 1. Ainsi la stochasticité de $\Pi^1_{\frac{t}{n}}$ et $\Pi^2_{\frac{t}{n}}$ est vérifiée. La même remarque que précédemment (voir A.1.2) peut s'appliquer : il est théoriquement possible d'obtenir des termes négatifs.

Troisième partie

Approximation de Poisson

Cette troisième partie est composé de trois articles écrits en collaboration avec Miguel Abadi. Les deux premiers présentent la théorie, développée par Miguel Abadi, des temps de retour et de l'approximation de Poisson dans les processus de mélange. Le dernier présente une application que j'ai développée autour de ces deux premiers articles : la recherche de mots exceptionnels.

Chapitre 1

Sharp error terms for return time statistics under mixing conditions

Summary

We describe the statistics of repetition times of a string of symbols in a stochastic process. We consider a string A of length n and prove : 1) The time elapsed until the process starting with A repeats A, denoted by τ_A, has a distribution which can be well approximated by a degenerated law at the origin and an exponential law. 2) The number of consecutive repetitions of A, denoted by S_A, has a distribution which is approximately a geometric law. We provide sharp error terms for each of these approximations. The errors we obtain are point-wise and allow to get also approximations for all the moments of τ_A and S_A. Our results hold for processes that verify the ϕ-mixing condition.

Keywords : Mixing, recurrence, rare event, return time, sojourn time.

1.1 Introduction

This paper describes the statistics of return times of a string of symbols in a mixing stochastic process with a finite alphabet. Generally speaking, the study of the time elapsed until the first occurrence of an event with small probability in dependent processes has a long history which can be traced out in [GS97]. The typical result is :

$$\lim_{n\to\infty} \mathbb{P}\left(\tau_{A_n} > t\ b_n \mid \mu_0\right) = e^{-t}\ , \tag{1.1}$$

where τ_{A_n} is the first time the process hits a given measurable set A_n, $n \in \mathbb{N}$ and such that the measure $\mathbb{P}(A_n)$ go to zero as $n \to \infty$, $\{b_n\}_{n\in\mathbb{N}}$ is a suitable re-scaling sequence of positive numbers and μ_0 is a given initial condition.

Recently and exhaustive analysis of these statistics was motivated for applications in different areas as entropy estimation, genome analysis, computer science, linguistic, among others. From the point of view of applications, a fundamental task is to know the rate of convergence of the above limit. A detailed review of such results appearing in the literature can be find in [AG01].

It is the purpose of this paper to present the following new results : For *any* n-string A
- A sharp upper bound for the above rate of convergence in general ϕ-mixing processes that holds when $\mu_0 = A$.
- A sharp upper bound between the law of the number of *consecutive* visits to A and a geometric law.

When μ_0 is taken as A, we refer to the distribution in (1.1) as the *return time*. In general it can not be well approximated by an exponential law. This was firstly noted in [Hir93], where it is proved the convergence of the number of visits to a small cylinder around a point to the Poisson law for axiom A diffeomorphisms. The result holds for *almost* every point. Then, it is proved that for periodic points, the asymptotic limit law of the return time differs of the one-level Poisson law, namely e^{-t}.

Our first result concerns the rate of convergence of limit in (1.1) when $\mu_0 = A$ for any n-string A. We prove that $\mathbb{P}(\tau_A > t/\mathbb{P}(A) \mid A)$ converges to a mixture of a Dirac law at the origin and an exponential law. Namely, for large n

$$\mathbb{P}\left(\tau_A > \frac{t}{\mathbb{P}(A)} \mid A\right) \approx (1-\zeta)\delta + \zeta e^{-\zeta t}\ .$$

δ is the Dirac measure at the origin. ζ is a parameter related to the self-repeating properties of the string A. It worth noting that the parameter of the exponential law is exactly the weight of the convex combination. So far, the self-repeating properties of a string appears as a major factor to describe the statistical properties of the return time. For instance, if a string admits to overlaps itself, then it will turn out in the sequel that $\zeta \neq 1$ and the return time distribution approximates the above mixture of laws. However, for a word which does not overlap itself, it will turn out that $\zeta = 1$ and the return time distribution approximates a purely exponential law.

It worth recalling at this point that when in equation (1.1) the initial condition is the equilibrium measure of the process, τ_A is called the *hitting time* of A. In [HSV99] it is proved a rate of convergence of the return time as function of the distance between the hitting time and return time laws. While this result applies only for cylinders around periodic points. Our result applies to *all* of them.

The great enhance of our work is that, contrarily to all the previous works which present bounds depending only on the string A, our error estimate decays exponentially fast in t for all $t > 0$. As a byproduct we obtain *explicit expressions* for *all the moments* of the return time. This also appears as a generalization of the famous Kac's lemma (see [Kac47]) which states that the *first* moment of the return time to a n-string A of positive measure is equal to $\mathbb{P}(A)^{-1}$ and the result in [Cha03] which presents conditions for the *existence* of the moments of return times. Further, [HSV99] proves that hitting and return times coincide if and only if the return time converges to the exponential law. We extend this result establishing that the laws of hitting and return time coincide if and only if the Dirac measure of the return time law is absent which is equivalent to consider a non-self-repeating string.

Our framework is the class of ϕ-mixing processes. For instance, irreducible and aperiodic finite state Markov chains are known to be ψ-mixing (and then ϕ-mixing) with exponential decay. Moreover, Gibbs states which have summable variations are ψ-mixing (see [Wal75]). They have exponential decay if they have Hölder continuous potential (see [Bow75]). However, sometimes the ψ-mixing condition is very restricted hypothesis difficult to test. We establish our result under the more general ϕ-mixing condition. Further examples of ϕ-mixing processes can be found in [LSV98]. The error term is explicitly expressed as a function of the mixing rate ϕ. We refer the reader to [Dou95] for a source of examples and definitions of the several kinds of mixing processes.

The base of our proof is a sharp approximation on the rate of convergence of the hitting time to an exponential law proved in [Aba04].

The self-repeating phenomena in the distribution of the return time leads us to consider the problem of the sojourn time. Our second result states that the law of the number of consecutive repetitions of the string A, denoted by S_A, converges to a geometric law. Namely

$$\mathbb{P}\left(S_A = k \mid A\right) \approx (1-\rho)\rho^k \ . \tag{1.2}$$

Again here, the parameter ρ depends on the self-repeating properties of the string. Furthermore we show that under suitable conditions one has $\rho \approx 1-\zeta$. As far as we know, this is the first result on this subject for dependent processes.

As in our previous result, the error bound we obtain decreases geometrically fast in k (see (1.2)). This decay on the error bound allows us to obtain an approximation for *all the moments* of S_A for those of a geometrically distributed random variable.

Our results are applied in a forthcoming paper : In [AHV05] the authors prove large deviations and fluctuations properties of the repetition time function introduced by Wyner and Ziv in [WZ89] and further by Ornstein and Weiss in [OW93], and get entropy estimators.

This paper is organized as follows. In section 2 we establish our framework. In section 3 we describe the self-repeating properties needed to state the return time result. In section 4 we establish the approximation for the return time law. This is Theorem 1.4.1. Finally, in section 5 we state and prove the geometric approximation for the consecutive repetitions of a string. This is Theorem 1.5.1.

1.2 Framework and Notations

Let $\mathcal{E}$ be a finite set. Put $\Omega = \mathcal{E}^{\mathbb{Z}}$. For each $x = (x_m)_{m\in\mathbb{Z}} \in \Omega$ and $m \in \mathbb{Z}$, let $X_m : \Omega \to \mathcal{E}$ be the m-th coordinate projection, that is $X_m(x) = x_m$. We denote by $T : \Omega \to \Omega$ the one-step-left shift operator, namely $(T(x))_m = x_{m+1}$.

We denote by $\mathcal{F}$ the σ-algebra over Ω generated by strings. Moreover we denote by $\mathcal{F}_I$ the σ-algebra generated by strings with coordinates in I, $I \subseteq \mathbb{Z}$.

For a subset $A \subseteq \Omega$ we say that $A \in \mathcal{C}_n$ if and only if

$$A = \{X_0 = a_0; \ldots; X_{n-1} = a_{n-1}\} \ ,$$

with $a_i \in \mathcal{E},\ i = 0, \ldots, n-1$.

We consider an invariant probability measure $\mathbb{P}$ over $\mathcal{F}$. We shall assume without loss of generality that there is no singleton of probability 0.

We say that the process $(X_m)_{m\in\mathbb{Z}}$ is ϕ-mixing if the sequence

$$\phi(l) = \sup |\mathbb{P}_B(C) - \mathbb{P}(C)| \ ,$$

converges to zero. The supremum is taken over B and C such that $B \in \mathcal{F}_{\{0,..,n\}}, n \in \mathbb{N}, \mathbb{P}(B) > 0, C \in \mathcal{F}_{\{m\geq n+l+1\}}$.

For two measurables V and W, we denote as usual $\mathbb{P}(V|W) = \mathbb{P}_W(V) = \mathbb{P}(V;W)/\mathbb{P}(W)$ the conditional measure of V given W. We write $\mathbb{P}(V;W) = \mathbb{P}(V \cap W)$. We also write $V^c = \Omega\backslash V$, for the complement of V.

We use the probabilistic notation : $\{X_n^m = x_n^m\} = \{X_n = x_n, \ldots, X_m = x_m\}$. For a n-string $A = \{X_0^{n-1} = x_0^{n-1}\}$ and $1 \leq w \leq n$, we write $A^{(w)} = \{X_{n-w}^{n-1} = x_{n-w}^{n-1}\}$ for the w-string belonging to the σ-algebra $\mathcal{F}_{\{n-w,\ldots,n-1\}}$ and consisting of the *last* w symbols of A.

The conditional mean of a r.v. X with respect to any measurable V will be denoted by $\mathbb{E}_V(X)$ and we put $\mathbb{E}(X)$ when $V = \Omega$. Wherever it is not ambiguous we will write C for different positive constants even in the same sequence of equalities/inequalities. For brevity we put $(a \vee b) = \max\{a, b\}$ and $(a \wedge b) = \min\{a, b\}$.

1.3 Periodicity

Definition 1.3.1 *Let $A \in \mathcal{C}_n$. We define the* periodicity *of A (with respect to T) as the number $\tau(A)$ defined as follows :*

$$\tau(A) = \min\left\{k \in \{1, \ldots, n\} \mid A \cap\ T^{-k}(A) \neq \emptyset\right\} \ .$$

Let us take $A \in \mathcal{C}_n$, and write $n = q\,\tau(A) + r$, with $q = [n/\tau(A)]$ and $0 \leq r < \tau(A)$. Thus

$$A = \left\{X_0^{\tau(A)-1} = X_{\tau(A)}^{2\tau(A)-1} = \ldots = X_{(q-1)\tau(A)}^{q\tau(A)-1} = a_0^{\tau(A)-1}\ ;\ X_{q\tau(A)}^{n-1} = a_0^{r-1}\right\} \ .$$

So, we say that A has *period* $\tau(A)$ and *rest* r. We remark that periods can be "read backward" (and for the purpose of section 5 it will be more useful to do it in this way), that is

$$\begin{aligned} A &= \left\{X_0^{r-1} = a_{n-r}^{n-1}; X_{n-q\tau(A)}^{n-(q-1)\tau(A)-1} = .. = X_{n-2\tau(A)}^{n-\tau(A)-1} = X_{n-\tau(A)}^{n-1} = a_{n-\tau(A)}^{n-1}\right\} \\ &= \left\{T^{q\tau(A)}A^{(r)}; T^{(q-1)\tau(A)}A^{(\tau(A))}; \ldots; T^{2\tau(A)}A^{(\tau(A))}; T^{\tau(A)}A^{(\tau(A))}; A^{(\tau(A))}\right\}. \end{aligned}$$

We recall the definition of $A^{(w)}, 1 \leq w \leq n$, at the end of section 2. For instance

$$A = (\overbrace{\texttt{aaaabb}}^{\text{period}}\overbrace{\texttt{aaaabb}}^{\text{period}}\overbrace{\texttt{aaa}}^{\text{rest}}) = (\underbrace{\overbrace{\texttt{aaa}}^{\text{rest}}}_{T^{12}A^{(3)}}\underbrace{\overbrace{\texttt{abbaaa}}^{\text{period}}}_{T^{6}A^{(6)}}\underbrace{\overbrace{\texttt{abbaaa}}^{\text{period}}}_{A^{(6)}}) \ . \tag{1.1}$$

In the middle of the above equality, periods are read forward while in the right hand side periods are read backward.

Consider the set of overlapping positions of A :

$$\left\{k \in \{1, \ldots, n-1\} \mid A \cap\ T^{-k}(A) \neq \emptyset\right\} = \{\tau(A), \ldots, [n/\tau(A)]\tau(A)\} \cup \mathcal{R}(A) \ ,$$

where

$$\mathcal{R}(A) = \left\{k \in \{[n/\tau(A)]\tau(A) + 1, \ldots, n-1\} \mid A \cap\ T^{-k}(A) \neq \emptyset\right\} \ .$$

The set $\{\tau(A), \ldots, [n/\tau(A)]\tau(A)\}$ is called the set of principal periods of A while $\mathcal{R}(A)$ is called the set of secondary periods of A. Furthermore, put $r_A = \#\mathcal{R}(A)$. Observe that one has $0 \leq r_A < n/2$. Returns before $\tau(A)$ are not possible, thus, $\mathbb{P}_A(\tau_A < \tau(A)) = 0$. Still, if A does not return at time $\tau(A)$, then it can not return at times $k\tau(A)$, with $1 \leq k \leq [n/\tau(A)]$, so one has

$$\mathbb{P}_A(\tau(A) < \tau_A \leq [n/\tau(A)]\tau(A)) = 0.$$

The first possible return after $\tau(A)$ is

$$n_A = \begin{cases} \min \mathcal{R}(A) & \mathcal{R}(A) \neq \emptyset \\ n_A = n & \mathcal{R}(A) = \emptyset \end{cases} .$$

Furthermore, by definition of $\mathcal{R}(A)$ one has $A \bigcap \mathcal{R}(A)^c = \emptyset$. Thus

$$\mathbb{P}_A \left(\{ [n/\tau(A)]\tau(A) + 1 \leq \tau_A \leq n-1 \} \cap \{ \tau_A \notin \mathcal{R}(A) \} \right) = 0.$$

We finally remark that $\{ T^{-i}A \cap T^{-j}A \mid i,j \in \mathcal{R}(A) \} = \emptyset$. Otherwise it would contradict the fact that the first return to A is $\tau(A)$ since for $i,j \in \mathcal{R}(A)$ one has $|i-j| < \tau(A)$. We conclude that

$$\mathbb{P}_A \left(T^{-i}A \cap T^{-j}A \mid i,j \in \mathcal{R}(A) \right) = 0. \tag{1.2}$$

1.4 Return Times

Given $A \in \mathcal{C}_n$, we define the *hitting time* $\tau_A : \Omega \to \mathbb{N} \cup \{\infty\}$ as the following random variable : For any $x \in \Omega$

$$\tau_A(x) = \inf\{ k \geq 1 : T^k(x) \in A \} .$$

The *return time* is the hitting time restricted to the set A, namely $\tau_A|_A$. We remark the difference between τ_A and $\tau(A)$: while $\tau_A(x)$ is the first time A appears in x, $\tau(A)$ is the first overlapping position of A.

For $A \in \mathcal{C}_n$ define

$$\zeta_A \stackrel{def}{=} \mathbb{P}_A(\tau_A \neq \tau(A)) = \mathbb{P}_A(\tau_A > \tau(A)) .$$

The equality follows by the comment at the end of the previous section.

It would be useful for the reader to note now that according to the comments of the previous section, one has

$$\tau_A|_A \ \in \ \{\tau(A)\} \cup \mathcal{R}(A) \cup \{ k \in \mathbb{N} \mid k \geq n \} . \tag{1.1}$$

Theorem 1.4.1 *Let $(X_m)_{m\in\mathbb{Z}}$ be a ϕ-mixing process. Then, there exist a strictly positive constant C_1 such that for any $A \in \mathcal{C}_n$, the following inequality holds for all t :*

$$\left| \mathbb{P}_A(\tau_A > t) - \mathbb{1}_{\{t<\tau(A)\}} - \mathbb{1}_{\{t\geq\tau(A)\}} \zeta_A e^{-\zeta_A \mathbb{P}(A)(t-\tau(A))} \right| \leq C_1 \epsilon(A) f(A,t), \tag{1.2}$$

where $f(A,t) = \mathbb{P}(A) t e^{-(\zeta_A - \epsilon(A))\mathbb{P}(A)t}$ and

$$\epsilon(A) \stackrel{def}{=} \inf_{0\leq w\leq n_A} \left[(r_A + n)\mathbb{P}(A^{(w)}) + \phi(n_A - w) \right] . \tag{1.3}$$

We postpone an example showing the sharpness of $\epsilon(A)$ after Lemma 1.4.2.

Remark 1.4.1 *$A^{(n_A)}$ is the part of the string A which does not overlap itself in $A \cap T^{-n_A}A$. Note that n_A is the position of the first possible return after $\tau(A)$. Recall that $r_A = \#\mathcal{R}(A)$ and $n_A = n$ if $\mathcal{R}(A) = \emptyset$. Thus $A^{(w)}$ with $1 \leq w \leq n_A$ is the part of the string $A^{(n_A)}$ after taking out its first $n_A - w$ letters (this will be to create a gap of length $n_A - w$ to use the mixing property).*

Remark 1.4.2 *When $\mathcal{R}(A) = \emptyset$, namely, A does not overlaps itself, the error of Theorem 1.4.1 reduces to $\epsilon(A) = \inf_{0\leq w\leq n} \left[n\mathbb{P}(A^{(w)}) + \phi(n-w) \right]$.*

Remark 1.4.3 *In the error term of the theorem, $\epsilon(A)$ provides a bound which shows the convergence uniform in t of the return time law to that mixture of laws as the length of the string growths. The factor $\mathbb{P}(A)t$ provides an extra bound for values of t smaller than $1/\mathbb{P}(A)$. The factor $e^{-(\zeta_A - \epsilon(A)))\mathbb{P}(A)t}$ provides an extra bound for values of t larger than $1/\mathbb{P}(A)$.*

Remark 1.4.4 *On one hand $\mathbb{P}(A) \leq Ce^{-cn}$ (see [Aba01a]). On the other hand, by construction $n_A > n/2$. Further $\phi(n) \to 0$ as $n \to \infty$. Taking for instance $w = n/4$ in (1.3) we warrant the smallness of $\epsilon(A)$ for large enough n.*

Corollary 1.4.1 *Let the process* $(X_m)_{m\in\mathbb{Z}}$ *be* ϕ*-mixing. Let* $\beta > 0$*. Then, for any* $A \in \mathcal{C}_n$*, the* β*-moment of the re-scaled time* $\mathbb{P}(A)\tau_A$ *converges, as* $n \to \infty$*, to* $\Gamma(\beta+1)/\zeta_A^{\beta-1}$*. Moreover*

$$\left| \mathbb{P}(A)^\beta \mathbb{E}_A(\tau_A^\beta) - \frac{\Gamma(\beta+1)}{\zeta_A^{\beta-1}} \right| \leq \epsilon^*(A) \frac{C\beta\ e^{2\epsilon(A)(\beta+1)/\zeta_A}}{\zeta_A^2}\ \frac{\Gamma(\beta+1)}{\zeta_A^{\beta-1}}\ , \tag{1.4}$$

where $\epsilon^*(A) = (\epsilon(A) \vee (n\mathbb{P}(A))^\beta)$*,* $C > 0$ *is a constant and* Γ *is the analytic gamma function.*

Remark 1.4.5 *In particular, the corollary establishes that all the moments of the return time are finite.*

Remark 1.4.6 *In the special case when* $\beta = 1$*, the above corollary establishes a weak version of Kac's Lemma (see [Kac47]).*

Remark 1.4.7 *For each* β *fixed and* n *large enough one has* $\beta\ e^{2\epsilon(A)(\beta+1)}/\zeta_A^2$ *is close to* β/ζ_A^2*. Thus in virtue of inequality (1.4), the corollary reads not just as a difference result but also as a ratio result.*

The next corollary extends Theorem 2.1 in [HSV99].

Corollary 1.4.2 *Let the process* $(X_m)_{m\in\mathbb{Z}}$ *be* ϕ*-mixing. There exists a constant* $C > 0$ *such that, for each* $A \in \mathcal{C}_n$*, the following conditions are equivalent :*

(a) $\left|\mathbb{P}_A(\tau_A > t) - e^{-\mathbb{P}(A)t}\right| \leq C\ \epsilon(A)\ f(A,t)$,
(b) $|\mathbb{P}_A(\tau_A > t) - \mathbb{P}(\tau_A > t)| \leq C\ \epsilon(A)\ f(A,t)$,
(c) $\left|\mathbb{P}(\tau_A > t) - e^{-\mathbb{P}(A)t}\right| \leq C\ \epsilon(A)\ f(A,t)$,
(d) $|\zeta_A - 1| \leq C\ \epsilon(A)$.

Moreover, if $\{A_n\}_{n\in\mathbb{N}}$ *is a sequence of strings such that* $\mathbb{P}(A_n) \to 0$ *as* $n \to \infty$*, then the following conditions are equivalent :*
(ã) the return time law of A_n *converges to a parameter one exponential law,*
(b̃) the return time law and the hitting time law of A_n *converge to the same law,*
(c̃) the hitting time law of A_n *converges to a parameter one exponential law,*
(d̃) The sequence $(\zeta_{A_n})_{n\in\mathbb{N}}$ *converges to one.*

1.4.1 Preparatory results

Here we collect a number of results that will be useful for the proof of Theorem 1.4.1. The next lemma is a useful way to use the ϕ-mixing property.

Lemma 1.4.1 *Let* $(X_m)_{m\in\mathbb{Z}}$ *be a* ϕ*-mixing process. Suppose that* $A \supseteq B \in \mathcal{F}_{\{0,\ldots,b\}}, C \in \mathcal{F}_{\{b+g,\infty\}}$ *with* $b, g \in \mathbb{N}$*. The following inequality holds :*

$$\mathbb{P}_A(B;C) \leq \mathbb{P}_A(B)\left(\mathbb{P}(C) + \phi(g)\right)\ .$$

Proof Since $B \subseteq A$, obviously we have $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(B \cap C)$. By the ϕ-mixing property $\mathbb{P}(B;C) \leq \mathbb{P}(B)\left(\mathbb{P}(C) + \phi(g)\right)$. Dividing the above inequality by $\mathbb{P}(A)$ the lemma follows. □

The following lemma says that returns over $\mathcal{R}(A)$ have small probability.

Lemma 1.4.2 *Let* $(X_m)_{m\in\mathbb{Z}}$ *be a* ϕ*-mixing process. For all* $A \in \mathcal{C}_n$*, the following inequality holds :*

$$\mathbb{P}_A(\tau_A \in \mathcal{R}(A)) \leq \epsilon(A)\ . \tag{1.5}$$

Proof For any w such that $1 \leq w \leq n_A$

$$\begin{aligned}
\mathbb{P}_A(\tau_A \in \mathcal{R}(A)) &= \mathbb{P}_A\left(\bigcup_{j\in\mathcal{R}(A)} T^{-j}A\right) \\
&\leq \mathbb{P}_A\left(\bigcup_{j\in\mathcal{R}(A)} T^{-j}A^{(w)}\right) \\
&\leq r_A\mathbb{P}\left(A^{(w)}\right) + \phi(n_A - w)\ .
\end{aligned} \tag{1.6}$$

The equality follows by (1.2). Since $\{T^{-j}A\} \subset \{T^{-j}A^{(w)}\}$, first inequality follows. Second one follows by the above lemma with $B = A$ and $C = \cup_{j\in\mathcal{R}(A)} T^{-j}A^{(w)}$. This ends the proof since w is arbitrary. □

Example 1.4.1 *Consider a process* $(X_m)_{m\in\mathbb{Z}}$ *defined on the alphabet* $\mathcal{E}=\{a,b\}$. *Consider the string introduced in (1.1) :*

$$A=\{(X_0...X_{14})=(\texttt{aaaabbaaaabbaaa})\}\ .$$

Then, $n=15$, $\tau(A)=6$, $\mathcal{R}(A)=\{13,14\}$, $r_A=2$ *and* $n_A=13$. *Thus*

$$A^{(13)}=((X_2...X_{14})=(\texttt{aabbaaaabbaaa}))\ .$$

The ϕ*-mixing property factorizes the probability*

$$\mathbb{P}_A\left(\bigcup_{j=13}^{14}T^{-j}A\right)=\mathbb{P}_A\left(\bigcup_{j=13}^{14}T^{-j}A^{(13)}\right)\leq\mathbb{P}_A\left(\bigcup_{j=13}^{14}T^{-j}A^{(w)}\right)\ .$$

In such case, a gap at $t=15$ *of length* w *with* $0\leq w\leq 13$ *is the best we can do to apply the* ϕ*-mixing property.*

The next lemma will be used to get the non-uniform factor $f(A,t)$ in the error term of Theorem 1.4.1.

Lemma 1.4.3 *Let* $(X_m)_{m\in\mathbb{Z}}$ *be a* ϕ*-mixing process. Let* $B\in\mathcal{F}_{\{kf,\infty\}}$, *with* $k\in\mathbb{N}$ *and* $f>n$. *Then, for all* $n<g<f$, *the following inequality holds :*

$$\mathbb{P}_A\left(\tau_A>kf\ ;\ B\right)\leq\left[\mathbb{P}\left(\tau_A>f-g\right)+\phi(g)\right]^{k-1}\left[\mathbb{P}\left(B\right)+\phi(g)\right]\ .$$

Proof First introduce a gap of length g, then use Lemma 1.4.1 to get the inequalities

$$\begin{aligned}\mathbb{P}_A\left(\tau_A>kf\ ;\ B\right)&\leq\mathbb{P}_A\left(\tau_A>kf-g\ ;\ B\right)\\&\leq\mathbb{P}_A\left(\tau_A>kf-g\right)\left[\mathbb{P}\left(B\right)+\phi(g)\right]\ .\end{aligned}\tag{1.7}$$

Apply the above procedure to $\{\tau_A>(k-1)f\}$ and $B=\left\{\tau_A\circ T^{(k-1)f}>f-g\right\}$ to bound $\mathbb{P}_A\left(\tau_A>kf-g\right)$ by

$$\mathbb{P}_A\left(\tau_A>(k-1)f-g\right)\left[\mathbb{P}\left(\tau_A>f-g\right)+\phi(g)\right]\ .$$

Iterate this procedure to bound $\mathbb{P}_A\left(\tau_A>kf-g\right)$ by

$$\mathbb{P}_A\left(\tau_A>f-g\right)\left[\mathbb{P}\left(\tau_A>f-g\right)+\phi(g)\right]^{k-1}\leq\left[\mathbb{P}\left(\tau_A>f-g\right)+\phi(g)\right]^{k-1}.$$

This ends the proof of the Lemma. □

The next proposition establishes a relationship between *hitting* and *return* times with an error *uniform* in t. In particular, (b) says that they coincide if and only if $\zeta_A=1$, namely, the string A is non-self-repeating.

