Numéro d'identification : 01EVRY0006

Université d'Evry Val d'Essonne

Thèse présentée pour le titre de docteur en mathématiques

Grandes déviations et chaînes de *Markov* pour l'étude des occurrences de mots dans les séquences biologiques

par Grégory Nuel

Soutenance le 6 juillet 2001 devant le jury suivant :

Mme.	de Turckheim	Élisabeth	Rapporteur
M.	Ben Arous	Gérard	Rapporteur
	Prum	Bernard	Directeur de thèse
	Robin	Stéphane	Rapporteur
	Tavaré	Simon	Président

Remerciements

Je tiens tout d'abord à remercier toute l'équipe du "laboratoire de statistiques et génomes" dont l'accueil chaleureux qu'elle m'a reservé dès les premières heures, ainsi que le soutien constant qu'elle m'a manifesté tout au long de ces trois années, méritent d'étre soulignés.

Un grand merci également aux membres de l'équipe "statistiques des séquences biologiques" pour l'aide précieuse que nos réunions ont pu m'apporter en contribuant à enrichir mes connaissances dans le domaine de la bioinformatique mais aussi pour le regard critique qu'ils ont su porter sur mon travail.

Merci, bien sûr, aux membres du jury et aux rapporteurs pour leurs conseils et commentaires grace auxquels j'ai pu élaborer la version finale de ce manuscrit et, comme la forme est indissociable du fond, merci à tous les contributeurs du merveilleux outil d'édition que consitue LaTeX et sans lequel la rédaction d'un document tel que celui-ci aurait relevé du véritable cauchemar.

Avant de commencer ma thèse, c'est avec Stéphane Robin que j'ai découvert la nature du travail du chercheur et il est clair que, sans l'expérience très positive que j'ai eue avec lui, je n'aurais certainement pas entrepris de doctorat. Merci à lui dont l'exemple a suscité chez moi sinon une vocation, la motivation nécessaire à la participation à une telle entreprise.

Je voudrais également faire part de ma plus grande admiration à Bernard Prum pour ses immenses qualités humaines et professionnelles, aussi bien en tant qu'enseignant qu'en tant que chercheur. Travailler sous sa direction à été pour moi une expérience exceptionnelle et je veux profiter de ces quelques lignes pour lui adresser ici mes plus sincères remerciements.

Merci enfin à ma famille et à mes amis pour leur soutien moral et je voudrais tout particulièrement remercier ma douce Sophie pour la gentillesse avec laquelle elle a su m'encourager, avec constance, tout au long de mes études.

A mes parents.

Grandes déviations et chaînes de *Markov* pour l'étude des mots exceptionnels dans les séquences biologiques

G. Nuel

30 octobre 2001

Table des matières

Liste des tableaux

Table des figures

En 1985, la découverte de la *PCR* (*Polymerase Chain Reaction*) qui permet l'amplification de fragments d'ADN (et son amélioration en 1988 avec la découverte de polymérases résistant aux hautes températures), ouvre la porte au séquençage c'est à dire à la lecture du patrimoine génétique des organismes au niveau moléculaire. En 1995, le séquençage complet de *Haemophilus influenzae* est achevé et depuis, des données similaires concernant de nombreux organismes ne cessent d'affluer en quantités toujours croissantes (un à deux nouveaux organismes chaque mois désormais).

Si l'accès à ces informations a représenté et représente toujours un défi technique d'envergure, c'est surtout la compréhension du message moléculaire qui constitue le véritable enjeu de la génomique. Comment donner du sens au texte écrit dans un alphabet à quatre lettres (`a`, `c`, `g` et `t`) et comportant plusieurs millions (voir milliards dans certains cas) de caractères, que représente la séquence d'ADN complète d'un organisme ?

De par leur capacité à détecter du signifiant, les statistiques constituent une approche naturelle pour un tel problème. En particulier, on sait que certains courts fragments d'ADN, des mots, jouent des rôles importants dans certains mécanismes moléculaires mais l'identification de ces fragments dans la masse gigantesque des combinaisons possibles est bien évidemment très difficile par l'expérience seule. Il est clair que les statistiques devraient être en mesure d'effectuer un traitement préalable de ces mots pour en extraire uniquement ceux ayant un comportement (nombre d'occurrences, répartition de ces occurrences) significativement différent des autres.

Le propos de ce travail de thèse est précisément la mise au point et l'utilisation de techniques statistiques permettant d'effectuer la détection de tels mots exceptionnels dans des séquences biologiques en se fondant pour cela sur l'étude des nombres d'occurrences de ces mots.

Dans la première partie de ce document, on propose tout d'abord une introduction au problème de la recherche de mots exceptionnels. Le chapitre 1 (page 15) présente de façon succincte les notions biologiques qui seront utilisées par la suite et donne quelques exemples concrets de mots jouant des rôles particuliers dans les mécanismes du vivant. Pour chacun de ces exemples, le caractère exceptionnel des mots (ou des motifs) est mis en relation avec leurs fréquences d'apparition dans les génomes considérés. Le chapitre 2 (page 27) définit les notations qui seront en vigueur dans le reste du document et formalise le problème du point de vue des statistiques en introduisant la notion de significativité d'un comptage par rapport à un modèle. Dans le chapitre 3 (page 47), on passe en revue les différentes solutions existantes permettant le calcul de cette significativité, tout en décrivant les avantages et inconvénients de chacune d'entre elles. Enfin, pour terminer cette partie introductive, on présente, au chapitre 4 (page 63), l'outil statistique des grandes déviations dont les atouts pour l'étude des événements rares sont soulignés.

On énonce dans la partie II les différents résultats théoriques concernant les grandes déviations de niveau 1 (déviations d'une moyenne) pour le nombre d'occurrences de mots dans des chaînes de *Markov* (chapitre 5 page 83) avant d'expliquer comment mettre en œuvre ces résultats d'un point de vue pratique (chapitre 6 page 91).

Un travail similaire est ensuite effectué dans la partie III pour les grandes déviations de niveau 2 (déviations d'une distribution) avec les résultats théoriques du chapitre 7 (page 107) et leurs utilisations au chapitre 8 (page 117).

Dans la partie IV, les résultats de cette nouvelle méthode sont examinés. Au chapitre 9 (page 127) tout d'abord, on valide l'approche en effectuant une comparaison de ces résultats avec ceux des approches existantes. Le chapitre 10 (page 149) reprend les exemples de mots exceptionnels pour statuer sur la validité opérationnelle des résultats quant à la détection de mots impliqués dans les mécanismes biologiques. Enfin, on propose au chapitre 11 (page 159) quelques utilisations un peu plus poussées de ces méthodes, notamment en ce qui concerne l'exploitation des résultats.

La partie V constitue enfin la conclusion de ce travail : on y effectue le bilan tant théorique que pratique de ce qui a été fait et on y présente les perspectives de développements futurs.

Notons qu'un soin tout particulier a été ici porté pour permettre un accès rapide à l'information dans ce document. Dans cet objectif, on trouvera en début de chaque chapitre un résumé de son contenu ainsi qu'une courte description et une indication de pré-requis nécessaires à sa lecture. De même, les références bibliographiques concernant ces chapitres figurent bien évidemment en fin d'ouvrage mais aussi à la fin de chaque chapitre. Toujours dans le souci de faciliter la lecture, on a préféré isoler en annexe la plupart des preuves mathématiques ou calculs techniques plutôt que de les faire figurer dans le corps du document. Enfin, signalons qu'un index est disponible en fin d'ouvrage et que son utilisation, couplée à celles de la table des matières, devrait permettre au lecteur de retrouver rapidement un point particulier de ce travail.

Première partie

Introduction

FIG. 1.1 – Représentation graphique des mots de longueur $h = 9$ chez *Haemophilus influenzae* dans le modèle $M1$. On consultera la section 11.1 page 160 pour plus de détails.

Chapitre 1

Mots exceptionnels

Pré-requis :

Aucun.

Description :

Présentation du contexte biologique dans lequel ce travail prend place et introduction de la notion de mots exceptionnels par le biais de plusieurs exemples.

Résumé :

Les êtres vivants sont constitués de cellules à l'intérieur desquelles le patrimoine génétique se trouve stocké sous la forme d'une longue molécule : l'ADN. Cette molécule est primordiale à bien des égards et intervient notamment dans la synthèse des protéines. Mais il existe également divers courts fragments d'ADN (des mots ou des motifs) jouant un rôle important dans certains mécanismes biologiques.

Les sites de restriction et les motifs *chi* participent par exemple à la défense des bactéries contre les agressions virales et les motifs *uptake* sont impliqués dans l'évolution des organismes dans lesquels on les trouve. Notons enfin que la pression de sélection lie, pour chacun de ces exemples, les nombres d'occurrence des mots et leurs caractères exceptionnels.

Contenu du chapitre

1.1 Introduction à la génomique

Voici tout d'abord une très courte introduction présentant les termes et notions de biologie qui seront utilisés par la suite.

1.1.1 Organisation des cellules vivantes

Les êtres vivants sont tous constitués d'un nombre plus ou moins important de cellules. Ce nombre peut être réduit à un dans le cas d'un organisme unicellulaire (bactéries, levures, ...) et aller jusqu'à plusieurs centaines de milliards (10^{14} cellules chez l'homme par exemple).

Chacune des cellules d'un organisme contient son patrimoine génétique et s'organise de deux manières différentes :
- chez les *procaryotes* : les cellules ne possèdent qu'un seul compartiment dans lequel on trouve tous les constituants de la cellule, acides aminés, protéines, ribosomes et surtout la molécule d'ADN.
- chez les *eucaryotes* : les cellules s'oganisent en de nombreux compartiments dont l'un, le *noyau*, contient l'ADN.

La structure très simple des cellules procaryotes est celle d'organismes assez spécialisés et relativement peu évolués comme les bactéries tandis que la structure complexe des cellules eucaryotes est caractéristique d'une évolution plus poussée et permet, dans les organismes, la mise en place de mécanismes de régulation sophistiqués ; les mammifères, les plantes, les levures sont des exemples d'organismes eucaryotes. On peut noter que si les procaryotes sont nécessairement des êtres unicellulaires l'inverse n'est pas vrai (exemple des levures notamment).

organisme	date	taille	description
Haemophilus influenzae	05/95	1.8 Mb	bacille infectieux
Mycoplasma genitalium	10/95	0.6 Mb	parasite des voies génitales
Methanococcus jannaschii	08/96	1.7 Mb	archébactérie
Saccharomyces cerevisiae	10/96	12.1 Mb	levure de bière
Escherichia coli	09/97	4.6 Mb	bacille modèle
Bacillus subtilis	10/97	1.9 Mb	bacille
Caenorhabditis elegans	12/98	97 Mb	ver nématode
Lactococcus lactis	07/99	2.4 Mb	bactérie fromagère
Neisseria meningitidis	03/00	2.2 Mb	bacille de la méningite
Drosophilia melanogaster	03/00	137 Mb	mouche à vinaigre
Mycobacterium leprae	04/00	3.2 Mb	bacille de la lèpre
Listeria monocytogenes	04/00	2.9 Mb	bacille de la listériose
Homo sapiens	06/00	3 100 Mb	l'homme
Xyllela fastodiosa	07/00	2.7 Mb	bactérie pathogène des agrumes
Vibrio Cholerae	08/00	4 Mb	bactérie du tube digestif
Pseudomonas aeruginosa	08/00	6.3 Mb	bactérie très résistante
Yersinia Pestis	10/00	4.7 Mb	bacille de la peste
Arabidopsis thaliana	12/00	120 Mb	plante modèle

TAB. 1.1 – Liste des organismes séquencés dans l'ordre chronologique, taille des génomes donnée en million de bases (Mb) d'après le numéro de novembre 2000 de la revue Biofutur.

1.1.2 Les séquences d'ADN

On trouve le patrimoine génétique des organismes au coeur de leurs cellules sous la forme d'une longue molécule appelée *ADN* pour *Acide Désoxyribo-Nucléique*. Cette molécule est une séquence linéaire et orientée, composée de "briques" élémentaires, les *acides nucléiques* (ou *bases*), qui sont au nombre de quatre : adénine (`a`), cytosine (`c`), guanine (`g`) et thymine (`t`). Ces acides nucléiques se succèdent pour former l'ADN que l'on décrit par conséquent naturellement comme une longue séquence écrite dans un alphabet à quatre lettres $\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$.

Les acides nucléiques possèdent une affinité chimique particulière les uns avec les autres qui permet l'apparition de liens forts entre adénine et thymine d'une part (lien `a-t`) et de liens faibles entre cytosine et guanine d'autre part (lien `c-g`). L'existence de ces liens fait qu'une molécule d'ADN se trouve presque toujours associée avec sa molécule *complémentaire inverse* (car les deux brins sont orientés en sens inverses) où chaque base est liée avec sa base complémentaire. La structure ainsi obtenue (ADN double brin) a une bien plus grande stabilité que l'ADN seul (simple brin) et possède la conformation spatiale en double hélice bien connue, et se replie sur elle-même à plusieurs reprise pour former des chromosomes ; les données contenues dans l'ensemble des chromosomes d'un organisme portent le nom de *génome*.

Séquencer un organisme, c'est trouver la séquence d'acides nucléiques correspondant à chacun de ses chromosomes. En laboratoire, il est possible de "lire" une séquence d'ADN lorsque la molécule étudiée ne fait pas plus de quelques centaines de bases de long. Toute la difficulté du séquençage consiste dans le fait que les organismes vivants ont des génomes dont la longueur oscille entre plusieurs centaines de milliers (580 000 pour *Mycoplasma genitalium*) à plusieurs milliards (75 milliards pour le gui). Pour parvenir tout de même à séquencer les organismes, on découpe leurs génomes en morceaux séquençables puis on s'efforce de reconstituer la séquence initiale à partir des nombreux morceaux obtenus (en fait, on découpe différemment plusieurs copies identiques du génomes pour pouvoir réordonner les différents morceaux en utilisant leurs chevauchements). On comprend bien ici l'immensité du travail à accomplir pour séquencer un organisme donné ce qui explique le nombre assez restreint d'espèces ayant déjà été séquencées dont on présente une liste dans la table 1.1.

Notons que les molécules d'ADN possèdent de nombreuses propriétés remarquables et notamment la capacité à se *répliquer* par complémentarité qui est, d'une certaine façon, caractéristique de la vie.

1.1.3 Les protéines

Ce qui constitue un organisme n'est cependant pas l'ADN lui-même mais les produits que cet organisme fabrique avec l'ADN : les protéines (membrane des cellules, hémoglobine, enzymes, ...). Dans une séquence d'ADN, il existe des zones *codantes* qui peuvent sous certaines conditions donner naissance à des molécules simple brin ressemblant à l'ADN, l'ARN messager (Acide Ribo-Nucléique) . Dans le cas des procaryotes, il s'agit d'une simple copie d'une partie de la séquence d'ADN initiale tandis que dans le cas des eucaryotes, il s'agit d'une copie dans laquelle on élimine certaines parties de la séquence initiale, les *introns*, pour concaténer les parties s'exprimant, les *exons*. Il existe une différence chimique notable entre l'ADN et l'ARN : la thymine y est remplacée par l'uracile (`u`). Dans un souci de simplification on notera cependant de façon indifférente `t` pour la thymine ou l'uracile.

Ces séquences d'ARN messager sont alors "lues" par les *ribosomes* qui associent à chaque codon (triplet de lettres) un acide aminé (voir la table 1.2 pour la liste des acides aminés) selon le code génétique présenté de la table 1.3. Cette étape de traduction prend fin lorsqu'un codon *stop* est rencontré,

acide aminé	**symboles**		
Alanine	`Ala`	`A`	neutres et hydrophobes
Valine	`Val`	`V`	
Leucine	`Leu`	`L`	
Isoleucine	`Ile`	`I`	
Proline	`Pro`	`P`	
Tryptophane	`Trp`	`W`	
Phenylalanine	`Phe`	`F`	
Méthionine	`Met`	`M`	
Glycine	`Gly`	`G`	neutres et polaires
Serine	`Ser`	`S`	
Threonine	`Thr`	`T`	
Tyrosine	`Tyr`	`Y`	
Cysteine	`Cys`	`C`	
Asparagine	`Asn`	`N`	
Glutamine	`Glu`	`Q`	
Lysine	`Lys`	`K`	basiques
Arginine	`Arg`	`R`	
Histidine	`His`	`H`	
Acide Aspartique	`Asp`	`D`	acides
Acide Glutamique	`Glu`	`E`	

TAB. 1.2 – Liste des 20 acides aminés regroupés en fonction de leurs propriétés physico-chimiques.

$1^{er}\backslash 2^{ème}$	t		c		a		g		$3^{ème}$
t	ttt	F	tct	S	tat	Y	tgt	C	t
	ttc		tcc		tac		tgc		c
	tta	L	tca		taa	*	tga	*	a
	ttg		tcg		tag		tgg	W	g
c	ctt	L	cct	P	cat	H	cgt	R	t
	ctc		ccc		cac		cgc		c
	cta		cca		caa	Q	cga		a
	ctg		ccg		cag		cgg		g
a	aat	I	act	T	aat	N	agt	S	t
	aac		acc		aac		agc		c
	aaa		aca		aaa	K	aga	R	a
	aag	M	acg		aag		agg		g
g	gtt	V	gct	A	gat	D	ggt	G	t
	gtc		gcc		gac		ggc		c
	gta		gca		gaa	E	gga		a
	gtg		gcg		gag		ggg		g

TAB. 1.3 – Code génétique. * désigne le codon stop. On utilise successivement les trois lettres du codon pour trouver l'acide aminé auquel il correspond. Exemple : le codon agc correspond à la Sérine (S) ; on trouve ce résultat dans la case ligne a et colonne g sur la ligne c.

on obtient alors une molécule formée par une succession d'acides aminés qui se replie selon une structure tridimensionnelle qui donne à la protéine ses propriétés physico-chimiques.

Les protéines sont en général constituées de quelques centaines d'acides aminés à quelques milliers pour les plus grosses et sont impliquées dans tous les mécanismes du vivant. L'un des défis majeurs de la biologie moléculaire consiste à pouvoir prédire la façon dont s'effectuent les repliements des protéines. Ces repliements sont en effet primordiaux pour pouvoir comprendre les fonctions des protéines étudiées et sont malheureusement très difficiles à étudier en laboratoire (cristallographie, résonance magnétique nucléaire). On ne dispose actuellement pas de méthode prédictive réellement performante pour aller des séquences aux formes tridimensionnelles.

enzyme	**site**
Bsu6633I	cgcg
Bsu8565I	ggatcc
Eco92I	ccgcgg
Eco120I	ggtctc

TAB. 1.4 – Exemples de couples enzyme/site de restriction issue de deux organismes : *Bacillus subtilis* (noms commençant par Bsu) et *Escherichia coli* (noms commençant par Eco). Informations issues de la base de données REBASE (voir [RM01]).

1.2 Exemples de mots exceptionnels

On a vu qu'il existe deux structures naturelles de séquences dans le vivant : les séquences d'ADN écrites dans un alphabet à quatre lettres

$$\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$$

et les séquence d'acides aminés écrites dans un alphabet à 20 lettres

$$\mathcal{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{D}, \mathtt{E}, \mathtt{F}, \mathtt{G}, \mathtt{H}, \mathtt{I}, \mathtt{K}, \mathtt{L}, \mathtt{M}, \mathtt{N}, \mathtt{P}, \mathtt{Q}, \mathtt{R}, \mathtt{S}, \mathtt{T}, \mathtt{V}, \mathtt{X}\}.$$

Bien que les molécules que ces séquences décrivent aient une structure tridimensionnelles et des propriétés qui y sont liées, certaines successions de leurs éléments, certains motifs (unidimensionnels) peuvent néanmoins, eux aussi, jouer un rôle particulier. Nous allons justement ici examiner quelques exemples de ce type.

1.2.1 Sites de restriction

On considère tout d'abord l'exemple des *sites de restriction*. Les bactéries doivent faire face à de constantes agressions virales. Les virus infectent en effet les bactéries en y introduisant leur génome (ADN le plus souvent simple brin au départ) et en utilisant la machinerie chimique de leur hôte pour se reproduire et proliférer jusqu'à épuisement des ressources ; jusqu'à la mort de la cellule.

Pour faire face à cette menace, les bactéries utilisent des enzymes dites de restriction dont le rôle est de découper l'ADN en un point précis : le site de restriction (voir la table 1.4 pour quelques exemples de couples enzymes/site de restriction). Lorsqu'un virus pénètre dans la cellule, il se retrouve exposé à une ou plusieurs de ces enzymes. S'il a le malheur de comporter dans sa

organisme	**chi**
Bacillus subtilis	`ccgct`
Escherichia coli	`gctggtgg`
Haemophilus influenzae	`g.tggtgg`

TAB. 1.5 – Exemples de motifs chi pour plusieurs organismes. Le symbole " . " signifie n'importe quelle lettre c'est à dire que le chi de *Haemophilus influenzae* correspond en fait à une famille des quatre mots obtenus en remplaçant le point par les quatre lettres possibles.

séquence certains des sites reconnus par ces enzymes, il est découpé en autant de parties et devient inoffensif.

Une telle méthode de défense est efficace mais présente une difficulté : comment éviter que les mêmes enzymes qui assurent la protection de la bactérie ne découpent leur propre génome y provoquant des dégâts irréparables et finalement là encore, la mort de la cellule ? Une seule solution très simple : éviter la présence des sites de restriction dans le génome de la bactérie.

En effet, la bactérie possède des mécanismes de réparations efficace (structure double brins en particulier) qui lui permettent de reconstruire des morceaux de l'ADN découpé par les enzymes de restrictions. Pour que ces mécanismes puissent faire face à l'action des enzymes il suffit que les sites de restriction soit suffisamment rares et c'est exactement ce qui se passe dans la réalité.

La sélection naturelle fait peu à peu disparaître les bactéries dans le génomes desquelles les sites de restrictions apparaissent trop fréquemment.