Proposition 1.4.1 *Let* $(X_m)_{m\in\mathbb{Z}}$ *be a* ϕ*-mixing processes. Let* $A\in\mathcal{C}_n$. *Then the following holds :*

(a) For all $M,M'\geq g\geq n$,

$$\begin{aligned}&\left|\mathbb{P}_A\left(\tau_A>M+M'\right)-\mathbb{P}_A\left(\tau_A>M\right)\mathbb{P}\left(\tau_A>M'\right)\right|\\\leq\ &\mathbb{P}_A\left(\tau_A>M-g\right)2\left[g\mathbb{P}(A)+\phi(g)\right],\end{aligned}$$

and similarly

$$\begin{aligned}&\left|\mathbb{P}_A\left(\tau_A>M+M'\right)-\mathbb{P}_A\left(\tau_A>M\right)\mathbb{P}\left(\tau_A>M'-g\right)\right|\\\leq\ &\mathbb{P}_A\left(\tau_A>M-g\right)\left[g\mathbb{P}(A)+2\phi(g)\right].\end{aligned}$$

(b) For all $t\geq\tau(A)\in\mathbb{N}$,

$$\left|\mathbb{P}_A\left(\tau_A>t\right)-\zeta_A\mathbb{P}\left(\tau_A>t\right)\right|\leq 2\epsilon(A)\ .\tag{1.8}$$

Proof To simplify notation, for $t\in\mathbb{Z}$ we write $\tau_A^{[t]}$ to mean $\tau_A\circ T^t$. We introduce a gap of length g after coordinate M to construct the following triangular inequality

$$\begin{aligned}&\left|\mathbb{P}_A\left(\tau_A>M+M'\right)-\mathbb{P}_A\left(\tau_A>M\right)\mathbb{P}\left(\tau_A>M'\right)\right|\\\leq\ &\left|\mathbb{P}_A\left(\tau_A>M+M'\right)-\mathbb{P}_A\left(\tau_A>M;\tau_A^{[M+g]}>M'-g\right)\right|&(1.9)\\+\ &\left|\mathbb{P}_A\left(\tau_A>M;\tau_A^{[M+g]}>M'-g\right)-\mathbb{P}_A\left(\tau_A>M\right)\mathbb{P}\left(\tau_A>M'-g\right)\right|&(1.10)\\+\ &\mathbb{P}_A\left(\tau_A>M\right)\left|\mathbb{P}\left(\tau_A>M'-g\right)-\mathbb{P}\left(\tau_A>M'\right)\right|\ .&(1.11)\end{aligned}$$

Term (1.9) is bounded as (1.7) by

$$\mathbb{P}_A\left(\tau_A > M; \tau_A^{[M]} \leq g\right) \leq \mathbb{P}_A\left(\tau_A > M - g\right)\left[g\mathbb{P}(A) + \phi(g)\right] \ .$$

Term (1.10) is bounded using the ϕ-mixing property by $\mathbb{P}_A\left(\tau_A > M\right)\phi(g)$. The modulus in (1.11) is bounded using stationarity by $\mathbb{P}\left(\tau_A \leq g\right) \leq g\mathbb{P}(A)$. This ends the proof of both inequalities of item (a).

Item (b) for $t \geq 2n$ is proved applying item (a) with $M = n$ and $M' = t - n$. Then, by stationarity $\mathbb{P}\left(\tau_A > t - n\right) - \mathbb{P}\left(\tau_A > t\right) \leq n\mathbb{P}(A)$. Further, $\mathbb{P}_A\left(\tau_A > \tau(A)\right) - \mathbb{P}_A\left(\tau_A > n\right) \leq \epsilon(A)$ by Lemma 1.4.2.

Consider now $\tau(A) \leq t < 2n$. Take any $1 \leq w \leq n_A$.

$$\begin{aligned}
\zeta_A - \mathbb{P}_A\left(\tau_A > t\right) &= \mathbb{P}_A\left(\tau(A) < \tau_A \leq t\right) \\
&= \mathbb{P}_A\left(\tau_A \in \mathcal{R}(A) \cup \left(n \leq \tau_A < 2n\right)\right) \\
&\leq (r_A + n)\mathbb{P}\left(A^{(w)}\right) + \phi(n_A - w) \ .
\end{aligned}$$

First and second equalities follow by the considerations of section 3. The inequality follows similarly to (1.6). □

The following two propositions are the key of the proof of Theorem 1.4.1.

Proposition 1.4.2 *Let $(X_m)_{m\in\mathbb{Z}}$ be a ϕ-mixing process. Let $n < g < f$. Then the following inequality holds :*

$$\begin{aligned}
&\left|\mathbb{P}_A\left(\tau_A > kf\right) - \mathbb{P}_A\left(\tau_A > f\right)\mathbb{P}\left(\tau_A > f - g\right)^{k-1}\right| \\
\leq\ & 2\left[g\mathbb{P}(A) + \phi(g)\right](k-1)\left(\mathbb{P}\left(\tau_A > f - g\right) + \phi(g)\right)^{k-2} \ .
\end{aligned}$$

Proof The left hand side of the above inequality is bounded by

$$\sum_{j=2}^{k} \left|\mathbb{P}_A\left(\tau_A > jf\right) - \mathbb{P}_A\left(\tau_A > (j-1)f\right)\mathbb{P}\left(\tau_A > f - g\right)\right| \mathbb{P}\left(\tau_A > f - g\right)^{k-j} .$$

The modulus in the above sum is bounded by

$$2\left[g\mathbb{P}(A) + \phi(g)\right]\mathbb{P}_A\left(\tau_A > (j-1)f - g\right) \ ,$$

due to Proposition 1.4.1 (a). The right-most factor is bounded using Lemma 1.4.3 by $\left[\mathbb{P}\left(\tau_A > f - g\right) + \phi(g)\right]^{j-2}$. The conclusion follows. □

Let us define

$$\delta(A) = \inf_{n \leq y \leq 1/\mathbb{P}(A)} \left(y\mathbb{P}(A) + \phi(y)\right) \ .$$

In the proof of Theorem 1.4.1 we will make use of the following version of Proposition 1.4.2 proved in [Aba04] for *hitting* times instead of *return* times as was done in Proposition 1.4.2. We quote it here for easy reference.

Proposition 1.4.3 (Abadi, 2004) *Let $(X_m)_{m\in\mathbb{Z}}$ be a ϕ-mixing process. There exists a finite constant $C > 0$, such that for any $f \in (4n, 1/(2\mathbb{P}(A))]$ such that*

$$\phi(f/4) \leq \mathbb{P}\left(\tau_A \leq f/4 \ ; \ \tau_A^{[f/4]} > f/2\right) \ ,$$

there exists a $\Delta = \Delta(f) > 0$, with $n < \Delta \leq f/4$, such that for all positive integers k, the following inequalities hold :

$$\left|\mathbb{P}\left(\tau_A > kf\right) - \mathbb{P}\left(\tau_A > f - 2\Delta\right)^k\right| \leq C\delta(A)k\mathbb{P}\left(\tau_A > f - 2\Delta\right)^k \ ,$$

and

$$\left|\mathbb{P}\left(\tau_A > kf\right) - \mathbb{P}\left(\tau_A > f\right)^k\right| \leq C\delta(A)k\mathbb{P}\left(\tau_A > f - 2\Delta\right)^k \ .$$

1.4.2 Proofs of Theorem 1.4.1 and corollaries

Proof of Theorem 1.4.1 We divide the proof according the different values of t : (i) $t < n$, (ii) $n \leq t \leq 1/(2\mathbb{P}(A))$ and (ii) $t > 1/(2\mathbb{P}(A))$. (Factor 2 is rather technical.)

Consider first $t < n$. If $t \leq \tau(A)$, (1.1) says that the left hand side of (1.2) is zero. If $\tau(A) < t \leq [n/\tau(A)]\tau(A)$, (1.1) also implies that the left hand side of equation (1.2) is $\zeta_A - \zeta_A e^{-\zeta_A \mathbb{P}(A)(t-\tau(A))} \leq \mathbb{P}(A)n$. If $[n/\tau(A)]\tau(A) < t < n$ it follows by (1.1) and Lemma 1.4.2 that the left hand side of (1.2) is bounded by $\epsilon(A)$.

Consider now $n \leq t \leq 1/(2\mathbb{P}(A))$. First write

$$\mathbb{P}_A(\tau_A > t) = \frac{\mathbb{P}_A(\tau_A > t)}{\mathbb{P}(\tau_A > t)} \mathbb{P}(\tau_A > t) = \rho_{t+1} \mathbb{P}(\tau_A > t) \ ,$$

and

$$\begin{aligned} \mathbb{P}(\tau_A > t) &= \prod_{i=\tau(A)+1}^{t} \mathbb{P}(\tau_A > i | \tau_A > i-1) \\ &= \prod_{i=\tau(A)+1}^{t} (1 - \mathbb{P}(T^{-i}A | \tau_A > i-1)) \\ &= \prod_{i=\tau(A)+1}^{t} (1 - \rho_i \mathbb{P}(A)) \ , \end{aligned}$$

where

$$\rho_i \stackrel{def}{=} \frac{\mathbb{P}_A(\tau_A > i-1)}{\mathbb{P}(\tau_A > i-1)} \ .$$

Further

$$\left| 1 - \rho_i \mathbb{P}(A) - e^{-\zeta_A \mathbb{P}(A)} \right| \leq |\rho_i - \zeta_A| \, \mathbb{P}(A) + \left| 1 - \zeta_A \mathbb{P}(A) - e^{-\zeta_A \mathbb{P}(A)} \right| \ . \tag{1.12}$$

Firstly, by Proposition 1.4.1 (b) and the fact that $\mathbb{P}(\tau_A > i) \geq 1/2$ since $i \leq 1/(2\mathbb{P}(A))$ we have

$$|\rho_i - \zeta_A| \leq \frac{2\epsilon(A)}{\mathbb{P}(\tau_A > i)} \leq 4\epsilon(A) \ ,$$

for all $i = \tau(A)+1, \ldots, 1/(2\mathbb{P}(A))$. Secondly, note that $|1 - x - e^{-x}| \leq x^2/2$ for $x > 0$ small enough. Apply it with $x = \zeta_A \mathbb{P}(A)$ to bound the most right term of (1.12) by $(\zeta_A \mathbb{P}(A))^2/2$. Further, since $|\prod a_i - \prod b_i| \leq (\#i) \max |a_i - b_i|$ for $0 \leq a_i, b_i \leq 1$, we conclude that $|\mathbb{P}(\tau_A > t) - e^{-\zeta_A \mathbb{P}(A)t}|$ and therefore the left hand side of (1.2) are both bounded by

$$(t - \tau(A)) \left(4\epsilon(A)\mathbb{P}(A) + \frac{\mathbb{P}(A)^2}{2} \right) \leq C\epsilon(A)\mathbb{P}(A)t \ e^{-\zeta_A \mathbb{P}(A)t} \ , \tag{1.13}$$

for all $\tau_A \leq t \leq 1/(2\mathbb{P}(A))$. The inequality follows since $e^{-\zeta_A \mathbb{P}(A)t} \geq e^{-1/2}$.

Finally, consider $t > 1/(2\mathbb{P}(A))$. The proof has two steps. First we prove for t of the form $t = (k + p/q) \ f$ with $f = 1/(2\mathbb{P}(A))$, $k \in \mathbb{N}$, p a positive integer and $1 \leq p \leq q$ with $q := 1/(2\delta(A))$. The basic tools are the Mean Value Theorem (MVT) and the ϕ-mixing property. Then we prove for the remaining t's. Basically we approximate such a t by one of the form $(k + p/q) \ f$.

Proof : t's of the form $t = \left(k + \frac{p}{q}\right) \ f$.

Let $t = (k + (p/q)) \ f$, with k, p, q and f as was just told. For brevity put $r = (p/q) \ f$. Let Δ be the one given by Proposition 1.4.3. Then

$$\begin{aligned} & \left| \mathbb{P}_A(\tau_A > t) - \zeta_A e^{-\zeta_A \mathbb{P}(A)t} \right| \\ = \ & \left| \mathbb{P}_A(\tau_A > kf + r) - \zeta_A e^{-(\zeta_A/2)t/f} \right| \\ \leq \ & |\mathbb{P}_A(\tau_A > kf + r) - \mathbb{P}_A(\tau_A > kf) \, \mathbb{P}(\tau_A > r)| \end{aligned}$$

$$
\begin{aligned}
&+ \left|\mathbb{P}_A\left(\tau_A > kf\right) - \mathbb{P}_A\left(\tau_A > f\right)\mathbb{P}\left(\tau_A > f-\Delta\right)^{k-1}\right| \mathbb{P}\left(\tau_A > r\right)\\
&+ \left|\mathbb{P}\left(\tau_A > f-\Delta\right)^{k-1} - \mathbb{P}\left(\tau_A > f-2\Delta\right)^{k-1}\right| \mathbb{P}_A\left(\tau_A > f\right)\mathbb{P}\left(\tau_A > r\right)\\
&+ \left|\mathbb{P}_A\left(\tau_A > f\right) - \zeta_A\mathbb{P}\left(\tau_A > f-2\Delta\right)\right| \mathbb{P}\left(\tau_A > f-2\Delta\right)^{k-1}\mathbb{P}\left(\tau_A > r\right)\\
&+ \left|\mathbb{P}\left(\tau_A > r\right) - \mathbb{P}\left(\tau_A > f-2\Delta\right)^{r/f}\right| \zeta_A\mathbb{P}\left(\tau_A > f-2\Delta\right)^{k}\\
&+ \left|\mathbb{P}\left(\tau_A > f-2\Delta\right)^{t/f} - e^{-(\zeta_A/2)t/f}\right| \zeta_A\ .
\end{aligned}
$$

The first term on the right hand side of the above inequality is bounded using first Proposition 1.4.1 (a) with $M = kf$, $M' = r$ and $g = \Delta$ and then Lemma 1.4.3 with $B = \{\tau_A > f - g\}$ by

$$2\left(\Delta\mathbb{P}(A) + \phi(\Delta)\right)\left(\mathbb{P}\left(\tau_A > f-\Delta\right) + \phi(\Delta)\right)^{k-1}\ .$$

The modulus in the second one is bounded using Proposition 1.4.2 by

$$2\left(\Delta\mathbb{P}(A) + \phi(\Delta)\right)(k-1)\left(\mathbb{P}\left(\tau_A > f-\Delta\right) + \phi(\Delta)\right)^{k-2}\ .$$

The modulus in the third one is bounded using the MVT by

$$\Delta\mathbb{P}(A)(k-1)\mathbb{P}\left(\tau_A > f-2\Delta\right)^{k-2}\ .$$

The modulus in the fourth one is bounded using Proposition 1.4.1 (b) by $2\epsilon(A)$. The modulus in the fifth one is bounded by

$$C\delta(A)\mathbb{P}\left(\tau_A > f-2\Delta\right)^{[k+(p/q)]/2}\ ,$$

as shown in the proof of Theorem 1 of [Aba04] (see p. 254). The modulus in the sixth one is bounded using the MVT and (1.13) with $t = f - 2\Delta$ by

$$\epsilon(A)\frac{t}{f}\left(\mathbb{P}\left(\tau_A > f-2\Delta\right) \vee e^{-(\zeta_A/2)}\right)^{(t/f)-1} \leq C\epsilon(A)\mathbb{P}(A)te^{-(\zeta_A-\epsilon(A))\mathbb{P}(A)t}\ .$$

Proof : A general t.

Now, let t be any positive real. We write $t = kf + r$, with k a positive integer and r such that $0 \leq r < f$. We can choose a $\bar{t}$ such that $\bar{t} > t$ and $\bar{t} = (k + (p/q))\ f$ with p, q as before. Then

$$
\begin{aligned}
&\left|\mathbb{P}_A\left(\tau_A > t\right) - \zeta_A e^{-\zeta_A\mathbb{P}(A)t}\right|\\
\leq\ &\left|\mathbb{P}_A\left(\tau_A > t\right) - \mathbb{P}_A\left(\tau_A > \bar{t}\ \right)\right|\\
+\ &\left|\mathbb{P}_A\left(\tau_A > \bar{t}\right) - \zeta_A\mathbb{P}\left(\tau_A > f-2\Delta\right)^{[k+(p/q)]/2}\right|\\
+\ &\zeta_A\left|\mathbb{P}\left(\tau_A > f-2\Delta\right)^{[k+(p/q)]/2} - e^{-\zeta_A\mathbb{P}(A)\bar{t}}\right|\\
+\ &\zeta_A\left|e^{-\zeta_A\mathbb{P}(A)\bar{t}} - e^{-\zeta_A\mathbb{P}(A)t}\right|\ .
\end{aligned}
$$

The first term on the right hand side of the above inequality is bounded applying Lemma 1.4.3

$$
\begin{aligned}
&\left|\mathbb{P}_A\left(\tau_A > t\right) - \mathbb{P}_A\left(\tau_A > \bar{t}\ \right)\right|\\
=\ &\mathbb{P}_A\left(\tau_A > t\ ;\ \tau_A^{[\bar{t}]} \leq \bar{t} - t\right)\\
\leq\ &\mathbb{P}_A\left(\tau_A > (k-1)f\ ;\ \tau_A^{[\bar{t}]} \leq \Delta\right)\\
\leq\ &\left(\mathbb{P}\left(\tau_A > f-\Delta\right) + \phi(\Delta)\right)^{k-2}\left(\Delta\mathbb{P}(A) + \phi(\Delta)\right)\ .
\end{aligned}
$$

For the third term, first note that $e^{-\zeta_A\mathbb{P}(A)\bar{t}} = e^{-\zeta_A[k+(p/q)]/2}$. Yet, by stationarity and (1.13)

$$\left|\mathbb{P}\left(\tau_A > f-2\Delta\right) - e^{-\zeta_A}\right| \leq C\epsilon(A)\ .$$

Therefore, the MVT implies that the third term is bounded by

$$
\begin{aligned}
&C\epsilon(A)\ \mathbb{P}(A)\bar{t}\ \left(\mathbb{P}\left(\tau_A > f-2\Delta\right) \vee e^{-\zeta_A}\right)^{\mathbb{P}(A)\bar{t}-1}\\
\leq\ &C\epsilon(A)\ \mathbb{P}(A)t\ e^{-(\zeta_A-\epsilon(A))\mathbb{P}(A)t}\ .
\end{aligned}
$$

The upper bound for the fourth term is obtained similarly by the MVT and the fact that $|t-\bar{t}| \leq \Delta$. Finally, the second term is bounded as in the first part of the proof. To end the proof we notice that

$$\mathbb{P}\left(\tau_A > f - \Delta\right) \leq \mathbb{P}\left(\tau_A > f - \Delta\right) + \phi(\Delta) = \mathbb{P}\left(\tau_A > f - 2\Delta\right) \ .$$

The equality follows since $\phi(\Delta) = \mathbb{P}\left(\tau_A \leq \Delta; \tau_A^{[\Delta]} > f - 2\Delta\right)$ (see [Aba04] p. 250.) Therefore

$$\mathbb{P}\left(\tau_A > f - 2\Delta\right)^{k-2} \leq Ce^{-(\zeta_A - \epsilon(A))\mathbb{P}(A)t} \ .$$

This ends the proof of the theorem. □

Proof of Corollary 1.4.1 Rewrite (1.2) as

$$\begin{aligned} &\left|\mathbb{P}_A(\mathbb{P}(A)\tau_A > t) - \mathbb{1}_{\{t<\mathbb{P}(A)\tau(A)\}} - \mathbb{1}_{\{t\geq\mathbb{P}(A)\tau(A)\}}\zeta_A e^{-\zeta_A(t-\mathbb{P}(A)\tau(A))}\right| \\ &\leq C_1\epsilon(A)f(A, t/\mathbb{P}(A)) \ . \end{aligned} \tag{1.14}$$

Let $Y = Y_1 + Y_2$ where

$$\mathbb{P}(Y_1 > t) = \mathbb{1}_{\{t<\mathbb{P}(A)\tau(A)\}} \ ,$$

and

$$\mathbb{P}(Y_2 > t) = \mathbb{1}_{\{t\geq\mathbb{P}(A)\tau(A)\}}\zeta_A e^{-\zeta_A(t-\mathbb{P}(A)\tau(A))} \ .$$

Integrating (1.14) we get

$$\begin{aligned} & \left|\mathbb{E}\left((\mathbb{P}(A)\tau_A)^\beta\right) - \mathbb{E}\left(Y^\beta\right)\right| \\ = & \left|\int_{\mathbb{P}(A)}^\infty \beta t^{\beta-1}\left(\mathbb{P}\left(\mathbb{P}(A)\tau_A > t\right) - \mathbb{P}\left(Y > t\right)\right)\right| \\ \leq & \int_{\mathbb{P}(A)}^\infty \beta t^{\beta-1}\left|\mathbb{P}\left(\mathbb{P}(A)\tau_A > t\right) - \mathbb{P}\left(Y > t\right)\right| \\ \leq & \ C_1\epsilon(A)\int_{\mathbb{P}(A)}^\infty \beta t^{\beta-1}f(A, t/\mathbb{P}(A))dt \ . \end{aligned}$$

Now we proside to compute $\mathbb{E}\left(Y^\beta\right) = \int_{\mathbb{P}(A)}^\infty \beta t^{\beta-1}\mathbb{P}(Y > t)$.

Since $\mathbb{P}(Y_1 > t)$ and $\mathbb{P}(Y_2 > t)$ have disjoint support, one has $\mathbb{E}(Y^\beta) = \mathbb{E}(Y_1^\beta) + \mathbb{E}(Y_2^\beta)$. On one hand $\mathbb{E}(Y_1^\beta) = (\mathbb{P}(A)\tau(A))^\beta$. On the other hand

$$\begin{aligned} \mathbb{E}(Y_2^\beta) &= \int_{\mathbb{P}(A)\tau_A}^\infty \beta t^{\beta-1}\zeta_A e^{-\zeta_A(t-\mathbb{P}(A)\tau_A)}dt \\ &= \zeta_A \ e^{\zeta_A\mathbb{P}(A)\tau_A}\left(\int_0^\infty - \int_0^{\mathbb{P}(A)\tau_A}\right)\beta t^{\beta-1}e^{-\zeta_A t}dt \ . \end{aligned}$$

Yet, since $\zeta_A\mathbb{P}(A)\tau_A \leq \mathbb{P}(A)n$ and $\mathbb{P}(A)$ decays exponentially fast on n we have $e^{\zeta_A\mathbb{P}(A)\tau_A} - 1 \leq C\mathbb{P}(A)n$. Further, the first integral is $\Gamma(\beta+1)/\zeta_A^\beta$. The second one is bounded by $(\mathbb{P}(A)\tau_A)^\beta$. We conclude that

$$\left|\mathbb{E}\left(Y^\beta\right) - \mathbb{E}\left(Y_2^\beta\right)\right| \leq Cn\mathbb{P}(A) + 2(n\mathbb{P}(A))^\beta) \leq C(n\mathbb{P}(A))^{(\beta\wedge 1)} \ .$$

Similar computations give

$$\begin{aligned} \int_{\mathbb{P}(A)}^\infty \beta t^{\beta-1}f\left(A, t/\mathbb{P}(A)\right)dt &\leq \frac{\beta}{\beta+1}\frac{\Gamma(\beta+2)}{(\zeta_A - \epsilon(A))^{\beta+1}} \\ &\leq \frac{\beta e^{2\epsilon(A)(\beta+1)/\zeta_A}}{\zeta_A^2}\frac{\Gamma(\beta+1)}{\zeta_A^{\beta-1}} \ . \end{aligned}$$

In the last inequality we used $x \leq 2(1-e^{-x})$ for small enough $x > 0$. This ends the proof of the corollary. □

Proof of Corollary 1.4.2. $(a) \Leftrightarrow (d)$. It follows directly from Theorem 1.4.1.

$(b) \Rightarrow (a), (c)$. It follows by Theorem 1.4.1 and Theorem 1 in [Aba04]

$(a) \Rightarrow (b)$ and $(c) \Rightarrow (b)$. They follow by Theorem 1.4.1, Theorem 1 in [Aba04] and (1.13). The corollary is proved. □

1.5 Sojourn Time

In this section we consider the number of consecutive visits to a fixed string A and prove that the distribution law of this number can be well approximated by a geometric law.

Definition 1.5.1 *Let $A \in \mathcal{C}_n$. We define the* sojourn time *on the set A as the r.v. $S_A : \Omega \to \mathbb{N} \cup \{\infty\}$*

$$S_A(x) = \sup\left\{k \in \mathbb{N} \mid x \in A \cap T^{-j\tau(A)}A \ ; \ \forall j = 1, \ldots, k\right\} \ ,$$

and $S_A(x) = 0$ if the supremum is taken over the empty set.

Before to state our main result we have to introduce the following definition about certain continuity property of the probability $\mathbb{P}$ conditioned to i occurrences of the string A.

Definition 1.5.2 *Given $A \in \mathcal{C}_n$, we define the sequence of probabilities $(p_i(A))_{i\in\mathbb{N}}$ as follows :*

$$p_i(A) \stackrel{def}{=} \mathbb{P}\left(A \mid \bigcap_{j=1}^{i} T^{j\tau(A)}A\right) \ .$$

If the limit $\lim_{n\to\infty} p_i(A)$ exists then we denote it by $\rho(A)$.

Remark 1.5.1 *By stationarity $p_1(A) = 1 - \zeta_A$.*

In the following examples, the sequence $(p_i(A))_{i\in\mathbb{N}}$ not just converges but even is constant.

Example 1.5.1 *For a i.i.d. Bernoulli process with parameter $0 < \theta = \mathbb{P}(X_i = 1) = 1 - \mathbb{P}(X_i = 0)$, and for the n-string $A = \{X_0^{n-1} = 1\}$, we have that $p_i(A) = 1 - \zeta_A = \theta$ for all $i \in \mathbb{N}$.*

Example 1.5.2 *Let $(X_m)_{m\in\mathbb{Z}}$ be a irreducible and aperiodic finite state Markov chain. For $A = \{X_0^{n-1} = a_0^{n-1}\} \in \mathcal{C}_n$, the sequence $(p_i(A))_{i\in\mathbb{N}}$ is constant. More precisely, by the Markovian property and for all $i \in \mathbb{N}$*

$$\begin{aligned} p_i(A) &= \mathbb{P}\left(X_{n-\tau(A)}^{n-1} = a_{n-\tau(A)}^{n-1} | X_{\tau(A)-1} = a_{\tau(A)-1}\right) \\ &= \prod_{j=\tau(A)}^{n-1} \mathbb{P}\left(X_j = a_j | X_{j-1} = a_{j-1}\right) \ . \end{aligned}$$

In the following theorem we assume that $(p_i(A))_{i\in\mathbb{N}}$ converges with velocity $d_i = d_i(A)$. Namely, there is a real number $\rho(A) \in (0,1)$ such that

$$|p_i(A) - \rho(A)| \leq d_i \qquad \text{for all } i \in \mathbb{N}, \tag{1.1}$$

where d_i is a sequence decreasing to zero.

Theorem 1.5.1 *Let $(X_m)_{m\in\mathbb{Z}}$ be a stationary process. Let $A \in \mathcal{C}_n$. Assume that (1.1) holds. Then, there is a constant $c \in [0,1)$, such that the following inequalities hold for all $k \in \mathbb{N}$:*

$$\left|\mathbb{P}_A\left(S_A = k\right) - (1 - \rho(A))\ \rho(A)^k\right| \leq c^k \sum_{i=1}^{k+1} d_i \leq c^k(k+1)d_1 \ .$$

We deduce immediately that the β-moments of S_A can be approximated by $\mathbb{E}(Y^\beta)$ where Y is a geometric random variable with parameter $\rho(A)$.