1.2.2 Chi

On considère ici un autre mécanisme de défense des bactéries contre les agressions virales avec l'utilisation des motifs *chi* (*Crossover Hotspot Investigation*). Les bactéries possèdent des protéines particulières appelées *nucléases* dont le rôle est la dégradation des molécules d'ADN en ses éléments simples : les acides nucléiques. Lorsqu'une nucléase "rencontre" un brin d'ADN, elle le parcourt à partir d'une extrémité en le détruisant jusqu'à ce qu'il n'en reste plus rien.

Un virus pénétrant dans la cellule sera donc rapidement dégradé par des nucléases et ne menacera ainsi pas la santé de son hôte. Comme dans le cas des sites de restriction, il y a cependant un problème : comment éviter que ces nucléases ne "dévorent" le génome de la bactérie dont elles sont issues ?

Un premier élément de réponse réside dans le fait que ces bactéries ont

organisme	**uptake**
Neisseria meningitidis	`gccgtctgaa`
Haemophilus influenzae	`aagtgcggt`

TAB. 1.6 – Exemples de motifs uptake pour plusieurs organismes.

un génome circulaire n'offrant pas de "prise" aux nucléases, cependant, ce génome va être amené à s'ouvrir à de nombreuses reprises au cours des cycles de la vie de l'organisme (réplication, synthèse des protéines, ...) s'exposant alors à l'action des nucléases.

Les bactéries ont donc mis au point une astuce, un mot de passe signifiant aux nucléases qu'elles se trompent de cible : le motif chi. On trouvera plusieurs exemples de motifs chi dans la table 1.5. Lorsqu'une nucléase rencontre ce motif sur l'ADN qu'elle dégrade, cela inhibe son action.

Afin de pouvoir survivre grâce à ses activités de réparation, la bactérie "doit" donc, là encore par pression de sélection, disposer suffisamment de motifs Chi tout au long de son génome (voir [EKBSG99]).

1.2.3 Uptake

On considère enfin un dernier exemple avec les séquences Uptake qui jouent un rôle dans les stratégies de survie des organismes pathogènes (voir [SGS99]). Ces organismes sont soumis à une très forte sélection car leurs cibles ne cessent de trouver des parades à leurs actions ce qui amène ces organismes pathogènes au choix suivant : être inventif ou mourir.

L'inventivité d'un organisme se mesure à sa capacité à transformer son génome par le biais de mutations (erreurs dans la réplication de l'ADN) ou encore de transferts horizontaux au cours desquels des morceaux entiers de génomes sont échangés entre deux génomes d'une même espèce voir même d'espèces différentes. Si de telles modifications ne sont bien souvent pas viables, elles permettent aussi parfois l'intégration dans le génome de nouvelles propriétés ou fonctionnalités.

Les séquences uptake (voir la table 1.6 pour quelques exemples) jouent un rôle important dans ces mécanismes de transferts horizontaux et s'affirment ainsi comme de nouveaux exemples de mots exceptionnels.

Nous avons donc donné des exemples de mots :
- dont le nombre d'occurrences est particulièrement faible pour des raisons biologiques (sites de restriction) ;
- dont le nombre d'occurrences est particulièrement élevé pour des raisons biologiques (chi et uptake).

Nous reviendrons sur ces différents exemples dans le chapitre 10 (page 149).

Références

[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.

[LBB+95] H. Lodish, D. Baltimore, A. Berk, S. L. Zipursky, P. Matsudaira, and J. Darnell. *Molecular Cell Biology.* Scientific American Book, Inc, 1995.

[RM01] R.J. Roberts and D. Macelis. Rebase - restriction enzymes and methylases. *Nucleic Acids Research*, 29 :268–269, 2001.

[SGS99] H.O. Smith, M.L. Gwinn, and Salzberg S.L. Dna uptake signal sequences in naturally transformable bacteria. *Res. Microbiol.*, 150(9-10) :603–616, Nov-Dec 1999.

FIG. 2.1 – Représentation graphique des mots de longueur $h = 9$ chez *Mycoplasma genitalium* dans le modèle $M1$. On consultera la section 11.1 page 160 pour plus de détails.

Chapitre 2

Modèles markoviens

Pré-requis :
Aucun.
Description :
Introduction des modèles markoviens et de la notion de significativité de comptages ainsi que présentation des différentes notations qui seront utilisées dans la suite.
Résumé :
A travers l'exemple du texte du "Cyrano de Bergerac" d'*Edmond Rostand*, on présente le problème de la détection des mots exceptionnels.

Cela nous permet d'introduire la notion de modèle statistique et en particulier de modèles markoviens avant de les formaliser.

On définit ensuite les notations concernant les comptages de mots ou de motifs et on explique la notion de significativité de ces comptages.

L'estimation des paramètres des modèles de *Markov* est ensuite abordée avant que d'autres exemples de modèles ne soient finalement brièvement présentés.

Contenu du chapitre

2.1 Motivations

Comme on l'a vu au travers des exemples du chapitre précédent, on peut relier le caractère exceptionnels de certains motifs nucléiques d'un point de vue biologique au caractère *inhabituel* de leur nombre d'occurrences dans la séquence étudiée. Cependant, pour pouvoir être en mesure de détecter une fréquence de motif inhabituelle ou *inattendue*, il convient de définir une "norme", un comportement typique de la séquence nous permettant d'évaluer le nombre d'occurrence *attendu* pour un motif donné et de le comparer avec le nombre observé.

La notion abordée ici n'est autre que celle de *modèle statistique* qui va être explorée au travers d'un exemple concret. Considérons la séquence `cyrano` composée du texte intégral en version française la comédie d'*Edmond Rostan*, "Cyrano de Bergerac". On considère que cette séquence est écrite dans un alphabet à 26 lettres en éliminant les espaces du texte initial et en remplaçant tous les caractères accentués par leurs versions sans accents et on obtient alors un texte de longueur $n = 154\,623$ (voir figure 2.2 pour un extrait).

On va s'intéresser à l'ensemble des mots de trois lettres de l'alphabet $\mathcal{A} = \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \ldots, \mathtt{x}, \mathtt{y}, \mathtt{z}\}$ (c'est à dire au 26^3 éléments de l'ensemble $\mathcal{A}^3 = \{\mathtt{aaa}, \mathtt{aab}, \mathtt{aac}, \ldots\}$) et tenter d'y détecter des éléments exceptionnels.

```
> Cyrano de Bergerac n = 154 623 lettres
edmondrostandcyranodebergeraccomedieheroiqueencinqactesenver
srepresenteeaparissurletheatredelaportesaintmartinledecembre
cestalamedecyranoquejevoulaisdediercepoememaispuisquelleapas
seenvouscoquelincestavousquejelededieerpersonnagescyranodebe
rgeracchristiandeneuvillettecomtedeguicheragueneaulebretcarb
ondecasteljalouxlescadetslignieredevalvertunmarquisdeuxiemem
arquistroisiememarquismontfleurybellerosejodeletcuigybrissai
lleunfacheuxunmousquetaireunautreunofficierespagnolunchevaul
egerleportierunbourgeoissonfilsuntirelaineunspectateurungard
ebertrandoulefifrelecapucindeuxmusicienslespoeteslespatissie
[...]
isjamaisahtevoilatoilasottisejesaisbienqualafinvousmemettrez
abasnimportejemebatsjemebatsjemebatsilfaitdesmoulinetsimmens
esetsarretehaletantouivousmarracheztoutlelaurieretlarosearra
chezilyamalgrevousquelquechosequejemporteetcesoirquandjentre
raichezdieumonsalutbalaieralargementleseuilbleuquelquechoseq
uesansunplisansunetachejemportemalgrevousilselancelepeehaute
etcestlepeesechappedesesmainsilchancelletombedanslesbrasdele
bretetderagueneauroxanesepenchantsurluietluibaisantlefrontce
stcyranorouvrelesyeuxlareconnaitetditensouriantmonpanacherid
eau
```

FIG. 2.2 – Extrait le la séquence `cyrano` au format FASTA dans l'alphabet $\mathcal{A} = \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \ldots, \mathtt{x}, \mathtt{y}, \mathtt{z}\}$
.

Pour cela on commence par examiner simplement les nombres d'occurrences de tous ces mots dans la séquence :

aaa	0	aba	33	...	...	zza	0
aab	0	abb	0	...	...	zzb	0
aac	1	abc	0	...	...	zzc	0
⋮		⋮				⋮	
aaz	0	abz	0	...	...	zzz	0

On remarque immédiatement qu'une grande majorité des mots de trois lettres ne sont pas présents dans la séquence : 13 428 mots sur les $26^3 = 17\ 576$ possibles ont un comptage nul (soit plus de 75% d'entre eux). Cela n'est guère étonnant car la langue française interdit de nombreuses combinaisons de lettres, mais, de ce fait, tous les mots de comptages non nuls ont un caractère exceptionnel.

Voici les 10 premiers mots de cette liste (ordre alphabétique) :

mot	`aac`	`aai`	`aal`	`aam`	`aar`	`aas`	`aau`	`aav`	`aba`	`abd`
comptage	1	1	4	1	3	1	5	1	33	1

Si certains d'entre eux ne semblent pas faire partie de la langue française, il ne faut pas s'en inquiéter outre mesure. En effet, la séquence étudiée ne comportant pas d'espaces (voir figure 2.2) il est possible que des combinaisons de lettres étranges voient le jour aux jonctions entre plusieurs mots habituellement séparés (ex : "`...il se leve et va a carbon de castel-jaloux...`" qui devient "`...ilseleveetvaacarbondecasteljaloux...`").

Une façon naturelle de distinguer entre eux les mots de comptages nuls peut consister à dire : "plus le nombre d'occurrences d'un mot est grand, plus ce mot est exceptionnel". On obtient ainsi un classement des mots dont voici le 10 premiers :

mot	`ant`	`ran`	`ent`	`ous`	`que`	`ano`	`yra`	`cyr`	`les`	`est`
comptage	1066	991	973	847	841	823	791	786	780	713

Certains des mots mis en évidence le doivent vraisemblablement à la fréquence de leur utilisation dans la langue française (`ant`, `ent`, `les`, ...) tandis que d'autres, plus exotiques, le doivent indéniablement à la spécificité du texte "Cyrano de Bergerac" (on pense ici à `ano`, `yra` et `cyr`, trois mots faisant partie du nom `cyrano` par exemple).

On peut s'étonner de constater que cette approche, malgré sa naïveté, met en évidence des propriétés de la langue française mais aussi des spécificités de la séquence étudiée et soulève donc, par la même occasion, l'épineux problème de la distinction de ces deux types d'informations. Sur ce point, on se contentera ici de dire qu'il est évidemment impossible d'effectuer ce distinguo au vu de la séquence seule, sans connaissance préalable particulière (connaissance de la langue française dans notre cas) ; dans la suite on ne cherchera donc pas de causes extérieures à la séquence considérée.

Sans l'avoir dit, on a ici implicitement supposé qu'aucun des mots de trois lettres n'était privilégié par rapport aux autres, ce qui correspond à un modèle d'indépendance dans lequel les lettres seraient identiquement distribuées. Ne serait-ce qu'à cause des fortes différences dans les fréquences d'utilisation des lettres de l'alphabet en français, une telle hypothèse ne correspond évidemment pas à la réalité dans le texte considéré. En effet, les trois lettres les plus fréquentes sont `e`, `a` et `s`, avec les comptages $n(\mathtt{e}) = 25\ 141$, $n(\mathtt{a}) = 13\ 184$ et $n(\mathtt{s}) = 11\ 753$, tandis que les trois les moins fréquentes sont `w`, `k` et `z`, avec les comptages $n(\mathtt{w}) = 0$, $n(\mathtt{k}) = 1$ et $n(\mathtt{z}) = 556$, il apparaît alors clairement qu'un traitement identique de toutes les lettres n'est pas approprié.

Avec de tels déséquilibres, certaines observations vont, à l'évidence, davantage surprendre que d'autres ; $n(\mathtt{aaa}) = n(\mathtt{zzz}) = 0$ mais si, pour `zzz`, cela n'est guère étonnant car la lettre `z` est rare, il est remarquable que la grande fréquence de la lettre `a` ne favorise pas l'apparition du mot `aaa`.

En considérant les fréquences des lettres, il apparaît clairement que le comptage nul du mot `aaa` fait indéniablement de celui-ci un mot exceptionnel. Cette information sans connaissance *a priori* autre que le texte lui-même parvient ici à rendre compte d'un phénomène caractéristique de la langue française : la lettre `a` y est fort rarement répétée et donc *a fortiori*, encore plus rarement triplée.

Cette dernière remarque soulève une nouvelle question : ne pourrait-on essayer de tenir compte de la rareté du mot `aa` avant de déclarer le mot `aaa` exceptionnel ? Plus généralement, y a-t-il un moyen de tenir compte des comptages des mots de deux lettres pour examiner les comptages des mots de trois lettres ?

Nous verrons dans la section suivante quelle réponse précise on peut apporter à cette question. Pour l'instant, on va se contenter d'un exemple assez parlant avec les mots `aaa` et `qur`. Ces deux mots sont de comptages nuls mais les mots de lettres les composant sont de fréquences très différentes : $n(\mathtt{aa}) = 17$ alors que $n(\mathtt{qu}) = 1627$ et $n(\mathtt{ur}) = 1661$. A la lecture de ces chiffres, on conclut sans mal que, si le comptage nul de `aaa` n'est guère étonnant, celui de `qur` l'est très certainement.

On met ainsi à nouveau en évidence une information concernant la langue française : un `r` ne doit pas (ou bien très rarement) suivre le mot `qu`. Si cette remarque est naturelle dans la mesure où le mot `qur` n'est pas présent dans la séquence, c'est, compte tenu des fréquences des mots de deux lettres, le caractère exceptionnel de son comptage qui a cependant permis de le remarquer dans la masse des mots de trois lettres de comptages nuls.

Avant de s'atteler aux définitions formelles de ces différentes notions, il est intéressant de considérer un dernier exemple avec les mots `ana` et `din`. On se place dans un modèle d'indépendance, et comme les comptages de ces deux mots sont égaux, $n(\texttt{ana}) = n(\texttt{din}) = 40$, on s'attend à leur trouver des caractères exceptionnels identiques. En fait, la structure d'*auto-recouvrement* du premier mot facilite l'apparition de davantage de ses occurrences. En effet, si le mot `ana` est présent dans la séquence, il suffit que le mot de deux lettres `na` soit présent à sa suite pour qu'une nouvelle occurrence du mot n'apparaisse comme cela est représenté sur le schéma suivant :

```
   ↓ ↓
...anana...
   ana
     ana
```

Cet exemple, pris parmi une multitude d'autres, met bien en évidence la nature complexe du problème considéré et justifie pleinement l'établissement de calculs de significativités plutôt que de simples écarts à l'attendu. De plus, il est clair que ce phénomène gagne en complexité et en importance avec la longueur des mots examinés et à mesure que le cardinal de l'alphabet considéré diminue. Par conséquent, son influence sera donc particulièrement forte dans le cas de l'étude de séquences d'ADN (écrites, comme on le sait, dans un alphabet à quatre lettres).

2.2 Définitions

2.2.1 Notations

On introduit tout d'abord quelques notations élémentaires qui resteront valables tout au long du document.

Notation 2.1 (Alphabet)
$\mathcal{A}$ désigne un alphabet fini de cardinal k dont on peut supposer, sans perte de généralité, qu'il est égal à $\{1, \ldots, k\}$.

Notation 2.2 (Séquence)
$x = x_1 \ldots x_n$ avec $x_i \in \mathcal{A}$ $\forall i \in \{1, \ldots, n\}$ désigne une séquence observée de longueur n. On notera $X = X_1 \ldots X_n$ avec, $\forall i \in \{1, \ldots, n\}$, X_i variable aléatoire à valeurs dans $\mathcal{A}$, une séquence aléatoire de longueur n (dont une séquence x est donc une réalisation).

Quitte à rajouter une lettre supplémentaire identique à la première en fin de séquence, on peut supposer que la première lettre de la séquence est *la même que la première ; que la séquence est circulaire. Une telle hypothèse change à peine la nature du problème (il s'agit d'un "effet de bord" que l'on peut d'autant plus facilement négliger que les longueurs des séquences considérées sont importantes) et simplifiera notablement les notations dans la suite.*

Notation 2.3 (Mot)
$W = w_1 \dots w_h$ *avec* $w_i \in \mathcal{A}\ \forall i \in \{1, \dots, h\}$ *désigne un mot de longueur* h *où un* h-mot.

On désigne par

$$N(W) = \sum_{i=1}^{n-h+1} Y_i$$

où $Y_i = \mathbb{I}_{X_i=w_1} \times \dots \times \mathbb{I}_{X_{i+h-1}=w_h}$ *est l'indicatrice de la présence du mot* W *à la position* i *dans* X.

Et on désigne par

$$n(W) = \sum_{i=1}^{n-h+1} y_i$$

où $y_i = \mathbb{I}_{x_i=w_1} \times \dots \times \mathbb{I}_{x_{i+h-1}=w_h}$ *est l'indicatrice de la présence du mot* W *à la position* i *dans* x.

Ces notations étant posées, on peut maintenant définir les différents modèles que nous allons utiliser.

2.2.2 Modèles indépendants

Le modèle le plus simple consiste à considérer que chaque lettre de la séquence est tirée indépendamment des autres selon une loi identique.

Définition 2.4 (Modèle $M0$)
$(X_i)_{i\in\{1,\dots,n\}}$ *est un échantillon de taille* n *de loi* μ *(*i.e. *les* n *variables aléatoires son indépendantes et identiquement distribuées* - i.i.d. - *selon* μ*).*

Ainsi, si $W = w_1 \dots w_h$ est un mot de longueur h, on peut définir son *nombre d'occurrences attendues* sous le modèle $M0$ par

$$\mathbb{E}[N(W)] = \mu(w_1) \times \dots \times \mu(w_h) \times (n - h + 1).$$

Dans le cas particulier où μ est la distribution uniforme sur $\mathcal{A}$ on est ramené à un modèle plus simple.

Définition 2.5 (Modèle $M00$)
$(X_i)_{i\in\{1,\dots,n\}}$ *est un échantillon de taille* n *de loi uniforme sur* $\mathcal{A}$.

Dans ce modèle très simple, le nombre d'occurrences attendues pour un mot donné ne dépend que de sa longueur h

$$\mathbb{E}[N(W)] = \frac{n-h+1}{k^h}.$$

En reprenant l'exemple de la section 2.1, ($n = 154\ 623$ et $k = 26$) on peut calculer $\mathbb{E}[N(W)] = 8.80$ pour un mot de longueur $h = 3$.

2.2.3 Modèles markoviens

Malgré leur faible degré de sophistication, les modèles indépendants peuvent souvent rendre de grands services. Ils ne peuvent cependant faire intervenir les comptages des sous-mots de longueurs supérieures ou égales à deux, dans l'examen d'un mot donné. Cette lacune, nous avait par exemple amenés dans la section précédente à juger le mot `aaa` très exceptionnel par son absence de la séquence `cyrano` alors que la lettre `a` y est précisément très fréquente, alors que le mot `aa` est lui-même peu fréquent ($n(\texttt{aa}) = 17$).

Pour remédier à ce problème, on introduit les modèles à dépendance markovienne qui sont tout à fait naturels dans un sens que l'on précisera en section 2.3.

Définition 2.6 (Modèle $M1$)
On suppose que $(X_i)_{i\in\{1,\ldots,n\}}$ est une chaîne de Markov *d'ordre* 1 *ce qui signifie que, $\forall i \in \{1,\ldots,n\}$ et $\forall x_1,\ldots,x_i \in \mathcal{A}$, on a*

$$\begin{aligned}\mathbb{P}(X_i = x_i|X_1 = x_1,\ldots,X_{i-1} = x_{i-1}) &= \mathbb{P}(X_i = x_i|X_{i-1} = x_{i-1})\\ &= \Pi(x_{i-1}, x_i).\end{aligned}$$

De plus, Π, matrice stochastique, s'appelle matrice de transition *de la chaîne de* Markov *et une distribution μ vérifiant $\mu\Pi = \mu$ est appelée* distribution stationnaire.

Ceci étant posé on peut alors définir le nombre d'occurrences attendues sous le modèle $M1$ d'un mot $W = w_1 \ldots w_h$ de longueur h par

$$\mathbb{E}[N(W)] = \mu(w_1) \times \Pi(w_1, w_2) \times \ldots \times \Pi(w_{h-1}, w_h) \times (n-h+1).$$

On peut généraliser ce modèle en

Définition 2.7 (Modèle Mm)
On suppose que $(X_i)_{i\in\{1,\ldots,n\}}$ est une chaîne de Markov *d'ordre m ce qui signifie que, $\forall i \in \{1,\ldots,n\}$ et $\forall x_1,\ldots,x_i \in \mathcal{A}$, on a*

$$\begin{aligned}\mathbb{P}(X_i = x_i|X_1 = x_1,\ldots,X_{i-1} = x_{i-1}) &= \mathbb{P}(X_i = x_i|X_{i-m} = x_{i-m},\ldots X_{i-1} = x_{i-1})\\ &= \Pi(x_{i-m},\ldots,x_{i-1}, x_i).\end{aligned}$$

Si Π gère les transitions de la chaîne, on note μ la distribution sur $\mathcal{A}^m$ vérifiant, $\forall x_1, \ldots, x_m, x_{m+1} \in \mathcal{A}$,

$$\sum_{x_1 \in \mathcal{A}} \mu(x_1, \ldots, x_m)\Pi(x_1, \ldots, x_{m+1}) = \mu(x_2, \ldots, x_{m+1}).$$

Si $m < h$, le nombre d'occurrences attendues du mot $W = w_1 \ldots w_h$ de longueur h sous le modèle Mm est alors défini par

$$\mathbb{E}[N(W)] = \mu(w_1, \ldots, w_m) \times \Pi(w_1, \ldots, w_{m+1}) \times \ldots \times \Pi(w_{h-m}, \ldots, w_h) \times (n-h+1).$$

Notons au passage que

Remarque 2.8 (Emboitement des modèles)
Soient M et M' deux modèles, on pose alors la relation suivante : $M \subset M'$ qui signifie que M peut être vu comme un cas particulier de M'. On a alors

$$M00 \subset M0 \subset M1 \subset \ldots \subset Mm \subset Mm+1 \subset \ldots$$

2.2.4 Significativité

Comme on l'a vu en fin de section 2.1 avec l'exemple du mot `ana`, l'écart entre nombre d'occurrences observées d'un mot et son nombre d'occurrences attendues ne peut, à lui seul, être un indicateur fiable du degré d'exceptionalité des observations par rapport au modèle.