Corollary 1.5.1 *Let Y be a r.v. with geometric distribution with parameter $\rho(A)$. Let β a positive integer. Then*

$$\left|\mathbb{E}_A\left(S_A^\beta\right) - \mathbb{E}(Y^\beta)\right| \leq C_\beta d_1 \ ,$$

where C_β is a constant that just depends on β.

In the proof of Theorem 1.5.1 we will use the following lemma.

Lemma 1.5.1 *Let $(l_i)_{i\in\mathbb{N}}$ be a sequence of real numbers such that $0 < l_i < 1$, for all $i \in \mathbb{N}$. Let $0 \leq l < 1$ be such that $|l_i - l| \leq d_i$ for all $i \in \mathbb{N}$ with $d_i \to 0$. Then, there is a constant $c \in [0,1)$, such that the following inequalities hold for all $k \in \mathbb{N}$:*

$$\left|\prod_{i=1}^{k} l_i - l^k\right| \leq c^{k-1} \sum_{i=1}^{k} d_i \leq k\ c^{k-1} d_1\ .$$

Proof

$$\begin{aligned}
\left|\prod_{i=1}^{k} l_i - l^k\right| &= \left|\prod_{i=1}^{k} l_i - \prod_{i=1}^{k-1} l_i l + \prod_{i=1}^{k-1} l_i l - \prod_{i=1}^{k-2} l_i l^2 + \prod_{i=1}^{k-2} l_i l^2 - \ldots - l^k\right| \\
&\leq \sum_{i=1}^{k} \left(\prod_{j=1}^{k-i} l_j\right) |l_{k-i+1} - l|\ l^{i-1} \leq c^{k-1} \sum_{i=1}^{k} d_i \\
&\leq k\ c^{k-1} d_1\ ,
\end{aligned}$$

where $c = \max(l_0, l)$. □

Proof of Theorem 1.5.1 For $k = 0$, we just note that $\mathbb{P}_A(S_A = 0) = 1 - p_1(A)$ and $|1 - p_1(A) - (1 - \rho(A))| \leq d_1$. Suppose $k \geq 1$. Therefore

$$\begin{aligned}
& \mathbb{P}_A(S_A = k) \\
= \ & \mathbb{P}_A\left(\bigcap_{j=0}^{k} T^{-j\tau(A)} A \ ;\ T^{-(k+1)\tau(A)} A^c\right) \\
= \ & \mathbb{P}\left(T^{-(k+1)\tau(A)} A^c \middle| \bigcap_{j=0}^{k} T^{-j\tau(A)} A\right) \prod_{i=1}^{k} \mathbb{P}\left(T^{-i\tau(A)} A \middle| \bigcap_{j=0}^{i-1} T^{-j\tau(A)} A\right) \\
= \ & (1 - p_{k+1}(A)) \prod_{i=1}^{k} p_i(A)\ .
\end{aligned}$$

Third equality follows by stationarity. Lemma 1.5.1 ends the proof of the theorem. □

Proof of Corollary 1.5.1 We use the inequality

$$\left|\mathbb{E}\left(X^\beta\right) - \mathbb{E}\left(Y^\beta\right)\right| \leq \sum_{k\geq 0} k^\beta \left|\mathbb{P}(X = k) - \mathbb{P}(Y = k)\right|\ ,$$

which holds for any pair of positive r.v. X, Y. We apply the above inequality with $X = S_A$ and Y geometrically distributed with parameter $\rho(A)$.

The exponential decay of the error term in Theorem 1.5.1 ends the proof of the corollary. □

Acknowledgments

MA thanks Capes for support during part of this work. Ministere de la Recherche et Ministere de l'Education and FAPESP The authors thank P. Ferrari and A. Galves for useful discussions.

Chapitre 2

Sharp error terms for point-wise Poisson approximations under mixing conditions : A new Approach

Summary

We describe the statistics of the number of occurrences of a string of symbols in a stochastic process : Chosen a string A of length n, we prove that the number of visits to A up to time t, denoted by N_t, has approximately a Poisson distribution. We provide a sharp error for this approximation. Contrarily to previous works who present uniform error terms based on the total variation distance, our error is point-wise. As a byproduct we obtain that all the moments of N_t are finite. Moreover, we obtain explicit approximations for all of them. Our result holds for processes that verify the ϕ-mixing condition. The error term is explicitly expressed as function of the rate function ϕ and then easily computable. We briefly extend our result to the weaker α-mixing case.

Keywords : Mixing, recurrence, rare event, number of visits, Poisson distribution.

2.1 Introduction

This paper describes the statistics of occurrence times of a string of symbols in a mixing stochastic process with a finite alphabet. For $n \in \mathbb{N}$, we consider a fixed string of n symbols. We prove an upper bound for the difference between the law of the number of occurrences of the string in a long sequence and a Poisson law. Our result stands for ϕ-mixing and strong or α-mixing processes (see definitions below), each one with its corresponding error.

The first result about the number of visits to a fixed set is obviously the convergence of the binomial distribution to the Poisson distribution. Recently, motivated by the statistical analysis of data sources coming from different areas such as physics, biology, computer science, linguistics among other there was a major interest to generalize this convergence in various sense :

(a) dependent process ;
(b) explicit rate of convergence ;
(c) different kind of observables.

The pioneer paper considering (a) is that of Doeblin ([Doe40]), who studied the Poisson approximation for the Gauss transformation. There is abundant literature on this subject in the dynamical systems context. See for instance Galves and Schmitt ([GS97]) and the references there in.

Probably the most used tool to attack (b) is the Chen-Stein method introduced by Chen ([Che75]). There is also abundant literature on this subject (see e.g. [AGG89], [AGG90], [BHJ92].) The principal feature of this method is that it provides only uniform bounds for the rate of convergence based on the total variation distance. As far as we know, this method was only implemented in processes that verify the Markov property. Whether it is useful in other context is an open question for us. We are aware of only one work which provides point-wise rate of convergence. Haydn and Vaienti ([HV04]) prove a rate of convergence using the method of factorial moments. The result holds for $(\psi - f)$-mixing processes. The bound decreases factorially fast on k but contrary to our, it holds only for values of k that do not exceed the inverse of some (positive) power of the measure of the n-string.

Our result tends to give bring some light over (a), (b), and (c).

With respect to (b), we prove an upper bound for the rate of convergence of the number of occurrences of a

fixed string to the Poisson law, namely,

$$\lim_{\mathbb{P}(A)\to 0} \mathbb{P}\left(N_{t/\mathbb{P}(A)} = k\right) = \frac{e^{-t}t^k}{k!} \ ,$$

where N_t is the number of visits of the process to the string A up to time t.

The striking point of our work tends to be the following. The error bound we obtain decreases factorially fast as a function of k for *all* values of k. This control on the tail of distribution of N_t allows us to obtain an approximation for *all the moments* of N_t by those of a Poisson random variable which are finite.

Our approach relies on a sharp result proved by Abadi ([Aba01a]) that states that for any string that does not overlap itself,

$$\mathbb{P}(N_{t/\mathbb{P}(A)} = 0) \approx e^{-t} \ .$$

A crucial point is that, if A is any string, $N_{t/\mathbb{P}(A)}$ could not be well approximated by a Poisson law. An example of this fact is shown in Hirata ([Hir93]), where it is proved that for periodic points, the asymptotic limit law of $\left\{N_{t/\mathbb{P}(A)} = 1\right\}$ (as function of t) differs of the one-level Poisson law. When this happens, Abadi and Vergne ([VA08], Theorem 2) show that the law of τ_A is different from the exponential. Moreover, Theorem 24 in the same paper shows that A occurs in clumps with geometric size, which says that N_t is not Poisson distributed.

Our result is established with its own error term. This error is explicitly expressed as function of the mixing rate. As we said, it turns out that the error term depends on the overlapping properties of A. We state some basic facts about overlapping useful to prove our theorem. More on that topic can be find in [VA08].

With respect to (a), we establish our result under the mixing conditions. Mixing is a large family of processes. For instance, irreducible and aperiodic finite state Markov chain are known to be ψ-mixing (and then ϕ-mixing) with exponential decay. Moreover, Gibbs states which have summable variations are ψ-mixing (see [Wal75]). They have exponential decay if they have Hölder continuous potential (see [Bow75]). However, the ψ-mixing condition is sometimes a very condition, difficult to test. We establish our result under the more general ϕ-mixing condition. Further examples of ϕ-mixing processes can be found in [LSV98]. The error term is explicitly expressed as a function of the mixing rate ϕ. We refer the reader to [Dou95] and [Bra05] for a source of examples and definitions of the several kinds of mixing processes. Those include ϕ-mixing and α-mixing with functions ϕ and α decreasing at any rate.

Even when the result is weaker, we find interesting to present also the α-mixing case. In this case we need to impose an extra condition in order for the theorem to hold : the string needs to be non-overlapping (up to some fraction of its own length). We remark that α-mixing is a quite general condition : they could have even zero entropy and strings decaying only polynomially fast.

With respect to (c), since any observable can be constructed as a union of strings, we focus our work on them.

Our result is applied in a forthcoming paper : In [VA08] the authors applied the Poisson approximation to develop a method for testing hypothesis to detect strings of high or low frequency in DNA and protein sequences. This method can not work with approximation in total variation distance or any other uniform distributions distance.

This paper is organized as follows. In section 2 we establish our framework. In section 3 we collect some definitions and properties of overlapping of strings. In section 4 we state and prove the convergence of the number of occurrences to a Poisson law. This is Theorem 2.4.1. The statement are for ϕ-mixing processes. Since the proof of the α-mixing case is similar and easily obtained from that one, we briefly treat this case section 5. This is Theorem 2.5.1.

2.2 Framework and Notations

Let $\mathcal{C}$ be a finite set. Put $\Omega = \mathcal{C}^{\mathbb{Z}}$. For each $x = (x_m)_{m\in\mathbb{Z}} \in \Omega$ and $m \in \mathbb{Z}$, let $X_m : \Omega \to \mathcal{C}$ be the m-th coordinate projection, that is $X_m(x) = x_m$. We denote by $T : \Omega \to \Omega$ the one-step-left shift operator, namely $(T(x))_m = x_{m+1}$.

We denote by $\mathcal{F}$ the σ-algebra over Ω generated by strings. Moreover we denote by $\mathcal{F}_I$ the σ-algebra generated by strings with coordinates in I, $I \subseteq \mathbb{Z}$.

For a subset $A \subseteq \Omega$ we say that $A \in \mathcal{C}_n$ if and only if

$$A = \{X_0 = a_0; \ldots; X_{n-1} = a_{n-1}\} \ ,$$

with $a_i \in \mathcal{C},\ i = 0, \ldots, n-1$.

We consider an invariant probability measure $\mathbb{P}$ over $\mathcal{F}$. We shall assume without loss of generality that there is no singleton of probability 0.

We say that the process $(X_m)_{m\in\mathbb{Z}}$ is ϕ-mixing if the sequence

$$\phi(l) = \sup |\mathbb{P}(C \mid B) - \mathbb{P}(C)| \ ,$$

converges to zero. The supremum is taken over B and C such that $B \in \mathcal{F}_{\{0,..,n\}}, n \in \mathbb{N}, \mathbb{P}(B) > 0, C \in \mathcal{F}_{\{m\geq n+l+1\}}$.

Similarly, we say that the process $(X_m)_{m\in\mathbb{Z}}$ is α-mixing if the sequence

$$\alpha(l) = \sup |\mathbb{P}(B \cap C) - \mathbb{P}(B)\mathbb{P}(C)| \ ,$$

converges to zero. The supremum is taken over B and C such that $B \in \mathcal{F}_{\{0,..,n\}}, n \in \mathbb{N}, C \in \mathcal{F}_{\{m\geq n+l+1\}}$.

For two measurables V and W, we denote as usual $\mathbb{P}_W(V) = \mathbb{P}(V|W) = \mathbb{P}(V;W)/\mathbb{P}(W)$ the conditional measure of V given W. We write $\mathbb{P}(V;W) = \mathbb{P}(V \cap W)$. We also write $V^c = \Omega\backslash V$, for the complement of V.

We use the probabilistic notation : $\{X_n^m = x_n^m\} = \{X_n = x_n, \ldots, X_m = x_m\}$. For a n-string $A = \{X_0^{n-1} = x_0^{n-1}\}$ and $1 \leq w \leq n$, we define the w-string

$$A^{(w)} = \{X_{n-w}^{n-1} = x_{n-w}^{n-1}\} \ .$$

It belongs to the σ-algebra $\mathcal{F}_{\{n-w,\ldots,n-1\}}$ and consisting of the *last* w symbols of A.

The mean of a r.v. X is denoted by $\mathbb{E}(X)$. Wherever it is not ambiguous we will write C and c for different positive constants even in the same sequence of equalities/inequalities. For brevity we put $(a \vee b) = \max\{a, b\}$ and $(a \wedge b) = \min\{a, b\}$.

2.3 Overlapping

In this section we describe some basic facts about overlapping of a string that are needed to establish our main result.

Definition 2.3.1 *Let $A \in \mathcal{C}_n$. We define the* periodicity *of A (with respect to T) as the number $\tau(A)$ defined as follows :*

$$\tau(A) = \min\left\{k \in \{1, \ldots, n\} \mid A \cap\ T^{-k}(A) \neq \emptyset\right\} \ .$$

Let us write $n = q\,p + r$, with $\tau(A) = p$, $q = [n/p]$ and $0 \leq r < p$. Thus

$$A = \left\{X_0^{p-1} = X_p^{2p-1} = \ldots = X_{(q-1)p}^{qp-1} = a_0^{p-1}\ ;\ X_{qp}^{n-1} = a_0^{r-1}\right\} \ .$$

For instance

$$A = (\overbrace{\texttt{aaaabb}}^{\text{period}}\overbrace{\texttt{aaaabb}}^{\text{period}}\overbrace{\texttt{aaa}}^{\text{rest}}) \ .$$

Thus, consider the set of overlapping positions of A :

$$\mathcal{O}(A) = \left\{k \in \{1, \ldots, n-1\} \mid A \cap\ T^{-k}(A) \neq \emptyset\right\} \ .$$

Split $\mathcal{O}(A)$ in a disjoint union of $\{\tau(A), \ldots, [n/\tau(A)]\tau(A)\}$ and $\mathcal{R}(A)$ where

$$\mathcal{R}(A) = \left\{k \in \{[n/\tau(A)]\tau(A) + 1, \ldots, n-1\} \mid A \cap\ T^{-k}(A) \neq \emptyset\right\} \ .$$

Put $r_A = \#\mathcal{R}(A)$. The cardinal of $\mathcal{O}(A)$ is then $\sigma(A) = [n/\tau(A)] + r_A \leq n$.

2.4 Poisson Approximation

2.4.1 Main Result

For $1 \leq t' < t$ integers, let

$$N_{t'}^t = \sum_{i=t'}^{t} \mathbb{1}_{T^{-i}A} \ .$$

So that, $N_{t'}^t$ counts the number of occurrences of A between t' and t. For the sake of simplicity we write $N_t = N_1^t$. With some abuse of notation we also put $(-1)! = 1$.

Theorem 2.4.1 *Let $(X_m)_{m\in\mathbb{Z}}$ be a ϕ-mixing process. There exists a constant $C>0$, such that for all $A\in\mathcal{C}_n$, and all non negative integer k, the following inequality holds :*

$$\left|\mathbb{P}\left(N_{t/\mathbb{P}(A)}=k\right)-\frac{e^{-t}t^k}{k!}\right|\leq Ce(A)g(A,k)\ ,$$

with $e(A)=e_1(A)+e_2(A)$,

$$e_1(A)\stackrel{def}{=}\inf_{1\leq w\leq\tau(A)}\left[(\sigma(A)+n)\mathbb{P}(A^{(w)})+\phi\left((\tau(A))-w\right)\right]\ ,$$

$$e_2(A)\stackrel{def}{=}\phi(n)+\inf_{n\leq\ell\leq 1/\mathbb{P}(A)}\left[\ell\mathbb{P}(A)+\frac{\phi\left(\ell\right)}{\mathbb{P}(A)}\right]\ ,$$

and

$$g(A,k)\stackrel{def}{=}\begin{cases}\frac{(2\lambda)^{k-1}}{(k-1)!} & k\notin\left\{\frac{\lambda}{e(A)},\ldots,\frac{t}{\mathbb{P}(A)}\right\}\\ \\ \frac{(2\lambda)^{k-1}}{\left(\frac{\lambda}{e(A)}\right)!\ \left(\frac{1}{e(A)}\right)^{k-\frac{1}{e(A)}-1}} & k\in\left\{\frac{\lambda}{e(A)},\ldots,\frac{t}{\mathbb{P}(A)}\right\}\end{cases}\ ,$$

where $\lambda\stackrel{def}{=}t\left[1+\frac{\phi(\ell_A)}{\mathbb{P}(A)}\right]$ and ℓ_A is the ℓ that defines $e_2(A)$.

We state several remarks to better understand the error term of the theorem below the next corollaries.

In the next corollary we show how the point-wise error term given in Theorem 2.4.1 allows us to estimate the moments of $N_{t/\mathbb{P}(A)}$ by those of a r.v. with Poisson distribution.

Corollary 2.4.1 *Let $(X_m)_{m\in\mathbb{Z}}$ be a ϕ-mixing process with summable sequence ϕ. Let $\beta>0$. Let Z be a r.v. with Poisson distribution of parameter $t>0$. Then*

$$\left|\mathbb{E}\left(N^{\beta}_{t/\mathbb{P}(A)}\right)-\mathbb{E}(Z^{\beta})\right|\leq C_{t,\beta}\ e(A)\ ,$$

where $C_{t,\beta}$ is a constant that just depends on t and β.

Corollary 2.4.2 *Let $(X_m)_{m\in\mathbb{Z}}$ be a ϕ-mixing process with summable sequence ϕ. Let Z be a r.v. with Poisson distribution of parameter $t>0$. Then*

$$\sup_{K\subseteq\mathbb{N}}\left|\mathbb{P}(N_{t/\mathbb{P}(A)}\in K)-\mathbb{P}(Z\in K)\right|\leq C_t\ e(A)\ ,$$

where C_t is a constant that just depends on t.

Remark 2.4.1 *Clearly $e(A)$ is the uniform error term and $g(A,k)$ is the error factor that provides the control on the tail of distribution.*

Remark 2.4.2 *$e_1(A)$ is the error that arises from the short correlations of the process while $e_2(A)$ is the error that arises from long ones.*

Remark 2.4.3 *$\mathbb{P}(A_n)\leq Ce^{-cn}$ (see [Aba01a]). $\phi(n)$ goes to zero by hypothesis. Therefore $e_1(A)$ is small if $\tau(A)$ is large enough to chose a w between 1 and $\tau(A)$ such that Ce^{-cw} and $\phi(\tau(A)-w)$ are small.*

Remark 2.4.4 *Take a sequence of n-strings A_n with n diverging. $e_1(A)\to 0$ if $\tau(A_n)$ also diverges with n faster than $\ln n$ (since $\mathbb{P}(A_n)$ decays exponentially fast).*

Remark 2.4.5 *$e_2(A)\to 0$ as $n\to\infty$ if the sequence $\phi(\ell)$ is summable.*

Remark 2.4.6 *Collet et al. ([CGS99]) proved that for exponentially ψ-mixing processes there exist positive constants C and c such that*

$$\mathbb{P}(A\in\mathcal{C}_n\ ;\ \tau(A)\leq n/3)\leq Ce^{-cn}\ .$$

Abadi ([Aba01a]) extended the above inequality to ϕ-mixing processes when $n/3$ is replaced with some $s\in(0,1)$. Abadi and Vaienti ([AV06]) proved the above inequality for ψ-mixing processes for any *value of s (with $c=c(s)$.) This shows that Theorem 2.4.1 holds for typical (in the sense of $\tau(A)$) strings. Taking limit on the length of the strings along infinite sequences, we get that the Poisson limit law holds almost everywhere.*

Remark 2.4.7 *When $\tau(A)$ is not large enough, the return time is better approximated by a mixture of a Dirac measure at the origin and an exponential law as shown by Abadi and Vergne ([VA08], Theorem 2). Therefore, the numbers of occurrences of the string can not be Poisson distributed.*

Remark 2.4.8 *When $e_2(A)$ is small, so is $\phi(\ell)/\mathbb{P}(A)$. Therefore λ is just the parameter of the Poisson law with a small correction factor $1+\phi(\ell)/\mathbb{P}(A)$. Thus $\lambda/e(A)$ is a large number (smaller or equal to $t/\mathbb{P}(A)$.)*

For $k \leq \lambda/e(A)$ or $k \geq t/\mathbb{P}(A)$ we get that $g(A,k)$ decays factorially fast. For k in the strip $\lambda/e(A)$ to $t/\mathbb{P}(A)$ we do not get $k!$ but something that we could call "truncated factorial" : just get $(1/e(A))!$ times $k-(1/e(A))$ factors $1/e(A)$.

2.4.2 Examples

Example 2.4.1 *Suppose that $(X_m)_{m\in\mathbb{Z}}$ are i.i.d. r.v. Then the process is ϕ-mixing with sequence $\phi(l)=0$ for all $l \in \mathbb{N}$. Then $\ell_A = n$ and $e_2(A) = n\mathbb{P}(A)$. Further, take $w = \tau_A$. Thus $e_1(A) \leq 2n\mathbb{P}(A^{(\tau_A)})$. Thus $e(A) \leq 3n\mathbb{P}(A^{(\tau_A)})$. Here $\mathbb{P}(A^{(\tau_A)})$ is the probability of the part of the string A that does not overlap A. In particular, if A does not overlap itself, then $e(A) \leq 3n\mathbb{P}(A)$.*

Example 2.4.2 *Suppose that $(X_m)_{m\in\mathbb{Z}}$ is an irreducible and aperiodic finite state Markov chain. Then a classical theorem of Markov chains said that the process is ϕ-mixing and there are positive constants C and M such that*

$$\phi(l) \leq Ce^{-Ml} \qquad \text{for all } l \in \mathbb{N} .$$

We recall that the measure of n-cylinders decays exponentially fast on n. Thus, take for instance $\ell_A = K_1 n$ with K_1 a positive constant large enough to make $\phi(K_1 n)/\mathbb{P}(A)$ small. Assume that $\tau(A) = K_2 n$. Take $w = \tau(A)/2$. Thus $e_1(A)$ and $e_2(A)$ decay exponentially fast on n. In particular, if A does not overlap itself, (as is typically the case, see Remark 2.4.6) then $\sigma(A) = 0$. Thus

$$e_1(A) = \inf_{1\leq w\leq n} \{n\mathbb{P}(A^{(w)}) + \phi(n-w)\} ,$$

and

$$e_2(A) = K_1 n\mathbb{P}(A) + \frac{\phi(K_1 n)}{\mathbb{P}(A)} + \phi(n) .$$

Example 2.4.3 *Suppose that $(X_m)_{m\in\mathbb{Z}}$ is ϕ-mixing with polynomial sequence ϕ such that $\phi(l) = l^{-\kappa}$ for some $\kappa > 1$. Then $l = \mathbb{P}(A)^{-2/(\kappa+1)}$. Thus, the first term in $e_2(A)$ is $\mathbb{P}(A)^{(\kappa-1)/(\kappa+1)}$ which decays exponentially fast on n. So, $e_2(A)$ is of order $n^{-\kappa}$. Assume A does not overlap itself, namely $\tau(A) = n$. Take $w = K_1 n$. Thus the first term of $e_1(A)$ is exponential and $e_1(A)$ is of order $(K_2 n)^{-\kappa}$ given by the second term.*

2.4.3 Preparatory Results

The next lemma says that the occurrence of two copies of A very close have small probability.

Lemma 2.4.1 *Let $(X_m)_{m\in\mathbb{Z}}$ be a ϕ-mixing process. Then, for all $A \in \mathcal{C}_n$ the following inequalities hold :*

$$\mathbb{P}_A\left(\bigcup_{j=1}^{2n-1} T^{-j}A\right) \leq e_1(A) ,$$

– *for all $\ell \geq 2n$*

$$\mathbb{P}_A\left(\bigcup_{j=2n}^{\ell} T^{-j}A\right) \leq \ell\mathbb{P}(A) + \phi(n) .$$

Proof By the overlapping properties of A one has

$$A \cap \bigcup_{j=1}^{2n-1} T^{-j}A = A \cap \left(\bigcup_{j\in\mathcal{O}(A)} \bigcup_{j=n}^{2n-1} T^{-j}A\right) .$$

Now since $T^{-j}A \subseteq T^{-j}A^{(w)}$ for any $1 \leq w \leq \tau(A)$, the first part of the lemma follows using the ϕ-mixing property with $B = A$ and

$$C = \bigcup_{j \in \mathcal{O}(A)} \bigcup_{j=n}^{2n-1} T^{-j} A^{(w)} \ .$$

Namely

$$\mathbb{P}_A \left(\bigcup_{j=1}^{2n-1} T^{-j} A \right) \leq \mathbb{P} \left(\bigcup_{j \in \mathcal{O}(A)} \bigcup_{j=n}^{2n-1} T^{-j} A^{(w)} \right) + \phi(\tau(A) - w) \ .$$

The first statement of the lemma follows since the cardinal of the union is $\sigma(A) + n$. The cardinal of the union in the second statement of the lemma is $\ell - n + 1$. The second part of the lemma follows using the ϕ-mixing property as in the previous case. □

Definition 2.4.1 *Given $A \in \mathcal{C}_n$, and $j \in \mathbb{N}$, we define the* j-occurrence time *of A as the r.v. $\tau_A^{(j)} : \Omega \to \mathbb{N} \cup \{\infty\}$, defined as follows : For any $x \in \Omega$, define $\tau_A^{(1)}(x) = \inf\{k \geq 1 : T^k(x) \in A\}$ and for $j \geq 2$*

$$\tau_A^{(j)}(x) = \inf\{k > \tau_A^{(j-1)}(x) : T^k(x) \in A\} \ .$$

The next proposition says that the measure of all the configurations where there are no two occurrences of A very close, is close to the product measure.

Proposition 2.4.1 *Let $(X_m)_{m \in \mathbb{Z}}$ be a ϕ-mixing process. Then, for all $A \in \mathcal{C}_n$, all $0 \leq t_1 < t_2 < \ldots < t_k \leq t$, and all $k \in \mathbb{N}$ for which*

$$\min_{2 \leq j \leq k} \{t_j - t_{j-1}\} > 2(\ell_A + n) \ ,$$

(ℓ_A defined in Theorem 2.4.1) the following inequality holds :

$$\begin{aligned} & \left| \mathbb{P} \left(\bigcap_{j=1}^{k} \tau_A^{(j)} = t_j ; \tau_A^{(k+1)} > t \right) - \mathbb{P}(A)^k \prod_{j=1}^{k+1} \mathbb{P}\left(t_j - t_{j-1} - 2(\ell_A + n)\right) \right| \\ \leq \ & 5k \left(\mathbb{P}(A) + \phi(\ell_A) \right)^k e(A) \ . \end{aligned}$$

Proof We prove the proposition by induction on k. For shorthand notation put $\overline{\ell}_A = 2(\ell_A + n)$, $\Delta_1 = t_1$, $\Delta_{k+1} = t - t_k$, $\Delta_i = t_i - t_{i-1}$ and $\mathcal{P}_i = \mathbb{P}\left(\tau_A > \Delta_i - \overline{\ell}_A\right)$; $i = 1, \ldots, k+1$.

For $k = 1$, the triangle inequality gives

$$\begin{aligned} & \left| \mathbb{P}\left(\tau_A = t_1 \ ; \ \tau_A^{(2)} > t\right) - \mathbb{P}(A) \prod_{j=1}^{2} \mathcal{P}_j \right| && (2.1) \\ \leq \ & \left| \mathbb{P}\left(\tau_A = t_1 \ ; \ \tau_A^{(2)} > t\right) - \mathbb{P}\left(\tau_A = t_1 \ ; \ N_{t_1+\ell_A+n}^{t} = 0\right) \right| && (2.2) \\ + \ & \left| \mathbb{P}\left(\tau_A = t_1 \ ; \ N_{t_1+\ell_A+n}^{t} = 0\right) - \mathbb{P}\left(\tau_A = t_1\right) \mathcal{P}_2 \right| && (2.3) \\ + \ & \left| \mathbb{P}\left(A \ ; \ \tau_A > t_1 - 1\right) - \mathbb{P}\left(A \ ; \ N_{n+\ell_A}^{t_1-1} = 0\right) \right| \ \mathcal{P}_2 && (2.4) \\ + \ & \left| \mathbb{P}\left(A \ ; \ N_{n+\ell_A}^{t_1-1} = 0\right) \mathcal{P}_2 - \mathbb{P}(A) \prod_{j=1}^{2} \mathcal{P}_j \right| \ . && (2.5) \end{aligned}$$

In (2.4) we used that by stationarity $\mathbb{P}\left(\tau_A = t\right) = \mathbb{P}\left(A ; \tau_A > t - 1\right)$. Term (2.2) is equal to

$$\begin{aligned} & \mathbb{P}\left(\tau_A = t_1 ; \bigcup_{i=t_1+1}^{t_1+\ell_A+n-1} T^{-i} A ; N_{t_1+\ell_A+n}^{t} = 0 \right) \\ \leq \ & \mathbb{P}\left(T^{-t_1} A ; \bigcup_{i=t_1+1}^{t_1+\ell_A+n-1} T^{-i} A \right) \\ = \ & \mathbb{P}\left(A ; \bigcup_{i=1}^{\ell_A+n-1} T^{-i} A \right) \ . && (2.6) \end{aligned}$$

We divide the above union in those sets with $1 \leq i < 2n$, and $2n \leq i \leq \ell_A + n$. Lemma 2.4.1 implies

$$\mathbb{P}\left(A; \bigcup_{i=1}^{2n-1} T^{-i}A\right) \leq \mathbb{P}(A)e_1(A) \ .$$

and,

$$\mathbb{P}\left(A; \bigcup_{i=2n+1}^{\ell_A+n} T^{-i}A\right) \leq \mathbb{P}(A)\left(\ell_A \mathbb{P}(A) + \phi(n)\right) \ .$$

Term (2.3) is bounded using the mixing property by $\phi(\ell_A)\mathbb{P}(A)$. Analogous computations are used to bound terms (2.4) and (2.5). This shows that (2.1) is bounded by $2e(A)\mathbb{P}(A)$.

Now let us suppose that the proposition holds for $k-1$ and let us prove it for k. We use a triangle inequality where the terms involved are defined below. We briefly comment the idea behind each term. For brevity denote for each non negative i, $\mathcal{S}_i = \left\{\tau_A^{(i)} = t_i\right\}$. Thus we have

$$\left|\mathbb{P}\left(\bigcap_{j=1}^{k} \mathcal{S}_j \ ; \ \tau_A^{(k+1)} > t\right) - \mathbb{P}(A)^k \prod_{j=1}^{k+1} \mathcal{P}_j\right| \leq I + II + III + IV + V \ .$$

In I we open a gap of length $\ell_A + n$ at the left of the k-th occurrence of A, namely, between coordinates $t_k - (\ell_A + n)$ and $t_k - 1$.

$$\begin{aligned} I \ &\stackrel{def}{=} \ \left|\mathbb{P}\left(\bigcap_{j=1}^{k} \mathcal{S}_j; \tau_A^{(k+1)} > t\right) - \mathbb{P}\left(\bigcap_{j=1}^{k-1} \mathcal{S}_j; N_{t_{k-1}+1}^{t_k-(\ell_A+n)} = 0; T^{-t_k}A; N_{t_k+1}^{t} = 0\right)\right| \\ &= \ \mathbb{P}\left(\bigcap_{j=1}^{k-1} \mathcal{S}_j; N_{t_{k-1}+1}^{t_k-(\ell_A+n)} = 0; \bigcup_{i=t_k-(\ell_A+n)+1}^{t_k-1} T^{-i}A; T^{-t_k}A; N_{t_k+1}^{t} = 0\right) \qquad (2.7) \\ &\leq \ \mathbb{P}\left(\bigcap_{j=1}^{k-1} T^{-t_j}A \ ; \bigcup_{i=t_k-(\ell_A+n)+1}^{t_k-1} T^{-i}A \ ; \ T^{-t_k}A\right) \ . \end{aligned}$$

As with (2.6) we split the above union in sets with $t_k-(\ell_A+n)+1 \leq i \leq t_k-2n$, $t_k-2n+1 \leq i \leq t_k-1$. We recall that by hypothesis $\Delta_i > \overline{\ell}_A$ for all $i = 1, \ldots, k$. As in Lemma 2.4.1 we have for $t_k - (\ell_A + n) + 1 \leq i \leq t_k - 2n$

$$\begin{aligned} &\mathbb{P}\left(\bigcap_{j=1}^{k-1} T^{-t_j}A; \bigcup_{i=t_k-(\ell_A+n)+1}^{t_k-2n} T^{-i}A; T^{-t_k}A\right) \\ \leq \ &\mathbb{P}\left(\bigcap_{j=1}^{k-1} T^{-t_j}A; \bigcup_{i=t_k-(\ell_A+n)+1}^{t_k-2n} T^{-i}A\right)\left(\mathbb{P}(A) + \phi(n)\right) \ . \end{aligned}$$

By the ϕ-mixing property over the left most factor in the right hand side of the above inequality, we get that it is bounded by

$$\mathbb{P}\left(\bigcap_{j=1}^{k-1} T^{-t_j}A\right)\left(\ell_A \mathbb{P}(A) + \phi(\ell_A)\right) \ .$$

Iterating this procedure we get

$$\mathbb{P}\left(\bigcap_{j=1}^{k-1} T^{-t_j}A\right) \leq \left(\mathbb{P}(A) + \phi(\ell_A)\right)^{k-1} \ .$$

Similarly, for $t_k - 2n + 1 \leq i \leq t_k - 1$

$$\mathbb{P}\left(\bigcap_{j=1}^{k-1} T^{-t_j}A; \bigcup_{i=t_k-2n+1}^{t_k-1} T^{-i}A; T^{-t_k}A\right) \leq \left(\mathbb{P}(A) + \phi(\ell_A)\right)^k e_1(A) \ .$$

In II we apply the ϕ-mixing property to factorize the probability in the right hand side of the modulus in I. Then we iterated the ϕ-mixing property to obtain the last inequality.

$$\begin{aligned} II &\stackrel{def}{=} \left| \mathbb{P}\left((\bigcap_{j=1}^{k-1} \mathcal{S}_j; N_{t_{k-1}+1}^{t_k-(\ell_A+n)} = 0); \left(T^{-t_k}A; N_{t_k+1}^{t} = 0\right)\right) - \right. \\ &\qquad \left. - \mathbb{P}\left(\bigcap_{j=1}^{k-1} \mathcal{S}_j \ ; \ N_{t_{k-1}+1}^{t_k-(\ell_A+n)} = 0 \right) \mathbb{P}\left(A \ ; \ N_1^{t-t_k} = 0\right) \right| \\ &\leq \mathbb{P}\left(\bigcap_{j=1}^{k-1} \mathcal{S}_j \ ; \ N_{t_{k-1}+1}^{t_k-(\ell_A+n)} = 0 \right) \phi(\ell_A) \\ &\leq \mathbb{P}\left(\bigcap_{j=1}^{k-1} T^{-t_j}A \right) \phi(\ell_A) \\ &\leq \left(\mathbb{P}(A) + \phi(\ell_A)\right)^k \ \frac{\phi(\ell_A)}{\mathbb{P}(A)} \ . \end{aligned}$$

In III we "fill-up" the gap we opened in I

$$\begin{aligned} III &\stackrel{def}{=} \left| \mathbb{P}\left(\bigcap_{j=1}^{k-1} \mathcal{S}_j; N_{t_{k-1}+1}^{t_k-(\ell_A+n)} = 0 \right) - \mathbb{P}\left(\bigcap_{j=1}^{k-1} \mathcal{S}_j; N_{t_{k-1}+1}^{t_k-1} = 0 \right) \right| \times \\ &\qquad \times \mathbb{P}\left(A; N_1^{t-t_k} = 0\right) \\ &\leq \mathbb{P}\left(\bigcap_{j=1}^{k-1} \mathcal{S}_j \ ; \ N_{t_{k-1}+1}^{t_k-(\ell_A+n)} = 0 \ ; \bigcup_{t_k-(\ell_A+n)+1}^{t_k-1} T^{-i}A \right) \mathbb{P}(A) \\ &\leq \mathbb{P}\left(\bigcap_{j=1}^{k-1} T^{-t_j}A; \bigcup_{i=t_k-(\ell_A+n)+1}^{t_k-1} T^{-i}A \right) \mathbb{P}(A) \\ &\leq \left(\mathbb{P}(A) + \phi(\ell_A)\right)^k 2\ell_A \mathbb{P}(A) \ . \end{aligned}$$

In IV we use the inductive hypothesis

$$\begin{aligned} IV &\stackrel{def}{=} \left| \mathbb{P}\left(\bigcap_{j=1}^{k-1} \mathcal{S}_j; N_{t_{k-1}+1}^{t_k-1} = 0 \right) - \mathbb{P}(A)^{k-1} \prod_{j=1}^{k} \mathcal{P}_j \right| \mathbb{P}\left(A; N_1^{t-t_k} = 0\right) \\ &\leq C(k-1) \left(\mathbb{P}(A) + \phi(\ell_A)\right)^{k-1} e(A)\mathbb{P}(A) \ . \end{aligned}$$

In V we use that the proposition is already proved for $k = 1$ to get

$$\begin{aligned} V &\stackrel{def}{=} \mathbb{P}(A)^{k-1} \prod_{j=1}^{k} \mathcal{P}_j \left| \mathbb{P}\left(A \ ; \ N_1^{t-t_k} = 0\right) - \mathbb{P}(A)\mathcal{P}_{k+1} \right| \\ &\leq \mathbb{P}(A)^k \ 2e(A) \ . \end{aligned}$$

Summing the bounds above we end the proof of the proposition. □

2.4.4 Proof of Theorem 2.4.1 and Corollary 2.4.1.

Proof of Theorem 2.4.1. Take $t \in \mathbb{N}$. Let us write for the sake of simplicity $N = N_t$. For $k = 0$ note that $\mathbb{P}(N = 0) = \mathbb{P}\left(\tau_A^{(1)} > t\right)$. By Theorem 1 in Abadi ([Aba04]) one has

$$\left| \mathbb{P}\left(\tau_A^{(1)} > t\right) e^{-\xi_A \mathbb{P}(A)t} \right| \leq e(A)(\mathbb{P}(A)t \vee 1) e^{-\xi_A \mathbb{P}(A)t} \ , \tag{2.8}$$

with a certain $\xi_A > 0$. Moreover, it follows in the proof of Theorem 2 in Abadi and Vergne ([AV08]) that $|\xi_A - \zeta_A| \leq e_1(A)$ where $\zeta_A = \mathbb{P}_A(\zeta_A > \tau(A))$. Finally $|\zeta_A - 1| = \mathbb{P}_A(\zeta_A = \tau(A)) \leq e_1(A)$ by Lemma 2.4.1. This concludes the proof for $k = 0$.

For $k > t$ we have that $\mathbb{P}(N = k) = 0$. Then

$$\begin{aligned} \left| \mathbb{P}(N = k) - \frac{e^{-t\mathbb{P}(A)}(t\mathbb{P}(A))^k}{k!} \right| &= \frac{e^{-t\mathbb{P}(A)}(t\mathbb{P}(A))^k}{k!} \\ &\leq \frac{(t\mathbb{P}(A))^{k-1}}{(k-1)!}\, \mathbb{P}(A) \ . \end{aligned}$$

To conclude just note that $\mathbb{P}(A) \leq e(A)$.

Let us consider now k with $1 \leq k \leq t$. The idea of the proof is the following : Consider a realization $x = (x_m)_{m\in\mathbb{Z}}$ of the process $(X_m)_{m\in\mathbb{Z}}$ such that the sequence $(x_1, \ldots, x_t)$ contains exactly k occurrences of A. These occurrences can appear in clusters or isolated one from each other. We prove that realizations with isolated A's give the approximation to the Poisson law and realizations with clustered A's have small measure. We now formalize this idea. Given $1 \leq t_1 < \ldots < t_k \leq t$, let us define the following measurable set :

$$\mathcal{T}(t_1, \ldots, t_k) = \bigcap_{j=1}^{k} \left\{ \tau_A^{(j)} = t_j \right\} \bigcap \left\{ \tau_A^{(k+1)} > t \right\} \ .$$

As in Proposition 2.4.1 we put $\Delta_j = t_j - t_{j-1}$, for $j = 2, \ldots, k$. Put also $\Delta_1 = t_1$ and $\Delta_{k+1} = t - t_k$. Define the minimum distance between two consecutive occurrences of A by

$$I(\mathcal{T}(t_1, \ldots, t_k)) = \min \{ \Delta_j \mid 2 \leq j \leq k \} \ .$$

As before put $\bar{\ell}_A = 2(\ell_A + n)$. Let us divide $\{N = k\}$ in two sets

$$B_k = \bigcup_{I(\mathcal{T}(t_1,\ldots,t_k)) < \overline{\ell}_A} \mathcal{T}(t_1, \ldots, t_k) \quad \text{and} \quad G_k = \bigcup_{I(\mathcal{T}(t_1,\ldots,t_k)) \geq \overline{\ell}_A} \mathcal{T}(t_1, \ldots, t_k) \ .$$

Since $\{N = k\} = B_k \cup G_k$, disjoint union, we have

$$\begin{aligned} &\left| \mathbb{P}(N = k) - \frac{e^{-t\mathbb{P}(A)} t^k \mathbb{P}(A)^k}{k!} \right| \\ \leq\ & \mathbb{P}(B_k) + \left| \mathbb{P}(G_k) - \frac{e^{-t\mathbb{P}(A)} t^k \mathbb{P}(A)^k}{k!} \right| \ . \end{aligned} \tag{2.9}$$

We will prove that both quantities in the right hand side of (2.9) are small.

Proof : configurations with clusters have small measure.

We will prove an upper bound for $\mathbb{P}(B_k)$. Let us start computing how many clusters there are in a given $\mathcal{T}(t_1, \ldots, t_k)$ with

$$C(\mathcal{T}(t_1, \ldots, t_k)) = \sum_{j=2}^{k} \mathbb{1}_{\{\Delta_j > \overline{\ell}_A\}} + 1 \ .$$

Suppose that $C(\mathcal{T}(t_1, \ldots, t_k)) = 1$ and fix the position t_1. Each occurrence inside the unique cluster (with the exception of the most left one which is fixed at t_1) can appear at distance d of the previous one, with $d \in \mathcal{O}(A)$ or $n \leq d \leq \overline{\ell}_A$. Firstly note that

$$\mathcal{T}(t_1, t_2, \ldots, t_k) \subseteq \bigcap_{j=1}^{k} T^{-t_j} A \ .$$

Then

$$\bigcup_{\substack{i=2,\ldots,k \\ t_i = t_{i-1}, \ldots, t_{i-1} + \overline{\ell}_A}} \mathcal{T}(t_1, t_2, \ldots, t_k) \subseteq \bigcup_{\substack{i=2,\ldots,k \\ t_i = t_{i-1}, \ldots, t_{i-1} + \overline{\ell}_A}} \bigcap_{j=1}^{k} T^{-t_j} A \ .$$

Therefore, the iterative argument of the ϕ-mixing property used to bound (2.7) leads to the bound

$$\begin{aligned} & \mathbb{P}\left(\bigcup_{\substack{i=2,\dots,k \\ t_i=t_{i-1},\dots,t_{i-1}+\overline{\ell}_A}} \bigcap_{j=1}^{k} T^{-t_j}A\right) \\ \leq\ & \mathbb{P}(A)\ \left(e_1(A)+\overline{\ell}_A\mathbb{P}(A)+\phi(n)\right)^{k-1} \\ \leq\ & \mathbb{P}(A)e(A)^{k-1}\ . \end{aligned} \tag{2.10}$$

Suppose now that $C(\mathcal{T}(t_1,\dots,t_k))=i$. Assume also that the most left occurrence of these i clusters occurs at $1\leq t(1)<\ldots<t(i)\leq t$ fixed. By the same argument used in (2.10), we have the inequalities

$$\begin{aligned} & \mathbb{P}\left(\bigcup_{\{t_1,\dots,t_k\}\setminus\{t(1),\dots,t(i)\}} \mathcal{T}(t_1,\dots,t_k)\right) \\ \leq\ & \mathbb{P}(A)\left(\mathbb{P}(A)+\phi(\ell_A)\right)^{i-1}e(A)^{k-i} \\ \leq\ & \left(\mathbb{P}(A)+\phi(\ell_A)\right)^{i}e(A)^{k-i}\ . \end{aligned} \tag{2.11}$$

In order to obtain an upper bound for $\mathbb{P}(B_k)$ we must sum the above bound over all $\mathcal{T}(t_1,\dots,t_k)$ such that $C(\mathcal{T}(t_1,\dots,t_k))=i$ with i that runs from 1 to $k-1$.

Fixed $C(\mathcal{T}(t_1,\dots,t_k))=i$, the locations of the most left occurrences of A of each one of the i clusters can be chosen at most in $\binom{t}{i}$ many ways.

The cardinality of each one of the i clusters can be arranged in $\binom{k-1}{i-1}$ many ways. (This corresponds to break the interval $(1/2,k+1/2)$ in i intervals at points chosen from $\{1+1/2,\dots,k-1/2\}$.)

Collecting these information and (2.11) we have that $\mathbb{P}(B_k)$ is bounded by

$$\sum_{i=1}^{k-1}\binom{t}{i}\binom{k-1}{i-1}\left(\mathbb{P}(A)+\phi(\ell_A)\right)^{i}e(A)^{k-i} \leq e(A)^k \max_{1\leq i\leq k-1}\left\{\frac{\gamma^i}{i!}\right\}\sum_{i=1}^{k-1}\binom{k-1}{i-1},$$

where $\gamma=t\mathbb{P}(A)\left[1+\phi(\ell_A)/\mathbb{P}(A)\right]/e(A)$. The maximum in the above expression is reached at $(k-1\wedge\gamma)$. The most right sum is bounded by 2^{k-1}. Therefore we have

$$\mathbb{P}(B_k)\leq e(A).\begin{cases} \frac{(2\gamma e(A))^{k-1}}{(k-1)!} & k-1<\gamma \\ \frac{2^{k-1}(\gamma e(A))^{\gamma}}{\gamma!\left(\frac{1}{e(A)}\right)^{k-\gamma-1}} & k\geq\gamma \end{cases}.$$

This ends the proof of the bound for $\mathbb{P}(B_k)$.

Proof : A's isolated provide the Poisson limit law.

We can bound the most right term on the right-hand side of (2.9) by the following triangular inequality :

$$\sum_{\mathcal{T}(t_1,\dots,t_k)\in G_k}\left|\mathbb{P}\left(\bigcap_{j=1}^{k}\tau_A^{(j)}=t_j\ ;\ \tau_A^{(k+1)}>t\right)-\mathbb{P}(A)^k\prod_{j=1}^{k+1}\mathcal{P}_j\right| \tag{2.12}$$

$$+\quad \mathbb{P}(A)^k\sum_{\mathcal{T}(t_1,\dots,t_k)\in G_k}\left|\prod_{j=1}^{k+1}\mathcal{P}_j-\prod_{j=1}^{k+1}e^{-(\Delta_j-\overline{\ell}_A)\mathbb{P}(A)}\right| \tag{2.13}$$

$$+\quad \mathbb{P}(A)^k\#G_k\left|e^{-(t-(k+1)\overline{\ell}_A)\mathbb{P}(A)}-e^{-t\mathbb{P}(A)}\right| \tag{2.14}$$

$$+\quad \left|\frac{\#G_k\ k!}{t^k}\ -1\right|\ \frac{e^{-t\mathbb{P}(A)}t^k\mathbb{P}(A)^k}{k!}\ . \tag{2.15}$$

By a simple combinatorial argument we get the bounds

$$\frac{(t-k(n+\overline{\ell}_A))^k}{k!}\leq\binom{t-k(n+\overline{\ell}_A-1)-1}{k}\leq\#G_k\leq\binom{t}{k}\leq t^k/k!\ . \tag{2.16}$$

Moreover, the leading term in (2.12) is bounded using Proposition 2.4.1. Thus (2.12) is bounded by

$$5\ \frac{t^k}{(k-1)!}\ (\mathbb{P}(A)+\phi(\ell_A))^k\ e(A)\ .$$

The difference between the leading factors in (2.13) is bounded as follows : again by (2.8)

$$|\mathcal{P}_j - e^{-\xi_A \mathbb{P}(A)(\Delta_j - \overline{\ell}_A)}| \leq C e_1(A)\ .$$

As stated at the beginning of the proof one has $|\xi_A - 1| \leq e_1(A)$. Therefore (2.13) is bounded by

$$\frac{t^k}{k!}\mathbb{P}(A)^k(k+1)\max_{1\leq j\leq k+1}\left|\mathcal{P}_j - e^{-(\Delta_j-\overline{\ell}_A)\mathbb{P}(A)}\right| \leq \frac{k+1}{k}\frac{(t\mathbb{P}(A))^k}{(k-1)!}Ce_1(A)\ .$$

(2.14) is bounded using the Mean Value Theorem by

$$\frac{t^k\mathbb{P}(A)^k}{k!}(k+1)\overline{\ell}_A\mathbb{P}(A) \leq \frac{k+1}{k}\frac{(t\mathbb{P}(A))^k}{(k-1)!}4\ell_A\mathbb{P}(A)\ .$$

The left hand side of (2.16) and the Mean Value Theorem provide a bound for the difference below

$$\left|\frac{\#G_k\ k!}{t^k}\ -1\right| \leq \left|\frac{(t-k(n+\overline{\ell}_A))^k}{t^k}-1\right| \leq \frac{k\ k(n+\overline{\ell}_A)}{t} \leq k\ .$$

So, (2.15) is bounded by

$$\frac{(t\mathbb{P}(A))^k}{(k-1)!}\ 4\ell_A\mathbb{P}(A)\ .$$

Summing the bounds obtained for (2.12), (2.13), (2.14) and (2.15) we get the desired bound for the difference in the right hand term of inequality (2.9). The exchange of variables $\widetilde{t} = t\mathbb{P}(A)$ ends the proof of the theorem. □

Proof of Corollary 2.4.1. By definition

$$\begin{aligned}\left|\mathbb{E}\left(N^\beta_{t/\mathbb{P}(A)}\right) - \mathbb{E}(Z^\beta)\right| &= \left|\sum_{k\geq 0}k^\beta\mathbb{P}\left(N_{t/\mathbb{P}(A)}=k\right) - \sum_{k\geq 0}k^\beta\frac{e^{-t}t^k}{k!}\right| \\ &\leq \sum_{k\geq 0}k^\beta\ \left|\mathbb{P}\left(N_{t/\mathbb{P}(A)}=k\right) - \frac{e^{-t}\ t^k}{k!}\right|\ .\end{aligned}$$

Since for ϕ summable

$$\sum_{k=0}^{\infty}k^\beta\ g(A,k) \leq C_{t,\beta} < \infty\ ,$$

the corollary follows. □

Proof of Corollary 2.4.2. This follows by the above corollary with $\beta = 1$. □

2.5 α-mixing processes

Theorem 2.5.1 *Let $(X_m)_{m\in\mathbb{Z}}$ be α-mixing process. Let C_1 be a positive constant, $C_1 \in (0,1)$. For all $A \in \mathcal{C}_n$ such that $\tau(A) \geq C_1 n$, the following inequality holds :*

$$\left|\mathbb{P}\left(N_{t/\mathbb{P}(A)}=k\right) - \frac{e^{-t}t^k}{k!}\right| \leq C_2 e_\alpha(A,k)g_\alpha(A,k),$$

with

$$e_\alpha(A,k) \stackrel{def}{=} \inf_{1\leq w\leq n_A}\left\{(1+C_1)n\mathbb{P}(A^{(w)}) + \frac{\alpha\,(n_A-w)}{\mathbb{P}(A)^k}\right\} + n\sqrt{\mathbb{P}(A)}\ ,$$

and

$$g_\alpha(A,k) \stackrel{def}{=} \begin{cases} \frac{2^{k-1}}{(k-1)!} & k \notin \left\{ \frac{t}{e_\alpha(A,1)}, \ldots, \frac{2t}{n\mathbb{P}(A)} \right\} \\ \frac{2^{k-1}}{\left(\frac{1}{e_\alpha(A,1)}\right)! \left(\frac{1}{e_\alpha(A,1)}\right)^{k-\frac{1}{e_\alpha(A,1)}-1}} & k \in \left\{ \frac{t}{e_\alpha(A,1)}, \ldots, \frac{2t}{n\mathbb{P}(A)} \right\} \end{cases} .$$

Furthermore, assume that for a fixed $k \in \mathbb{N}$, $e_\alpha(A_n,k) \to 0$ as $n \to \infty$. Then, $N_{t/\mathbb{P}(A_n)}$ converges in distribution to a Poisson law for $N_{t/\mathbb{P}(A_n)} = j$, for all $0 \le j \le k$.

Remark 2.5.1 *The condition $\alpha(n_A - w)/\mathbb{P}(A_n)^k \to 0$ as $n \to \infty$ for all k means that we need a faster convergence of the sequence α to get convergence for larger k. However, we usually are interested in the convergence for not too large values of k, say $k \le C$, for a certain positive constant C. In that case, we need a limited rate of convergence of the sequence α.*

Proof The proof follows the steps of the proof of Theorem 2.4.1. We only indicate briefly the modifications needed to prove the α-mixing case.

In this case we only consider error of the type $e_1(A)$, namely we are going to take a gap of order n and $e_2(A)$ will be zero. We now show an inequality analogous to that of Lemma 2.4.1. Firstly, for any positive integer Δ the α-mixing property gives the upper bound

$$\mathbb{P}(A; T^{-\Delta}A) \le \begin{cases} \mathbb{P}(A)\mathbb{P}(A^{(w)}) + \alpha(\Delta - w) & \text{for } \Delta < 2n; 1 \le w \le (n \wedge \Delta) \\ \mathbb{P}(A)^2 + \alpha(\Delta - n) & \text{for } \Delta \ge 2n \end{cases} . \tag{2.1}$$

Now, since $\tau(A) \ge C_1 n$, one has

$$A \cap \bigcup_{i=1}^{2n-1} T^{-i}A = A \cap \bigcup_{i=C_1 n}^{2n-1} T^{-i}A .$$

Applying the above inequality one has

$$\mathbb{P}_A\left(\bigcup_{i=1}^{2n-1} T^{-i}A \right) \le (1+C_1) n \mathbb{P}(A^{(w)}) + \frac{\alpha(C_1 n - w)}{\mathbb{P}(A)} .$$

Now take a realization with k occurrences of A at $t_1, \ldots, t_k$. Call $\Delta_i = t_i - t_{i-1}$. Choose w_i as w in (2.1) for $i = 2, \ldots, k$. Put $w_{\min} = \min\{w_i \mid 2 \le i \le k\}$. Iterating the above procedure of the α-mixing property one has

$$\begin{aligned} \mathbb{P}\left(\bigcap_{i=1}^{k} T^{-t_i}A \right) &\le \mathbb{P}(A) \prod_{i=2}^{k} \mathbb{P}(A^{(w_i)}) + \sum_{j=0}^{k-1} \alpha(\Delta_j - w_j) \prod_{i=0}^{j} \mathbb{P}(A^{(w_i)}) \\ &\le \mathbb{P}(A)\mathbb{P}(A^{(w_{\min})})^{k-1} + \alpha(n_A - w_{\min}) \sum_{j=0}^{k-1} \mathbb{P}(A^{(w_{\min})})^j \\ &\le \mathbb{P}(A)\mathbb{P}(A^{(w_{\min})})^{k-1} + C\alpha(n_A - w_{\min}) . \end{aligned}$$

This bound can be used to prove a similar result to Proposition 2.4.1 in the α-mixing context.