Pour résoudre ce problème on introduit ici la notion de *significativité statistique* d'un événement qui est, tout simplement, la probabilité de cet événement dans le modèle choisi.

Si on considère une séquence aléatoire de longueur $n = 154\ 623$ tirée suivant le modèle $M00$ et qu'on examine l'événement :

$$\begin{aligned} A &= \{\text{on compte 40 occurrences du mot } \texttt{ana} \text{ dans la séquence}\} \\ &= \{N(\texttt{ana}) = 40\}. \end{aligned}$$

Comme dans la séquence `cyrano` de même longueur n on observe $n(\texttt{ana}) = 40$ occurrences du mot `ana`, la significativité de ce comptage dans le modèle $M00$ sera $\mathbb{P}(A)$.

Ainsi, d'une manière générale, la significativité du comptage observé d'un mot W par rapport à un modèle donné sera définie comme la probabilité suivante :

$$\mathbb{P}\big(N(W) = n(W)\big)$$

(on pourra consulter la proposition 4.7 page 70 pour obtenir des précisions sur la nature des relations entre $\mathbb{P}\big(N(W) \geqslant n(W)\big)$ et la probabilité précédente) et c'est cette quantité que l'on va dans toute la suite chercher à calculer ou à approcher.

2.3 Estimation des paramètres

Si le modèle $M00$ ne requiert aucun autre paramètre que la taille k de l'alphabet considéré, les autres modèles présentés en section 2.2.3 nécessitent de nombreux paramètres.

Comme on souhaite évaluer la significativité des comptages par rapport aux données que l'on possède, on va, en général, estimer ces paramètres sur des séquences et sur la séquence étudiée en particulier.

On commence ici par traiter les cas simples des modèles $M0$ et $M1$ avant de considérer le modèle Mm en toutes généralités.

2.3.1 Modèle $M0$

Dans le cas du modèle $M0$ il faut disposer d'une loi μ sur $\mathcal{A}$, c'est à dire estimer k paramètres.

Proposition 2.9 *L'estimateur $\widehat{\mu}$ défini par*

$$\widehat{\mu}(x) = \frac{n(x)}{n} \quad x \in \mathcal{A}$$

maximise la vraisemblance de l'observation.

Preuve. la *log-vraisemblance* de l'observation $x = x_1 \dots x_n$ sous la loi μ est

$$\begin{aligned} L &= \sum_{i=1}^{n} \log \mu(x_i) \\ &= \sum_{x \in \mathcal{A}} n(x) \log \mu(x) \end{aligned}$$

En appliquant le lemme 2.10 présenté ci-après, on constate que le maximum de la log-vraisemblance L (vue comme une fonction de μ est atteint en le $\widehat{\mu}$ de la proposition. ■

Lemme 2.10 *On considère la fonction*

$$\begin{array}{cccc} f : & \mathbb{R}^d & \to & \mathbb{R} \\ & (x_1, \dots, x_d) & \mapsto & \sum_{i=1}^{d} c_i \log x_i \end{array}$$

où $(c_1, \dots, c_d) \in (\mathbb{R}^+)^d$ avec $C = \sum_{i=1}^{d} c_i \neq 0$, alors

$$\inf_{x \in \Gamma} f(x) = f\left(\frac{c_1}{C}, \dots, \frac{c_d}{C}\right).$$

Preuve. La fonction f est continue sur le compact Γ et atteint donc son minimum sur Γ en un certain $a \in \Gamma$. On pose

$$\begin{array}{cccc} g: & \mathbb{R}^d & \to & \mathbb{R} \\ & (x_1, \ldots, x_d) & \mapsto & \sum_{i=1}^d x_i - 1 \end{array}$$

ce qui permet d'écrire $\Gamma = \{x \in (\mathbb{R}^+)^d, g(x) = 0\}$ et on sait alors, grâce au théorème des extrémas liés, que $\exists \lambda \in \mathbb{R}$ tel que $df_a = \lambda dg_a$ c'est à dire tel que

$$\frac{\partial f}{\partial x_i}(a) = \lambda \frac{\partial g}{\partial x_i}(a) \quad \forall i \iff \frac{c_i}{a_i} = \lambda \quad \forall i$$

or comme $a = (a_1, \ldots, a_d) \in \Gamma$, il est clair que $\lambda = C$ ce qui donne le résultat recherché. ■

On peut utiliser la formule de la proposition pour estimer les paramètres du modèle $M0$ sur la séquence `cyrano`, après quoi on est en mesure d'examiner les comptages attendus de certains mots. On peut ainsi reprendre d'un point de vue numérique les remarques qualitatives effectuées en section 2.1.

On a par exemple

$$\mathbb{E}[N(\texttt{aaa})] = 95.85 \text{ et } \mathbb{E}[N(\texttt{zzz})] = 0.01$$

alors que les comptages de ces deux mots sont nuls. On confirme ainsi le caractère exceptionnel de `aaa` dans ce modèle tandis que l'observation concernant `zzz` semble très naturelle.

2.3.2 Modèle $M1$

Dans le cas du modèle $M1$, la matrice stochastique Π d'ordre k doit être estimée ce qui donne k^2 paramètres.

Proposition 2.11 *L'estimateur $\widehat{\Pi}$ défini par*

$$\widehat{\Pi}(x, y) = \frac{n(xy)}{\sum_{z \in \mathcal{A}} n(xz)} \quad x, y \in \mathcal{A}$$

maximise la vraisemblance de l'observation.

Preuve. On écrit la log-vraisemblance de l'observation dans l'ensemble des séquences commençant par x_1 et qui suivent le modèle $M1$

$$\begin{aligned} L &= \sum_{i=1}^{n-1} \log \Pi(x_i, x_{i+1}) \\ &= \sum_{x,y\in\mathcal{A}} n(xy) \log \Pi(x,y) \\ &= \sum_{x\in\mathcal{A}} \underbrace{\sum_{y\in\mathcal{A}} n(xy) \log \Pi(x,y)}_{L(x)} . \end{aligned}$$

Comme les termes $L(x)$ ne dépendent que des $\Pi(x,y)$ $y \in \mathcal{A}$, on peut les maximiser indépendamment les uns des autres et, en utilisant le lemme 2.10, on obtient aisément le résultat souhaité. ■

Pour pouvoir utiliser la proposition 2.11, il faut que

$$\sum_{z\in\mathcal{A}} n(xz) \neq 0 \quad \forall x \in \mathcal{A}.$$

Si on constate que cela n'est pas le cas pour un certain $x \in \mathcal{A}$, cela signifie en particulier que cette lettre n'apparaît pas dans la séquence observée (ou alors en dernière position ce qui revient alors au même en enlevant cette lettre). Il suffit alors de retirer x de $\mathcal{A}$ pour pouvoir à nouveau travailler.

Dans le cas de l'estimation des paramètres du modèle $M1$ sur la séquence `cyrano` on est ainsi amené à supprimer la lettre `w` de l'alphabet initial. Une fois les paramètres estimés, on peut alors à nouveau effectuer des calculs de comptages attendus et on trouve $\mathbb{E}[N(\texttt{aaa})] = 0.10$ (à mettre en relation avec le 95.85 du modèle $M0$), ce qui montre que le comptage nul du mot `aaa` est naturel dans le modèle $M1$. En revanche, on a $n(\texttt{qur}) = 0$ et $\mathbb{E}[N(\texttt{qur})] = 1210.75$, ce qui montre vraisemblablement que le mot `qur` est exceptionnel dans le modèle $M1$.

2.3.3 Modèle Mm

On peut généraliser l'estimation dans le modèle $M1$ à celle des k^{m+1} paramètres de Π dans le modèle Mm. On obtient alors la

Proposition 2.12 *L'estimateur $\widehat{\Pi}$ défini par*

$$\widehat{\Pi}(x_1, \ldots, x_{m+1}) = \frac{n(x_1 \ldots x_{m+1})}{\sum_{y\in\mathcal{A}} n(x_1 \ldots x_m y)} \quad x_1, \ldots, x_m \in \mathcal{A}$$

maximise la vraisemblance de l'observation.

Preuve. Il s'agit ici, d'une simple généralisation de la preuve dans le cas du modèle $M1$. On écrit la log-vraisemblance de l'observation dans l'ensemble des séquences commençant par $x_1 \dots x_m$ et qui suivent le modèle Mm

$$\begin{aligned} L &= \sum_{i=1}^{n-m} \log \Pi(x_i, \dots, x_{i+m+1}) \\ &= \sum_{x_1,\dots,x_{m+1} \in \mathcal{A}} n(x_1 \dots x_{m+1}) \log \Pi(x_1, \dots, x_{m+1}) \\ &= \sum_{x_1,\dots,x_m \in \mathcal{A}} \underbrace{\sum_{y \in \mathcal{A}} n(x_1 \dots x_m y) \log \Pi(x_1, \dots, x_m, y)}_{L(x_1,\dots,x_m)} . \end{aligned}$$

Comme les termes $L(x_1, \dots, x_m)$ ne dépendent que des $\Pi(x_1, \dots, x_m, y)$ $y \in \mathcal{A}$, on peut les maximiser indépendamment les uns des autres et, en utilisant le lemme 2.10, on obtient aisément le résultat souhaité. ■

On précise tout d'abord le commentaire du début de la section 2.2.3 par la

Remarque 2.13 *Les comptages des mots de longueur $m + 1$ interviennent dans l'estimation des paramètres du modèle Mm de sorte que, si l'on désire prendre en compte des sous-mots d'un mot de longueur h, il est naturel de se placer dans le modèle markovien d'ordre $m = h - 2$.*

D'autre part, on effectue la

Remarque 2.14 *Si $m = h - 1$ alors l'estimation des paramètres fait intervenir les comptages des mots de longueur h et il est alors aisé de montrer que*

$$\mathbb{E}[N(W)] \sim n(W)$$

lorsque la longueur n des séquences considérées est grande et, dans de telles conditions, le comptage observé du mot W ne peut être exceptionnel (remarquons que l'égalité ne peut être obtenue à cause des "effets de bords" au début et en fin de séquence).

Si $m > h - 1$ les comptages des mots de longueur h interviennent indirectement dans ceux des mots de longueur $h + 1$ et les mêmes conséquences que dans le cas précédent sont observées.

ce qui nous amène à supposer dans la suite que

$$m \leqslant h - 2.$$

2.4 Autres modèles markoviens

Il existe d'autres types de modèles classiquement utilisés pour l'étude des séquences génomiques. Ces modèles sont présentés ici de manière succincte.

2.4.1 Modèles périodiques

Comme on l'a vu dans le chapitre 1 (page 15), l'ADN codant à une structure périodique naturelle de période trois : les codons. Lorsque l'on étudie des occurences de mots dans des séquences codantes, on peut souhaiter tenir compte de la phase de lecture, c'est à dire de la position modulo trois des lettres dans la séquence.

Définition 2.15 (modèle Mm_3)
Le modèle Mm_3, selon la notation introduite par [Sch95], est ainsi défini :

$$\mathbb{P}(X_i = x_i | X_1 = x_1, \dots, X_{i-1} = x_{i-1}) = \Pi_{\varphi(i)}(x_{i-m}, \dots, x_{i-1}, x_i)$$

où

$$\varphi(i) = \begin{cases} 1 & si\ i \equiv 1[3] \\ 2 & si\ i \equiv 2[3] \\ 3 & si\ i \equiv 0[3] \end{cases}$$

désigne la phase dans laquelle la $i^{\text{ème}}$ lettre de la séquence se trouve.

Dans un tel modèle, on a donc trois fois plus de paramètres que dans un modèle Mm. En suivant le même principe que dans le cas des modèles précédents, il est alors possible d'effectuer une estimation de ces paramètres en utilisant encore une fois les comptages des mots de longueur $m + 1$ à ceci près qu'il faut traiter séparément les mots se terminant dans les trois phases.

Notons qu'il est possible de ramener ce modèle au cas markovien simple en utilisant un alphabet plus grand où chacune des lettres de l'alphabet initial existe en trois versions différentes selon la phase dans laquelle on se trouve. Il ne reste plus alors qu'à interdire les transitions faisant se succéder des "lettres" dans des phases incompatibles.

A titre d'exemple, considérons la séquence écrite dans l'alphabet $\mathcal{A} = \{\texttt{a}, \texttt{c}, \texttt{g}, \texttt{t}\}$:

$$x = \texttt{atgctcgatctcggtcgagcgcaa}$$

cette séquence se réécrit

$$x = \texttt{a}_1\texttt{t}_2\texttt{g}_3\texttt{c}_1\texttt{t}_2\texttt{c}_3\texttt{g}_1\texttt{a}_2\texttt{t}_3\texttt{c}_1\texttt{t}_2\texttt{c}_3\texttt{g}_1\texttt{g}_2\texttt{t}_3\texttt{c}_1\texttt{g}_2\texttt{a}_3\texttt{g}_1\texttt{c}_2\texttt{g}_3\texttt{c}_1\texttt{a}_2\texttt{a}_3$$

dans l'alphabet $\mathcal{A}_3 = \{\texttt{a}_1, \texttt{c}_1, \texttt{g}_1, \texttt{t}_1, \texttt{a}_2, \texttt{c}_2, \texttt{g}_2, \texttt{t}_2, \texttt{a}_3, \texttt{c}_3, \texttt{g}_3, \texttt{t}_3\}$ où il est bien clair qu'aucune occurrence du mot $\texttt{g}_1\texttt{c}_2\texttt{t}_1$ (par exemple) ne pourra jamais survenir.

2.4.2 Chaînes de *Markov* cachées

Tous les modèles présentés jusqu'à présent font une hypothèse très forte sur les séquences qu'ils sont sensé aider à étudier : l'hypothèse d'homogénéité. Bien évidemment, les longues séquences biologiques présentent bien souvent des segmentations naturelles en différents types (codant, non codant, codant sens direct, codant sens indirect, ...).

Les modèles de chaînes de *Markov* cachées permettent de prendre en compte cette segmentation en utilisant différents modèles markoviens pour les transitions à l'intérieur de chaque type de segment, et gèrent les transitions d'un type de segment à un autre à l'aide d'un autre modèle markovien (on pourra par exemple se reporter à [BP66], [Rab89] ou encore [Mur97] pour plus de détail sur le sujet).

Formellement, on considère $X = X_1 \dots X_n$ avec, pour tout i, $X_i \in \mathcal{A}$ une séquence aléatoire et $S = S_1 \dots S_n$, $S_i \in \{1, \dots, s\}$ la suite des types de segment ; des états cachés (cachés car ces états ne sont pas connus lorsque l'on observe une réalisation x de X). On a alors la définition suivante :

Définition 2.16 (modèle $M1 - Mm$)
La séquence S est générée par une chaîne de Markov *d'ordre* 1 *sur l'espace* $\{1, \dots, s\}$ *des états cachés et on a s modèles markoviens d'ordre m dont l'usage dépend de l'état dans lequel on se trouve :*

$$\mathbb{P}(X_i = x_i, S_i = s_i | X_1 = x_1, \dots, X_{i-1} = x_{i-1}) = \Pi_{s_i}(x_{i-m}, \dots, x_{i-1}, x_i).$$

Avec ce type de modèles, la grosse difficulté consiste à être capable d'estimer correctement les paramètres à partir d'une observation où les états cachés ne sont pas visibles. Il existe une abondante littérature traitant de ce problème et, plus généralement, des domaines d'application de ces modèle très utiles.

Remarquons qu'une fois l'estimation des paramètres effectuée, il est à nouveau possible de ramener ce modèle à un modèles markovien simple et cela de deux manières différentes :

(1) On augmente la taille de l'alphabet en considérant une version des lettres de l'alphabet pour chaque état caché ;

(2) On traite séparément chaque état caché comme un modèle markovien simple.

Si cette dernière méthode à l'avantage de la simplicité et est numériquement la plus avantageuse, elle a néanmoins le défaut d'interdire l'étude d'événements faisant intervenir plusieurs états cachés en même temps.

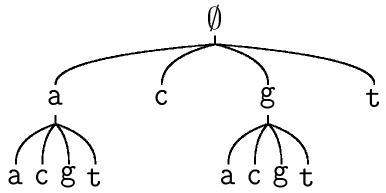


FIG. 2.3 – Exemple d'arbre de contexte complet dans l'alphabet $\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$.

2.4.3 Modèles à dépendance variable

On a vu en section 2.3 que l'estimation des paramètres d'un modèle markovien d'ordre m nécessite la connaissance des comptages des mots de longueur $m + 1$. Il est évident que si certains mots apparaissent rarement dans la séquence considérée, la qualité de l'estimation des paramètres correspondants va s'en ressentir. Pour résoudre ce problème, on introduit des modèles à dépendances variables qui permettent d'effectuer une économie de paramètres en ne conservant que ceux nécessaires à une bonne modélisation des données.

On représente de tels modèles avec des arbres de contexte (voir [WST95] et [Ver01] pour plus de détails). Dans l'exemple de la figure 2.3, on considère un modèle dans lequel les occurrences des lettres suivant un `a` ou `g` sont réglées par un modèle $M2$ tandis que celles qui suivent les lettres `c` ou `t` le sont par un modèle $M1$ pour un total final de 40 paramètres (au lieu des 64 paramètres du modèle $M2$ complet). Dans le cas de l'arbre représenté en figure 2.4, il n'est même plus possible de décrire le modèle résultant avec les modèles précédents. Ici, il ne reste plus que les 24 paramètres correspondant aux probabilités suivantes :

$$\begin{cases} \mathbb{P}(X_i = x | X_{i-2} = \mathtt{c}, X_{i-1} = \mathtt{a}) \\ \mathbb{P}(X_i = x | X_{i-2} \neq \mathtt{c}, X_{i-1} \neq \mathtt{a}) \\ \mathbb{P}(X_i = x | X_{i-1} = \mathtt{c}) \\ \mathbb{P}(X_i = x | X_{i-2} = \mathtt{a}, X_{i-1} = \mathtt{g}) \\ \mathbb{P}(X_i = x | X_{i-2} = \mathtt{g}, X_{i-1} = \mathtt{g}) \\ \mathbb{P}(X_i = x | X_{i-2} \neq \mathtt{a}, X_{i-2} \neq \mathtt{g}, X_{i-1} = \mathtt{g}) \\ \mathbb{P}(X_i = x | X_{i-1} = \mathtt{t}) \end{cases}$$

avec $x \in \mathcal{A}$.

En fait, on utilise ces arbres pour connaître le contexte pertinent en une position dans la séquence. Pour cela, on regarde d'abord la lettre précédant la position étudiée : si cette lettre ne possède pas de branche dans l'arbre menant à un nouveau nœud depuis la racine $\emptyset$ alors on s'arrête simplement ;

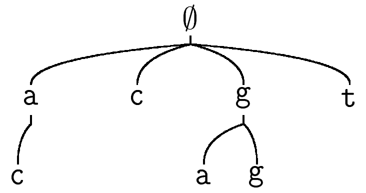


FIG. 2.4 – Exemple d'arbre de contexte incomplet dans l'alphabet $\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$.

sinon, on réitère le procédé en traitant le nœud comme une nouvelle racine. Le dernier nœud désigné par cet algorithme constitue le contexte pertinent recherché ; c'est à dire les éléments du "passé" qu'il faut considérer pour émettre un caractère à la nouvelle position.

De tels modèles représentent l'extension naturelle des modèles markoviens classiques. Remarquons cependant qu'il est possible de décrire ces modèles à l'aide de simples modèles markoviens dès que l'on est prêt à introduire de nombreux paramètres inutiles (il suffit pour cela de se placer dans le modèle de *Markov* dont l'ordre correspond à la plus grande taille de contexte). Une telle approche permet d'utiliser les résultats valables pour les modèles markoviens classiques dans le cadre de modèles à dépendance variable dont on aura, évidemment, pris soin d'estimer les paramètres au préalable.

Références

[BP66] L. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The Annals of Mathematicals Statistics.*, (37) :1554–1563, 1966.

[CMU+98] R. Cole, J. Mariani, H. Uszkoreit, G. B. Varile, A. Zaenen, and A. Zampolli. *Survey of the state of the art in human language technology.* Cambridge University Press, 1998.

[Mur97] F. Muri. *Comparaison d'algorithmes d'identification de chaînes de Markov cachées et applications à la détection de régions homogènes dans les séquences d'ADN.* PhD thesis, Université René Descartes, Paris V, 1997.

[Rab89] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEE.*, (77) :257–286, 1989.

[Ros97] E. Rostand. *Cyrano de Bergerac.* 1897.

[Sch95] S. Schbath. *Etude asymptotique du nombre d'occurrences d'un mot dans une chaîne de Markov et application à la recherche de mots de fréquence exceptionnelle dans les séquences d'ADN*. PhD thesis, PARIS V, 1995.

[SWW75] G. Salton, A. Wong, and C. S. Wang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11) :613–620, 1975.

[Ver01] J-Ph. Vert. *Méthodes satitistiques pour la modélisation du langage naturel.* PhD thesis, Université Paris 6, 2001.

[WST95] F. M. J. Willems, Y. M. Shtarkov, and Tj. J. Tjalkens. The context-tree weighting method : basic properties. *IEEE Trans. Inform. Theory*, IT-41 :653–664, 1995.

FIG. 3.1 – Représentation graphique des mots de longueur $h = 9$ chez *Methanococcus jannaschii* dans le modèle $M1$. On consultera la section 11.1 page 160 pour plus de détails.

Chapitre 3

Méthodes existantes

Pré-requis :

Notations du chapitre 2.

Description :

Présentation des différentes méthodes existantes pour traiter le problème de la significativité d'un comptage de mots ou de motifs dans une chaîne de *Markov*.

Résumé :

Le calcul de la significativité d'un comptage peut se faire par le biais d'approximations asymptotiques. En calculant les moments d'ordres 1 et 2 des comptages, on est en mesure d'approcher la loi de ces comptages par des distributions gaussiennes. Après avoir introduit la notion de train de mots, il est également possible d'utiliser les méthodes de *Chein-Stein* pour approcher la loi du comptage par des compositions de lois de *Poisson*.

Une autre approche consiste à utiliser les séries génératrices pour calculer les lois exactes des comptages. On peut ici effectuer les calculs par deux moyens, le premier mettant en œuvre des méthodes purement analytiques tandis que le second utilise les automates.

Contenu du chapitre

3.1 Méthodes asymptotiques

3.1.1 Introduction

On considère un mot $W = w_1 \dots w_h$ dans un alphabet fini $\mathcal{A}$ et une séquence aléatoire $X = X_1 \dots X_n$ dans ce même alphabet suivant un modèle Mm. On s'intéresse aux comportements de la variable aléatoire $N(W)$ comptant le nombre d'occurrences de W dans X qui, d'après la section 2.2.1 (page 32) est définie par

$$N(W) = \sum_{i=1}^{n-h+1} Y_i$$

où $Y_i = \mathbb{I}_{X_i=w_1} \times \dots \times \mathbb{I}_{X_{i+h-1}=w_h}$ est l'indicatrice de la présence du mot W à la position i dans X.

Si on note $p(W)$ la probabilité d'observer le mot W en une position i quelconque (cette probabilité ne dépend pas de i grâce à l'hypothèse d'homogénéité sur la chaîne de *Markov*), alors

$$Y_i \sim \mathcal{B}[p(W)]$$

et la variable aléatoire $N(W)$ est une somme de telles v.a. de *Bernoulli* qui ne sont malheureusement pas indépendantes.

Afin d'obtenir une bonne intuition de ce qui se passe, supposons cependant que ces variables de *Bernoulli* sont indépendantes. Il est alors clair que $N(W)$ est binomiale :

$$N(W) \sim \mathcal{B}[n, p(W)]$$

et l'on peut alors classiquement approcher asymptotiquement cette loi :

- par une loi normale si $np(W)$ tend vers l'infini (voir section 3.1.2) ;
- par une loi de *Poisson* si $np(W)$ tend vers une constante (voir 3.1.3).

On pourra consulter [Sch95] ou [Sch97] pour une synthèse concernant ces approches pour le problème de la détection des mots exceptionnels. Signalons également que le programme R'MES (pour *Recherche de Mots ExceptionnelS*) propose une implémentation de ces approches et est disponible sur le web (`http://www.inra.fr/bia/J/AB/genome/`).

3.1.2 Approximations gaussiennes

L'article [PRT95] propose l'idée de travailler conditionnellement à la *statistique exhaustive* du modèle considéré et donne les résultats dans le cas du modèle $M1$. Le travail effectué dans [Sch95] généralise ces résultats aux modèles Mm (et même aux modèles périodiques évoqués en section 2.4.1 40).

Si on choisit un début de séquence $x_1, \ldots, x_m$ et que l'on se donne les comptages des mots de longueur $m+1$ alors on note

$$\mathcal{S}_m = \left\{ x_1, \ldots, x_m \text{ et } \big(n(W')\big)_{W' \in \mathcal{A}^{m+1}} \right\}$$

la statistique exhaustive correspondant au modèle Mm (c'est à dire les données qu'il suffit de connaître pour pouvoir calculer la vraisemblance d'une observation dans le modèle Mm).

On pose alors

$$U_m(W) = \frac{N(W) - \mathbb{E}[N(W)|\mathcal{S}_m]}{\mathbb{V}ar[N(W)|\mathcal{S}_m]^{\frac{1}{2}}}$$

et on a

Théorème 3.1

$$U_m(W) \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Pour pouvoir utiliser pratiquement ce résultat, il ne reste plus qu'à calculer les quantités $\mathbb{E}[N(W)|\mathcal{S}_m]$ et $\mathbb{V}ar[N(W)|\mathcal{S}_m]$, ou pour le moins, leurs équivalents asymptotiques.

Pour l'espérance on obtient le résultat naturel suivant

Proposition 3.2

$$\mathbb{E}[N(W)|\mathcal{S}_m] \xrightarrow[n \to +\infty]{} \frac{\prod_{j=1}^{h-m} N(w_j \ldots w_{j+m})}{\prod_{j=2}^{h-m} N(w_j \ldots w_{j+m-1})} \quad p.s.$$

en revanche, le résultat concernant la variance est beaucoup plus compliqué. Pour pouvoir l'énoncer, il nous faut d'abord introduire la notion de periodicité pour les mots :

Définition 3.3 (période)
On dit qu'un mot $W = w_1 \dots w_h$ est périodique *de période d si ses $h-1-d$ dernières lettres de W sont les mêmes que les $h-1-d$ premières et on note alors $W^{(d)}W$ le mot composé de deux occurrences chevauchantes de W :*

$$W^{(d)}W = w_1 \dots w_d w_1 \dots w_h$$

et on définit également l'indicatrice suivante :

$$\delta_d(W) = \begin{cases} 1 & \text{si } d \text{ est une période de } W \\ 0 & \text{sinon} \end{cases} .$$

On a le résultat suivant :

Proposition 3.4

$$\begin{aligned} \mathbb{V}ar[N(W)|\mathcal{S}_m] \xrightarrow[n\to+\infty]{} & \; p(W) + 2\sum_{d=1}^{h-m-1} \delta_d(W) p(W^{(d)}W) \\ + & \; p(W)^2 \left(\sum_{a_1,\dots,a_m} \frac{\sum_b n(a_1 \dots a_m b)}{p(a_1 \dots a_m)} \right. \\ & \left. - \sum_{a_1,\dots,a_{m+1}} \frac{n(a_1 \dots a_{m+1})}{p(a_1 \dots a_{m+1})} + \frac{1 - 2\sum_b n(w_1 \dots w_m b)}{p(w_1 \dots w_m)} \right) \quad p.s. \end{aligned}$$

où, si $V = v_1 \dots v_l$ est un mot de taille $l \geqslant m+2$ alors

$$p(V) = \frac{1}{n} \times \frac{\prod_{j=1}^{l-m} N(v_j \dots v_{j+m})}{\prod_{j=2}^{l-m} N(v_j \dots v_{j+m-1})}.$$

En utilisant ces formules, on est capable de calculer la valeurs de la statistique $U_m(W)$ pour un mot donné de longueur h dans l'alphabet $\mathcal{A}$ fini de cardinal k en $O(k^{m+1})$ en temps et en espace (la longueur h du mot considéré intervient également mais de manière négligeable par rapport à l'ordre m du modèle).

3.1.3 Approximations poissonniennes

On propose dans [Sch95] d'approcher les comptages de mots par des variables aléatoires de *Poisson* dans le cas où $np(W)$ tend vers une constante lorsque n tend vers l'infini. Ces résultats utilisant largement la méthode de *Chen-Stein* on commence ici par rappeler en quoi elle consiste.

Théorème 3.5 (***Chen-Stein***)
On considère I un ensemble d'indice fini ou dénombrable et soit

$$N = \sum_{i \in I} Y_i$$

une variable aléatoire vérifiant $\mathbb{E}[N] = \lambda = \sum_i p_i \in]0, +\infty[$ et de sorte que $p_i = \mathbb{P}(Y_i = 1) > 0$ pour tout i. On se donne également pour tout $i \in I$ un sous-ensemble $B_i \subset I$, voisinage de i (c'est à dire que $i \in B_i$). Alors la distance en variation totale vérifie

$$d\big(\mathcal{L}(N), \mathcal{P}(\lambda)\big) \leqslant 2(b_1 + b_2 + b_3)$$

avec

$$\begin{aligned} b_1 &= \sum_{i \in I} \sum_{j \in B_i} p_i p_j \\ b_2 &= \sum_{i \in I} \sum_{j \in B_i \setminus \{i\}} \mathbb{E}[Y_i Y_j] \\ b_3 &= \sum_{i \in I} \mathbb{E}\left| \mathbb{E}\big(Y_i - p_i | \sigma(Y_j, j \in B_i^c)\big) \right| \end{aligned}$$

où $\mathcal{P}(\lambda)$ désigne la loi de Poisson *de paramètre λ.*

Lorsqu'un mot ne possède aucune période (voir définition 3.3), on peut montrer que la variable aléatoire $N(W)$ suit asymptotiquement une loi de *Poisson*. Ce résultat ne tient cependant plus si le mot possède des périodes.

Pour un tel mot, il arrive que plusieurs occurrences de ce mot surviennent les unes à la suite des autres en se chevauchant. On appelle *q-train* d'un mot un train de q de ses occurrences.

Voici un exemple d'occurrence d'un 4-train du mot `aacaaa` qui possède 4 et 5 comme périodes (le 4-train est souligné et chaque occurrence du mot est indiquée par une flèche) :

```
       ↓   ↓    ↓   ↓
...attgaacaaacaaaacaaacaaaatgc...
       aacaaa
           aacaaa
                aacaaa
                    aacaaa
```

On note $\mathcal{C}_q(W)$ l'ensemble de tous les q-trains possibles de W et on note $\widetilde{N}(W)$ le nombre de trains de W dans la séquence. On peut alors montrer que cette variable aléatoire suit une loi de *Poisson* et on a bien ainsi généralisé les premiers résultats (si un mot W ne possède pas de périodes, $N(W) = \widetilde{N}(W)$).

On utilise ce résultat préliminaire pour montrer le

Théorème 3.6 *Si on suppose que l'on se place dans le modèle* $M1$ *alors la distance en variation totale*

$$d\big(\mathcal{L}(N(W)), \mathcal{CP}(\Lambda)\big) \xrightarrow[n\to+\infty]{} 0$$

où $\mathcal{CP}(\Lambda)$ *désigne la* loi de Poisson composée *de paramètres* $\Lambda = (\lambda_i)_{i\geqslant 1}$ *c'est à dire la loi de la somme de variables indépendantes*

$$\sum_{q\geqslant 1} qZ_q$$

avec $Z_q \sim \mathcal{P}(\lambda_q)$ *et où les* λ_q *sont définies par*

$$\lambda_q = \sum_{C\in\mathcal{C}_q} p(C) - 2\sum_{C\in\mathcal{C}_{q+1}} p(C) + \sum_{C\in\mathcal{C}_{q+2}} p(C)$$

Du point de vue numérique, on trouvera dans [JKK92] (ainsi que dans [RS01]) une formule explicite permettant un calcul rapide des probabilités de la loi de *Poisson* composée du théorème 3.6.

Pour généraliser ce résultat au cas des modèles Mm, la technique utilisée ici est celle du changement d'alphabet qui permet, moyennant une astuce d'écriture, de se ramener au cas du modèle $M1$. Cette technique sera vue en détails dans le chapitre 6 (page 91) si bien que nous ne l'étudions pas ici plus avant.

Les articles [RS98a] et [RS99] proposent également des résultats concernant l'approximation du nombre d'occurrences d'une famille finie de mots par une loi de *Poisson* composée, mais on ne développera pas davantage ce point.

3.2 Méthodes exactes

3.2.1 Introduction

Les méthodes exactes se proposent pour leur part de calculer les lois des comptages de mots ou de motifs via l'outil des *séries génératrices* (voir définition 3.7 ci-après). La détermination de ces séries peut se faire essentiellement selon deux approches : une première approche analytique (voir section 3.2.2) et la seconde utilisant les automates (voir section 3.2.3).

À partir de ces séries, il est possible d'obtenir facilement les moments des variables aléatoires étudiées et donc d'en tirer des approximations asymptotiques, ou bien on peut chercher à calculer les lois exactes à distance finie mais cela peut impliquer des calculs complexes.

L'article [RS01] présente une intéressante comparaison des résultats obtenus avec les méthodes présentées dans cette section et celles qui le sont dans la section précédente ; on pourra s'y référer pour choisir l'approche qui convient à un problème donné, aussi bien en terme de précision que de complexité algorithmique.

Définition 3.7 (série génératrices)
On appelle série génératrice ordinaire sur le corps K, l'objet formel

$$f(z) = \sum_{i \geqslant 0} a_i z^i$$

où $(a_i)_i$ est une suite de K qui sera, en général, le corps des réels ou bien encore celui des fractions rationnelles. Il s'agit en fait d'une simple généralisation des polynômes qui présente cependant l'intérêt de permettre l'utilisation des méthodes et écritures applicables aux séries entières (sans se soucier des problèmes de convergence).

Ainsi on note :

$$\exp(z) = \sum_{i \geqslant 0} \frac{1}{i!} z^i \quad \textit{ou bien} \quad \frac{1}{1-z} = \sum_{i \geqslant 0} z^i$$

et on définit des opérations naturelles sur les séries génératrices de sorte que les versions séries entières des équations correspondantes soient justes (à l'intérieur des disques de convergence). Ainsi si $f(z)$, $g(z)$ et $h(z)$ sont trois séries génératrices respectivement associées aux suites $(a_i)_i$, $(b_i)_i$ et $(c_i)_i$ alors si $h(z) = f(z) + g(z)$ on a $c_i = a_i + b_i$ pour tout $i \geqslant 0$ et si $h(z) = f(z) \times g(z)$ alors

$$c_i = \sum_{j=0}^{i} a_j b_{i-j} \quad \forall i \geqslant 0.$$

Notons qu'il est également possible de définir des séries génératrices multivariées avec un formalisme similaire :

$$f(z_1, \ldots, z_r) = \sum_{i_1, \ldots, i_r \geqslant 0} a_{i_1, \ldots, i_r} z_1^{i_1} \ldots z_r^{i_r}$$

Avant d'aller plus loin, il nous faut également introduire les notions de *langage* et d'opérations sur les langages :

Définition 3.8 (langage)
On appelle langage d'un alphabet fini $\mathcal{A}$ un ensemble $\mathcal{L}$ quelconque de mots écrits avec les lettres de $\mathcal{A}$. Un langage peut être fini ou infini et peut contenir des mots inclus les uns dans les autres ou se chevauchant ; on dira cependant

d'un langage qu'il est simple *s'il ne vérifie aucune des deux dernières propriétés (inclusions et chevauchements).*

Il existe trois opérations classiques sur les langages :

- Union (+) : *si* $\mathcal{L}$ *et* $\mathcal{L}'$ *sont deux langages alors*

$$\mathcal{L} + \mathcal{L}' = \mathcal{L} \cup \mathcal{L}';$$

- Concaténation ($\cdot$) : *si* $\mathcal{L}$ *et* $\mathcal{L}'$ *sont deux langages alors*

$$\mathcal{L} \cdot \mathcal{L}' = \{WW', W \in \mathcal{L}, W' \in \mathcal{L}'\}\,;$$

- Etoile (*) : *si* $\mathcal{L}$ *est un langage alors on a la définition suivante :*

$$\mathcal{L}^* = \varepsilon + \mathcal{L} + \mathcal{L}^2 + \mathcal{L}^3 + \ldots$$

où, de façon récursive

$$\mathcal{L}^* = (\varepsilon + \mathcal{L}) \cdot \mathcal{L}^*$$

avec ε *désignant le langage vide.*

Les langages définis à partir de ces trois opérations sont les langages rationnels.

3.2.2 Approches analytiques

Dans l'article [Rég00], on considère dans l'alphabet $\mathcal{A}$ fini de cardinal k, r mots $W_1, \ldots, W_r$ de longueurs respectives $h_1, \ldots, h_r$ et on souhaite déterminer la série génératrice

$$T(z, u_1, \ldots, u_r) = \sum_{n, n_1, \ldots, n_r \geqslant 0} p_n(n_1, \ldots, n_r) z^n u_1^{n_1} \ldots u_r^{n_r}$$

où

$$p_n(n_1, \ldots, n_r) = \mathbb{P}_n\big(N(W_1) = n_1, \ldots, N(W_r) = n_r\big)$$

avec $\mathbb{P}_n$ désignant la loi d'une séquence de longueur n générée selon le modèle $M1$. Les formules obtenues sont assez complexes mais on peut mettre en évidence deux matrices formelles intervenant dans les résultats : $\mathbb{Z}(z)$ et $\mathbb{A}(z)$.

La première de ces deux matrices est de dimensions (k, k) et est caractéristique du modèle considéré. En effet, si Π désigne la matrice de transition du modèle $M1$ considéré et que l'on note μ la distribution stationnaire de cette chaîne de *Markov* (vecteur-ligne) alors

$$\mathbb{Z}(z) = \big(Id - (\Pi - M)z\big)^{-1}$$

où M désigne la matrice formée par k vecteurs μ et Id la matrice identité. Dans le cas particulier où le modèle $M1$ est en fait un modèle $M0$, remarquons que l'on a $\Pi = M$ et donc $\mathbb{Z}(z) = Id$. Dès que l'on souhaite travailler dans un modèle Mm, on se ramène au cas du modèle $M1$ par la technique du changement d'alphabet (voir chapitre 6 page 91) dont le cardinal devient k^m. En conséquence, l'inversion formelle de la matrice d'ordre k^m intervenant dans la définition de $\mathbb{Z}(z)$ devient très rapidement d'une trop grande complexité numérique.

La seconde matrice, $\mathbb{A}(z)$ de dimensions (r, r), contient une information concernant la famille $W_1, \ldots, W_r$ des mots considérés ; c'est la matrice de leurs corrélations. Si on note $\mathcal{A}(W, W')$ l'ensemble (éventuellement vide) des motifs obtenus en faisant se chevaucher une occurrence du premier avec une occurrence du second alors

$$\left[\mathbb{A}(z)\right](i, j) = \mathbb{I}_{i=j} + \sum_{W \in \mathcal{A}(W_i, W_j)} \mathbb{P}(W) z^{|W|}$$

en particulier, pour un mot W on considère l'ensemble des mots obtenus en chevauchant deux occurrences $\mathcal{A}(W, W) = \{W^{(d)}W, \delta_d(W) = 1\}$ et on appelle le polynôme correspondant ($i^{\text{ème}}$ terme de la matrice $\mathbb{A}(z)$ si W est W_i) *polynôme d'auto-corrélation* du mot W.

Exemple 3.9 *Si* $\mathcal{A} = \{\mathtt{a}, \mathtt{b}\}$ *et* $r = 2$, $W_1 = \mathtt{babb}$ *et* $W_2 = \mathtt{bba}$ *alors les chevauchements suivants sont possibles :*

$\mathbb{A}$	babb	bba
babb	babbabb	babbba babba
bba	bbabb	$\emptyset$

ce qui donne la matrice de corrélation suivante :

$$\mathbb{A}(z) = \begin{pmatrix} 1 + \mathbb{P}(\mathtt{babbabb})z^7 & \mathbb{P}(\mathtt{babba})z^5 + \mathbb{P}(\mathtt{babbba})z^6 \\ \mathbb{P}(\mathtt{bbabb})z^5 & 1 \end{pmatrix}.$$

Une fois calculée, la matrice $\mathbb{A}(z)$ intervient sous la forme de son inverse formelle dans certains résultats si bien que la taille de l'ensemble de mots considérés ne doit pas être trop grande sous peine de voir la complexité numérique augmenter de façon très importante.

Dans tous les cas, on peut utiliser les dérivées de la fonction génératrice T pour calculer aisément les différents moments des variables aléatoires

$\big(N(W_i)\big)_i$ et en tirer des approximations asymptotiques. D'un autre coté, il est également possible de calculer les coefficients de *Taylor* de la fonction T à un ordre suffisant pour pouvoir identifier un terme $p_n(n_1,\ldots,n_r)$ donné. Cette dernière approche peut cependant s'avérer très gourmande en puissance de calcul et n'est en général pas celle qui est choisie, elle reste cependant la seule valable dans le cas de l'étude de courtes séquences sortant du cadre des approximations asymptotiques.

Il est à signaler que l'on peut également obtenir un résultat de grandes déviations par ce type d'approche. L'article [RS98b] considère en effet le cas d'un mot W de longueur h et montre que son comptage $N(W)$ sur une séquence de longueur n tirée selon le modèle $M1$ vérifie :

$$\lim_{n\to+\infty}\frac{1}{n}\log\mathbb{P}(N(W)\geqslant an)=-I(a)$$

(voir le chapitre 4 pour plus de précision sur cette forme de résultats) avec

$$I(a)=a\omega_a+\rho(\omega_a)$$

où ρ est une fonction implicite définie comme la racine d'une fraction rationnelle faisant intervenir des quantités formelles similaires aux termes des matrices $\mathbb{Z}(z)$ et $\mathbb{A}(z)$.

On peut enfin noter que les articles [RD00] et [RD99] proposent d'utiliser des techniques similaires pour établir les lois des distances entre occurrences de mots, ce qui est particulièrement intéressant pour mettre en évidence des zones de comportement exceptionnel pour un mot donné. Les algorithmes qui y sont établis font actuellement l'objet d'une implémentation qui devrait bientôt déboucher sur la mise à disposition d'un programme utilisable : DEMOS.

3.2.3 Approches par automates

Les auteurs de [NSF99] proposent de leur côté une approche différente du problème par le biais des automates. Un *automate* sert à reconnaître les mots d'un langage ; les automates finis correspondant aux langages rationnels. Celui de la figure 3.2 s'interesse par exemple au langage $\mathcal{L}_3=\mathcal{A}^*\cdot\mathtt{aba}$ (avec $\mathcal{A}=\{\mathtt{a},\mathtt{b}\}$ désignant l'alphabet dans lequel on travaille, ce langage reconnait les textes se terminant par le mot `aba`) et une transformation de cet automate (figure 3.3) permet de compter les occurrences du mot `aba`.