Theorem 1 in [Aba04] used in the proof of Theorem 2.4.1 can be replaced by Theorem 17 in the same paper which establishes that

$$\sup_{t \ge 0} |\mathbb{P}(\tau_A > t) - e^{-\lambda(A)\mathbb{P}(A)t}| \le n\sqrt{\mathbb{P}(A)} ,$$

with a certain parameter $\lambda(A)$. It is easy to follow the proof of Lemma 19 in the same paper to show that $|\lambda(A) - \zeta_A| \le e_1^\alpha(A)$ (recall that $\zeta_A = \mathbb{P}_A(\zeta_A > \tau(A))$). The condition $\tau(A) \ge C_1 n$ implies that $|\zeta_A - 1| \le e_1^\alpha(A)$.

The rest of the proof follows as in the ϕ-mixing case. □

Acknowledgments

MA is partially supported by CNPq grant grant 308250/2006-0. The authors are beneficiaries of Capes, Brasil - Cofecub, France grant. MA thanks Ministere de la Recherche et Ministere de l'Education for support during part of this work. The authors thank P. Ferrari, A. Galves, B. Prum and P. Shields for useful discussions.

Chapitre 3

Poisson approximation for search of rare words in DNA sequences

Summary

Using recent results on the occurrence times of a string of symbols in a stochastic process with mixing properties, we present a new method for the search of rare words in biological sequences modelled by a Markov chain. We obtain a bound on the error between the distribution of the number of occurrences of a word in a sequence and its Poisson approximation. A global bound is already given by a Chen-Stein method. Our approach, the ψ-mixing method, gives local bounds. Since we only need the error in the tails of distribution, the global uniform bound of Chen-Stein is too large and it is a better way to consider local bounds. It is the first time that local bounds are devised for Poisson approximation. We search for two thresholds on the number of occurrences from which we can regard a studied word as an over-represented or an under-represented one. A biological role is suggested for these over- or under-represented words. Our method gives such thresholds for a panel of words much broader than the Chen-Stein method which cannot give any result in a great number of cases where our method works. Comparing the methods, we observe a better accuracy for the ψ-mixing method for the bound of the tails of distribution. Our method can obviously be used in domains other than biology. We also present the software `PANOW`[9] dedicated to the computation of the error term and the thresholds for a studied word. **Keywords :** Poisson approximation, Chen-Stein method, mixing processes, Markov chains, rare words, DNA sequences.

3.1 Introduction

Modelling DNA sequences with stochastic models and developing statistical methods to analyse the enormous set of data that results from the multiple projects of DNA sequencing are challenging questions for statisticians and biologists. Many DNA sequence analysis are based on the distribution of the occurrences of patterns having some special biological function. The most popular model in this domain is the Markov chain model that gives a description of the local behaviour of the sequence (see [Alm83, Bla85, PAI87, GKP92]). An important problem is to determine the statistical significance of a word frequency in a DNA sequence. [NDV02] discuss about this relevance of finding over- or under-represented words. The naive idea is the following : a word may have a significant low frequency in a DNA sequence because it disrupts replication or gene expression, whereas a significantly frequent word may have a fundamental activity with regard to genome stability. Well-known examples of words with exceptional frequencies in DNA sequences are biological palindromes corresponding to restriction sites avoided for instance in *E. coli* ([KBM92]), the Cross-over Hotspot Instigator sites in several bacteria ([SKS$^+$81, EKBSG99]), and uptake sequences ([SGS99]) or polyadenylation signals ([vHdOPO00]).

The exact distribution of the number of a word occurrences under the Markovian model is known and some softwares are available ([RD99, RÓ0]) but, because of numerical complexity, they are often used to compute expectation and variance of a given count (and thus use, in fact, Gaussian approximations for the distribution). In fact these methods are not efficient for long sequences or if the Markov model order is larger than 2 or 3. For such cases, several approximations are possible : Gaussian approximations ([PRdT95]), Binomial or Poisson approximations ([vHACV98, God91]), compound Poisson approximations ([RS98]), or large deviations approach ([Nue04]). In this paper we only focus on the Poisson approximation. For the first time, we give a local bound for the Poisson approximation. We approximate $\mathbb{P}(N(A) = k)$ by $\exp(-t\mathbb{P}(A))[t\mathbb{P}(A)]^k(k!)^{-1}$ where $\mathbb{P}(N(A) = k)$ is the

[9] available at `http ://stat.genopole.cnrs.fr/sg/software/panow/`

stationary probability under the Markov model that the number of occurrences $N(A)$ of word A is equal to k, $\mathbb{P}(A)$ is the probability that word A occurs at a given position, and t is the length of the sequence. Intuitively, a binomial distribution could be used to approximate the distribution of occurrences of a particular word. Length t of the sequence is large, $\mathbb{P}(A)$ is small if A is large. Thus, we use the more numerically convenient Poisson approximation. Our aim is to bound the error between the distribution of the number of occurrences of word A and its Poisson approximation. In [RS98], the authors prove an upper bound for a compound Poisson approximation. They use a Chen-Stein method, which is the usual method in this purpose. This method has been developed by Chen on Poisson approximations ([Che75]) after a work of Stein on normal approximations ([Ste72]). Its principle is to bound the difference between the two distributions in total variation distance for all subsets of the definition domain. Since we are interested in under- or over-represented words, we are only interested in this difference for the tails of the distributions. Then, the uniform bound given by the Chen-Stein method is too large for our purpose. We present here a new method, based on the property of mixing processes. Our method has the useful particularity to give a bound on the error at each point of the distribution. More precisely, it offers an error term ϵ, for the number of occurrences k, of word A :

$$\left| \mathbb{P}(N(A) = k) - \frac{e^{-t\mathbb{P}(A)}(t\mathbb{P}(A))^k}{k!} \right| \leq \epsilon(A, k).$$

Moreover, $\epsilon(A, k)$ decays factorially fast with respect to k.

[Aba01a, Aba04] presents lower and upper bounds for the exponential approximation of the first occurrence time of a rare event, also called *hitting time*, in a stationary stochastic process on a finite alphabet with α- or ϕ-mixing property. (Abadi and Vergne, in preparation) describe the statistics of *return times* of a string of symbols in such a process. In (Abadi and Vergne, in preparation), the authors prove a Poisson approximation for the distribution of occurrence times of a string of symbols in a ϕ-mixing process. The first part of our present work is to determine some constants not explicitly computed in the results of the above mentioned articles but necessary for the proof of our theorem and moreover for its practical use. Theoritical constants are useless in the way of numerical tests, that is why we have to determine these constants. Our work is complementary to all these articles, in the sense that it relies on them for preliminary results and it adapts them to ψ-mixing processes. Since Markov chains are mixing processes, all these results established for mixing processes also apply to Markov chains which model biological sequences.

This paper is organised in the following way. In section 3.2, we introduce the Chen-Stein method. In section 3.3, we define a ψ-mixing process and state some preliminary notations, mostly on the properties of a word. We also present in this section the principal result of our work : the Poisson approximation (Theorem 3.3.1). In section 3.4, we state preliminary results. Mainly, we recall results of [Aba04], but computing all the necessary constants and we present lemmas and propositions necessary for the proof of Theorem 3.3.1. In section 3.5, we establish the proof of our main result : Theorem 3.3.1 on Poisson approximation. Using ψ-mixing properties and preliminary results, we prove an upper bound for the difference between the exact distribution of the number of occurrence of word A and the Poisson distribution of parameter $t\mathbb{P}(A)$. Section 3.6 is dedicated to numerical results. For the search of over-represented words, we show how our method is better than the Chen-Stein method on both synthetic and biological data. In this section, we also present results obtained by a similar method, the ϕ-mixing method. We end the paper presenting some examples of biological applications, and some conclusions and perspectives of future works.

3.2 The Chen-Stein Method

3.2.1 Total Variation Distance

Definition 3.2.1 *For any two random variables X and Y with values in the same discrete space E, the total variation distance between their probability distributions is defined by*

$$d_{\mathrm{TV}}(\mathcal{L}(X), \mathcal{L}(Y)) = \frac{1}{2} \sum_{i \in E} |\mathbb{P}(X = i) - \mathbb{P}(Y = i)| .$$

We remark that for any subset S of E

$$|\mathbb{P}(X \in S) - \mathbb{P}(Y \in S)| \leq d_{\mathrm{TV}}(\mathcal{L}(X), \mathcal{L}(Y)).$$

3.2.2 The Chen-Stein Method

The Chen-Stein method is used to bound the error between the distribution of the number of occurrences of a word A in a sequence X and the Poisson distribution with parameter $t\mathbb{P}(A)$ where t is the length of the sequence and $\mathbb{P}(A)$ the stationary measure of A. The Chen-Stein method for Poisson approximation has been developed by [Che75] ; a friendly exposition is in [AGG89] and a description with many examples can be found in [AGG90] and [BCL92]. We will use Theorem 1 in [AGG90] with an improved bound by [BCL92] (Theorem 1.A and Theorem 10.A).

First, we will fix a few notations. Let $\mathcal{A}$ be a finite set (for example, in the DNA case $\mathcal{A} = \{a, c, g, t\}$). Put $\Omega = \mathcal{A}^{\mathbb{Z}}$. For each $x = (x_m)_{m\in\mathbb{Z}} \in \Omega$, we denote by X_m the m-th coordinate of the sequence x : $X_m(x) = x_m$. We denote by $T : \Omega \to \Omega$ the one-step-left shift operator : so we will have $(T(x))_m = x_{m+1}$. We denote by $\mathcal{F}$ the σ-algebra over Ω generated by strings and by $\mathcal{F}_I$ the σ-algebra generated by strings with coordinates in I with $I \subseteq \mathbb{Z}$. We consider an invariant probability measure $\mathbb{P}$ over $\mathcal{F}$. Consider a stationary Markov chain $X = (X_i)_{i\in\mathbb{Z}}$ on the finite alphabet $\mathcal{A}$. Let us fix a word $A = (a_1, \ldots, a_n)$. For $i \in \{1, 2, \cdots, t-n+1\}$, let Y_i be the following random variable

$$\begin{aligned} Y_i = Y_i(A) &= \mathbb{1}\{\text{word } A \text{ appears at position } i \text{ in the sequence}\} \\ &= \mathbb{1}\{(X_i, \ldots, X_{i+n-1}) = (a_1, \ldots, a_n)\}, \end{aligned}$$

where $\mathbb{1}\{F\}$ denotes the indicator function of set F. We put $Y = \sum_{i=1}^{t-n+1} Y_i$, the random variable corresponding to the number of occurrences of a word, $\mathbb{E}(Y_i) = m_i$ and $\sum_{i=1}^{t-n+1} m_i = m$. Then, $\mathbb{E}(Y) = m$. Let Z be a Poisson random variable with parameter m : $Z \sim \mathcal{P}(m)$. For each i, we arbitrarily define a set $V(i) \subset \{1, 2, \cdots, t-n+1\}$ containing the point i. The set $V(i)$ will play the role of a neighbourhood of i.

Theorem 3.2.1 ([AGG90, BCL92]) *Let I be an index set. For each $i \in I$, let Y_i be a Bernoulli random variable with $p_i = \mathbb{P}(Y_i = 1) > 0$. Suppose that, for each $i \in I$, we have chosen $V(i) \subset I$ with $i \in V(i)$. Let $Z_i, i \in I$, be independent Poisson variables with mean p_i. The total variation distance between the dependent Bernoulli process $\underline{Y} = \{Y_i, i \in I\}$ and the Poisson process $\underline{Z} = \{Z_i, i \in I\}$ satisfies*

$$d_{\mathrm{TV}}(\mathcal{L}(\underline{Y}), \mathcal{L}(\underline{Z})) \leq b_1 + b_2 + b_3$$

where

$$b_1 = \sum_i \sum_{j\in V(i)} \mathbb{E}(Y_i)\mathbb{E}(Y_j),$$

$$b_2 = \sum_i \sum_{j\in V(i), j\neq i} \mathbb{E}(Y_i Y_j),$$

$$b_3 = \sum_i \mathbb{E}\left|\mathbb{E}(Y_i - p_i | Y_j, j \notin V(i))\right|.$$

Moreover, if $W = \sum_{i\in I} Y_i$ and $\lambda = \sum_{i\in I} p_i < \infty$, then

$$d_{TV}(\mathcal{L}(W), \mathcal{P}(\lambda)) \leq \frac{1 - e^{-\lambda}}{\lambda}(b_1 + b_2) + \min\left(1, \sqrt{\frac{2}{\lambda e}}\right) b_3.$$

We think of $V(i)$ as a neighbourhood of strong dependence of Y_i. Intuitively, b_1 describes the contribution related to the size of the neighbourhood and the weights of the random variables in that neighbourhood ; if all Y_i had the same probability of success, then b_1 would be directly proportional to the neighbourhood size. The term b_2 accounts for the strength of the dependence inside the neighbourhood ; as it depends on the second moments, it can be viewed as a "second order interaction" term. Finally, b_3 is related to the strength of dependence of Y_i with random variables outside its neighbourhood. In particular, note that $b_3 = 0$ if Y_i is independent of $\{Y_j | j \notin V(i)\}$.

One consequence of this theorem is that for any indicator function of an event, i.e. for any measurable functional h from Ω to $[0, 1]$, there is an error bound of the form $|\mathbb{E}h(\underline{Y}) - \mathbb{E}h(\underline{Z})| \leq d_{TV}(\mathcal{L}(\underline{Y}), \mathcal{L}(\underline{Z}))$. Thus, if $S(\underline{Y})$ is a test statistic then, for all $t \in \mathbb{R}$,

$$\mathbb{P}(S(\underline{Y}) \geq t) - \mathbb{P}(S(\underline{Z}) \geq t) \leq b_1 + b_2 + b_3,$$

which can be used to construct confidence intervals and to find p-values for tests based on this statistic.

3.3 Preliminary Notations and Poisson Approximation

3.3.1 Preliminary Notations

We focus on Markov processes in our biological applications (see 3.6) but the theorem given in the following subsection is established for more general mixing processes : the so called ψ-mixing processes.

Definition 3.3.1 *Let $\psi = (\psi(\ell))_{\ell \geq 0}$ be a sequence of real numbers decreasing to zero. We say that $(X_m)_{m\in\mathbb{Z}}$ is a ψ-mixing process if for all integers $\ell \geq 0$, the following holds*

$$\sup_{n\in\mathbb{N}, B\in\mathcal{F}_{\{0,.,n\}}, C\in\mathcal{F}_{\{n\geq 0\}}} \frac{|\mathbb{P}(B\cap T^{-(n+\ell+1)}(C)) - \mathbb{P}(B)\mathbb{P}(C)|}{\mathbb{P}(B)\mathbb{P}(C)} = \psi(\ell),$$

where the supremum is taken over the sets B and C, such that $\mathbb{P}(B)\mathbb{P}(C) > 0$.

For a word A of Ω, that is to say a measurable subset of Ω, we say that $A \in \mathcal{C}_n$ if and only if

$$A = \{X_0 = a_0, \ldots, X_{n-1} = a_{n-1}\},$$

with $a_i \in \mathcal{A}, i = 1, \ldots, n$. Then, the integer n is the length of word A. For $A \in \mathcal{C}_n$, we define the hitting time $\tau_A : \Omega \to \mathbb{N} \cup \{\infty\}$, as the random variable defined on the probability space (Ω,$\mathcal{F}$,$\mathbb{P}$) :

$$\forall x \in \Omega, \quad \tau_A(x) = \inf\{k \geq 1 : T^k(x) \in A\}.$$

τ_A is the first time that the process hits a given measurable A. We also use the classical probabilistic shorthand notations. We write $\{\tau_A = m\}$ instead of $\{x \in \Omega : \tau_A(x) = m\}$, $T^{-k}(A)$ instead of $\{x \in \Omega : T^k(x) \in A\}$ and $\{X_r^s = x_r^s\}$ instead of $\{X_r = x_r, ..., X_s = x_s\}$. Also we write for two measurable subsets A and B of Ω, the conditional probability of B given A as $\mathbb{P}(B|A) = \mathbb{P}_A(B) = \mathbb{P}(B\cap A)/\mathbb{P}(A)$ and the probability of the intersection of A and B by $\mathbb{P}(A \cap B)$ or $\mathbb{P}(A; B)$. For $A = \{X_0^{n-1} = x_0^{n-1}\}$ and $1 \leq w \leq n$, we write $A^{(w)} = \{X_{n-w}^{n-1} = x_{n-w}^{n-1}\}$ for the event consisting of the *last* w symbols of A. We also write $a \vee b$ for the supremum of two real numbers a and b. We define the periodicity p_A of $A \in \mathcal{C}_n$ as follows :

$$p_A = \inf\{k \in \mathbb{N}^* | A \cap T^{-k}(A) \neq \emptyset\}.$$

p_A is called the principal period of word A. Then, we denote by $\mathcal{R}_p = \mathcal{R}_p(n)$ the set of words $A \in \mathcal{C}_n$ with periodicity p and we also define $\mathcal{B}_n$ as the set of words $A \in \mathcal{C}_n$ with periodicity less than $[n/2]$, where $[.]$ defines the integer part of a real number :

$$\mathcal{R}_p = \{A \in \mathcal{C}_n | p_A = p\}, \mathcal{B}_n = \bigcup_{p=1}^{\left[\frac{n}{2}\right]} \mathcal{R}_p.$$

$\mathcal{B}_n$ is the set of words which are self-overlapping before half their length (see Example 3.3.1). We define $\mathcal{R}(A)$ the set of return times of A which are not a multiple of its periodicity p_A :

$$\mathcal{R}(A) = \left\{k \in \{[n/p_A]p_A + 1, \ldots, n-1\} | A \cap T^{-k}(A) \neq \emptyset\right\}.$$

Let us denote $r_A = \#\mathcal{R}(A)$, the cardinality of the set $\mathcal{R}(A)$. Define also $n_A = \min \mathcal{R}(A)$ if $\mathcal{R}(A) \neq \emptyset$ and $n_A = n$ otherwise. $\mathcal{R}(A)$ is called the set of secondary periods of A and n_A is the smallest secondary period of A. Finally, we introduce the following notation. For an integer $s \in \{0, \ldots, t-1\}$, let $N_s^t = \sum_{i=s}^t \mathbb{1}\{T^{-i}(A)\}$. The random variable N_s^t counts the number of occurrences of A between s and t (we omit the dependence on A). For the sake of simplicity, we also put $N^t = N_0^t$.

Example 3.3.1 *Consider the word $A = aaataaataaa$. Since $p_A = 4$, we have $A \in \mathcal{B}_n$ where $n = 11$. See the following figure to note that $\mathcal{R}(A) = \{9; 10\}$, $r_A = 2$ and $n_A = 9$.*

0	1	2	3	**4**	5	6	7	**8**	**9**	**10**									
a	*a*	*a*	*t*	*a*	*a*	*a*	*t*	*a*	*a*	*a*									
				a	*a*	*a*	*t*	*a*	*a*	*a*	*t*	*a*	*a*	*a*					
								a	*a*	*a*	*t*	*a*	*a*	*a*	*t*	*a*	*a*	*a*	
									a	*a*	*a*	*t*	*a*	*a*	*a*	*t*	*a*	*a*	*a*
										a	*a*	*a*	*t*	*a*	*a*	*a*	*t*	*a*	*a*
											a	*a*	*a*	*t*	*a*	*a*	*a*	*t*	*a*

3.3.2 The Mixing Method

We present a theorem that gives an error bound for the Poisson approximation. Compared to the Chen-Stein method, it has the advantage to present non uniform bounds that strongly control the decay of the tail distribution of N^t.

Theorem 3.3.1 (ψ-mixing approximation) *Let $(X_m)_{m\in\mathbb{Z}}$ be a ψ-mixing process. There exists a constant $C_\psi = 254$, such that for all $A \in \mathcal{C}_n \setminus \mathcal{B}_n$ and all non negative integers k and t, the following inequality holds :*

$$\left|\mathbb{P}(N^t = k) - \frac{e^{-t\mathbb{P}(A)}(t\mathbb{P}(A))^k}{k!}\right| \leq C_\psi e_\psi(A) e^{-(t-(3k+1)n)\mathbb{P}(A)} g_\psi(A,k)$$

$$\text{where } g_\psi(A,k) = \begin{cases} \dfrac{(2\lambda)^{k-1}}{(k-1)!} & k \notin \{\frac{\lambda}{e_\psi(A)}, ..., \frac{2t}{n}\} \\ \dfrac{(2\lambda)^{k-1}}{\left(\frac{\lambda}{e_\psi(A)}\right)!\left(\frac{1}{e_\psi(A)}\right)^{k-\frac{\lambda}{e_\psi(A)}-1}} & k \in \{\frac{\lambda}{e_\psi(A)}, ..., \frac{2t}{n}\} \end{cases},$$

$$e_\psi(A) = \inf_{1\leq w\leq n_A}\left[(r_A+n)\mathbb{P}\left(A^{(w)}\right)(1+\psi(n_A-w))\right],$$

$$\text{and } \lambda = t\mathbb{P}(A)(1+\psi(n)).$$

This result is at the core of our study. It shows an upper bound for the difference between the distribution of the number of occurrences of word A in a sequence of length t and the Poisson distribution of parameter $t\mathbb{P}(A)$. Proof is postponed in Section 3.5.

3.4 Calculation of the Constants

Our goal is to compute a bound as small as possible to control the error between the Poisson distribution and the distribution of the number of occurrences of a word. Thus, we determine the global constant C_ψ appearing in Theorem 3.3.1 by means of intermediary bounds appearing in the proof. General bounds are interesting asymptotically in n, but for biological applications, n is approximately between 10 or 20, which is too small. Then along the proof, we will indicate the intermediary bounds that we compute. Before establishing the proof of that Theorem 3.3.1, we point out here, for easy references, some results of [Aba04], and some other useful results. In [Aba04], these results are given only in the ϕ-mixing context. Moreover exact values of the constants are not given, while these are necessary for practical use of these methods. We provide the values of all the constants appearing in the proofs of these results.

Proposition 3.4.1 (Proposition 11 in [Aba04]) *Let $(X_m)_{m\in\mathbb{Z}}$ be a ψ-mixing process. There exist two finite constants $C_a > 0$ and $C_b > 0$, such that for any n, any word $A \in \mathcal{C}_n$, and any $c \in \left[4n, \frac{1}{2\mathbb{P}(A)}\right]$ satisfying*

$$\psi(c/4) \leq \mathbb{P}\left(\{\tau_A \leq c/4\} \cap \{\tau_A \circ T^{c/4} > c/2\}\right),$$

there exists Δ, with $n < \Delta \leq c/4$, such that for all positive integers k, the following inequalities hold :

$$\left|\mathbb{P}(\tau_A > kc) - \mathbb{P}(\tau_A > c-2\Delta)^k\right| \leq C_a\varepsilon(A)\, k\mathbb{P}(\tau_A > c-2\Delta)^k, \quad (3.1)$$

$$\left|\mathbb{P}(\tau_A > kc) - \mathbb{P}(\tau_A > c)^k\right| \leq C_b\varepsilon(A)\, k\mathbb{P}(\tau_A > c-2\Delta)^k, \quad (3.2)$$

$$\text{with } \varepsilon(A) = \inf_{n\leq\ell\leq\frac{1}{\mathbb{P}(A)}}[\ell\mathbb{P}(A) + \psi(\ell)].$$

Both inequalities provide an approximation of the hitting time distribution by a geometric distribution at any point t of the form $t = kc$. The difference between these distributions is that in 3.1, the geometric term inside the modulus is the same as in the upper bound, while in 3.2, the geometric term inside the modulus is larger than the one in the upper bound. That is, the second bound gives a larger error. We will use both in the proof of Theorem 3.4.1.

Proposition 3.4.2 *We have $C_a = 24$ and $C_b = 25$.*

Proof 1 *For the details of the proof of Proposition 3.4.1, we refer to Proposition 11 in [Aba04]. For any $c \in \left[4n, \frac{1}{2\mathbb{P}(A)}\right]$ and $\Delta \in [n, c/4]$, we denote $\mathcal{N}_j^i = \left\{\tau_A \circ T^{ic+j\Delta} > c - j\Delta\right\}$ and $\mathcal{N} = \{\tau_A > c - 2\Delta\}$ for the sake of simplicity. [Aba04] obtains the following bound :*

$$\forall k \geq 2, \left|\mathbb{P}\left(\tau_A > kc\right) - \mathbb{P}\left(\mathcal{N}\right)^k\right| \leq (a) + (b) + (c), \text{ with}$$

$$(a) = \sum_{j=0}^{k-2} \mathbb{P}\left(\mathcal{N}\right)^j \left|\mathbb{P}\left(\tau_A > (k-j)\,c\right) - \mathbb{P}\left(\tau_A > (k-j-1)\,c; \mathcal{N}_2^{k-j-1}\right)\right|,$$

$$(b) = \sum_{j=0}^{k-2} \mathbb{P}\left(\mathcal{N}\right)^j \left|\mathbb{P}\left(\tau_A > (k-j-1)\,c; \mathcal{N}_2^{k-j-1}\right) - \mathbb{P}\left(\tau_A > (k-j-1)\,c\right)\mathbb{P}\left(\mathcal{N}_2^0\right)\right|,$$

$$(c) = \mathbb{P}\left(\mathcal{N}\right)^{(k-1)} \left|\mathbb{P}\left(\tau_A > c\right) - \mathbb{P}\left(\mathcal{N}\right)\right|.$$

First, for any measurable $B \in \mathcal{F}_{\{(\ell+1)c,(\ell+2)c+n-1\}}$, we have $\mathbb{P}(B) + \psi(\Delta) \leq 3\psi(\Delta) \leq \frac{3}{2}\varepsilon(A)$. We can also remark that $\mathbb{P}(\mathcal{N}) \geq 1/2$. Then, by iteration of the mixing property, we have the following inequality for all $\ell \in \mathbb{N}$:

$$\mathbb{P}\left(\bigcap_{i=0}^{\ell} \mathcal{N}_1^i; B\right) \leq 6\mathbb{P}\left(\mathcal{N}\right)^{\ell+1} \varepsilon\left(A\right).$$

We apply this bound in the inequalities (14) and (15) of [Aba04] to get

$$(a) \leq \sum_{j=0}^{k-2} \mathbb{P}\left(\mathcal{N}\right)^j \left(6\mathbb{P}\left(\mathcal{N}\right)^{k-j-2+1} \varepsilon\left(A\right)\right) = 6(k-1)\varepsilon\left(A\right)\mathbb{P}\left(\mathcal{N}\right)^{(k-1)},$$

$$(b) \leq \sum_{j=0}^{k-2} \mathbb{P}\left(\mathcal{N}\right)^j \left(6\mathbb{P}\left(\mathcal{N}\right)^{k-j-2+1} \varepsilon\left(A\right)\right) = 6(k-1)\varepsilon\left(A\right)\mathbb{P}\left(\mathcal{N}\right)^{(k-1)}.$$

We also have $(c) \leq \mathbb{P}\left(\mathcal{N}\right)^{k-1} \mathbb{P}\left(\mathcal{N}; \tau_A \circ T^{c-2\Delta} \leq 2\Delta\right) \leq \varepsilon\left(A\right)\mathbb{P}\left(\mathcal{N}\right)^{k-1}$.

We obtain (3.1) : $\left|\mathbb{P}\left(\tau_A > kc\right) - \mathbb{P}\left(\mathcal{N}\right)^k\right| \leq 24k\varepsilon\left(A\right)\mathbb{P}\left(\mathcal{N}\right)^k$.

We deduce (3.2) : $\left|\mathbb{P}\left(\tau_A > kc\right) - \mathbb{P}\left(\tau_A > c\right)^k\right| \leq 25k\varepsilon\left(A\right)\mathbb{P}\left(\mathcal{N}\right)^k$.

Then, $C_a = 24$ and $C_b = 25$.

Theorem 3.4.1 (Theorem 1 in [Aba04]) *Let $(X_m)_{m\in\mathbb{Z}}$ be a ψ-mixing process. Then, there exist constants $C_h > 0$ and $0 < \Xi_1 < 1 \leq \Xi_2 < \infty$, such that for all $n \in \mathbb{N}$ and any $A \in \mathcal{C}_n$, there exists $\xi_A \in [\Xi_1, \Xi_2]$, for which the following inequality holds for all $t > 0$:*

$$\left|\mathbb{P}\left(\tau_A > \frac{t}{\xi_A}\right) - e^{-t\mathbb{P}(A)}\right| \leq C_h\varepsilon(A)f_1(A,t),$$

$$\text{with } \varepsilon(A) = \inf_{n \leq \ell \leq \frac{1}{\mathbb{P}(A)}} \left[\ell\mathbb{P}(A) + \psi(\ell)\right] \text{ and } f_1(A,t) = (t\mathbb{P}(A) \vee 1)e^{-t\mathbb{P}(A)}.$$

We prove an upper bound for the distance between the rescaled hitting time and the exponential law of expectation equal to one. The factor $\varepsilon(A)$ in the upper bound shows that the rate of convergence to the exponential law is given by a trade off between the length of this time and the velocity of loosing memory of the process.

Proposition 3.4.3 *We have $C_h = 105$.*

Proof 2 *We fix $c = \frac{1}{2\mathbb{P}(A)}$ and Δ given by Proposition 3.4.1. We define*

$$\xi_A = \frac{-\log \mathbb{P}(\tau_A > c - 2\Delta)}{c\mathbb{P}(A)}.$$

There are three steps in the proof of the theorem. First, we consider t of the form $t = kc$ with k a positive integer. Secondly, we prove the theorem for any t of the form $t = (k + p/q)c$ with k, p positive integers and $1 \leq p \leq q$ with $q = \frac{1}{2\varepsilon(A)}$. We also put $r = (p/q)c$. Finally, we consider the remaining cases. Here, for the sake of simplicity , we do not detail the two first steps (for that, see [Aba04]), but only the last one. Let t be any positive real number.

We write $t = kc + r$, with k a positive integer and r such that $0 \leq r < c$. We can choose a $\bar{t}$ such that $\bar{t} < t$ and $\bar{t} = (k + p/q)c$ with p, q as before. [Aba04] obtains the following bound :

$$\begin{aligned}\left|\mathbb{P}\left(\tau_A > t\right) - e^{-\xi_A \mathbb{P}(A) t}\right| &\leq \left|\mathbb{P}\left(\tau_A > t\right) - \mathbb{P}\left(\tau_A > \bar{t}\right)\right| + \left|\mathbb{P}\left(\tau_A > \bar{t}\right) - e^{-\xi_A \mathbb{P}(A)\bar{t}}\right| \\ &+ \left|e^{-\xi_A \mathbb{P}(A)\bar{t}} - e^{-\xi_A \mathbb{P}(A) t}\right|.\end{aligned}$$

The first term in the triangular inequality is bounded in the following way :

$$\begin{aligned}\left|\mathbb{P}\left(\tau_A > t\right) - \mathbb{P}\left(\tau_A > \bar{t}\right)\right| &= \mathbb{P}\left(\tau_A > \bar{t}; \tau_A \circ T^{\bar{t}} \leq t - \bar{t}\right) \\ &\leq \mathbb{P}\left(\tau_A > kc; \tau_A \circ T^{\bar{t}} \leq \Delta\right) \\ &\leq \mathbb{P}\left(\mathcal{N}\right)^{k-2}\left(\Delta \mathbb{P}(A) + \psi(\Delta)\right)) \\ &\leq 4\mathbb{P}\left(\mathcal{N}\right)^{k} \varepsilon(A) \\ &\leq 4\varepsilon(A) e^{-\xi_A \mathbb{P}(A) t}.\end{aligned}$$

The second term is bounded like in the two first steps of the proof in [Aba04]. We apply inequalities (3.1) and (3.2) to obtain

$$\left|\mathbb{P}\left(\tau_A > \bar{t}\right) - e^{-\xi_A \mathbb{P}(A)\bar{t}}\right| \leq (3 + C_a t \mathbb{P}(A) + C_a + 2C_b)\varepsilon(A) e^{-\xi_A \mathbb{P}(A) t}.$$

Finally, the third term is bounded using the Mean Value Theorem (see for example [Dou96])

$$\left|e^{-\xi_A \mathbb{P}(A)\bar{t}} - e^{-\xi_A \mathbb{P}(A) t}\right| \leq \xi_A \mathbb{P}(A)\left(r - \frac{p}{q}c\right) e^{-\xi_A \mathbb{P}(A)\bar{t}} \leq \varepsilon(A) e^{-\xi_A \mathbb{P}(A) t}.$$

Thus we have $\left|\mathbb{P}\left(\tau_A > t\right) - e^{-\xi_A \mathbb{P}(A) t}\right| \leq 105\varepsilon(A) f_1(A, \xi_A t)$ and the theorem follows by the change of variables $\tilde{t} = \xi_A t$. Then $C_h = 105$.

Lemma 3.4.1 *$(X_m)_{m\in\mathbb{Z}}$ be a ψ-mixing process. Suppose that $B \subseteq A \in \mathcal{F}_{\{0,\ldots,b\}}$, $C \in \mathcal{F}_{\{b+g,\ldots,\infty\}}$ with $b, g \in \mathbb{N}$. The following inequality holds :*

$$\mathbb{P}_A(B \cap C) \leq \mathbb{P}_A(B)\mathbb{P}(C)(1 + \psi(g)).$$

Proof 3 *Since $B \subseteq A$, obviously $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(B \cap C)$. By the ψ-mixing property $\mathbb{P}(B \cap C) \leq \mathbb{P}(B)(\mathbb{P}(C) + \psi(g))$. We divide the above inequality by $\mathbb{P}(A)$ and the lemma follows.*

For all the following propositions and lemmas, we recall that

$$e_\psi(A) = \inf_{1 \leq w \leq n_A}\left[(r_A + n)\mathbb{P}\left(A^{(w)}\right)\left(1 + \psi\left(n_A - w\right)\right)\right].$$

Proposition 3.4.4 *Let $(X_m)_{m\in\mathbb{Z}}$ be a ψ-mixing process. Let $A \in \mathcal{R}_p(n)$. Then the following holds :*
(a) For all $M, M' \geq g \geq n$,

$$\begin{aligned}&\left|\mathbb{P}_A\left(\tau_A > M + M'\right) - \mathbb{P}_A\left(\tau_A > M\right)\mathbb{P}\left(\tau_A > M'\right)\right| \\ \leq\ &\mathbb{P}_A\left(\tau_A > M - g\right) 2g\mathbb{P}(A)\left[1 + \psi(g)\right],\end{aligned}$$

and similarly

$$\begin{aligned}&\left|\mathbb{P}_A\left(\tau_A > M + M'\right) - \mathbb{P}_A\left(\tau_A > M\right)\mathbb{P}\left(\tau_A > M' - g\right)\right| \\ \leq\ &\mathbb{P}_A\left(\tau_A > M - g\right)\left[g\mathbb{P}(A) + 2\psi(g)\right].\end{aligned}$$

(b) For all $t \geq p \in \mathbb{N}$, with $\zeta_A = \mathbb{P}_A(\tau_A > p_A)$,

$$\left|\mathbb{P}_A\left(\tau_A > t\right) - \zeta_A \mathbb{P}\left(\tau_A > t\right)\right| \leq 2e_\psi(A).$$

The above proposition establishes a relation between hitting and return times with an error bound uniform with respect to t. In particular, (b) says that these times coincide if and only if $\zeta_A = 1$, namely, the string A is non-self-overlapping.

Proof 4 *In order to simplify notation, for $t \in \mathbb{Z}$, $\tau_A^{[t]}$ stands for $\tau_A \circ T^t$. We introduce a gap of length g after coordinate M to construct the following triangular inequality*

$$\begin{aligned} & \left|\mathbb{P}_A\left(\tau_A > M+M'\right) - \mathbb{P}_A\left(\tau_A > M\right)\mathbb{P}\left(\tau_A > M'\right)\right| \\ \leq\ & \left|\mathbb{P}_A\left(\tau_A > M+M'\right) - \mathbb{P}_A\left(\tau_A > M; \tau_A^{[M+g]} > M'-g\right)\right| \qquad (3.3) \\ +\ & \left|\mathbb{P}_A\left(\tau_A > M; \tau_A^{[M+g]} > M'-g\right) - \mathbb{P}_A\left(\tau_A > M\right)\mathbb{P}\left(\tau_A > M'-g\right)\right| \qquad (3.4) \\ +\ & \mathbb{P}_A\left(\tau_A > M\right)\left|\mathbb{P}\left(\tau_A > M'-g\right) - \mathbb{P}\left(\tau_A > M'\right)\right|. \qquad (3.5) \end{aligned}$$

Term (3.3) is bounded with Lemma 3.4.1 by

$$\mathbb{P}_A\left(\tau_A > M; \tau_A^{[M]} \leq g\right) \leq \mathbb{P}_A\left(\tau_A > M-g\right) g\mathbb{P}(A)\left[1+\psi(g)\right].$$

Term (3.4) is bounded using the ψ-mixing property by $\mathbb{P}_A\left(\tau_A > M\right)\psi(g)$. The modulus in (3.5) is bounded using stationarity by $\mathbb{P}\left(\tau_A \leq g\right) \leq g\mathbb{P}(A)$. This ends the proof of both inequalities of item (a).
Item (b) for $t \geq 2n$ is proven similarly to item (a) with $t = M+M'$, $M = p$, and $g = w$ with $1 \leq w \leq n_A$. Consider now $p \leq t < 2n$.

$$\zeta_A - \mathbb{P}_A\left(\tau_A > t\right) = \mathbb{P}_A\left(p < \tau_A \leq t\right) = \mathbb{P}_A\left(\tau_A \in \mathcal{R}(A) \cup \left(n \leq \tau_A \leq t\right)\right) \leq e_\psi(A).$$

First and second equalities follow by definition of τ_A and $\mathcal{R}(A)$. The inequality follows by Lemma 3.4.1.

Let $\zeta_A = \mathbb{P}_A(\tau_A > p_A)$ and $h = 1/(2\mathbb{P}(A)) - 2\Delta$, then $\xi_A = -2\log\mathbb{P}(\tau_A > h)$.

Lemma 3.4.2 *Let $(X_m)_{m\in\mathbb{Z}}$ be a ψ-mixing process. Then the following inequality holds :*

$$|\xi_A - \zeta_A| \leq 11 e_\psi(A).$$

Hence, we have

$$\zeta_A - 11e_\psi(A) \leq \xi_A \leq \zeta_A + 11e_\psi(A).$$

Proof 5

$$\begin{aligned} \mathbb{P}\left(\tau_A > h\right) &= \prod_{i=1}^{h} \mathbb{P}\left(\tau_A > i | \tau_A > i-1\right) = \prod_{i=1}^{h}\left(1 - \mathbb{P}\left(T^{-i}(A) | \tau_A > i-1\right)\right) \\ &= \prod_{i=1}^{h}\left(1-\rho_i\mathbb{P}(A)\right), \end{aligned}$$

where $\rho_i \stackrel{def}{=} \dfrac{\mathbb{P}_A\left(\tau_A > i-1\right)}{\mathbb{P}\left(\tau_A > i-1\right)}$. Therefore

$$\begin{aligned} & \left|\xi_A + 2\sum_{i=1}^{p_A}\log(1-\rho_i\mathbb{P}(A)) - 2\sum_{i=p_A+1}^{h}\zeta_A\mathbb{P}(A)\right| \\ \leq\ & 2\sum_{i=p_A+1}^{h}\left|-\log(1-\rho_i\mathbb{P}(A)) - \zeta_A\mathbb{P}(A)\right|. \end{aligned}$$

The above modulus is bounded by

$$\left|-\log(1-\rho_i\mathbb{P}(A)) - \rho_i\mathbb{P}(A)\right| + \left|\rho_i - \zeta_A\right|\mathbb{P}(A).$$

Now note that $|y - (1-e^{-y})| \leq (1-e^{-y})^2$ for $y > 0$ small enough. Apply it with $y = -\log(1-\rho_i\mathbb{P}(A))$ to bound the most left term of the above expression by $(\rho_i\mathbb{P}(A))^2$. Further by Proposition 3.4.4 (b) and the fact that $\mathbb{P}\left(\tau_A > h\right) \geq 1/2$ we have

$$|\rho_i - \zeta_A| \leq \frac{2e_1(A)}{\mathbb{P}\left(\tau_A > h\right)} \leq 4e_\psi(A).$$

for all $i = p_A+1, \ldots, h$. Yet as before

$$-\sum_{i=1}^{p_A}\log(1-\rho_i\mathbb{P}(A)) \leq p_A\left(\rho_i\mathbb{P}(A) + (\rho_i\mathbb{P}(A))^2\right) \leq e_\psi(A).$$

Finally, by definition of h

$$\left| 2 \sum_{i=p_A+1}^{h} \zeta_A \mathbb{P}(A) - \zeta_A \right| \leq 4\Delta\mathbb{P}(A) + 2p_A\mathbb{P}(A) \leq 6e_\psi(A).$$

This ends the proof of the lemma.

Proposition 3.4.5 *Let $(X_m)_{m\in\mathbb{Z}}$ be a ψ-mixing process. Then the following inequality holds :*

$$|\mathbb{P}(\tau_A > t) - e^{-t\mathbb{P}(A)}| \leq C_p e_\psi(A)(t\mathbb{P}(A) \vee 1)e^{-(\zeta_A - 11e_\psi(A))t\mathbb{P}(A)}.$$

Proof 6 *We bound the first term with Theorem 3.4.1 and the second with Lemma 3.4.2 :*

$$\begin{array}{rcl} |\mathbb{P}(\tau_A > t) - e^{-t\mathbb{P}(A)}| & \leq & |\mathbb{P}(\tau_A > t) - e^{-\xi_A t\mathbb{P}(A)}| + |e^{-\xi_A t\mathbb{P}(A)} - e^{-t\mathbb{P}(A)}| \\ |\mathbb{P}(\tau_A > t) - e^{-\xi_A t\mathbb{P}(A)}| & \leq & C_h\varepsilon(A)e^{-\xi_A t\mathbb{P}(A)} \leq C_h e_\psi(A)e^{-(\zeta_A - 11e_\psi(A))t\mathbb{P}(A)} \\ |e^{-\xi_A t\mathbb{P}(A)} - e^{-t\mathbb{P}(A)}| & \leq & t\mathbb{P}(A)|\xi_A - 1|e^{-\min\{1,\xi_A\}t\mathbb{P}(A)} \\ & \leq & 11t\mathbb{P}(A)e_\psi(A)e^{-(\zeta_A - 11e_\psi(A))t\mathbb{P}(A)}. \end{array}$$

This ends the proof of the proposition with $C_p = C_h + 11$.

Definition 3.4.1 *Given $A \in \mathcal{C}_n$, we define for $j \in \mathbb{N}$, the j-th occurrence time of A as the random variable $\tau_A^{(j)} : \Omega \to \mathbb{N} \cup \{\infty\}$, defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ as follows : for any $x \in \Omega$, $\tau_A^{(1)}(x) = \tau_A(x)$ and for $j \geq 2$,*

$$\tau_A^{(j)}(x) = \inf\{k > \tau_A^{(j-1)}(\omega) : T^k(x) \in A\}.$$

Proposition 3.4.6 *Let $(X_m)_{m\in\mathbb{Z}}$ be a ψ-mixing process. Then, for all $A \notin \mathcal{B}_n$, all $k \in \mathbb{N}$, and all $0 \leq t_1 < t_2 < ... < t_k \leq t$ for which $\min_{2\leq j\leq k}\{t_j - t_{j-1}\} > 2n$, there exists a positive constant C_1 independent of A, n, t and k such that*

$$\begin{array}{rcl} & & \left| \mathbb{P}\left(\bigcap_{j=1}^{k} \left(\tau_A^{(j)} = t_j\right); \tau_A^{(k+1)} > t \right) - \mathbb{P}(A)^k \prod_{j=1}^{k+1} \mathcal{P}_j \right| \\ & \leq & C_1 k(\mathbb{P}(A)(1+\psi(n)))^k e_\psi(A)e^{-(t-(3k+1)n)\mathbb{P}(A)} \end{array}$$

where $\mathcal{P}_j = \mathbb{P}(\tau_A > (t_j - t_{j-1}) - 2n)$.

Proof 7 *We will show this proposition by induction on k. We put $\Delta_j = t_j - t_{j-1}$ for $j = 2, ..., k$, $\Delta_1 = t_1$ and $\Delta_{k+1} = t - t_k$. Firstly, we note that by stationarity*

$$\mathbb{P}(\tau_A = t) = \mathbb{P}(A; \tau_A > t - 1).$$

For $k = 1$, by a triangular inequality we obtain

$$\begin{array}{rcll} & & \left| \mathbb{P}\left(\tau_A = t_1; \tau_A^{(2)} > t\right) - \mathbb{P}(A)\prod_{j=1}^{2}\mathcal{P}_j \right| & \\ & \leq & \left| \mathbb{P}\left(\tau_A = t_1; \tau_A^{(2)} > t\right) - \mathbb{P}\left(\tau_A = t_1; N_{t_1+2n}^t = 0\right) \right| & (3.6) \\ & + & \left| \mathbb{P}\left(\tau_A = t_1; N_{t_1+2n}^t = 0\right) - \mathbb{P}\left(\tau_A = t_1\right)\mathcal{P}_2 \right| & (3.7) \\ & + & \left| \mathbb{P}(A; \tau > t_1 - 1) - \mathbb{P}\left(A; N_{2n}^{t_1-1} = 0\right) \right| \mathcal{P}_2 & (3.8) \\ & + & \left| \mathbb{P}\left(A; N_{2n}^{t_1-1} = 0\right)\mathcal{P}_2 - \mathbb{P}(A)\prod_{j=1}^{2}\mathcal{P}_j \right|. & (3.9) \end{array}$$

Term (3.6) is equal to $\mathbb{P}\left(\tau_A = t_1; \bigcup_{i=t_1+1}^{t_1+2n} T^{-i}(A); N_{t_1+2n}^t = 0\right)$ and then

$$(3.6) = \mathbb{P}\left(A; \bigcup_{i\in\mathcal{R}(A)\cup i=1}^{2n} T^{-i}(A); N_{2n}^t = 0\right).$$

Since $A \notin \mathcal{B}_n$, for $1 \leq i < p_A$, the above probability is zero. Thus, using mixing property

$$\begin{aligned}(3.6) \quad &\leq \mathbb{P}\left(A; \bigcup_{i \in \mathcal{R}(A) \cup i = p_A}^{2n} T^{-i}(A); N_{2n}^t = 0\right) \\ &\leq 2\mathbb{P}(A)\mathbb{P}(A)(r_A + n)(1 + \psi(n))\mathbb{P}\left(N_{2n}^t = 0\right) \\ &\leq 2\mathbb{P}(A)e_\psi(A)e^{-(t-(3k+1)n)\mathbb{P}(A)}.\end{aligned}$$

Term (3.7) is bounded using ψ-mixing property

$$\begin{aligned}(3.7) \quad &\leq \psi(n)(1 + \psi(n))\mathbb{P}(A)\mathcal{P}_1\mathcal{P}_2 \\ &\leq \psi(n)\mathbb{P}(A)e_\psi(A)e^{-(t-(3k+1)n)\mathbb{P}(A)}.\end{aligned}$$

Analogous computations are used to bound terms (3.8) and (3.9).

Now, let us suppose that the proposition holds for $k-1$ and let us prove it for k. We put $\mathcal{S}_i = \{\tau_A^{(i)} = t_i\}$. We use a triangular inequality again to bound the term in the left hand side of the inequality of the proposition by a sum of five terms :

$$\left|\mathbb{P}\left(\bigcap_{j=1}^{k}\left(\tau_A^{(j)} = t_j\right); \tau_A^{(k+1)} > t\right) - \mathbb{P}(A)^k \prod_{j=1}^{k+1}\mathcal{P}_j\right| \leq I + II + III + IV + V.$$

$$\begin{aligned}I &= \left|\mathbb{P}\left(\bigcap_{j=1}^{k}\mathcal{S}_j; \tau_A^{(k+1)} > t\right) - \mathbb{P}\left(\bigcap_{j=1}^{k-1}\mathcal{S}_j; N_{t_{k-1}+1}^{t_k-2n} = 0; T^{-t_k}(A); N_{t_k+1}^t = 0\right)\right| \\ &= \mathbb{P}\left(\bigcap_{j=1}^{k-1}\mathcal{S}_j; N_{t_{k-1}+1}^{t_k-2n} = 0; \bigcup_{i=t_k-2n+1}^{t_k-1} T^{-i}(A); T^{-t_k}(A); N_{t_k+1}^t = 0\right) \\ &\leq \left(\mathbb{P}(A)(1+\psi(n))\right)^k(1-\psi(n))\left(np_A + (r_A+n)\mathbb{P}(A^{(w)})\right)e^{-(t-(3k+1)n)\mathbb{P}(A)}, \\ II &= \left|\mathbb{P}\left(\bigcap_{j=1}^{k-1}\mathcal{S}_j; N_{t_{k-1}+1}^{t_k-2n} = 0; T^{-t_k}(A); N_{t_k+1}^t = 0\right)\right. \\ &\quad \left. - \mathbb{P}\left(\bigcap_{j=1}^{k-1}\mathcal{S}_j; N_{t_{k-1}+1}^{t_k-2n} = 0\right)\mathbb{P}\left(A; N_1^{t-t_k} = 0\right)\right| \\ &\leq \mathbb{P}\left(\bigcap_{j=1}^{k-1}; N_{t_{k-1}+1}^{t_k-2n} = 0\right)\mathbb{P}\left(A; N_1^{t-t_k} = 0\right)\psi(n) \\ &\leq \left(\mathbb{P}(A)(1+\psi(n))\right)^k\psi(n)e^{-(t-(3k+1)n)\mathbb{P}(A)}, \\ III &= \left|\mathbb{P}\left(\bigcap_{j=1}^{k-1}\mathcal{S}_j; N_{t_{k-1}+1}^{t_k-2n} = 0\right) - \mathbb{P}\left(\bigcap_{j=1}^{k-1}\mathcal{S}_j; N_{t_{k-1}+1}^{t_k-1} = 0\right)\right|\mathbb{P}\left(A; N_1^{t-t_k} = 0\right) \\ &\leq \mathbb{P}\left(\bigcap_{j=1}^{k-1}\mathcal{S}_j; N_{t_{k-1}+1}^{t_k-2n} = 0; \bigcup_{t_k-2n+1}^{t_k-1} T^{-i}(A)\right)\mathbb{P}(A) \\ &\leq 2\mathbb{P}(A)\left(\mathbb{P}(A)(1+\psi(n))\right)^k e^{-(t-(3k+1)n)\mathbb{P}(A)}.\end{aligned}$$

We use the inductive hypothesis for the term IV and the case with $k = 1$ for the term V.

$$\begin{aligned}IV &= \left|\mathbb{P}\left(\bigcap_{j=1}^{k-1}\mathcal{S}_j; N_{t_{k-1}+1}^{t_k-1} = 0\right) - \mathbb{P}(A)^{k-1}\prod_{j=1}^{k}\mathcal{P}_j\right|\mathbb{P}\left(A; N_1^{t-t_k} = 0\right) \\ &\leq C_1(k-1)\left(\mathbb{P}(A)(1+\psi(n))\right)^k e_\psi(A)e^{-(t-(3k+1)n)\mathbb{P}(A)}, \\ V &= \mathbb{P}(A)^{k-1}\prod_{j=1}^{k}\mathcal{P}_j\left|\mathbb{P}\left(A; N_1^{t-t_k} = 0\right) - \mathbb{P}(A)\mathcal{P}_{k+1}\right| \\ &\leq 2\left(\mathbb{P}(A)(1+\psi(n))\right)^k e_\psi(A)e^{-(t-(3k+1)n)\mathbb{P}(A)}.\end{aligned}$$

Finally, we obtain

$$I + II + III + IV + V \leq (3 + C_1(k-1) + 2)\left(\mathbb{P}(A) + \psi(n)\right)^k e_\psi(A).$$

To conclude the proof, it is sufficient that $C_1 k = 3 + C_1(k-1) + 2$, therefore $C_1 = 5$. This ends the proof of the proposition.

3.5 Proof of Theorem 3.3.1

In this section, we prove the main result of our work (see Section 3.3.2) : an upper bound for the difference between the exact distribution of the number of occurrences of word A and the Poisson distribution of parameter $t\mathbb{P}(A)$. Throughout the proof, we will note in italic the terms computed by our software PANOW (see Section 3.6.1).

Proof 8 *For $k = 0$, the result comes from Proposition 3.4.5 ($\mathbb{P}(N^t = 0) = \mathbb{P}(\tau_A > t)$).*
For $k > 2t/n$, since $A \notin \mathcal{B}_n$, we have $\mathbb{P}(N^t = k) = 0$. Hence,

$$\begin{aligned}\left|\mathbb{P}(N^t = k) - \frac{e^{-t\mathbb{P}(A)}(t\mathbb{P}(A))^k}{k!}\right| &= \frac{e^{-t\mathbb{P}(A)}(t\mathbb{P}(A))^k}{k!} \\ &\leq \frac{(t\mathbb{P}(A))^{k-1}}{(k-1)!}\frac{t\mathbb{P}(A)}{k} \\ &\leq \frac{1}{2}\frac{(t\mathbb{P}(A))^{k-1}}{(k-1)!}e_\psi(A).\end{aligned}$$

Indeed, since $\frac{t}{k} < \frac{n}{2}$ then $\frac{t\mathbb{P}(A)}{k} < \frac{n\mathbb{P}(A)}{2} \leq \frac{e_\psi(A)}{2}$.