On se donne un langage $\mathcal{L}$ quelconque et on suppose que l'on dispose d'un automate déterministe (il part de chaque état exactement une transition par symbole possible) capable de compter les occurrences des mots du langage dans un texte donné.

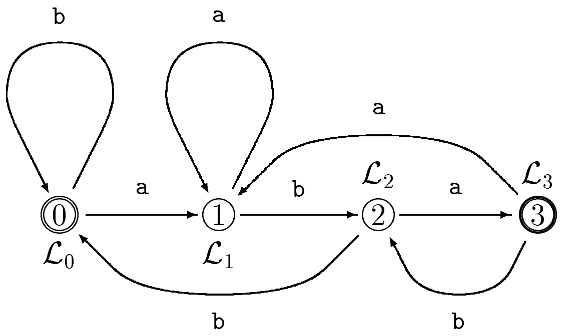


FIG. 3.2 – Exemple d'automate : détection des mots du langage $\mathcal{L}_3 = \mathcal{A}^* \cdot \mathtt{aba}$ avec $\mathcal{A} = \{\mathtt{a}, \mathtt{b}\}$. L'état initial est l'état 0, l'état final est l'état 3. A titre d'exemple, cet automate parcours la séquence babbabaababb en passant successivement dans les états 0012012312320 ; on identifie ainsi deux occurrences de aba dans la séquence (en comptant les passage par l'état 3).

On pose la définition suivante :

Définition 3.10 (série génératrice d'un langage) *Si $\mathcal{L}$ est un langage dans un alphabet $\mathcal{A} = \{a_1, \ldots, a_k\}$, alors sa série génératrice multivariée associée est*

$$F_{\mathcal{L}}(a_1, \ldots, a_k) = \sum_{W \in \mathcal{L}} a_1^{n_{a_1}(W)} \ldots a_k^{n_{a_k}(W)}$$

où $n_a(W)$ désigne pour $a \in \mathcal{A}$ le nombre de fois où le caractère a est présent dans le mot W.

Exemple 3.11 *La série génératrice du langage $\mathcal{L} = \{\varepsilon, \mathtt{aa}, \mathtt{ab}, \mathtt{ba}, \mathtt{aaab}\}$ (où ε désigne mot nul) est :*

$$F_{\mathcal{L}}(\mathsf{a}, \mathsf{b}) = 1 + \mathsf{a}^2 + 2\mathsf{ab} + \mathsf{a}^3\mathsf{b}.$$

On a la

Proposition 3.12 *Sous certaines conditions, les opérations sur les langages correspondent à des opérations sur les séries génératrices associées :*

- Union : *Si $\mathcal{L}$ et $\mathcal{L}'$ sont deux langages disjoints alors*

$$F_{\mathcal{L}+\mathcal{L}'}(a_1, \ldots, a_k) = F_{\mathcal{L}}(a_1, \ldots, a_k) + F_{\mathcal{L}'}(a_1, \ldots, a_k);$$

- Produit : *Si $\mathcal{L}$ et $\mathcal{L}'$ sont deux langages dont le produit est non ambigu (c'est à dire que chaque élément de $\mathcal{L} \cdot \mathcal{L}'$ correspond la concaténation d'un unique couple d'éléments de $\mathcal{L}$ et $\mathcal{L}'$) alors*

$$F_{\mathcal{L} \cdot \mathcal{L}'}(a_1, \ldots, a_k) = F_{\mathcal{L}}(a_1, \ldots, a_k) \times F_{\mathcal{L}'}(a_1, \ldots, a_k);$$

– Opérateur * : *Si* $\mathcal{L}$ *est un langage dont le produit avec lui-même est non ambigu alors*

$$F_{\mathcal{L}^*}(a_1, \ldots, a_k) = \frac{1}{1 - F_{\mathcal{L}}(a_1, \ldots, a_k)}.$$

La méthode de *Chomski-Schützenberger* permet de traduire un *automate déterministe* donné (à un symbole donné correspond au plus à une transition par état de l'automate) en opérations sur les langages des différents états. Comme, ces opérations vérifient nécessairement les conditions de la proposition 3.12 dans le cas d'un automate déterministe, il est alors possible de traduire directement ces opérations en opérations sur les séries génératrices associées aux langages.

Exemple 3.13 *Dans le cas de l'automate de la figure 3.2 on obtient les équations suivantes par la méthode de* Chomski-Schützenberger *:*

$$\begin{cases} \mathcal{L}_0 = \mathtt{a}\mathcal{L}_1 + \mathtt{b}\mathcal{L}_0 \\ \mathcal{L}_1 = \mathtt{b}\mathcal{L}_2 + \mathtt{a}\mathcal{L}_1 \\ \mathcal{L}_2 = \mathtt{a}\mathcal{L}_3 + \mathtt{b}\mathcal{L}_0 \\ \mathcal{L}_3 = \mathtt{a}\mathcal{L}_1 + \mathtt{b}\mathcal{L}_2 + \varepsilon \end{cases}$$

relations qui se traduisent par le système

$$\begin{cases} F_{\mathcal{L}_0} = \mathtt{a}F_{\mathcal{L}_1} + \mathtt{b}F_{\mathcal{L}_0} \\ F_{\mathcal{L}_1} = \mathtt{b}F_{\mathcal{L}_2} + \mathtt{a}F_{\mathcal{L}_1} \\ F_{\mathcal{L}_2} = \mathtt{a}F_{\mathcal{L}_3} + \mathtt{b}F_{\mathcal{L}_0} \\ F_{\mathcal{L}_3} = \mathtt{a}F_{\mathcal{L}_1} + \mathtt{b}F_{\mathcal{L}_2} + 1 \end{cases}$$

que l'on peut résoudre pour trouver (par exemple)

$$F_{\mathcal{L}_0}(\mathtt{a}, \mathtt{b}) = \frac{\mathtt{a}^2\mathtt{b}}{1 - \mathtt{a} - \mathtt{b}}.$$

On considère $\mathcal{M}$ un langage correspondant à un motif donné. Pour étudier ce motif, on va étudier l'automate déterministe reconnaissant le langage $\mathcal{A}^* \cdot \mathcal{M}$ en remarquant que chaque détection d'un élément de ce langage (c'est à dire d'une occurrence du motif) correspond au passage dans l'état final. On place donc, dans la séquence à étudier, une marque `m` après chaque occurrence du motif et on considère l'automate effectuant cette tâche (qui s'obtient très simplement à partir de l'automate initial).

Dans la figure 3.3, on peut examiner un exemple des modifications à apporter par rapport à l'automate de la figure 3.2 pour compter les occurrences du mot `aba`.

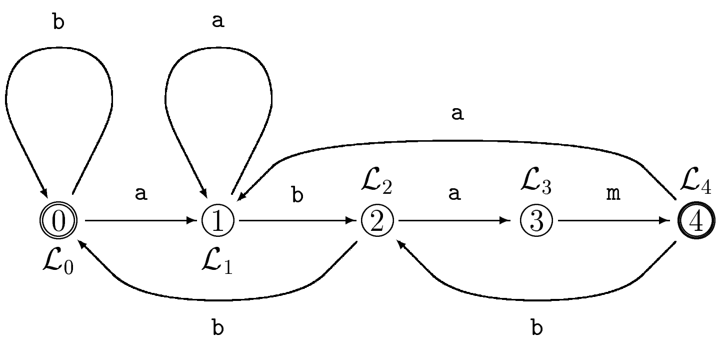


FIG. 3.3 – Exemple d'automate : détection des éléments du langage $\mathcal{L}_4 = \mathcal{A}^* \cdot \mathtt{aba} \cdot \mathtt{m}$ avec $\mathcal{A} = \{\mathtt{a}, \mathtt{b}\}$. L'état initial est l'état 0, l'état final est l'état 4. A titre d'exemple, cet automate parcours la séquence `babbabamabambb` en passant successivement dans les états 0012012341223420 ; on identifie ainsi deux occurrences de `abam` dans la séquence (une pour chaque passage dans l'état 3).

Si on se place dans le modèle $M0$ et qu'on note μ la distribution des lettres de $\mathcal{A}$, il est alors possible d'injecter ces paramètres dans la série génératrice obtenue par la méthode de *Chomski-Schützenberger* pour calculer

$$F(z,u) = \sum_{n,m \geqslant 0} p_{n,m} u^m z^n$$

où $p_{n,m}$ désigne la probabilité d'obtenir m occurrences du motif dans un texte de longueur n. Il suffit pour cela d'effectuer une substitution formelle dans la série génératrice du langage $\mathcal{A}^* \cdot \mathcal{M} \cdot \mathtt{m}$ en remplaçant chaque lettre $a \in \mathcal{A}$ par $\mu(a)z$ et $\mathtt{m}$ par u. La série génératrice ainsi obtenue peut dès lors être utilisée pour calculer des moments ou bien des probabilités.

Exemple 3.14 *Dans le cas de la figure 3.3, la méthode de* Chomski-Schützenberger *permet de calculer*

$$F(\mathtt{a},\mathtt{b},\mathtt{m}) = \frac{1 + \mathtt{ab}(1-\mathtt{m})}{1 - \mathtt{a} - \mathtt{b} - \mathtt{ab}(1-\mathtt{m}) - \mathtt{ab}^2(1-\mathtt{m})}$$

et donc

$$F(z,u) = \frac{1 + \mu(a)\mu(b)z^2(1-u)}{1 - \mu(a)s - \mu(b)z + \mu(a)\mu(b)z^2(1-u) - \mu(a)\mu(b)^2z^3(1-u)}.$$

Une telle approche peut bien évidemment se généraliser aux cas des modèles de *Markov* en utilisant des changements d'alphabets (voir le chapître 6 page 91 pour les détails concernant cette technique).

Notons que cette méthode, combinée avec l'approche exacte par les automates, a donné naissance à REGEXPCOUNT, une boîte à outils performante pour le logiciel de calcul symbolique MAPLE.

Si cette approche est très intéressante lorsque l'on dispose d'un automate déterministe identifiant le motif étudié, il faut néanmoins tenir compte des difficultés algorithmiques que l'on peut rencontrer pour obtenir cet automate. Dans le cas de certains motifs complexes, il est très long, voire impossible, de construire cet automate. L'article [Nic01] propose donc pour remédier à ce problème une méthode plus directe pour obtenir la série génératrice d'un motif mais qui n'est qu'une approximation du véritable résultat dès que le motif en question ne vérifie pas l'une des deux hypothèses suivantes :

- les éléments du motif ne sont pas inclus les uns dans les autres ;
- les éléments du motif ne se chevauchent pas.

Références

[JKK92] N. L. Johnson, S. Kotz, and A. W. Kemp. *Univariate discrete distributions*. Wiley : New-York, 1992.

[KB92] J. Kleffe and M. Borodovsky. First and second moment of counts of words in random text generated by markov chains. *Comp. Applic. Biosci.*, (8) :433–441, 1992.

[LMS] M.-Y. Leung, G. M. Marsh, and T. P. Speed. Over- and underrepresentation of short dna words in herpesvirus genomes. *J. Comp. Bio.*, (3) :345–360.

[Nic01] P. Nicodème. Fast approximate motif statistics. *J. Comp. Biol.*, 2001. to appear.

[NSF99] P. Nicodème, B. Salvy, and P. Flajolet. Motif statistics. *Theoretical Computer Science*, 1999. to appear.

[PRT95] B. Prum, F. Rodolphe, and E. de Turckheim. Finding words with unexpected frequencies in dna sequences. *J. R. Statist. Soc. B.*, 11 :190–192, 1995.

[RD99] S. Robin and J.J. Daudin. Exact distribution of word occurrences in a random sequence of letters. *J. App. Prob.*, 36 :179–193, 1999.

[RD00] S. Robin and J.J. Daudin. Exact distribution of the distances between any occurrences of a set of words. *Ann. Inst. Statist. Math.*, 2000.

[Rég00] M. Régnier. A unified approach to word occurrence probabilities. *Discrete applied mathematics*, 104(1) :259–280, 2000.

[RS98a] G Reinert and S. Schbath. Compound poisson and poisson process approximations for occurrences of multiple words in markov chains. *J. Comp. Biol.*, 5 :223–254, 1998.

[RS98b] M. Régnier and W. Szpankowski. On pattern frequency occurrences in a markovian sequence. *Algorithmica*, 22(4) :631–649, 1998.

[RS99] G. Reinert and S. Schbath. Compound poisson approximations for occurrences of multiple words. *Statistics in Molecular Biology and Genetics*, 33 :257–275, 1999.

[RS01] S. Robin and S. Schbath. Numerical comparison of several approximations of the word count distribution in random sequences. *J. Comp. Biol.*, 2001. To appear.

[Sch95] S. Schbath. *Etude asymptotique du nombre d'occurrences d'un mot dans une chaîne de Markov et application à la recherche de mots de fréquence exceptionnelle dans les séquences d'ADN*. PhD thesis, PARIS V, 1995.

[Sch97] S. Schbath. An efficient statistic to detect over- and under- represented words in dna sequences. *J. Comp. Biol.*, 4 :189–192, 1997.

[Sch00] S. Schbath. An overview on the distribution of word counts in markov chains. *J. Comp. Biol.*, (7) :193–202, 2000.

FIG. 4.1 – Représentation graphique des mots de longueur $h = 9$ chez *Saccharomyces cerevisiae* dans le modèle $M1$. On consultera la section 11.1 page 160 pour plus de détails.

Chapitre 4

Grandes déviations

Pré-requis :

Chapitre 2.

Description :

Présentation de l'outil mathématique des grandes déviations et des différentes techniques et notions qui lui sont liées.

Pour une définition des principaux termes (*principe de grandes déviation*, *fonction de taux*, ...) et résultats (théorème de *Gärtner-Ellis*, lemme de *Varadhan*, ...) concernant la théorie générale des grandes déviations, on se reportera davantage à l'annexe A (page 179) qu'à cette partie.

Résumé :

On s'attache tout d'abord à motiver l'approche *grandes déviations* par rapport aux approches statistiques classiques (*loi des grands nombres* ou encore *théorème de la limite centrale*) à travers un exemple simple (jeu de pile ou face) qui sera filé tout au long de l'argument.

On suppose ensuite que les variables aléatoires considérées sont à valeurs dans un espace fini et on se place dans le cas où ces mêmes variables sont indépendantes et identiquement distribuées.

Dans ce cadre très restreint mais suffisant pour notre propos, on présente en premier lieu le théorème de *Cramér-Chernov* concernant les *déviations des moyennes* et dans la démonstration duquel la technique clé du *changement de probabilités* est mise en exergue.

Enfin les résultats concernant les *déviations des distributions empiriques* des singletons, puis des paires, sont établis.

Contenu du chapitre

4.1 Introduction

Cette partie s'inspire largement de [dH00].

On considère une variable aléatoire réelle (v.a.r.) X dont on note $(X_i)_{1\leqslant i\leqslant n}$ un échantillon de taille n (c'est à dire n réalisations indépendantes de X).

On pose $S_n = X_1 + \ldots + X_n$ et on va s'intéresser au comportement de cette variable aléatoire lorsque n devient grand.

Un premier résultat fondamental confirme l'intuition selon laquelle la moyenne de ces n réalisations de X doit être "proche" de la moyenne attendue de X ($\mathbb{E}[X]$) ; il s'agit du

Théorème 4.1 (Loi forte des grands nombres)
Si $\mathbb{E}[X] = m \in \mathbb{R}$ *alors*

$$\frac{1}{n}S_n \xrightarrow[n\to+\infty]{} m \quad \textit{presque sûrement (}\text{p.s.}\textit{)}.$$

Il est cependant naturel que des écarts ou *déviations* de la somme (S_n) par rapport à l'attendu (mn) apparaissent. Bien que ce résultat soit moins intuitif que le précédent, il se trouve que la taille de ces déviations sera de l'ordre de $\sqrt{n}$ comme l'annonce le

Théorème 4.2 (Limite centrale)
Si $\mathbb{E}[X] = m \in \mathbb{R}$ *et* $\mathbb{V}ar[X] = \sigma^2 \in]0, +\infty[$ *alors*

$$\frac{1}{\sigma\sqrt{n}}(S_n - mn) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1) \quad \textit{(}\text{en loi}\textit{)}$$

$\mathcal{N}(0,1)$ *désignant un loi normale centrée réduite.*

En effet, si on considère un réel ν et qu'on examine les déviations par rapport à l'attendu supérieures à $\nu\sigma\sqrt{n}$ on constate aisément que la probabilité de ces déviations converge avec n vers une valeur finie :

$$\mathbb{P}(S_n \geqslant mn + \nu\sigma\sqrt{n}) \xrightarrow[n\to+\infty]{} \mathbb{P}\left(\mathcal{N}(0,1) \geqslant \nu\right).$$

Par opposition à ces *déviations "normales"* on peut définir les *grandes déviations* comme étant les écarts non plus de l'ordre de $\sqrt{n}$ mais de l'ordre de n. On s'intéressera ainsi typiquement à des événements de la forme $\{S_n \geqslant mn + \nu n\}$ ou plus simplement de la forme $\{S_n \geqslant an\}$ avec $a > m$ par exemple.

En utilisant le théorème 4.2 on obtient *intuitivement* l'équivalence suivante en l'infini :

$$\mathbb{P}(S_n \geqslant an) \sim \mathbb{P}\left(\mathcal{N}(0,1) \geqslant \frac{(a-m)}{\sigma}\sqrt{n}\right) \sim e^{-I(a)n}$$

où $I(a)$ est un constante positive et ce résultat peut alors se reformuler sous cette forme classique :

$$\lim_{n\to+\infty} \frac{1}{n} \log\left[\mathbb{P}(S_n \geqslant an)\right] = -I(a).$$

Tous les résultats de grandes déviations que nous énoncerons par la suite s'écrivant sous une forme identique ou proche de celle présentée ci-dessus, il convient d'insister sur le sens de cette écriture : *l'événement de grandes déviations voit sa probabilité décroître exponentiellement vite avec* n *et à une vitesse déterminée par le paramètre* $I(a)$.

On en profite pour introduire la notion suivante

Définition 4.3 (*équivalence logarithmique*)
On considère deux suites réelles (α_n) *et* (β_n) *strictement positives et on dit que ces suites sont logarithmiquement équivalentes en* $+\infty$ *si*

$$\alpha_n \overset{\log}{\sim} \beta_n \iff \lim_{n\to+\infty} \frac{1}{n}(\log \alpha_n - \log \beta_n) = 0.$$

Il est facile de vérifier que la relation $\overset{\log}{\sim}$ est bien une relation d'équivalence et on a le résultat

Proposition 4.4 *Soient* (α_n) *et* (β_n) *deux suites strictement positives alors :*

(i) $\alpha_n + \beta_n \overset{\log}{\sim} \max(\alpha_n, \beta_n)$

(ii) *si* $\alpha_n + \beta_n > 0$ *pour tout* n *alors* $\alpha_n - \beta_n \overset{\log}{\sim} \max(\alpha_n, \beta_n)$

Preuve. On a

$$\begin{aligned}\frac{1}{n}\big(\log(\alpha_n + \beta_n) - \log\max(\alpha_n, \beta_n)\big) &= \frac{1}{n}\log\left(1 + \frac{\min(\alpha_n, \beta_n)}{\max(\alpha_n, \beta_n)}\right)\\ &= \frac{1}{n}\log(1+u)\end{aligned}$$

avec $u = \frac{\min(\alpha_n,\beta_n)}{\max(\alpha_n,\beta_n)} \in]0,1[$ dès que $(\alpha_n) \neq (\beta_n)$ (ce que l'on peut supposer car la preuve est triviale dans le cas où $(\alpha_n) = (\beta_n)$). Il ne reste plus alors qu'à utiliser un développement de *Taylor* de

$$\log(1+u) = u + o(u)$$

on conclut la preuve de **(i)** en passant à la limite en $n \to +\infty$.

Pour démontrer **(ii)** on procède de la même façon à ceci près que l'on considère le développement de *Taylor*

$$\log(1-u) = -u + o(u)$$

avec $u = \frac{\beta_n}{\alpha_n}$. ■

4.2 Exemple simple : le jeu de pile ou face

On considère que notre variable aléatoire X suit une loi $\mathcal{B}(p)$, c'est à dire une loi de *Bernoulli* de paramètre p :

$$\begin{cases} \mathbb{P}(X=1) = p & \text{(pile)} \\ \mathbb{P}(X=0) = 1-p & \text{(face)} \end{cases}.$$

On a alors $\mathbb{E}[X] = p$ et $\mathbb{V}ar[X] = p(1-p)$ et on suppose que $p \in]0,1[$ si bien que les conditions des théorèmes 4.1 et 4.2 sont vérifiées.

Soit a un réel supérieur à p et on considère l'événement de

grandes déviations $\{S_n \geqslant an\}$. Pour simplifier l'écriture des calculs, on va supposer que $an \in \mathbb{N}$ (sinon il suffit de travailler avec la partie entière de an), on a alors :

$$\begin{aligned}\mathbb{P}(S_n \geqslant an) &= \sum_{k=na}^{+\infty} \mathbb{P}(S_n = k)\\ &= \sum_{k=na}^{+\infty} C_n^k p^k (1-p)^{n-k}\end{aligned}$$

n	10^2	10^3	10^4	10^5	10^6	10^7	10^8
N	1	3	10	32	100	316	1001
$\lvert N'-N\rvert/N$	8.10^{-1}	1.10^{-1}	2.10^{-2}	2.10^{-3}	2.10^{-4}	2.10^{-5}	2.10^{-6}
$\lvert N_{\mathrm{TCL}}-N\rvert/N$	1.10^{-1}	1.10^{-2}	2.10^{-4}	8.10^{-4}	8.10^{-4}	8.10^{-4}	8.10^{-4}

TAB. 4.1 – Comparaison des probabilités exactes et des approximations par le théorème central limite pour le jeu de pile ou face de paramètre $p = 0.5$ et pour $a = 0.55$. $\mathbb{P}(S_n \geqslant an) = \mathbb{P}(\mathcal{N} \geqslant N)$, $\mathbb{P}(S_n = na) = \mathbb{P}(\mathcal{N} \geqslant N')$ et N_{TCL} est défini par (4.2).

dont on peut calculer numériquement la valeur en utilisant dans les nombres de combinaison les valeurs exactes de $n!$ ou l'approximation par la formule de *Stirling* lorsque n est trop grand ($n \geqslant 10$) : $n! \sim n^n e^{-n}\sqrt{2\pi n}$.

De son côté, le théorème 4.2 nous fournit une approximation asymptotique de la même probabilité. En effet :

$$\mathbb{P}(S_n \geqslant an) = \mathbb{P}\left(\frac{S_n - pn}{\sqrt{p(1-p)n}} \geqslant \frac{(a-p)n}{\sqrt{p(1-p)n}}\right)$$

et donc

$$\mathbb{P}(S_n \geqslant an) \sim \mathbb{P}(\mathcal{N} \geqslant N_{TCL}) \tag{4.1}$$

avec $\mathcal{N} \sim \mathcal{N}(0,1)$ et

$$N_{TCL} = \sqrt{n}\frac{(a-p)}{\sqrt{p(1-p)}}. \tag{4.2}$$

Les probabilités considérées étant très petites dès que n est grand, on adopte à partir de maintenant une *représentation gaussienne* des valeurs numériques de ces probablités en manipulant les réels N et N' vérifiant respectivement $\mathbb{P}(S_n \geq an) = \mathbb{P}(\mathcal{N} \geqslant N)$ et $\mathbb{P}(S_n = an) = \mathbb{P}(\mathcal{N} \geqslant N')$ plutôt que les probabilités elles-mêmes.

On se place dans le cadre d'un jeu de pile ou face équilibré ($p = 0.5$) et que l'on considère l'événement {obtenir plus de 55% de "pile" en n coups} ($= \{S_n \geqslant an\}$ avec $a = 0.55$). Le tableau 4.2 compare les résultats exactes et ceux issus de l'approximation proposée en (4.1) pour différentes valeurs de n.

Il est tout d'abord essentiel de saisir la nature des probabilités manipulées. Pour $n = 100$ on obtient $N = 1$ ce qui correspond à la probabilité $\mathbb{P}(\mathcal{N} \geqslant 1) = 1.6 \times 10^{-1}$, mais ces chiffres deviennent rapidement beaucoup plus difficiles à appréhender : pour $n = 10^6$ on a $\mathbb{P}(\mathcal{N} \geqslant 100) = 1.3 \times 10^{-2174}$ et

n	10^2	10^3	10^4	10^5	10^6	10^7	10^8
N	10	31	99	315	995	3145	9949
$\lvert N' - N\rvert/N$	5.10^{-4}	5.10^{-5}	5.10^{-6}	5.10^{-7}	5.10^{-8}	5.10^{-9}	5.10^{-10}
$\lvert N_{\mathrm{TCL}} - N\rvert/N$	8.10^{-2}	9.10^{-2}	9.10^{-2}	1.10^{-1}	1.10^{-1}	1.10^{-1}	1.10^{-1}

TAB. 4.2 – Comparaison des probabilités exactes et des approximations par le théorème central limite pour le jeu de pile ou face de paramètre $p = 0.5$ et pour $a = 0.95$. $\mathbb{P}(S_n \geqslant na) = \mathbb{P}(\mathcal{N} \geqslant N)$, $\mathbb{P}(S_n = na) = \mathbb{P}(\mathcal{N} \geqslant N')$ et N_{TCL} est défini par (4.2).

pour $n = 10^8$ on a $\mathbb{P}(\mathcal{N} \geq 1001) = 7.0 \times 10^{-217\,586}$. On comprend ainsi mieux l'utilité de

la repésentation gaussienne précédemment introduite.

Les résultats numériques montrent que l'erreur relative faite en utilisant l'approximation gaussienne commence par décroître lorsque n augmente, ce qui est parfaitement naturel pour une approximation asymptotique. Par la suite cependant, la qualité de l'approximation atteint un minimum (pour $n = 10^4$) avant de commencer à stagner ce qui est pour le moins singulier.

Encore plus étrange, la qualité de l'approximation gaussienne ne cesse de se dégrader dans la table 4.2 et où on s'intéresse cette fois-ci à l'événement $\{$obtenir plus de 95% de "pile" en n coups$\}$.

Ces comportements sont en fait totalement prévisibles dans la mesure où plus un événement est rare, plus l'approximation gaussienne nécessite de grandes valeurs de n pour être valide. Or ici, la probabilité des événements considérés décroît rapidement lorsque n augmente et c'est ce phénomène, typique des grandes déviations, qui est la cause de l'imprécision des approximations gaussiennes.

Il est dès lors clair que les outils probabilistes classiques permettant l'étude des parties "centrales" des distributions et les

approximations gaussiennes en particulier ne peuvent être adaptés à l'étude précise des événements de grandes déviations pour lesquels des outils et méthodes spécifiques doivent être mises au point.

Remarquons enfin dans les deux tableaux la grande proximité des valeurs N et N' ce qui signifie que les probabilités $\mathbb{P}(S_n \geqslant an)$ et $\mathbb{P}(S_n = an)$ sont très proches dès lors que n est "assez grand" relativement à la déviation considérée. Cette remarque assez vague pourra être précisée par la suite (voir proposition 4.7) mais nous rassure sur les faibles conséquences du choix arbitraire et systématique de l'événement $\{S_n \geqslant an\}$ plutôt qu'un autre dans les résultats à venir.

4.3 Théorème de *Cramér-Chernov*

Le résultat qui suit est dû à *Cramér* ([Cra38]) qui pressent bien son caractère fondamental en le décrivant comme un *nouveau théorème-limite* ou encore comme une *nouvelle loi des grands nombres.*

4.3.1 Enoncé du théorème

Théorème 4.5 (*Cramér-Chernov*)
Soit $(X_i)_{1\leqslant i\leqslant n}$ *un échantillon de* X*, variable aléatoire à valeurs dans* $\mathcal{A} \subset \mathbb{R}$ *qui vérifie :*

$$\varphi(t) = \mathbb{E}[e^{tX}] < \infty, \forall t \in \mathbb{R}. \tag{4.3}$$

On pose $S_n = \sum_{i=1}^n X_i$ *et on considère* $a > \mathbb{E}[X]$ *alors :*

$$\lim_{n\to+\infty} \frac{1}{n} \log \mathbb{P}(S_n \geqslant an) = -I(a)$$

avec

$$I(z) = \sup_{t\in R}[zt - \log\varphi(t)].$$

Posons quelques termes et définitions intervenant dans cet énoncé et qui pourront nous être utiles par la suite : si X est une v.a.r. alors on définit sa *transformée de Laplace* comme la fonction $t \mapsto \mathbb{E}[e^{tX}]$ et sa *log-laplace* comme la fonction $t \mapsto \log\left(\mathbb{E}[e^{tX}]\right)$. D'autre part, si Λ est une fonction réelle on définit sa *duale de Legendre* par $\Lambda^*(z) = \sup_{t\in R}[zt - \Lambda(t)]$; on pourra se référer à l'annexe A (page 179) pour découvrir quelques propriétés et caractéristiques de la duale de Legendre.

Preuve. Voici un résumé de l'argument utilisé dans la preuve complète du théorème 4.5 voir la section D.1.1 page 211.

On commence par simplifier la preuve en se ramenant par translation au cas où $a = 0$ et $\mathbb{E}[X] < 0$.

On effectue une majoration de la $\overline{\lim}$ par une simple utilisation de l'inégalité de *Markov.*

La minoration de la $\underline{\lim}$ est pour sa part beaucoup plus délicate et fait intervenir la notion clée de *changement de probabilité* qui sera précisée en section 4.4. Notons que l'utilisation du théorème central limite est nécessaire pour achever la preuve en particulier grâce à la symétrie de la loi normale. ■

Corollaire 4.6 *Sous les mêmes hypothèses que précédemment, on considère* $a < \mathbb{E}[X]$ *et on a alors :*

$$\lim_{n\to+\infty} \frac{1}{n} \log \mathbb{P}(S_n \leqslant an) = -I(a)$$

n	10^2	10^3	10^4	10^5	10^6	10^7	10^8
$\lvert N_{\mathrm{TCL}} - N\rvert/N$	1.10^{-1}	1.10^{-2}	2.10^{-4}	8.10^{-4}	8.10^{-4}	8.10^{-4}	8.10^{-4}
$\lvert N_{\mathrm{GD}} - N\rvert/N$	1	2.10^{-1}	3.10^{-2}	4.10^{-3}	5.10^{-4}	7.10^{-5}	8.10^{-6}

TAB. 4.3 – Comparaison des approximations obtenues par le théorème central limite et par les
grandes déviations pour le jeu de pile ou face de paramètre $p = 0.5$ et pour $a = 0.55$. $\mathbb{P}(S_n \geqslant an) = \mathbb{P}(\mathcal{N} \geqslant N)$, N_{TCL} est défini par (4.2) et N_{GD} par (4.4)

Preuve. Il suffit d'appliquer le théorème 4.5 a $X' = -X$ pour obtenir le résultat. ■

Proposition 4.7 *On suppose que l'hypothèse (4.3) est vérifiée alors, $\forall a, a' \in \mathbb{R}$ tels que $\mathbb{E}[X] < a < a'$ on a*

$$\lim_{n\to+\infty} \frac{1}{n} \log \mathbb{P}(a'n \geqslant S_n \geqslant an) = -I(a).$$

Preuve. En effet, le théorème 4.5 nous donne :

$$\mathbb{P}(S_n \geqslant an) \overset{\log}{\sim} -I(a) \text{ et } \mathbb{P}(S_n \geqslant a'n) \overset{\log}{\sim} -I(a')$$

et comme

$$\mathbb{P}(a'n \geqslant S_n \geqslant an) = \mathbb{P}(S_n \geqslant an) - \mathbb{P}(S_n \geqslant a'n)$$

en utilisant la proposition 4.4 on obtient

$$\mathbb{P}(a'n \geqslant S_n \geqslant an) \overset{\log}{\sim} \max\big(-I(a), -I(a')\big).$$

La croissance de I sur $[\mathbb{E}[X], +\infty[$ nous permet alors de conclure. ■

Remarque 4.8 *[DZ98] permet d'affaiblir l'hypothèse (4.3) en*

$$0 \in \overset{\circ}{\mathcal{D}_\varphi} \quad \textit{avec } \mathcal{D}_\varphi = \{t \in \mathbb{R}, \varphi(t) < +\infty\}$$

où $\overset{\circ}{\mathcal{D}_\varphi}$ désigne l'intérieur de $\mathcal{D}_\varphi$.

4.3.2 Application au jeu de pile ou face

On va maintenant utiliser le résultat du théorème 4.5 dans le cadre de l'exemple déjà développé en section 4.2.

Pour cela il faut tout d'abord effectuer le calcul de la transformée de *Laplace* d'une variable aléatoire X de loi de *Bernoulli* de paramètre p ($\mathcal{B}(p)$) :

$$\begin{aligned} \varphi(t) &= \mathbb{E}[e^{tX}] \\ &= pe^t + (1-p) \end{aligned}$$

qui est bien finie pour tout $t \in \mathbb{R}$; la condition (4.3) est donc vérifiée et on peut appliquer le théorème 4.5.

Il se trouve par ailleurs que, dans ce cas simple, on dispose d'une formule explicite pour $I(a)$ avec la

Proposition 4.9

$$I(a) = a \log \frac{a}{p} + (1-a) \log \frac{(1-a)}{(1-p)}$$

Preuve. on pose $f_a(t) = at - \log(pe^t + (1-p))$ que l'on cherche à maximiser. Pour cela on calcule t tel que $f_a'(t) = 0$:

$$\begin{aligned} a = \frac{pe^t}{pe^t + (1-p)} &\Leftrightarrow ape^t + a(1-p) = pe^t \\ &\Leftrightarrow e^t = \frac{a(1-p)}{(1-a)p} \\ &\Leftrightarrow t = t_a \end{aligned}$$

avec

$$t_a = \log \frac{a(1-p)}{(1-a)p}.$$

Ainsi on a :

$$\begin{aligned} I(a) &= at_a - \log\left(pe^{t_a} + (1-p)\right) \\ &= a \log \frac{a(1-p)}{(1-a)p} - \log\left(p\frac{a(1-p)}{(1-a)p} + (1-p)\right) \\ &= a \log \frac{a(1-p)}{(1-a)p} - [a + (1-a)] \log \frac{(1-p)}{(1-a)} \end{aligned}$$

ce qui donne bien le résultat. ■

n	10^2	10^3	10^4	10^5	10^6	10^7	10^8
$\lvert N_{\mathrm{TCL}} - N\rvert/N$	8.10^{-2}	9.10^{-2}	9.10^{-2}	1.10^{-1}	1.10^{-1}	1.10^{-1}	1.10^{-1}
$\lvert N_{\mathrm{GD}} - N\rvert/N$	2.10^{-2}	3.10^{-3}	4.10^{-4}	5.10^{-5}	6.10^{-6}	7.10^{-7}	8.10^{-8}

TAB. 4.4 – Comparaison des approximations obtenues par le théorème central limite et par les
grandes déviations pour le jeu de pile ou face de paramètre $p = 0.5$ et pour $a = 0.95$. $\mathbb{P}(S_n \geqslant an) = \mathbb{P}(\mathcal{N} \geqslant N)$, N_{TCL} est défini par (4.2) et N_{GD} par (4.4)

En utilisant cette formule on peut utiliser l'approximation suivante :

$$\mathbb{P}(S_n \geqslant an) \sim e^{-nI(a)} = \mathbb{P}(\mathcal{N} \geqslant N_{GD}) \tag{4.4}$$

dont on va pouvoir comparer les valeurs avec celles obtenues précédemment à l'aide de l'approximation gaussienne.

On observe dans la table 4.3 et une constante amélioration de la qualité de l'approximation obtenue par les

grandes déviations lorsque n augmente ce qui n'était pas le cas de l'approximation gaussienne (voir section 4.2). Cette dernière est néanmoins de meilleure qualité pour de "petites" valeurs de n, mais dès que n dépasse 10^6 c'est l'approximation des

grandes déviations qu'il faut lui préférer. Dans la table 4.4 qui concerne un événement beaucoup plus rare (les déviations sont plus importantes), le phénomène observé précédemment s'accentue : les grandes déviations fournissent une approximation de meilleure qualité dès que n dépasse 10^2.

On peut retenir de ces observations les deux règles suivantes :

- *plus un événement est rare, meilleure est l'approximation de sa probabilité par les grandes déviations ;*
- *la précision de cette approximation croît d'autant plus vite avec n que la déviation considérée est importante.*

4.4 Changement de probabilités

La technique de changement de probabilité intervenant dans la preuve du théorème de *Cramér-Chernov* est d'une telle importance qu'elle mérite d'être examinée plus attentivement.

4.4.1 Corollaire de *Cramér-Chernov*

On se place sous les même hypothèses que dans le théorème 4.5 : X est une variable aléatoire de loi μ à valeurs dans $\mathcal{A} \subset \mathbb{R}$ de cardinal fini et

$$\varphi(t) = \mathbb{E}[e^{tX}] < \infty, \forall t \in \mathbb{R}.$$

Corollaire 4.10 *Soit $\widehat{\mu}$ défini sur $\mathcal{A}$ par*

$$\widehat{\mu}(x) = \frac{1}{\varphi(t_a)} e^{t_a x} \mu(x) \tag{4.5}$$

et soit t_a vérifiant $\Lambda'(t_a) = a$ avec $\Lambda = \log \varphi$ désignant la log-laplace de X.

On a alors :

(i) *$\widehat{\mu}$ est une loi sur $\mathcal{A}$;*

(ii) *si $\widehat{X}$ est une variable aléatoire de loi $\widehat{\mu}$ alors $\mathbb{E}[\widehat{X}] = a$*

Preuve. L'assertion **(i)** est évidente en utilisant la définition de φ dans la mesure où

$$\sum_{x \in \mathcal{A}} \widehat{\mu}(x) = \frac{1}{\varphi(t_a)} \sum_{x \in \mathcal{A}} e^{t_a x} \mu(x) = \frac{\mathbb{E}[e^{t_a X}]}{\varphi(t_a)} = 1$$

et on peut alors calculer la transformée de *Laplace* $\widehat{\varphi}$ de $\widehat{X}$ dont la dérivée en 0 nous donne l'espérance de $\widehat{X}$ (voir annexe A page 179).

$$\begin{aligned}
\widehat{\varphi}(t) &= \sum_{x \in \mathcal{A}} e^{tx} \widehat{\mu}(x) \\
&= \frac{1}{\varphi(t_a)} \sum_{x \in \mathcal{A}} e^{tx} e^{t_a x} \mu(x) \\
&= \frac{\varphi(t + t_a)}{\varphi(t_a)}
\end{aligned}$$

et donc

$$\widehat{\varphi}'(0) = \frac{\varphi'(t_a)}{\varphi(t_a)} = \Lambda'(t_a)$$

ce qui achève la preuve de **(ii)**. ■

L'espérance de la nouvelle probabilité définie en (4.5) étant a, la loi des grands nombres nous assure que la moyenne empirique d'un échantillon de $\widehat{X}$ sera de l'ordre de a. $\widehat{\mu}$ va donc pouvoir nous servir de "loupe" afin d'examiner ce qui se passe quand S_n est au voisinage de an ce qui n'arrive que très rarement sous la loi μ mais est très fréquent sous $\widehat{\mu}$.

On peut ainsi utiliser des échantillons générés sous $\widehat{\mu}$ pour calculer par simulation diverses quantités. A chaque fois le procédé consiste à ré-écrire la quantité en question sous la forme de l'espérance d'une fonctionnelle d'échantillons de la nouvelle loi.

La loi des grands nombres nous permet alors d'estimer cette quantité par un nombre réduit de simulations étant donnée la nature très "centrale" de l'événement considéré.

Dans le cas de la probabilité de l'événement de grandes déviations on utilise la formule de la proposition suivante :

Proposition 4.11 *On a*

$$\mathbb{P}(S_n \geqslant an) = \sum_{x_1,\ldots,x_n} \mathbb{I}_{s_n \geqslant an} e^{-s_n t_a} \varphi(t_a)^n \widehat{\mu}(x_1)\ldots\widehat{\mu}(x_n) \tag{4.6}$$

avec $s_n = x_1 + \ldots + x_n$.

4.4.2 Simulations pour le jeu de pile ou face

On va appliquer les résultats précédents à notre exemple. On rappelle deux résultats de la preuve de la proposition 4.9 :

$$t_a = \log\frac{a(1-p)}{(1-a)p} \quad \text{et} \quad \varphi(t_a) = \frac{1-p}{1-a}$$

En injectant ces formules dans (4.5) on obtient :

$$\begin{array}{rcl} \widehat{\mu}(x) & = & \frac{1-a}{1-p} e^{x \log \frac{a(1-p)}{(1-a)p}} \mu(x) \\ & = & \left\{ \begin{array}{ll} \frac{1-a}{1-p} \times \frac{a(1-p)}{(1-a)p} \times p & \text{si } x = 1 \\ \frac{1-a}{1-p} \times (1-p) & \text{si } x = 0 \end{array} \right. \\ & = & \left\{ \begin{array}{ll} a & \text{si } x = 1 \\ 1-a & \text{si } x = 0 \end{array} \right. \end{array}$$

qui correspond bien à la nouvelle probabilité qu'on attend.

La formule (4.6) s'écrit alors :

$$\mathbb{P}(S_n \geqslant an) = \mathbb{E}\left[f\left(\widehat{X}_1,\ldots,\widehat{X}_n\right)\right]$$

avec

$$f(x_1,\ldots,x_n) = \mathbb{I}_{s_n \geqslant an} \exp\left(n \log\frac{1-p}{1-a} - s_n \log\frac{a(1-p)}{(1-a)p}\right) \tag{4.7}$$

n	10^2	10^3	10^4	10^5
$\mathbb{P}(S_n \geqslant an)$	$1.4\ 10^{-1}$	$8.7\ 10^{-4}$	$7.8\ 10^{-24}$	$4.3\ 10^{-220}$
$\mathbb{P}_{\widehat{\mu}}(10^3)$	$1.3\ 10^{-1}$	$8.8\ 10^{-4}$	$6.1\ 10^{-24}$	$3.2\ 10^{-220}$
$\mathbb{P}_{\mu}(10^3)$	$1.4\ 10^{-1}$	0	0	0
$\mathbb{P}_{\mu}(10^5)$	$1.4\ 10^{-1}$	$9.0\ 10^{-4}$	0	0

TAB. 4.5 – Comparaison des approximations obtenues par $M = 10^3$ simulations sous $\widehat{\mu}$ ($\mathbb{P}_{\widehat{\mu}}(10^3)$ via la formule (4.8)) et $M = 10^3$ et $M = 10^5$ sous μ ($\mathbb{P}_{\mu}(10^3)$ et $\mathbb{P}_{\mu}(10^5)$ selon l'approche directe), pour le jeu de pile ou face de paramètre $p = 0.5$ et pour $a = 0.55$.

et on peut dès lors effectuer des simulations.