Now, let us consider $1 \leq k \leq 2t/n$. We consider a sequence which contains exactly k occurrences of A. These occurrences can be isolated or can be in clumps. We define the following set :

$$\mathcal{T} = \mathcal{T}(t_1, t_2, ..., t_k) = \left\{ \bigcap_{j=1}^{k} (\tau_A^{(j)} = t_j); \tau_A^{(k+1)} > t \right\}.$$

We recall that we put $\mathcal{P}_j = \mathbb{P}(\tau_A > (t_j - t_{j-1}) - 2n)$, $\Delta_j = t_j - t_{j-1}$ for $j = 2, ..., k$, $\Delta_1 = t_1$ and $\Delta_{k+1} = t - t_k$. Define $I(\mathcal{T}) = \min_{2\leq j\leq k}\{\Delta_j\}$. We say that the occurrences of A are isolated if $I(\mathcal{T}) \geq 2n$ and we say that there exists at least one clump if $I(\mathcal{T}) < 2n$. We also denote

$$B_k = \{\mathcal{T} | I(\mathcal{T}) < 2n\} \quad \text{and} \quad G_k = \{\mathcal{T} | I(\mathcal{T}) \geq 2n\}.$$

The set $\{N^t = k\}$ is the disjoint union between B_k and G_k, then

$$\mathbb{P}(N^t = k) = \mathbb{P}(B_k) + \mathbb{P}(G_k),$$

$$\left|\mathbb{P}(N^t = k) - \frac{e^{-t\mathbb{P}(A)}(t\mathbb{P}(A))^k}{k!}\right| \leq \mathbb{P}(B_k) + \left|\mathbb{P}(G_k) - \frac{e^{-t\mathbb{P}(A)}(t\mathbb{P}(A))^k}{k!}\right|.$$

We will prove an upper bound for the two quantities on the right hand side of the above inequality to conclude the proof of the theorem.

We prove an upper bound for $\mathbb{P}(\mathbf{B_k})$. *Define $C(\mathcal{T}) = \sum_{j=2}^{k} \mathbb{1}_{\{\Delta_j > 2n\}} + 1$. $C(\mathcal{T})$ computes how many clusters there are in a given $\mathcal{T}$. Suppose that $\mathcal{T}$ is such that $C(\mathcal{T}) = 1$ and fix the position t_1 of the first occurrence of A. Further, each occurrence inside the cluster (with the exception of the most left one which is fixed at t_1) can appear at distance d of the previous one, with $p_A \leq d \leq 2n$. Therefore, the ψ-mixing property leads to the bound*

$$\begin{aligned}\mathbb{P}\left(\bigcup_{t_2,\ldots,t_k} \mathcal{T}(t_1, t_2, \ldots, t_k)\right) &\leq \mathbb{P}\left(\bigcap_{j=1}^{k} \bigcup_{\substack{n/2\leq t_{i+1}-t_i\leq 2n; \\ i=2,\ldots,k}} T^{-t_j}(A)\right) \\ &\leq \mathbb{P}(A)e_\psi(A)^{k-1}e_\psi(A)e^{-(t-(3k+1)n)\mathbb{P}(A)}.\end{aligned} \tag{3.1}$$

Suppose now that $\mathcal{T}$ is such that $C(\mathcal{T}) = i$. Assume also that the most left occurrence of the i clusters of $\mathcal{T}$ occurs at $t(1), \ldots, t(i)$, with $1 \leq t(1) < \ldots < t(i) \leq t$ fixed. By the same argument used above, we have the inequalities

$$\begin{aligned}&\mathbb{P}\left(\bigcup_{\{t_1,\ldots,t_k\}\setminus\{t(1),\ldots,t(i)\}} \mathcal{T}(t_1, \ldots, t_k)\right) \\ \leq\ &(\mathbb{P}(A)(1 + \psi(n)))^{i-1}e_\psi(A)^{k-i}e^{-(t-(3k+1)n)\mathbb{P}(A)}.\end{aligned}$$

To obtain an upper bound for $\mathbb{P}(B_k)$ we must sum the above bound over all $\mathcal{T}$ such that $C(\mathcal{T}) = i$ with i running from 1 to $k-1$. Fixed $C(\mathcal{T}) = i$, the locations of the most left occurrences of A of each one of the i clusters can be chosen in at most C_t^i many ways. The cardinality of each one of the i clusters can be arranged in C_{k-1}^{i-1} many ways. (This corresponds to breaking the interval $(1/2, k+1/2)$ in i intervals at points chosen from $\{1+1/2, \ldots, k-1/2\}$.) Collecting these informations, we have that $\mathbb{P}(B_k)$ is bounded by

$$\begin{aligned}
&\sum_{i=1}^{k-1} C_t^i C_{k-1}^{i-1} (\mathbb{P}(A)(1+\psi(n)))^i e_\psi(A)^{k-i} e^{-(t-(3k+1)n)\mathbb{P}(A)} \\
\leq\ & e^{-(t-(3k+1)n)\mathbb{P}(A)} e_\psi(A)^k \max_{1\leq i\leq k-1} \frac{(\lambda/e_\psi(A))^i}{i!} \sum_{i=1}^{k-1} C_{k-1}^{i-1} \\
\leq\ & e^{-(t-(3k+1)n)\mathbb{P}(A)} e_\psi(A) \begin{cases} \dfrac{(2\lambda)^{k-1}}{(k-1)!} & k < \frac{\lambda}{e_\psi(A)} \\ \dfrac{(2\lambda)^{k-1}}{\left(\frac{\lambda}{e_\psi(A)}\right)! \left(\frac{\lambda}{e_\psi(A)}\right)^{k-1-\frac{\lambda}{e_\psi(A)}}} & k \geq \frac{\lambda}{e_\psi(A)} \end{cases}.
\end{aligned}$$

This ends the proof of the bound for $\mathbb{P}(B_k)$.
We compute $\mathbb{P}(B_k) \leq \sum_{i=1}^{k-1} C_t^i C_{k-1}^{i-1} (\mathbb{P}(A)(1+\psi(n)))^i e_\psi(A)^{k-i} e^{-(t-(3k+1)n)\mathbb{P}(A)}$.

We prove an upper bound for $\left|\mathbb{P}(G_k) - \dfrac{e^{-t\mathbb{P}(A)}(t\mathbb{P}(A))^k}{k!}\right|$***.*** *It is bounded by four terms by the triangular inequality*

$$\sum_{T\in G_k} \left| \mathbb{P}\left(\bigcap_{j=1}^{k} \left(\tau_A^{(j)} = t_j\right); \tau_A^{(k+1)} > t \right) - \mathbb{P}(A)^k \prod_{j=1}^{k+1} \mathcal{P}_j \right| \tag{3.2}$$

$$+ \sum_{T\in G_k} \mathbb{P}(A)^k \left| \prod_{j=1}^{k+1} \mathcal{P}_j - \prod_{j=1}^{k+1} e^{-(\Delta_j - 2n)\mathbb{P}(A)} \right| \tag{3.3}$$

$$+ \sum_{T\in G_k} \mathbb{P}(A)^k \left| e^{-(t-2(k+1)n)\mathbb{P}(A)} - e^{-t\mathbb{P}(A)} \right| \tag{3.4}$$

$$+ \left| \frac{\#G_k k!}{t^k} \frac{e^{-t\mathbb{P}(A)}(t\mathbb{P}(A))^k}{k!} - \frac{e^{-t\mathbb{P}(A)}(t\mathbb{P}(A))^k}{k!} \right|. \tag{3.5}$$

We will bound these terms to obtain Theorem 3.3.1.
First, we bound the cardinal of G_k

$$\#G_k \leq C_t^k \leq \frac{t^k}{k!}.$$

Term (3.2) is bounded with Proposition 3.4.6

$$(3.2) \leq C_1 \frac{t^k}{(k-1)!} (\mathbb{P}(A)(1+\psi(n)))^k e_\psi(A) e^{-(t-(3k+1)n)\mathbb{P}(A)}.$$

Term (3.3) is bounded with Proposition 3.4.5

$$\begin{aligned}
(3.3) \quad &\leq \frac{t^k}{k!} \mathbb{P}(A)^k \sum_{j=1}^{k+1} \prod_{i=1}^{j-1} \mathcal{P}_i \left| \mathcal{P}_j - e^{-(\Delta_j - 2n)\mathbb{P}(A)} \right| \prod_{i=j+1}^{k+1} e^{-(\Delta_i - 2n)\mathbb{P}(A)} \\
&\leq \frac{t^k}{k!} \mathbb{P}(A)^k (k+1) C_p e_\psi(A) e^{-(\zeta_A - 11 e_\psi(A))t\mathbb{P}(A)} \\
&\leq 2C_p \frac{(t\mathbb{P}(A))^k}{(k-1)!} e_\psi(A) e^{-(\zeta_A - 11 e_\psi(A))t\mathbb{P}(A)}
\end{aligned}$$

where C_p is defined in Proposition 3.4.5.
We compute

$$(3.3) \quad \leq \quad \frac{(t\mathbb{P}(A))^k}{(k-1)!} \frac{k+1}{k}$$

$$\left[(8 + C_a t\mathbb{P}(A) + C_a + 2C_b)\varepsilon(A) + 11t\mathbb{P}(A)e_\psi(A)\right] e^{-(\zeta_A - 11e_\psi(A))t\mathbb{P}(A)}.$$

Term (3.4) is bounded by

$$(3.4) \leq \frac{t^k}{k!}\mathbb{P}(A)^k(k+1)2n\mathbb{P}(A)e^{-t\mathbb{P}(A)}e^{2(k+1)n\mathbb{P}(A)}.$$

To bound term (3.5), we bound the following difference

$$\left|\frac{\#G_k k!}{t^k} - 1\right| \leq \left|\frac{(t-k(4n))^k}{t^k} - 1\right| \leq \frac{k\,(k+4n)}{t}.$$

Then, we have

$$(3.5) \leq \frac{k\,(k+4n)}{t}\frac{e^{-t\mathbb{P}(A)}(t\mathbb{P}(A))^k}{k!}.$$

Now, we just have to add the five bounds to obtain the theorem with the constant $C_\psi = 1 + C_1 + 2C_p + 8 + 8$. Proposition 3.4.6 shows that $C_1 = 5$ and Proposition 3.4.5 with Theorem 3.4.1 that $C_p = 116$. Then, we prove the theorem with $C_\psi = 254$.

3.6 Biological Applications

With the explicit value of the constant C_ψ of Theorem 3.3.1, and more particularly thanks to all the intermediary bounds given in the proof of this theorem, we can develop an algorithm to apply this formula to the study of rare words in biological sequences. In order to compare different methods, we also compute the bounds corresponding to a ϕ-mixing, process for which a proof of Poisson approximation is given in (Abadi and Vergne, in preparation). Let us recall the definition of such a mixing process.

Definition 3.6.1 *Let $\phi = (\phi(\ell))_{\ell \geq 0}$ be a sequence decreasing to zero. We say that $(X_m)_{m\in\mathbb{Z}}$ is a ϕ-mixing process if for all integers $\ell \geq 0$, the following holds*

$$\sup_{n\in\mathbb{N}, B\in\mathcal{F}_{\{0,..,n\}}, C\in\mathcal{F}_{\{n\geq 0\}}} \frac{|\mathbb{P}(B\cap T^{-(n+\ell+1)}(C)) - \mathbb{P}(B)\mathbb{P}(C)|}{\mathbb{P}(B)} = \phi(\ell),$$

where the supremum is taken over the sets B and C, such that $\mathbb{P}(B) > 0$.

Note that obviously, ψ-mixing implies ϕ-mixing. Then, we obtain two new methods for the detection of over- or under-represented words in biological sequences and we compare them to the Chen-Stein method.

We recall that Markov models are ψ-mixing processes and then also ϕ-mixing processes. Then, we first need to know the functions ψ and ϕ for a Markov model. It turns out that we can use

$$\psi(\ell) = \phi(\ell) = K\nu^\ell \text{ with } K > 0 \text{ and } 0 < \nu < 1,$$

where K and ν have to be estimated (see [MT93]). There are several estimations of K and ν. We choose ν equal to the second eigenvalue of the transition matrix of the model and $K = \left(\inf_{j\in\{1,\ldots,|\mathcal{A}|^k\}} \mu_j\right)^{-1}$ where $|\mathcal{A}|$ is the alphabet size, k the order of the Markov model and μ the stationary distribution of the Markov model.

We recall that we aim at guessing a relevant biological role of a word in a sequence using its number of occurrences. Thus we compare the number of occurrences expected in the Markov chain that models the sequence and the observed number of occurrences. It is recommended to choose a degree of significance s to quantify this relevance. We fix arbitrarily a degree of significance and we want to calculate the smallest number of occurrences u necessary for $\mathbb{P}(N > u) < s$, where N is the number of occurrences of the studied word. If the number of occurrences counted in the sequence is larger than this u, we can consider the word to be relevant with a degree of significance s. We have

$$\mathbb{P}(N > u) \leq \sum_{k=u}^{+\infty} (\mathbb{P}_{\mathcal{P}}(N = k) + Error(k))$$

where $\mathbb{P}_{\mathcal{P}}(N = k)$ is the probability under the Poisson model that N is equal to k and $Error(k)$ is the error between the exact distribution and its Poisson approximation, bounded using Theorem 3.3.1. Then, we search the smallest threshold u such that

$$\sum_{k=u}^{+\infty} (\mathbb{P}_{\mathcal{P}}(N = k) + Error(k)) < s. \tag{3.1}$$

Then, we have $\mathbb{P}(N > u) < s$ and we consider the word relevant with a degree of significance s if it appears more than u times in the sequence.

In order to compare the different methods, we compare the thresholds that they give. Obviously, the smaller the degree of significance, the more relevant the studied word is. But for a fixed degree of significance, the best method is the one which gives the smallest threshold u. Indeed, to give the smallest u is equivalent to give the smallest error in the tail of the distribution between the exact distribution of the number of occurrences of word A and the Poisson distribution with parameter $t\mathbb{P}(A)$.

3.6.1 Software Availability

We developed `PANOW`, dedicated to the determination of threshold u for given words. This software is written in ANSI `C++` and developed on x86 GNU/Linux systems with GCC 3.4, and successfully tested with GCC latest versions on Sun and Apple Mac OSX systems. It relies on `seq++` library ([MBR+05]).

Compilation and installation are compliant with the GNU standard procedure. It is available at http://stat.genopole.cnrs.fr/sg/software/panow/. On-line documentation is also available. `PANOW` is licensed under the GNU General Public License (http://www.gnu.org).

3.6.2 Comparisons between the three different Methods

Comparisons using synthetic Data.

We can compare the mixing methods and the Chen-Stein method through the values of threshold u obtained with `PANOW` using (Abadi and Vergne, in preparation) in the first case and [RS98] in the second one. We recall that the method which gives the smallest threshold u is the best method for a fixed degree of significance. Table 3.1 offers a good outline of the possibilities and limits of each method. It displays some results on different words randomly selected (no biological meaning for any of these words). Table 3.1 has been obtained with an order

TAB. 3.1 – Table of thresholds u obtained by the three methods (sequence length t equal to 10^6). For each one of the three methods and for each word, we compute the threshold which permits to consider the word as an over-represented word or not, for degree of significance s equal to 0.1 or 0.01. IMP means that the method can not return a result.

Words	$t = 10^6$					
	$s = 0.1$			$s = 0.01$		
	CS	ϕ	ψ	CS	ϕ	ψ
cccg	IMP	IMP	IMP	IMP	IMP	IMP
aagcgc	IMP	1301	378	IMP	1304	392
cgagcttc	18	38	18	IMP	40	22
ttgggctg	14	27	14	18	29	17
gtgcggag	16	32	16	22	34	20
agcaaata	19	39	19	IMP	41	23

one Markov model using a random transition matrix and for a degree of significance of 0.1 and 0.01. IMP means that the method can not return a result. There are several reasons for that and we explain them in the following paragraph. Analysing many results, we notice some differences between the methods.

Firstly, none of the methods gives us a result in all the cases. We recall that the Chen Stein method gives a bound (CS) using the total variation distance. If the degree of significance s that we choose is smaller than the bound of Chen-Stein, we never find a threshold u such that

$$CS + \sum_{k=u}^{+\infty} \mathbb{P}_{\mathcal{P}}(N = k) < s.$$

Then, each time that the given bound is higher than the significance degree, use of the Chen Stein method is impossible. Therefore there are many examples that we can not study with this method. Obviously, it is interesting to have a small degree of significance s and that may be impossible by this restriction of the Chen-Stein method. For example, this problem appears for the words `aagcgc` and `cgagcttc` in Table 3.1. For this second word, the Chen-Stein bound is equal to 0.0107954. Hence, we can use this method for a significance degree s equal to 0.1 but not for a significance degree of 0.01. The same phenomena appears for the word `agcaaata` (the Chen-Stein bound is equal to 0.0120193).

The ϕ- and ψ-mixing methods are not based on the total variation distance. Then, whatever the degree of significance s and if the studied word satisfies the three following weak properties, we always give a threshold u, contrary to the Chen Stein method. In spite of these three conditions, our methods enable us to study a much broader panel of words than the Chen-Stein method. Indeed, for these two methods, the only problematic cases arise either when function e_ψ (see Theorem 3.3.1) is larger than 1 or for a "high" parameter of the Poisson distribution ("high" means larger than 500) or when the word periodicity is smaller than half its length (see assumptions in Theorem 3.3.1 : $A \notin \mathcal{B}_n$). In fact, the first case does not occur very frequently (in any case in Table 3.1). The reason why the function e_ψ (or a similar function in the ϕ-mixing case) has to be smaller than 1 is that, for numerical reasons, the error term has to be decreasing with the number of occurrences k and without this condition on e_ψ we can not ensure this decrease. We have to compute error terms for a finite number of values of k but in order to reduce the computation time, when error term becomes smaller than a certain value (we choose 10^{-300}), we suppose all the following error terms equals to this value. That is why error term has to be decreasing. The second problem, a "high" parameter of the Poisson distribution, is just a computational difficulty and once again it does not occur very frequently (only for the word `cccg` in Table 3.1 for instance). We would like to insist on the main advantage of our methods : we can fix any significance degree s and, except in the very rare cases mentioned above, we will find a threshold u, contrary to the Chen-Stein method.

Also, we can use our methods for any Markov chain order. Indeed, `PANOW` runs fast enough contrary to the R program used to compute the Chen-Stein bound of [RS98]. Note that, in program `PANOW`, we give another method to compute the Chen-Stein bound (see [Aba01b]) and this method gives approximately the same Chen-Stein bound.

The second main observation we can make is that, when it works, the Chen-Stein method gives either a similar threshold u than the ψ-mixing method, or a smaller one. This means that the ψ-mixing method out-performs the Chen-Stein method.

Thirdly we notice that the ψ-mixing method is always better than the ϕ-mixing one. Obviously, this result was expected by the definitions of these mixing processes and also by the theorems because of the extra factor $e^{-(t-(3k+1)n)\mathbb{P}(A)}$ (see Theorem 3.3.1 and Theorem 2 in (Abadi and Vergne, in preparation). We are interested by the real impact of this factor on the threshold u : it is significantly better in the case of a ψ-mixing process.

Finally, let us remember you that Chen-Stein method give any result in a great number of cases where our method works. And it is more the case when our model of interest is a Markov model of order greater than 2. Indeed, Chen-Stein bounds for Markov model of order greater than 2 are very high and then cannot give any result whereas our local method works easily.

Biological Comparisons.

Now, we present a few results obtained on real biological examples with order one Markov models. There are many categories of words which have relevant biological functions (promoters, terminators, repeat sequences, chi sites, uptake sequences, bend sites, signal peptides, binding sites, restriction sites, ...). Some of them are highly present in the sequence, some others are almost absent. Then, it turns out to be interesting to consider the over or the under-representation of words to find words biologically relevant.

In this section, we test our methods on words already known to be relevant. We focus our study on Chi sites or uptake sequences. Chi sites of bacterias protect the genome by stopping its degradation performed by a particular enzyme. The function of this enzyme is to destroy viruses which could appear into the bacteria. Viruses do not contain Chi sites and then are exterminated. It turns out that Chi sites are highly present in the bacterial genome. Uptake sequences are abundant sequence motifs, often located downstream of ORFs, that are used to facilitate the within-species horizontal transfer of DNA.

Example 1

First, we consider the Chi of *Escherichia coli*, `gctggtgg`, (see Table 3.2), for different degrees of significance. We use complete sequence of *Escherichia coli K12* ([BPB+97]). Sequence length is equal to 4639221. We recall that for a fixed significance degree, the smaller the threshold u, the best the method is. Then, we can conclude that the ψ-mixing method gives the most interesting results. Chi of *E. coli* could be considered as an over-represented one from 99 occurrences for a significance degree s of 0.0001. Because Chen-Stein bound is equal to 0.067726, Chen-Stein method does not permit to conclude for significance degrees of 0.01 and 0.001. Moreover, it is well known that Chi of *E. coli* is a very relevant word in this bacteria. Then, we expect a very small significance degree for this word. Unfortunately, the minimal significance degree which could be obtained by Chen-Stein method is, in fact, the Chen-Stein bound : 0.067726. Our method allows to obtain very small significance degree and the minimal significance degree for which Chi of *E. coli* is considered as an over-represented word by the ψ-mixing method, is given at the last line of Table 3.2 : it is equal to 10^{-239}. Note also that the thresholds u increase with the significance degrees s. To understand this fact, it is sufficient to look at inequality (3.1). But they increase slowly while significance degrees s decreases. It could be surprising but it is due to the error term which decreases very fast from a certain number of occurrences.

TAB. 3.2 – Table of thresholds u obtained by the three methods for the Chi of *Escherichia coli* : `gctggtgg` (sequence length t equal to 4639221). For each one of the three methods we compute the threshold which permits to consider the word as an over-represented word or not, for degree of significance s. IMP means that the method can not return a result. "counts" correspond to the number of occurrences observed in the sequence.

s	Chen-Stein	ϕ-mixing	ψ-mixing	counts
0.1	87	193	83	499
0.01	IMP	195	92	499
0.0001	IMP	197	99	499
10^{-239}	IMP	549	498	499

Example 2
Second, we consider the Chi of *Haemophilus influenzae* and its uptake sequence (see Table 3.3), for a significance degree s equal to 0.01. We use complete sequence of *Haemophilus influenzae* ([FAW+95]). Sequence length is equal to 1830138. We observe that in all the cases the ψ-mixing method is the best one because it gives the smallest u,

TAB. 3.3 – Table of thresholds u obtained by the three methods for the Chi and the uptake sequence of *Haemophilus influenzae* (sequence length t equal to 1830138). For each one of the three methods and for each word, we compute the threshold which permits to consider the word as an over-represented word or not, for degree of significance equal to 0.01. IMP means that the method can not return a result. "counts" correspond to the number of occurrences observed in the sequence.

Words	Chen-Stein	ϕ-mixing	ψ-mixing	counts
gatggtgg (chi)	23	36	22	20
gctggtgg (chi)	21	32	20	44
ggtggtgg (chi)	16	IMP	IMP	57
gttggtgg (chi)	30	45	26	37
aagtgcggt (uptake)	13	17	13	737

except for the word `ggtggtgg` which has a periodicity less than $\left[\frac{n}{2}\right]$ (and then we can not study it : see assumptions in Theorem 3.3.1). We can not assume the good significance of the first Chi (`gatggtgg`) because we count only 20 occurrences in the sequence, whereas 23 occurrences are necessary to consider this word as exceptional. On the other hand, the uptake sequence is very significant (and then very relevant). Indeed, we could fix a significance degree equal to 10^{-224} and consider it as an over-represented word from 736 occurrences with the ψ-mixing method. As `aagtgcggt` is counted 737 times in the sequence, we obtain the well-known fact that this word is biologically relevant.

3.7 Conclusions and Perspectives

To conclude this paper, we recall the advantages of our new methods. We give an error valid for all the values k of the random variable N^t corresponding to the number of occurrences of word A in a sequence of length t. Then, we can find a minimal number of occurrences to consider a word as biologically relevant for a very large number of words and for all degrees of significance. That is the main advantage of our methods on the Chen-Stein one which is based on the total variation distance and for which small degrees of significance can not be obtained. Results of our ψ-mixing method and the Chen-Stein method remain similar but our method has less limitations. Note that our methods provide performing results for general modelling processes such as Markov chains as well as every ϕ- and ψ-mixing processes.

In terms of perspectives, as we expect more significant results, we hope to improve these methods adapting them directly to Markov chains instead of ψ- or ϕ-mixing. Moreover, it is well-known that a compound Poisson approximation is better for self-overlapping words (see [RSW00] and [RS98]). An error term for the compound Poisson approximation for self-overlapping words can be easily derived from our results.

Acknowledgments

The authors would like to thank Bernard Prum for his support and his useful comments. The authors would like to thank Sophie Schbath for her program, Vincent Miele for his very relevant help in the conception of the software and Catherine Matias for her invaluable advices.

Bibliographie

[Aba01a] M. Abadi. Exponential approximation for hitting times in mixing processes. *Mathematical Physics Electronic Journal*, 7, 2001.

[Aba01b] M. Abadi. *Instantes de ocorrência de eventos raros em processos misturadores.* PhD thesis, Universidade de São paulo, 2001. available at `http ://www.ime.unicamp.br/~miguel`.

[Aba04] M. Abadi. Sharp error terms and necessary conditions for exponential hitting times in mixing processes. *Annals of Probability*, 32 :243–264, 2004.

[AG01] M. Abadi and A. Galves. Inequalities for the occurrence times of rare events in mixing processes. The state of the art. *Markov Proc. Relat. Fields*, 7 :97–112, 2001.

[AGG89] R. Arratia, L. Goldstein, and L. Gordon. Two moments suffice for Poisson approximations : the Chen-Stein method. *Ann. Prob.*, 17 :9–25, 1989.

[AGG90] R. Arratia, L. Goldstein, and L. Gordon. Poisson approximation and the Chen-Stein method. *Statist. Sci.*, 5 :403–434, 1990.

[AHV05] M. Abadi, N. Haydn, and S. Vaienti. Statistics properties of repetition times. Preprint, 2005.

[Aka74] H. Akaike. A new look at the statistical identification model. *IEEE Transactions on Automatic Control*, 19 :716–723, 1974.

[Alm83] H. Almagor. A Markov analysis of DNA sequences. *J.Theor. Biol.*, 104 :633–645, 1983.

[AV06] M. Abadi and S. Vaienti. Large deviation for short recurrence. To appear in Discrete and Continuous Dynamical Systems, 2006.

[AV08] M. Abadi and N. Vergne. Sharp error terms for return time statistics under mixing conditions. Submitted, 2008.

[BCL92] A.D. Barbour, L.H.Y. Chen, and W.L. Loh. Compound Poisson approximation for nonnegative random variables via Stein's method. *Ann. Prob.*, 20 :1843–1866, 1992.

[BE94] T.L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol*, 2 :28–36, 1994.

[Bel98] E.D. Belsley. Rates of convergence of random walk on distance regular graphs. *Probab. Theory Related Fields*, 112 :493–533, 1998.

[Ber93] G. Bernardi. The vertebrate Genome : Isochores and Evolution. *Mol. Biol. Evol.*, 10 :186–204, 1993.

[Ber01a] G. Bernardi. Misunderstandings about isochores. *Gene*, 276 :3–13, 2001.

[Ber01b] A. Bernot. *Analyse de Génomes, Transcriptomes et Protéomes.* DUNOD, 3 edition, 2001.

[BHJ92] A.D. Barbour, L. Holst, and S. Janson. *Poisson approximation.* Oxford, New York, university press edition, 1992.