On considère $(x_1^i, \ldots, x_n^i)_{1 \leqslant i \leqslant M}$ M tirages de $\left(\widehat{X}_1, \ldots, \widehat{X}_n\right)$. On calcule la valeur de la fonction définie par (4.7) pour chacun de ces tirages et on en fait la moyenne pour obtenir une approximation de $\mathbb{P}(S_n \geqslant an)$ par

$$\mathbb{P}_{\widehat{\mu}}(M) = \frac{1}{M} \sum_{i=1}^{M} f(x_1^i, \ldots, x_n^i). \tag{4.8}$$

La table 4.5 montre que même si l'approche par simulations directes donne des résultats acceptables pour l'estimation de probabilités importantes ($n = 10^2$ par exemple), son efficacité décroît rapidement lorsque les probabilités des événements étudiés diminuent. On peut pallier cette déficience en augmentant le nombre M de simulations ce qui permet ici, par exemple, d'obtenir une nouvelle estimation pour $n = 10^3$, mais cette approche, très coûteuse en temps de calcul, ne peut être numériquement viable.

D'un autre côté, les simulations effectuées sous la loi $\widehat{\mu}$ donnent de très bons résultats quelle que soit la "rareté" de l'événement considéré ce qui était précisément le but du changement de probabilité effectué.

Notons au passage qu'une telle approche pourrait révéler tout son intérêt dans le cas d'applications sortant du cadre des approximations asymptotiques (petites valeurs de n par exemple) mais mettant néanmoins en jeu des événements de trop faibles probabilités pour des simulations directes.

4.5 Mesure empiriques

On s'est intéressé dans ce qui précède aux *grandes déviations de niveau 1* c'est à dire aux déviations d'une moyenne par rapport à un attendu.

Dans ce qui va suivre on va considérer des *grandes déviations de niveau 2* : celles des distributions empiriques par rapport aux distributions théoriques.

En assimilant un échantillon de taille n de X à un *texte aléatoire* dans l'*alphabet fini* $\mathcal{A}$, on utilisera les termes de *lettres* et *mots* pour désigner respectivement les éléments de $\mathcal{A}$ et les successions de plusieurs de ces éléments.

On suppose sans perte de généralité de $\mathcal{A} = \{1, \ldots, k\}$ dont le cardinal est, par conséquent, noté k.

4.5.1 Les singletons

On considère ici un échantillon $(X_1, \ldots, X_n)$ de X et on définit la distribution empirique $L_n^1 = L_n = (L_n(1), \ldots, L_n(k))$ de ses lettres de la façon suivante :

$$L_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{X_i = x} \quad \forall x \in \mathcal{A}. \tag{4.9}$$

Notons que l'équation (4.9) définit clairement une mesure de probabilité sur $\mathcal{A}$ et donc $L_n \in \mathcal{M}_1(\mathcal{A})$ (ce dernier symbole désigne l'ensemble des mesures de probabilités su $\mathcal{A}$).

Si μ désigne la loi de X, il est clair que $\mathbb{E}[\mathbb{I}_{X=x}] = \mu(x)$ pour tout $j \in \mathcal{A}$ et la loi des grands nombres nous permet alors d'obtenir

$$d(L_n, \mu) \xrightarrow[n \to +\infty]{} 0 \quad \text{p.s.}$$

avec d désignant la *distance en variation totale* définie par

$$d(\nu, \nu') = \frac{1}{2} \sum_{x \in \mathcal{A}} |\nu(x) - \nu'(x)| \quad \nu, \nu' \in \mathcal{M}_1(\mathcal{A}).$$

Le résultat dû a *Sanov* ([San61]) s'intéresse aux déviations de L_n par rapport à μ constitue l'analogue du théorème de *Chérnov* pour les distributions empiriques des lettres.

On suppose que $\mu(x) > 0$ pour tout $x \in \mathcal{A}$ sans perte de généralité (on peut s'y ramener en travaillant avec $\mathcal{A}' = \{x \in \mathcal{A} \text{ tq } \mu(x) > 0\}$) et on a alors :

Théorème 4.12 (*Sanov*)
On a les 2 résultats suivants :

(MAJ) *pour tout $F \subset \mathcal{M}_1(\mathcal{A})$ fermé on a*

$$\varlimsup_{n \to +\infty} \frac{1}{n} \log \mathbb{P}\left(L_n \in F\right) \leq - \inf_{\nu \in F} I(\nu);$$

(MIN) *et pour tout $O \subset \mathcal{M}_1(\mathcal{A})$ ouvert on a*

$$\overline{\lim_{n\to+\infty}} \frac{1}{n} \log \mathbb{P}\left(L_n \in O\right) \geq - \inf_{\nu \in O} I(\nu)$$

où

$$I(\nu) = H(\nu|\mu) = \sum_{x\in\mathcal{A}} \nu(x) \log\left(\frac{\nu(x)}{\mu(x)}\right)$$

*désigne l'*entropie relative *de ν par rapport à μ (voir la section A.4 page 184 pour plus de détails).*

Preuve. Voici un résumé de la preuve du théorème 4.12. Pour l'argument complet on se reportera à la section D.1.2 (page 215).

On va se contenter de montrer **(MAJ)** et **(MIN)** pour les boules ouvertes avant d'utiliser un petit argument topologique pour obtenir les résultats dans toute leur généralité.

Comme $\mathcal{A}$ est de cardinal fini, il est possible de calculer formellement la loi de L_n pour un n donné. La démonstration consiste simplement à utiliser ces formules pour obtenir un encadrement de $\mathbb{P}\left(L_n \in B_\rho^c\right)$ qui donne l'équivalent logarithmique attendu en l'infini.

On conclut finalement en utilisant un simple argument de densité. ■

Remarque 4.13 *Si Γ est une partie "assez régulière" de $\mathcal{M}_1(\mathcal{A})$ on obtient facilement par continuité de I*

$$\lim_{n\to+\infty} \frac{1}{n} \log \mathbb{P}\left(L_n \in \Gamma\right) = - \inf_{\nu\in\Gamma} I(\nu)$$

en appliquant successivement **(MAJ)** *à $\overline{\Gamma}$, l'adhérence de Γ et* **(MIN)** *à $\mathring{\Gamma}$, son intérieur.*

4.5.2 Les paires

On va s'intéresser ici aux distributions empiriques non plus des lettres mais des paires de lettres ; des mots de deux lettres.

Afin de simplifier l'écriture des formules, on va supposer à partir de maintenant que $X_{n+1} = X_1$. Cette hypothèse, classique dans l'étude de séquences de lettres, induit une perte de généralités que l'on négligera néanmoins dans la suite étant donné le faible impact de cet "effet de bord" sur l'étude d'une longue séquence.

On commence par définir $L_{n,2}$ sur $\mathcal{M}_1(\mathcal{A}^2)$, l'ensemble des mesures de probabilités sur les mots de deux lettres, par une formule analogue à (4.9) :

$$L_{n,2}(x,y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{X_i X_{i+1} = xy} \quad \forall (x,y) \in \mathcal{A}^2. \tag{4.10}$$

Si X et X' sont deux variables indépendantes de loi μ, il est clair que $\mathbb{E}[\mathbb{I}_{XX'=xy}] = \mu(x)\mu(y)$ et donc

$$d(L_{n,2}, \mu \times \mu) \xrightarrow[n\to+\infty]{} 0 \quad \text{p.s.}$$

par la loi des grands nombres (comme en section 4.5.1, d désigne la distance en variation totale mais cette distance est ici définie sur $\mathcal{M}_1(\mathcal{A}^2)$).

Par analogie avec le théorème 4.12 on établit alors le

Théorème 4.14 (Paires)
On a les 2 résultats suivants :

(MAJ) *pour tout $F \subset \mathcal{M}_1(\mathcal{A}^2)$ fermé on a*

$$\overline{\lim_{n\to+\infty}} \frac{1}{n} \log \mathbb{P}\left(L_{n,2} \in F\right) \leq -\inf_{\nu\in F} I_2(\nu);$$

(MIN) *et pour tout $O \subset \mathcal{M}_1(\mathcal{A}^2)$ ouvert on a*

$$\overline{\lim_{n\to+\infty}} \frac{1}{n} \log \mathbb{P}\left(L_{n,2} \in O\right) \geq -\inf_{\nu\in O} I_2(\nu)$$

avec

$$I_2(\nu) = \begin{cases} \sum\limits_{x,y\in\mathcal{A}} \nu(x,y) \log \frac{\nu(x,y)}{\overline{\nu}(x)\mu(y)} & \text{si } \nu \in \mathcal{S} \\ +\infty & \text{sinon} \end{cases}$$

où $\mathcal{S}$ désigne l'ensemble des éléments ν de $\mathcal{M}_1(\mathcal{A}^2)$ qui sont invariants par translations (shift-invariants) *c'est à dire vérifiant :*

$$\overline{\nu}(x) = \sum_{y\in\mathcal{A}} \nu(x,y) = \sum_{y\in\mathcal{A}} \nu(y,x). \tag{4.11}$$

(on pourra se reporter à la définition 7.3 page 110 pour plus de précisions sur cette notion).

Preuve. Voici un résumé de la preuve du théorème 4.14. Pour l'argument complet on se reportera à la section D.1.3 (page 219).

Calquant notre démonstration sur celle du théorème 4.12 on se contente là encore de de montrer **(MAJ)** et **(MIN)** pour les boules ouvertes.

On va, là aussi, utiliser le fait que $\mathcal{A}$ est de cardinal fini, pour calculer la loi de L_n^2 pour un n donné.

Une nouvelle difficulté surgit cependant : il faut trouver une condition nécessaire et suffisante à l'existence de séquences de longueur n respectant les comptages de mots de deux lettres donnés, puis être en mesure de dénombrer ces séquences.

Dans le cas où les comptages vérifient une condition d'invariance par translation similaire à celle posée en (4.11), *Whittle* ([Whi55]) donne la solution à ce problème (voir section B.2 page 195). C'est en utilisant sa formule que l'on obtient finalement les encadrements puis les équivalents logarithmiques attendus.

La conclusion s'effectue par un nouvel argument de densité. ■

Remarque 4.15 *Lorsque $\nu \in \mathcal{S}$, il est clair que*

$$I_2(\nu) = H(\nu|\overline{\nu} \otimes \mu)$$

forme sous laquelle les propriétés de la fonction de taux peuvent paraître plus naturelles.

Références

[Cra38] H. Cramér. Sur un nouveau théorème-limite dans la théorie des probabilités. *Actualités Scientifiques et Industrielles*, pages 5–23, 1938.

[dH00] F. den Hollender. *Large Deviations.* American Mathematical Society, 2000.

[DZ98] A. Dembo and O. Zeitouni. *Large deviations techniques and applications.* Springer-Verlag, 1998.

[San61] I.N. Sanov. On the probability of large deviations of random variables. *Mat. Sb. 42 (en Russe). Traduction en englais dans : Selected translations in Mathematical Statistics and Probability I*, pages 213–244, 1961.

[Whi55] P. Whittle. Some distribution and moment fomulæ for the markov chain. *J. R. Statist. Soc. B.*, 17 :235–242, 1955.

Deuxième partie

Grandes déviations de niveau 1

FIG. 5.1 – Représentation graphique des mots de longueur $h = 9$ chez *Eschericia Coli* dans le modèle $M1$. On consultera la section 11.1 page 160 pour plus de détails.

Chapitre 5

Théorie

Pré-requis :

Chapitre 2 et chapitre 4.

Description :

Résultats théoriques des grandes déviations de niveau 1 du nombre d'occurrences de mots sur une chaîne de *Markov*.

Résumé :

On établit tout d'abord quelques propriétés de la fonction Λ, dont les calculs des dérivées première et seconde sont effectués.

L'utilisation de ces calculs permet d'établir la preuve d'un résultat de grandes déviations de type *Cramér-Chernov* pour le comptage des mots de deux lettres sur une chaîne de *Markov* irréductible d'ordre 1 en utilisant la technique classique du changement de probabilité.

En invoquant les théorèmes de *Gärtner-Ellis* et de *Perron-Frobénius*, on montre enfin l'existence plus générale d'un principe de grandes déviations pour la même quantité.

Contenu du chapitre

5.1 Propriétés de Λ

On introduit ici un grand nombre de notations ainsi que les hypothèses qui vont servir dans tout le chapitre. On considère une matrice stochastique Π sur l'espace d'état $\mathcal{A}$ de cardinal k. On suppose que cette matrice est *irréductible*, c'est à dire que $\forall x, y \in \mathcal{A}$ deux états $\exists n(x, y) \in \mathbb{N}$ tel que $\Pi^n(x, y) > 0$ (il existe un "chemin" joignant x à y).

On considère également une fonction déterministe quelconque

$$f : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$$

et on définit $\forall \theta \in \mathbb{R}$ la matrice Π_θ par

$$\Pi_\theta(x, y) = \Pi(x, y)e^{\theta f(x,y)} \quad \forall x, y \in \mathcal{A}. \tag{5.1}$$

Il est clair que Π_θ est irréductible quelle que soit la valeur de θ et on peut donc appliquer le théorème de *Perron-Frobénius* (voir théorème B.5 page 193) à cette matrice positive dont on note alors $\rho(\Pi_\theta)$ la plus grande valeur propre.

On va dans la suite s'intéresser aux propriétés de la fonction Λ définie par

$$\Lambda(\theta) = \log \rho(\Pi_\theta) \quad \forall \theta \in \mathbb{R} \tag{5.2}$$

ainsi qu'à celles de Λ^* sa duale de *Legendre* (voir section A.1 page 179).

On a alors la

Proposition 5.1 *Les propriétés suivantes sont vérifiées :*

(i) $\Lambda(0) = 0$ *;*

(ii) Λ *est convexe ;*

(iii) Λ *est analytique ;*

(iv) $\Lambda^* \geq 0$ *;*

(v) $\Lambda^*(\Lambda'(0)) = 0$.

Preuve.

(i) Comme $\Pi_0 = \Pi$ est une matrice stochastique, elle admet le vecteur colonne $\mathbf{1} = [1 \dots 1]'$ comme vecteur propre associé à la valeurs propre 1. Comme $\mathbf{1} \gg 0$, la remarque B.8 (page 195) permet de conclure que $\rho(\Pi) = 1$ ce qui achève la preuve de **(i)**.

(ii) La proposition 5.4 (à venir) nous assure que $\Lambda'' \geq 0$ donc que Λ est convexe. On peut aussi constater, via le lemme 5.7, que Λ, est convexe comme limite de fonctions convexes.

(iii) $\rho(\Pi_\theta)$ étant une racine d'un polynôme à coefficients analytiques en θ (le polynôme caractéristique), l'analycité de Λ est alors clairement établie par le théorème des fonctions implicites.

(iv) Comme $\Lambda(0) = 0$, $\forall x \in \mathbb{R}$ on a $\Lambda^*(x) \geq [tx - \Lambda(t)]_{t=0} = 0$.

(v) D'après la proposition A.3 (page 180), il est clair que $\Lambda^*(\Lambda'(0)) = [\tau x - \Lambda(\tau)]_{\tau=0} = 0$.

■

Le théorème de *Perron-Frobénius* nous assure également l'existence de v_θ (resp. w_θ) vecteur propre à droite (resp. à gauche) de Π_θ associé à la valeur propre $\rho(\Pi_\theta)$. On peut alors définir pour tout $\theta \in \mathbb{R}$ la *matrice tiltée* $\widetilde{\Pi}_\theta$ par

$$\widetilde{\Pi}_\theta(x, y) = e^{-\Lambda(\theta)} \frac{v_\theta(y)}{v_\theta(x)} \Pi_\theta(x, y) = e^{-\Lambda(\theta)} \frac{v_\theta(y)}{v_\theta(x)} \Pi(x, y) e^{\theta f(x,y)}.$$

Proposition 5.2 *La matrice $\widetilde{\Pi}_\theta$ est irréductible, stochastique et admet (à un facteur de normalisation près)*

$$q_\theta = [v_\theta(1) w_\theta(1) \dots v_\theta(k) w_\theta(k)] \tag{5.3}$$

comme distribution stationnaire (i.e. $q_\theta \widetilde{\Pi}_\theta = q_\theta$).

Preuve. Il est évident que $\widetilde{\Pi}_\theta$ est irréductible car Π l'est, il faut montrer la stochasticité. Soit $x \in \mathcal{A}$,

$$\begin{aligned}
\sum_{y \in \mathcal{A}} \widetilde{\Pi}_\theta(x, y) &= \frac{e^{-\Lambda(\theta)}}{v_\theta(x)} \sum_{y \in \mathcal{A}} \Pi_\theta(x, y) v_\theta(y) \\
&= \frac{e^{-\Lambda(\theta)}}{v_\theta(x)} \rho(\Pi_\theta) v_\theta(x) \\
&= 1.
\end{aligned}$$

En ce qui concerne la distribution stationnaire, quitte à multiplier v_θ et w_θ par des constantes, on peut supposer que q_θ, comme défini en (5.3), est une mesure de probabilité sur $\mathcal{A}$. Soit $y \in \mathcal{A}$, on calcule

$$\begin{aligned}
\sum_{x\in\mathcal{A}} q_\theta(x)\widetilde{\Pi}_\theta(x,y) &= e^{-\Lambda(\theta)}\sum_{x\in\mathcal{A}} \frac{v_\theta(x)w_\theta(x)v_\theta(y)}{v_\theta(x)}\Pi_\theta(x,y) \\
&= e^{-\Lambda(\theta)}v_\theta(y)\sum_{x\in\mathcal{A}} w_\theta(x)\Pi_\theta(x,y) \\
&= e^{-\Lambda(\theta)}v_\theta(y)\rho(\Pi_\theta)w_\theta(y) \\
&= q_\theta(y).
\end{aligned}$$

■

On considère $(X_i)_i$ une chaîne de *Markov* d'ordre 1 sur $\mathcal{A}$ et on peut alors énoncer les deux propositions suivantes :

Proposition 5.3 (dérivées premières)
Pour tout $\theta \in \mathbb{R}$ on a

$$\Lambda'(\theta) = \mathbb{E}_{q_\theta}^{\widetilde{\Pi}_\theta}\left[f(X_1,X_2)\right]$$

($\mathbb{E}_{q_\theta}^{\widetilde{\Pi}_\theta}$ désigne l'espérance sous la loi de chaîne de Markov *de distribution initiale q_θ et de matreice de transition $\widetilde{\Pi}_\theta$).*

Preuve. La preuve très calculatoire consiste simplement à dériver par rapport à θ l'équation

$$\sum_{y\in\mathcal{A}} \widetilde{\Pi}_\theta(x,y) = 1$$

et à utiliser les définitions des vecteurs propres v_θ et w_θ.

Pour une preuve complète on se référera à la section D.2 (page 221). ■

Proposition 5.4 (dérivées secondes)
Pour tout $\theta \in \mathbb{R}$ on a

$$\begin{aligned}
\Lambda''(\theta) &= \sum_{n\in\mathbb{Z}} \mathbb{C}ov_{q_\theta}^{\widetilde{\Pi}_\theta}\left[f(X_1,X_2), f(X_n,X_{n+1})\right] \\
&= \lim_{n\to+\infty}\frac{1}{n}\mathbb{V}ar_{q_\theta}^{\widetilde{\Pi}_\theta}\left[\sum_{i=1}^{n} f(X_i,X_{i+1})\right].
\end{aligned}$$

Preuve. Les calculs sont plus compliqués que pour la preuve de la proposition 5.3 mais procèdent des mêmes méthodes.

On trouvera l'ensemble de ces calculs en section D.2 (page 221). ■

5.2 Théorème de *Cramér-Chernov*

On considère $(X_i)_{i=1,\ldots,n}$ une chaîne de *Markov* d'ordre 1 sur $\mathcal{A}$, de matrice de transition Π irréductible et de distribution stationnaire μ.

L'hypothèse d'irréductibilité assure l'*ergodicité* de la chaîne de *Markov* pour laquelle on a dès lors :

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i, X_{i+1}) \xrightarrow[n\to+\infty]{} \mathbb{E}_{\mu}^{\Pi}[f(X_1, X_2)] \quad \text{p.s.}$$

On va donc désormais alors s'intéresser aux déviations de

$$S_n = \sum_{i=1}^{n} f(X_i, X_{i+1})$$

qui sont gérées par le

Théorème 5.5 (*Cramér-Chernov* pour les chaînes de *Markov*) *Soit $a > \mathbb{E}_{\mu}^{\Pi}[f(X_1, X_2)]$ alors*

$$\lim_{n\to+\infty} \frac{1}{n} \log \mathbb{P}(S_n \geq an) = -\Lambda^*(a)$$

où Λ^* désigne la duale de Legendre *de Λ (définie en section 5.1).*

Preuve. Pour la preuve complète de ce théorème on se reportera à la section D.3.1 (page 228).

Le procédé reste ici le même que dans le cas i.i.d., on commence par simplifier la preuve en se ramenant par translation au cas $a = 0$ et $\mathbb{E}_{\mu}^{\Pi}[f(X_1, X_2)] < 0$. On effectue une majoration de la $\overline{\lim}$ en utilisant l'inégalité de *Markov*. La minoration de la $\underline{\lim}$ utilise un changement de probabilité ainsi qu'un résultat de limite centrale. ■

Comme dans le cas i.i.d., on obtient aisément le

Corollaire 5.6 *Soit $a < \mathbb{E}_{\mu}^{\Pi}[f(X_1, X_2)]$ alors*

$$\lim_{n\to+\infty} \frac{1}{n} \log \mathbb{P}(S_n \leq an) = -\Lambda^*(a).$$

en appliquant le théorème 5.5 à $(-X_i)_i$.

5.3 Application de *Gärtner-Ellis*

Il est également possible d'utiliser le théorème de *Gärtner-Ellis* (voir section A.3 page 182) pour établir un résultat plus général qu'en section 5.2.