[Bla85] B.E. Blaisdell. Markov chain analysis finds a significant influence of neighboring bases on the occurrence of a base in eucaryotic nuclear DNA sequences both protein-coding and noncoding. *J. Mol. Evol.*, 21 :278–288, 1985.

[Bow75] R. Bowen. Equilibrium states and the ergodic theory of Anosov diffeomorphisms. *Lecture Notes in Math*, 470, 1975.

[BPB+97] F.R. Blattner, G.3rd Plunkett, C.A. Bloch, N.T. Perna, V. Burland, M. Riley, J. Collado-Vides, J.D. Glasner, C.K. Rode, Mayhew G.F., J. Gregor, Davis N.W., H.A. Kirkpatrick, M.A. Goeden, D.J. Rose, B. Mau, and Y. Shao. The complete genome sequence of escherichia coli k-12. *Science*, 277 :1453–74, 1997.

[BR02] A. Berchtold and A.E. Raftery. The Mixture Transition Distribution Model for High-Order Markov Chains and Non-Gaussian Time Series. *Statistical Science*, 17(3) :328–356, 2002.

[Bra05] R. C. Bradley. Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys*, 2 :107–144, 2005.

[BW99] P. Buhlmann and A.J. Wyner. Variable length Markov chains. *Annals of Statistics*, 27(2) :480–513, 1999.

[CDvM+06] F.D. Coccarelli, T. Doerks, C. von Mering, C.J. Creevey, B. Snel, and P. Bork. Toward Automatic Reconstruction of a Highly Resolved Tree of Life. *Science*, 311(5765) :1283–1287, 2006.

[CGS99] P. Collet, A. Galves, and B. Schmitt. Repetition times for gibbsian sources. *Nonlinearity*, 12 :1225–1237, 1999.

[CHA79] E. CHARGAFF. How Genetics Got a Chemical Education. *Annals of the New York Academy of Sciences*, 325(1 The Origins of Modern Biochemistry : A Retrospect on Proteins) :345–362, 1979.

[Cha03] J.-R. Chazottes. Hitting and returning to non-rare events in mixing dynamical systems. *Nonlinearity*, 16 :1017–1034, 2003.

[Che75] L.H.Y. Chen. Poisson approximation for dependant trials. *Ann. Prob.*, 3 :534–545, 1975.

[Chu89] G. Churchill. Stochastic models for heterogeneous DNA sequences. *Bull. Math. Biol.*, 268 :8–14, 1989.

[Con01] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409 :860–921, 2001.

[dB78] C. de Boor. *A Practical Guide to Splines*. Springer-Verlag, 1978.

[DGK01] I. Dimatteo, C.R. Genovese, and R.E. Kass. Bayesian curve-fitting with free-knot splines. *Biometrika*, 88(4) :1055–1071, 2001.

[Die95] P. Dierckx. *Curve and Surface Fitting with Splines*. Oxford University Press, USA, 1995.

[DK02] F. Dardel and F. Képès. *Bioinformatique Génomique et post-génomique*. ellipses, 2002.

[DMS98] D.G.T Denison, B.K. Mallick, and A.F.M. Smith. Bayesian MARS. *Statistics and Computing*, 8(4) :337–346, 1998.

[Doe40] W. Doeblin. Remarques sur la théorie métrique des fractions continues. *Compositio Math.*, 7 :353–371, 1940.

[Dou95] P. Doukhan. Mixing. Properties and examples. *Lecture Notes in Statistics*, 85, 1995.

[Dou96] S.A. Douglass. *Introduction to Mathematical Analysis*, chapter 8. Addison-Wesley, Boston, 1996.

[DSHS05] E. Defez, L. Soler, A. Hervás, and C. Santamaría. Numerical solution of matrix differential models using cubic matrix splines. *Computers and Mathematics with Applications*, 50(5-6) :693–699, 2005.

[DSP+02] L. Duret, M. Semon, G. Piganeau, D. Mouchiroud, and N. Galtier. Vanishing GC-Rich Isochores in Mammalian Genomes. *Genetics*, 162 :1837–1847, 2002.

[EKBSG99] M. El Karoui, V. Biaudet, S. Schbath, and A. Gruss. Characteristics of Chi distribution on different bacterial genomes. *Res. Microbiol.*, 150 :579–587, 1999.

[EM05] P.H.C. Eilers and B.D. Marx. Splines, knots and penalties. *J. Comput. Graph. Statist*, 2005.

[Eub88] R.L. Eubank. *Spline Smoothing and Nonparametric Regression*. Marcel Dekker, 1988.

[FAW+95] R.D. Fleischmann, M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J. Tomb, B.A. Dougherty, J.M. Merrick, K. McKenney, G.G. Sutton, W. FitzHugh, C.A. Fields, J.D. Gocayne, J.D. Scott, R. Shirley, L.I. Liu, A. Glodek, J.M. Kelley, J.F. Weidman, C.A. Phillips, T. Spriggs, E. Hedblom, M.D. Cotton, T. Utterback, M.C. Hanna, D.T. Nguyen, D.M. Saudek, R.C. Brandon, L.D. Fine, J.L. Fritchman, J.L. Fuhrmann, N.S. Geoghagen, C.L. Gnehm, L.A. McDonald, K.V. Small, C.M. Fraser, H.O. Smith, and J.C. Venter. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science*, 269 :496–512, 1995.

[Fil91] J.A. Fill. Eigenvalue bounds on convergence to stationarity for nonreversible Markov chains, with an application to the exclusion process. *Ann. Appl. Probab.*, 1 :62–87, 1991.

[FS89] J.H. Friedman and B.W. Silverman. Flexible Parsimonious Smoothing and Additive Modeling. *Technometrics*, 31(1) :3–21, 1989.

[FTW92] J. W. Fickett, D. C. Torney, and D. R. Wolf. Base compositional Structure of Genomes. *Genomics*, 13 :1056–1064, 1992.

[GFM+03] D. Goldstein, C. Fondrat, F. Muri, G. Nuel, P. Saragueta, A.S. Tocquet, and B. Prum. Short inverse complementary amino acid sequences generate protein complexity. *Comptes rendus Biologies*, 326(3) :339–348, 2003.

[GKP92] M.S. Gelfand, C.G. Kozhukhin, and Pevzner P.A. Extendable words in nucleotide sequences. *Bioinformatics*, 8 :129–135, 1992.

[GMSP00] D. Goldstein, F. Muri, P. Saragueta, and B. Prum. Inverse complementary homologues of short cysteine signatures. *C. R. Acad. Sci. III*, 323 :167–172, 2000.

[God91] A.P. Godbole. Poisson approximations for runs and patterns of rare events. *Adv. Appl. Prob.*, 23 :851–865, 1991.

[GS97] A. Galves and B. Schmitt. Inequalities for hitting times in mixing dynamical systems. *Random Comput. Dyn*, 5 :337–348, 1997.

[Her99] GZ Hertz. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15(7) :563–577, 1999.

[Hir93] M. Hirata. Poisson law for axiom a diffeomorphism. *Ergod. Th. Dyn. Sys.*, 13 :533–556, 1993.

[HSV99] M. Hirata, B. Saussol, and S. Vaienti. Statistics of return times : a general framework and new applications. *Comm. Math. Phys.*, 206 :33–55, 1999.

[HV04] N. Haydn and S. Vaienti. The limiting distribution and error terms for return times of dynamical systems. *Discrete Contin. Dyn. Syst.*, 10(3) :589–616, 2004.

[Jup78] D.L.B. Jupp. Approximation to Data by Splines with Free Knots. *SIAM Journal on Numerical Analysis*, 15(2) :328–343, 1978.

[Kac47] M. Kac. On the notion of recurrence in discrete stochastic processes. *Bull. Amer. Math. Soc.*, 53 :1002–1010, 1947.

[KBM92] S. Karlin, C. Burge, and A. M.Campbell. Statistical analyses of counts and distributions of restriction sites in DNA sequences. *Nucl. Acids Res.*, 20 :1363–1370, 1992.

[KMH94] A. Krogh, L.S. Mian, and D. Haussler. A hidden Markov model that finds genes in *escherichia coli* DNA. *Nucl. Acids Res.*, 22 :4768–4778, 1994.

[LC67] HJ Lin and E. Chargaff. On the denaturation of deoxyribonucleic acid. II. Effects of concentration. *Biochim Biophys Acta*, 145(2) :398–409, 1967.

[Lee02] T. Lee. On Algorithms for Ordinary Least Squares Regression Spline Fitting : a Comparative Study. *Journal of Statistical Computation and Simulation*, 72(8) :647–663, 2002.

[Li97] W.H. Li. *Molecular Evolution*. Sunderland, 1997.

[Lin99] M.J. Lindstrom. Penalized Estimation of Free-Knot Spline. *Journal of Computational and Graphical Statistics*, 1999.

[Liu96] J.S. Liu. Metropolized Independant Sampling with Comparisons Rejection Sampling and Importance Sampling. *Stat. Comput.*, 6 :113–119, 1996.

[Lob99] J.R Lobry. Genomic landscapes. *Microbiol. Today*, 26 :164–165, 1999.

[Lob00] J.R. Lobry. Oriloc : prediction of replication boundaries in unannotated bacterial chromosomes. *Bioinformatics*, 16 :560–561, 2000.

[Lou96] W.Y.W. Lou. On runs and longest run tests : A method of finite markov chain imbedding. *J. Am. Statis. Assoc.*, 91 :373–380, 1996.

[LSAR+03] B. La Scola, S. Audic, C. Robert, L. Jungang, X. de Lamballerie, M. Drancourt, R. Birtles, J.M. Claverie, and D. Raoult. A giant virus in amoebae. *Science*, 299(5615) :2033, 2003.

[LSV98] C. Liverani, B. Saussol, and S. Vaienti. Conformal measures and decay of correlations for covering weighted systems. *Ergod. Theeor. dynam. Sys.*, 18 :1399–1420, 1998.

[MBR+05] V. Miele, P.Y. Bourguignon, D. Robelin, G. Nuel, and H. Richard. seq++ : analyzing biological sequences with a range of Markov-related models. *Bioinformatics*, 21 :2783–2784, 2005.

[MDD01] N. Molinari, J.P. Daurès, and J.F. Durand. Regression splines for threshold selection in survival data analysis. *Statistics in Medicine*, 20(2) :237–247, 2001.

[MKM+03] E.S. Miller, E. Kutter, G. Mosig, F. Arisaka, T. Kunisawa, and W. Rüger. Bacteriophage T4 genome. *Microbiology and molecular biology reviews*, 67(1) :86–156, 2003.

[MMCD02] N. Molinari, M. Morena, J.P. Cristol, and J.P. Daurès. Free knot splines for biochemical data. *Computer Methods and Programs in Biomedicine*, 67(3) :163–167, 2002.

[MNSS89] G. Meinardus, G. Nurnberger, M. Sommer, and H. Strauss. Algorithms for Piecewise Polynomials and Splines with Free Knots. *Mathematics of Computation*, 53(187) :235–247, 1989.

[MT93] S.P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, Heidelberg, 1993.

[Mur97] F. Muri. *Comparaisons d'algorithmes d'identification de chaînes de Markov cachées et application à la détection de régions homogènes dans les séquences d'ADN*. PhD thesis, Université Paris V, 1997. 156–194.

[NBM+02] P. Nicolas, L. Bize, F. Muri, M. Hoebeke, F. Rodolphe, S.D. Ehrlich, B. Prum, and P. Bessières. Mining *bascillus subtilis* chromosome heterogeneity using hidden Markov models. *Nucl. Acids Res.*, 30 :1418–1426, 2002.

[NDV02] P. Nicodème, T. Doerks, and M. Vingron. Proteome analysis based on motif statistics. *Bioinformatics*, 18(Suppl. 2) :5161–5171, 2002.

[Nic03] P. Nicolas. *Mise au point et utilisation de modèles de Markov cachées pour l'étude des séquences ADN*. PhD thesis, Université d'Evry Val d'Essonne, 2003.

[NL00] A. Nekrutenki and W.H. Li. Assessment of Compositional Heterogeneity Within and Between Eukaryotic Genomes. *Genome Research*, 10(12) :1986–1995, 2000.

[NP07] G. Nuel and B. Prum. *Analyse statistique des séquences biologiques : modélisation markovienne, alignements et motifs*. Hermes, 2007.

[Nue04] G. Nuel. LD-SPatt : Large Deviations Statistics for Patterns on Markov chains. *Comp. Biol.*, 11 :1023–1033, 2004.

[Nue06a] G. Nuel. Effective p-value computations using Finite Markov Chain Imbedding (FMCI) : application to local score and to pattern statistics. *Algorithms for Molecular Biology*, 1(5), 2006.

[Nue06b] G. Nuel. Numerical Solutions for Patterns Statistics on Markov Chains. *Statistical Applications in Genetics and Molecular Biology*, 5, 2006.

[OBGCRR01] J.L. Oliver, P. Bernaola-Galván, P. Carpena, and R. Román-Roldán. Isochore chromosome maps of eukaryotic genomes. *Gene*, 276 :47–56, 2001.

[Oeh92] G.W. Oehlert. A note on the delta method. *American Statistician*, 46(1) :27–29, 1992.

[OW93] D. Ornstein and B. Weiss. Entropy and data compression schemes. *IEEE Trans. Inform. Theory*, 39(1) :78–83, 1993.

[PAI87] G.J. Phillips, J. Arnold, and R. Ivarie. The effect of codon usage on the oligonucleotide composition of the e. coli genome and identification of over- and underrepresented sequences by Markov chain analysis. *Nucl. Acids Res.*, 15 :2627–2638, 1987.

[Pit02] J. Pittman. Adaptive Splines and Genetic Algorithms. *Journal of Computational & Graphical Statistics*, 11(3) :615–638, 2002.

[PRdT95] B. Prum, F. Rodolphe, and E. de Turckheim. Finding words with unexpected frequencies in DNA sequences. *J. R. Statis. Soc. B*, 11 :190–192, 1995.

[RÓ0] M. Régnier. A unified approach to word occurrence probabilities. *Discr. Appl. Math.*, 104 :259–280, 2000.

[Raf85] A.E. Raftery. A model for high-order Markov chains. *Journal of the Royal Statistical Society B*, 47(3) :528–539, 1985.

[RD99] S. Robin and J.J. Daudin. Exact distribution of word occurrences in a random sequence of letters. *J. Appl. Prob.*, 36, 1999.

[Rob05] D. Robelin. *Détection de courts segments inversés dans les génomes : méthodes et applications*. PhD thesis, Universités Paris XI et Evry-Val-d'Essonne, 2005.

[Ros95] J.S. Rosenthal. Convergence rates of markov chains. *SIAM*, 37 :387–405, 1995.

[RRP03] D. Robelin, H. Richard, and B. Prum. Sic : a tool to detect short inverted segments in a biological sequence. *Nucleic Acids Research*, 31(13) :3669–3671, 2003.

[RRS03] S. Robin, F. Rodolphe, and S. Schbath. *ADN, mots et modèles*. Belin, 2003.

[RS98] G. Reinert and S. Schbath. Compound Poisson and Poisson process approximations for occurrences of multiple words in Markov chains. *J. Comput. Biol.*, 5 :223–253, 1998.

[RSW00] G. Reinert, S. Schbath, and M.S. Waterman. Probabilistic and Statistical Properties of Words : An Overview. *J. Comput. Biol.*, 7, 2000.

[Rup02] D. Ruppert. Selecting the Number of Knots for Penalized Splines. *Journal of Computational & Graphical Statistics*, 11(4) :735–757, 2002.

[Sch78] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6 :461–464, 1978.

[Sch07] L.L Schumaker. *Splines Functions : Basic Theory.* Cambridge University Press, 3 edition, 2007.

[SGS99] H.O. Smith, M.L Gwinn, and S.L. Salzberg. DNA uptake signal sequences in naturally transformable bacteria. *Res. Microbiol.*, 150 :603–616, 1999.

[SKL+98] R.S. Stephens, S. Kalman, C. Lammel, J. Fan, R. Marathe, L. Aravind, W. Mitchell, L. Olinger, R.L. Tatusov, Q. Zhao, E.V. Koonin, and R.W. Davis. Genome sequence of an obligate intracellular pathogen of humans : Chlamydia trachomatis. *Science*, 282 :754–759, 1998.

[SKS+81] G.R. Smith, S.M. Kunes, D.W. Schultz, A. Taylor, and K.L. Triman. Structure of chi hotspots of generalized recombination. *Cell*, 24 :429–36, 1981.

[SPdT95] S. Schbath, B. Prum, and E. de Turckheim. Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. *Journal of Computational Biology*, 2 :417–437, 1995.

[SS95] H. Schwetlick and T. Schütze. Least squares approximation by splines with free knots. *BIT Numerical Mathematics*, 35(3) :361–384, 1995.

[Ste72] C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Proc. Sixth Berkeley Symp. Math. Statist. Probab.*, 2 :583–602, 1972. University of California Press.

[SW03] M. Stanke and S. Waack. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, 19 :215–225, 2003.

[SYM05] G. Simons, Y.C. Yao, and G. Morton. Global Markov models for eukaryote nucleotide data. *J. Statist. Plann. Inference*, 130 :251–275, 2005.

[TLB+05] M. Tompa, N. Li, T.L. Bailey, G.M. Church, B. De Moor, E. Eskin, A.V. Favorov, M.C. Frith, Y. Fu, W.J. Kent, et al. Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*, 23 :137–144, 2005.

[TSDR+08] F. Touzain, S. Schbath, I. Debled-Rennesson, B. Aigle, G. Kucherov, and P. Leblond. SIGffRid : a tool to search for sigma factor binding sites in bacterial genomes using comparative approach and biologically driven statistics. *BMC Bioinformatics*, 9(1) :73, 2008.

[VA08] N. Vergne and M. Abadi. Poisson approximation for search of rare words in DNA sequences. Submitted, 2008.

[vHACV98] J. van Helden, B. André, and J. Collado-Vides. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, 281 :872–842, 1998.

[vHdOPO00] J. van Helden, M. del Olmo, and J.E. Pérez-Ortín. Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. *Nucl. Acids Res.*, 28 :1000–1010, 2000.

[Wah90] G. Wahba. *Splines Models for Observational Data.* SIAM, 1990.

[Wal75] P. Walters. Ruelle's operator theorem and g-measures. *Trans. Amer. Math. Soc.*, 214 :375–387, 1975.

[Wan00] M.P. Wand. A Comparison of Regression Spline Smoothing Procedures. *Computational Statistics*, 15(4) :443–462, 2000.

[Wil99] E.L. Wilmer. *Exact Rates of Convergence for Some Simple Non-Reversible Markov Chains.* PhD thesis, Harvard university, Cambridge, Massachussets, 1999.

[WT71] R. Wu and E. Taylor. Nucleotide sequence analysis of DNA. II. Complete nucleotide sequence of the cohesive ends of bacteriophage lambda DNA. *J Mol Biol.*, 57 :491–511, 1971.

[WZ89] A. Wyner and J. Ziv. Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression. *IEEE Trans. Inform. Theory*, 35(6) :1250–1258, 1989.

[ZCB96] S. Zoubak, O. Clay, and G. Bernardi. The gene distribution of the human genome. *Gene*, 174 :95–102, 1996.

Épilogue

L'espoir

L'espoir est dans la rue
La victoire au bout de la fleur
Dans ton ventre pousse un arc en ciel
Avec la gueule ouverte
Droit devant, droit devant
Il ne te font plus peur
Les voleurs, les imposteurs
Les assassins de Joie
Vous vous tiendrez la main
Jusque là, jusque là
Jusqu'à tomber les murs
Avec vos 17 ans
Et puis ce nouveau jour
Juste au bout de vos doigts
L'espoir c'est cette boule délicieuse qui vous bouffe le ventre
C'est une arme au soleil quand j'entends près de toi
Tout ces milliers qui chantent
Aux étoiles, aux étoiles
Je viendrai avec toi
Patiner vers l'amour
Je viendrai pour toujours
Chercher mes 17 ans
Au coeur de ton espoir
Frais comme un petit jour
L'espoir est là partout
L'espoir
Et je viendrai sourire
Et pleurer près de toi
Avec le point serré
Je goûterai la joie
En allant piocher
Dans tes yeux plein d'amour
Et plein de 17 ans
Tu es l'Espoir
Ne lâche rien, jamais

Ils plieront, effrayés
Sous ta beauté, sous ton rien
Et sous tes cris
Qui montent de la rue
Jusqu'à l'éternité
Qui monteront toujours

Et je renifle, heureux
Comme un chien magnifique
La poussière d'étoile
Que tu sèmes, si fière
Dans mon coeur
Et tout autour
L'Espoir
L'Espoir est là toujours

À genoux l'aurore

À genoux
Les voleurs de Joie
La jeunesse est bien là
Et tu dois t'effacer
Ce jour n'est plus à toi
L'Espoir est un drapeau planté dans tes entrailles

L'Espoir est dans la rue
La victoire au bout de la fleur
Dans ton ventre pousse un arc en ciel
Avec la gueule ouverte
Droit devant, droit devant

Ils ne te font plus peur
Les voleurs, les imposteurs
Les assassins de Joie
Vous vous tiendrez la main
Jusque là, jusque là
Jusqu'à tomber les murs
Avec vos 17 ans
Et puis ce nouveau jour juste au bout de vos doigts

Cali

Abstract

The statistical analysis of biological sequence such as nucleotidic sequences (DNA and RNA) or amino-acids (proteins) needs the conception of different models according to the study. Since the way the nucleotides succeed one another in DNA sequences is dependant, Markov models are widely used for this purpose. The problem of these models is to consider the homogeneity of biological sequences. But, biological sequences are not homogeneous. A well-known example is the gc percent : along a sequence, gc-rich regions and gc-poor regions succeed one another. In order to take into account this heterogeneity, other models are used : the hidden Markov models (HMM). The sequence is divided in some homogeneous regions. There is a lot of applications to HMM, such as search of coding regions. But, all biological particularities can not appear under these models, that is why we develop new models : the drifting Markov models (DMM). Instead of fitting a transition matrix on a whole sequence (classical Markov model) or different transition matrices on different homogeneous parts of the sequence (HMM), we allow the transition matrix to vary (to drift) from the beginning to the end of the sequence. At each position t, we obtain a different transition matrix $\Pi_{\frac{t}{n}}$ (where n is the sequence length). Thus, our models are constrained heterogeneous Markov models. We give two ways to constrain models : polynomial DMM and polynomial splines DMM. For instance, for a degree 1 DMM (linear drift), we fix a transition matrix Π_0 at the beginning of the sequence and transition matrix Π_1 at the end of the sequence and we allow the transition matrix to vary linearly from Π_0 to Π_1 :

$$\Pi_{\frac{t}{n}} = \left(1 - \frac{t}{n}\right)\Pi_0 + \frac{t}{n}\Pi_1.$$

Such a model could correspond to a soft evolution between two hidden states of an HMM, for which transitions could appear too sudden. DMM can be seen as a competitive model to the HMM one but it over all can be understood as a complementary tool : the hidden models of an HMM, usually fixed Markov chains can be replaced by DMM.

Along this work, we consider polynomial drift or drift by polynomial splines (in the way to make them more flexible than the polynomial ones). We estimate our models by different ways, evaluate their qualities and used them in biological applications such as the search of rare words. We develop the software DRIMM (soon available at http://stat.genopole.cnrs.fr/sg/software/drimm/), dedicated to estimation of DMM. This program provide all the possibilities of DMM, such as computation of transition matrix in each position, computation of stationary laws... Use of this program for the search of rare words is proposed in auxiliary programs (available on request).

This work provides some perspectives. Instead of allowing the transition matrix to vary only with the position t, we could take into account covariables such as, hydrophobicity degree, gc-percent, an indicator of the protein structure (α-helix, β-sheet,...). But the main perspective stay the possibility to combine HMM and DMM, with DMM in the role of hidden states.

Résumé

L'analyse statistique des séquences biologiques telles les séquences nucléotidiques (l'ADN et l'ARN) ou d'acides aminés (les protéines) nécessite la conception de différents modèles s'adaptant chacun à un ou plusieurs cas d'étude. Étant donnée la dépendance de la succession des nucléotides dans les séquences d'ADN, les modèles généralement utilisés sont des modèles de Markov. Le problème de ces modèles est de supposer l'homogénéité des séquences. Or, les séquences biologiques ne sont pas homogènes. Un exemple bien connu est la répartition en `gc` : le long d'une même séquence, alternent des régions riches en `gc` et des régions pauvres en `gc`. Pour rendre compte de l'hétérogénéité des séquences, d'autres modèles sont utilisés : les modèles de Markov cachés. La séquence est divisée en plusieurs régions homogènes. Les applications sont nombreuses, telle la recherche des régions codantes. Certaines particularités biologiques ne pouvant apparaître suivant ces modèles, nous proposons de nouveaux modèles, les chaînes de Markov régulées (DMM pour *drifting Markov model*). Au lieu d'ajuster une matrice de transition sur une séquence entière (modèle de Markov homogène classique) ou différentes matrices de transition sur différentes régions de la séquence (modèles de Markov cachés), nous permettons à la matrice de transition de varier (*to drift*) du début à la fin de la séquence. À chaque position t dans la séquence, nous avons une matrice de transition $\Pi_{\frac{t}{n}}$ (où n est la longueur de la séquence) éventuellement différente. Nos modèles sont donc des modèles de Markov hétérogènes contraints. Dans cette thèse, nous donnerons essentiellement deux manières de contraindre les modèles : la modélisation polynomiale et la modélisation par splines. Par exemple, pour une modélisation polynomiale de degré 1 (une dérive linéaire), nous nous donnons une matrice de départ Π_0 et une matrice d'arrivée Π_1 puis nous passons de l'une à l'autre en fonction de la position t dans la séquence :

$$\Pi_{\frac{t}{n}} = \left(1 - \frac{t}{n}\right)\Pi_0 + \frac{t}{n}\Pi_1.$$

Cette modélisation correspond à une évolution douce entre deux états. Par exemple cela peut traduire la transition entre deux régimes d'un chaîne de Markov cachée, qui pourrait parfois sembler trop brutale. Ces modèles peuvent donc être vus comme une alternative mais aussi comme un outil complémentaire aux modèles de Markov cachés.

Tout au long de ce travail, nous avons considéré des dérives polynomiales de tout degré ainsi que des dérives par splines polynomiales : le but de ces modèles étant de les rendre plus flexibles que ceux des polynômes. Nous avons estimé nos modèles de multiples manières puis évalué la qualité de ces estimateurs avant de les utiliser en vue d'applications telle la recherche de mots exceptionnels. Nous avons mis en oeuvre le software `DRIMM` (bientôt disponible à http://stat.genopole.cnrs.fr/sg/software/drimm/), dédié à l'estimation de nos modèles. Ce programme regroupe toutes les possibilités offertes par nos modèles, tels le calcul des matrices en chaque position, le calcul des lois stationnaires, des distributions de probabilité en chaque position... L'utilisation de ce programme pour la recherche des mots exceptionnels est proposée dans des programmes auxiliaires (disponibles sur demande).

Plusieurs perspectives à ce travail sont envisageables. Nous avons jusqu'alors décidé de faire varier la matrice seulement en fonction de la position, mais nous pourrions prendre en compte des covariables tels le degré d'hydrophobicité, le pourcentage en *gc*, un indicateur de la structure des protéines (hélice α, feuillets β...). Nous pourrions aussi envisager de mêler HMM et variation continue, où sur chaque région, au lieu d'ajuster un modèle de Markov, nous ajusterions un modèle de chaînes de Markov régulées.