On considère pour cela une fonction déterministe quelconque

$$f : \mathcal{A} \times \mathcal{A} \to \mathbb{R}^d$$

et on va s'intéresser aux déviations de la variable aléatoire S_n à valeurs dans $\mathbb{R}^d$ définie par

$$S_n = \sum_{i=1}^{n} f(X_i, X_{i+1}).$$

Pour tout $n \in \mathbb{N}$ et pour tout $\theta \in \mathbb{R}$ on pose

$$\Lambda_n(\theta) = \log \mathbb{E}[e^{\langle \theta, S_n \rangle}]$$

et on étend la définition (5.1) et (5.2) en posant $\forall \theta \in \mathbb{R}$

$$\Pi_\theta(x, y) = \Pi(x, y) e^{\langle \theta, f(x,y) \rangle} \quad \forall x, y \in \mathcal{A}$$

avec $\langle , \rangle$ désignant un produit scalaire sur $\mathbb{R}^d$.

On a alors le résultat suivant :

Lemme 5.7 $\forall \theta \in \mathbb{R}$ *on a*

$$\lim_{n \to +\infty} \frac{1}{n} \Lambda_n(\theta) = \Lambda(\theta)$$

Preuve.

$$\begin{aligned}
\mathbb{E}[e^{\langle \theta, S_n \rangle}] &= \sum_{x_2, \ldots, x_{n+1}} e^{\langle \theta, f(s, x_2) \rangle} \ldots e^{\langle \theta, f(x_n, x_{n+1}) \rangle} \Pi(s, x_2) \ldots \Pi(x_n, x_{n+1}) \\
&= \sum_{x_2, \ldots, x_{n+1}} \Pi_\theta(s, x_2) \ldots \Pi_\theta(x_n, x_{n+1}) \\
&= \sum_{x} \Pi_\theta^n(s, x)
\end{aligned}$$

si bien qu'en utilisant le corollaire B.7 (page 194) on obtient le résultat. ∎

Comme Λ est par ailleurs différentiable (et même $\mathcal{C}^\infty$) sur $\mathcal{D}_\Lambda$ les conditions d'application du théorème A.6 (page 183) sont réunies ; on a donc

Proposition 5.8 $(\frac{S_n}{n})$ *suit un PGD de bonne fonction de taux* Λ^*.

Notons que le théorème 5.5 peut être vu comme un corollaire de ce dernier résultat.

Références

[Buc90] J. A. Bucklew. *Large deviation techniques in decision, simulation, and estimation.* Wiley, 1990.

[dH00] F. den Hollender. *Large Deviations.* American Mathematical Society, 2000.

[DZ98] A. Dembo and O. Zeitouni. *Large deviations techniques and applications.* Springer-Verlag, 1998.

[Tor98] N. Torrent. *Application des grandes déviations et de la loi d'Erdös-Réyni pour des variables indépendantes ou de dépendance markovienne.* PhD thesis, Université PARIS VII, 1998.

FIG. 6.1 – Représentation graphique des mots de longueur $h = 9$ chez *Bacillus Subtilis* dans le modèle $M1$. On consultera la section 11.1 page 160 pour plus de détails.

Chapitre 6

Mise en oeuvre pratique

Pré-requis :

Chapitre 5

Description :

Utilisation pratique des résultats théoriques de niveau 1 pour le calcul de la significativité de comptages de mots et de motifs.

Résumé :

En utilisant la technique du changement d'alphabet, il est possible de ramener dans le cadre des résultats précédents le problème concernant un mot de longueur h sur une chaîne de *Markov* d'ordre quelconque.

Le traitement de motifs s'effectue de la même façon, la seule difficulté consistant à choisir les valeurs de la fonction déterministe permettant le comptage des occurrences du motif.

Dans les deux cas, il est nécessaire de calculer le minimum d'une fonction dont chaque évaluation demande le calcul de la plus grande valeur propre d'une matrice (creuse) d'ordre k^{h-1}.

On remarque qu'il est possible d'expliciter le changement de probabilité intervenant dans les grandes déviations et d'en utiliser les paramètres pour effectuer des simulations.

Enfin on présente la mise en œuvre des calculs impliqués dans ces méthodes du point de vue algorithmique.

Contenu du chapitre

6.1 Un mot

6.1.1 Introduction et notations

On reprend ici les notations de la section 2.2.1 (page 32) : on considère $X = X_1 \dots X_n$ une chaîne de *Markov* d'ordre m et $W = w_1 \dots w_h$ un mot de longueur h et on s'intéresse à la variable aléatoire $N(W)$ du nombre d'occurrences de W observées dans X (voir figure 6.2 pour un exemple).

On se donne également une séquence *observée* $x = x_1 \dots x_n$ et on cherche à calculer la significativité du comptage $n(W)$ du mot W dans la séquence x par rapport au modèle markovien d'ordre m :

$$\mathbb{P}\big(N(W) = n(W)\big). \tag{6.1}$$

Comme on l'a vu en section 2.3 (page 36), si l'on utilise la séquence x pour estimer les paramètres du modèle Mm, alors, dès que $m \geq h-1$, les comptages observés des mots de longueur h interviennent dans l'estimation des paramètres, et donc en particulier $n(W)$, si bien que la probabilité (6.1) est nécéssairement égale à 1, résultat fort peu informatif.

Dans la suite on se placera sous les hypothèse suivantes :

$$\begin{cases} h \geq 2 \\ 0 \leq m \leq h-2 \\ n \gg h \end{cases} .$$

```
gccccagcccatctggccgccgttcagcgcctgctggtcggcagaatgagcaatcgccag
cttaccgaaatcagcgccgttacgcgcctgatcgacaatggcgcgcgcctggctttccgc
ttcgttcacctgatcagaggtcgggttttccggcagcgggatcaggatgtggctcaggtt
cagctcagtgctggcgtcgttttggttacccacctgctgcgccagggattcgacttcctg
```

FIG. 6.2 – Exemple de séquence x dans l'alphabet $\mathcal{A} = \{\texttt{a}, \texttt{c}, \texttt{g}, \texttt{t}\}$. Les trois occurrences du mot $W = \texttt{ctgc}$ sont soulignées.

et on va s'efforcer d'évaluer la quantité (6.1).

Pour cela, on souhaite utiliser les résultats développés au chapitre précédent. Ces résultats concerne les déviations de la quantité

$$S_n = \sum_{i=1}^{n} f(X_i, X_{i+1})$$

où $f : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ est une fonction déterministe arbitraire et lorsque $X = (X_i)_i$ est une chaîne de *Markov* irréductible d'ordre 1.

6.1.2 Cas d'un mot de deux lettres

Dans ce cas simple, les résultats théoriques peuvent être utilisés directement. Si $h = 2$ et $m = 1$, alors $W = w_1 w_2$ et il suffit en effet de poser

$$f(y, z) = \mathbb{I}_{y=w_1} \times \mathbb{I}_{z=w_2}$$

pour que $S_n = N(W)$.

On a alors le résultat suivant

Proposition 6.1 *Si* Π *est irréductible alors on a l'approximation*

$$\mathbb{P}\big(N(W) = n(W)\big) \sim e^{-n\Lambda^*(a)}$$

avec

$$\Lambda^*(a) = \sup_{\theta \in \mathbb{R}} \{\theta a - \Lambda(\theta)\}$$

duale de Legendre *en a de la fonction Λ définie en section 5.1 (page 84).*

De plus, $\exists \tau \in \mathbb{R}$ tel que

$$\Lambda^*(a) = \lim_{\theta \to \tau} \{\theta a - \Lambda(\theta)\}$$

et, si $\tau > 0$ (resp. $\tau < 0$) alors W est sur-représenté (resp. sous-représenté).

Preuve. On pose
$$a = \frac{n(W)}{n}$$
et on suppose que $a > \mathbb{E}_\mu^\Pi[f(X_1, X_2)] = \mu(w_1)\Pi(w_1, w_2)$ alors, en appliquant le théorème 5.5 (page 87), on obtient
$$\lim_{n\to+\infty} \frac{1}{n} \log \mathbb{P}\big(N(W) \geq n(W)\big) = -\Lambda^*(a).$$

En procédant de la même façon que dans la preuve de la proposition 4.7 (page 70) on obtient aisément $\forall \varepsilon > 0$
$$\lim_{n\to+\infty} \frac{1}{n} \log \mathbb{P}\big(n(W) + \varepsilon n > N(W) \geq n(W)\big) = -\Lambda^*(a).$$
ce qui, à distance finie, donne bien l'approximation de la proposition en faisant $\varepsilon \to 0$.

En utilisant les résultats des sections 5.1 (page 84) et A.1 (page 179) on obtient les résultats suivants : $\Lambda'(\mathbb{R}) =]\alpha, \beta[$ ($\alpha, \beta \in \overline{\mathbb{R}}$), Λ' croissante et $\Lambda'(0) = \mathbb{E}_\mu^\Pi[f(X_1, X_2)]$ (μ désignant la distribution stationnaire de Π). On peut alors établir que :

cas 1 si $a \leq \alpha$ alors $a - \Lambda'(\theta) \leq 0$ pour tout $\theta \in \mathbb{R}$ et donc en particulier $a \leq \Lambda'(0)$ si bien que
$$\Lambda^*(a) = \lim_{\theta\to-\infty} \{\theta a - \Lambda(\theta)\}$$
et la preuve est donc achevée en posant $\tau = -\infty$.

cas 2 si $\alpha < a < \beta$ alors $\exists \tau \in \mathbb{R}$ tel que $\Lambda'(\tau) = a$ si bien que
$$\Lambda^*(a) = \tau a - \Lambda(\tau).$$
Et de plus, comme Λ' est croissante, $\tau > 0 \Rightarrow a > \Lambda'(0)$ et $\tau < 0 \Rightarrow a < \Lambda'(0)$.

cas 3 si $a \leq \beta$ alors on procède comme dans le **cas 1** en posant cette fois-ci $\tau = +\infty$ pour terminer la preuve.

■

6.1.3 Cas général

Changement d'alphabet

Si $h > 2$, il n'est plus possible de définir f pour relier le comptage de W à S_n. Il existe cependant une astuce (très classique dans le cas des chaînes

```
[gcc][ccc][ccc][cca][cag][agc][gcc][ccc][cca][cat][atc][tct]
[ctg][tgg][ggc][gcc][ccg][cgc][gcc][ccg][cgt][gtt][ttc][tca]
[cag][agc][gcg][cgc][gcc][cct][ctg][tgc][gct][ctg][tgg][ggt]
[gtc][tcg][cgg][ggc][gca][cag][aga][gaa][aat][atg][tga][gag]
[agc][gca][caa][aat][atc][tcg][cgc][gcc][cca][cag][agc][gct]
[ctt][tta][tac][acc][ccg][cga][gaa][aaa][aat][atc][tca][cag]
[agc][gcg][cgc][gcc][ccg][cgt][gtt][tta][tac][acg][cgc][gcg]
[cgc][gcc][cct][ctg][tga][gat][atc][tcg][cga][gac][aca][caa]
[aat][atg][tgg][ggc][gcg][cgc][gcg][cgc][gcg][cgc][gcc][cct]
[ctg][tgg][ggc][gct][ctt][ttt][ttc][tcc][ccg][cgc][gct][ctt]
[ttc][tcg][cgt][gtt][ttc][tca][cac][acc][cct][ctg][tga][gat]
[atc][tca][cag][aga][gag][agg][ggt][gtc][tcg][cgg][ggg][ggt]
[gtt][ttt][ttt][ttc][tcc][ccg][cgg][ggc][gca][cag][agc][gcg]
[cgg][ggg][gga][gat][atc][tca][cag][agg][gga][gat][atg][tgt]
[gtg][tgg][ggc][gct][ctc][tca][cag][agg][ggt][gtt][ttc][tca]
[cag][agc][gct][ctc][tca][cag][agt][gtg][tgc][gct][ctg][tgg]
[ggc][gcg][cgt][gtc][tcg][cgt][gtt][ttt][ttt][ttg][tgg][ggt]
[gtt][tta][tac][acc][ccc][cca][cac][acc][cct][ctg][tgc][gct]
[ctg][tgc][gcg][cgc][gcc][cca][cag][agg][ggg][gga][gat][att]
[ttc][tcg][cga][gac][act][ctt][ttc][tcc][cct][ctg]
```

FIG. 6.3 – Séquence $x^{(3)}$: séquence x de la figure 6.2 écrite dans $\mathcal{A}^3$. Les trois occurrences du mot $W^{(3)} =$ `[ctg][tgc]` sont soulignées

de *Markov*) qui consiste à considérer le problème dans un nouvel alphabet : l'alphabet $\mathcal{A}^{h-1}$. Dans cet alphabet, les lettres correspondent aux anciens mots de $h-1$ lettres et se chevauchent les unes les autres.

Si on note $[y_1 \dots y_{h-1}] \in \mathcal{A}^{h-1}$ ($y_i \in \mathcal{A}$) la "lettre" du nouvel alphabet correspondant au mot $y_1 \dots y_{h-1}$ dans l'ancien, alors :

$$W^{(h-1)} = [w_1 \dots w_{h-1}][w_2 \dots w_h]$$

est un mot de longueur 2 dans $\mathcal{A}^{h-1}$ correspondant au mot W (de longueur h) dans $\mathcal{A}$ et

$$X^{(h-1)} = [X_1 \dots X_{h-1}][X_2 \dots X_h] \dots [X_{n-h+2} \dots X_n]$$

de longueur $n-h+2$ correspond à la séquence X (qui était de longueur n).

Avec ces notations, en posant

$$f([y_1 \dots y_{h-1}], [z_1 \dots z_{h-1}]) = \mathbb{I}_{[y_1 \dots y_{h-1}]=[w_1 \dots w_{h-1}]} \times \mathbb{I}_{[z_1 \dots z_{h-1}]=[w_2 \dots w_h]}$$

on a bien

$$N(W) = N\left(W^{(h-1)}\right) = S_n = \sum_{i=1}^{n-h+1} f\big([X_i \dots X_{i+h-1}], [X_{i+1} \dots X_{i+h}]\big)$$

comme on pourra s'en convaincre en examinant l'exemple de la figure 6.3.

Soit $\Pi^{(h-1)}$ la matrice de transition de $X^{(h-1)}$ la chaîne de *Markov* d'ordre 1 sur $\mathcal{A}^{h-1}$. La structure chevauchante des "lettres" du nouvel alphabet donne la formule suivante pour un terme générique de la matrice :

$$\Pi^{(h-1)}\big([y_1 \dots y_{h-1}], [z_1 \dots z_{h-1}]\big) = \begin{cases} 0 & \text{si } [y_2 \dots y_{h-1}] \neq [z_1 \dots z_{h-2}] \\ \mathbb{P}(X_h = z_h | X_1 = y_1, \dots X_{h-1} = y_{h-1}) & \text{sinon} \end{cases}$$

La matrice $\Pi^{(h-1)}$ d'ordre k^{h-1} contient donc ainsi, au maximum, k^h termes non nuls ; c'est une matrice *creuse.*

A titre d'exemple, voici un extrait du contenu d'un matrice $\Pi^{(3)}$ dans le cas où $\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$

$\Pi^{(3)}$	[aaa]	[aac]	[aag]	[aat]	[aca]	...
[aaa]	$\mathbb{P}$(a|aaa)	$\mathbb{P}$(c|aaa)	$\mathbb{P}$(g|aaa)	$\mathbb{P}$(t|aaa)	0	...
[aac]	0	0	0	0	$\mathbb{P}$(a|aac)	...
[aag]	0	0	0	0	0	...
⋮	⋮	⋮	⋮	⋮	⋮	

Résultats

Ces notations étant posées, il suffit de considérer la fonction f définie $\forall y = [y_1, \ldots, y_{h-1}], z = [z_1, \ldots, z_{h-1}] \in \mathcal{A}^{h-1}$ par

$$f(y,z) = \mathbb{I}_{[y_1,\ldots,y_{h-1}]=[w_1,\ldots,w_{h-1}]} \times \mathbb{I}_{[z_1,\ldots,z_{h-1}]=[w_2,\ldots,w_h]}$$

et, $\forall \theta \in \mathbb{R}$, de poser

$$\Pi_\theta^{(h-1)}(y,z) = \Pi^{(h-1)}(y,z)e^{\theta f(y,z)} \quad \forall y, z \in \mathcal{A}^{h-1}.$$

pour obtenir le

Théorème 6.2 *Si $\Pi^{(h-1)}$ est irréductible, alors on a l'approximation*

$$\mathbb{P}\big(N(W) = N(W)_{obs}\big) \sim e^{-n\Lambda^*(a)}$$

avec

$$\Lambda^*(a) = \sup_{\theta \in \mathbb{R}}\{\theta a - \Lambda(\theta)\}$$

où $\Lambda(\theta) = \log \rho\left(\Pi_\theta^{(h-1)}\right)$, $\rho\left(\Pi_\theta^{(h-1)}\right)$ désignant la valeur propre de Perron-Frobénius *de $\Pi_\theta^{(h-1)}$.*

De plus, $\exists \tau \in \overline{\mathbb{R}}$ tel que

$$\Lambda^*(a) = \lim_{\theta \to \tau} \theta a - \Lambda(\theta)$$

et, si $\tau > 0$ (resp. $\tau < 0$) alors W est sur-représenté (resp. sous-représenté).

Preuve. Ce résultat s'obtient simplement en appliquant la proposition 6.1 à la chaîne de *Markov* d'ordre 1 sur l'alphabet fini $\mathcal{A}^{(h-1)}$ pour le mot de 2 lettres $W^{(h-1)}$. ∎

Irréductibilité

Pour pouvoir appliquer ce résultat, il convient cependant de déterminer sous quelles conditions on obtient l'irréductibilité de $\Pi^{(h-1)}$.

On commence par effectuer la remarque suivante :

Proposition 6.3 *Si $\mathbb{P}(X_h = y_h | X_1 = y_1, \ldots X_{h-1} = y_{h-1}) > 0$ pour tout $y_1, \ldots, y_h \in \mathcal{A}$ alors $\Pi^{(h-1)}$ à la puissance $h-1$ ne possède que des termes strictement positifs (et donc en particulier $\Pi^{(h-1)}$ est irréductible).*

Preuve. Soient $y = [y_1, \ldots, y_{h-1}]$ et $z = [z_1, \ldots, z_{h-1}]$ deux éléments de $\mathcal{A}^{(h-1)}$. On considère le produit suivant :

$$\begin{aligned} P = \mathbb{P}(z_1|y_2, \ldots, y_h) \times \mathbb{P}(z_2|y_3, \ldots, y_h, z_1) \times \ldots \\ \ldots \times \mathbb{P}(z_{h-1}|y_h, z_1 \ldots, z_{h-2}) \times \mathbb{P}(z_h|z_1 \ldots, z_{h-1}). \end{aligned}$$

$P > 0$ comme produit de $h - 1$ termes strictement positif et P décrit précisément un chemin reliant y à z ce qui achève la preuve. ■

Ce résultat nous permet alors d'établir le

Corollaire 6.4 *Si les paramètres de $\Pi^{(h-1)}$ correspondent à ceux obtenus par l'estimateur du maximum de vraisemblance à partir d'une séquence observée x pour un modèle markovien d'ordre m alors*

$$\Pi^{(h-1)} \textit{ irréductible} \iff \forall y_1, \ldots, y_{m+1} \in \mathcal{A} \quad N(y_1, \ldots, y_{m+1}) > 0. \quad (6.2)$$

Preuve. On se donne une séquence observée $x = x_1 \ldots x_n$ à partir de laquelle on estime les paramètres de $\Pi^{(h-1)}$ en fonction de l'ordre m du modèle choisi. On reprend ici les résultats de la section 2.3 (page 36) pour trouver que :

$$\widehat{\mathbb{P}}(X_h = y_h|X_1 = y_1, \ldots X_{h-1} = y_{h-1}) = \frac{N(y_{h-m} \ldots y_{h-1} y_h)}{\sum_{z \in \mathcal{A}} N(y_{h-m} \ldots y_{h-1}, z)}$$

dès que le dénominateur est non nul (sans quoi on prend 0 comme estimateur).

En utilisant la proposition 6.3 et l'équation (6.2), on montre ainsi l'irréductibilité de $\Pi^{(h-1)}$ dès que les comptages des mots de longueur $m+1$ sont tous non nuls.

Pour la réciproque, on considère $y = y_1 \ldots y_{m+1}$ tel que $N(y) = 0$. L'équation (6.2) nous permet alors d'affirmer que

$$\mathbb{P}(y_h|z_1, \ldots, z_{h-m-1}, y_{h-m}, \ldots, y_{h-1}) = 0 \quad \forall z_1, \ldots, z_{h-m-1} \in \mathcal{A}.$$

si bien que le mot $z_2 \ldots z_{h-m-1} y_{h-m} \ldots h_m$ ne peut jamais être atteint (les seules transitions possibles provenant d'un mot de la forme $z_1 \ldots z_{h-m-1} y_{h-m}$) ce qui prouve bien la non-irréductibilité de la matrice. ■

Si cette condition n'est pas très astreignante lorsque m est faible ($m = 0$ ou $m = 1$ par exemple), elle peut rapidement être mise en défaut dès que la valeur de m augmente ce qui pose un grave problème.

Fort heureusement on peut user d'une astuce pour traiter les cas où certains comptages de mots de longueur $m+1$ sont nuls. Elle consiste à supprimer de l'alphabet $\mathcal{A}^{(h-1)}$ toutes les "lettres" contenant l'un de ces sous-mots ; on obtient ainsi un nouvel alphabet $\mathcal{A}_{bis}^{(h-1)}$ et on peut définir $\Pi_{bis}^{(h-1)}$ la restriction de $\Pi^{(h-1)}$ à ce nouvel alphabet.

Proposition 6.5 *Si les paramètres de $\Pi^{(h-1)}$ correspondent à ceux obtenus par l'estimateur du maximum de vraisemblance à partir d'une séquence observée x pour un modèle markovien d'ordre m alors $\Pi_{bis}^{(h-1)}$ est irréductible.*

Preuve. On considère deux éléments distincts $[y_1 \dots y_{h-1}]$ et $[z_1 \dots z_{h-1}]$ de $\mathcal{A}_{bis}^{(h-1)}$ (c'est à dire qu'aucun de leurs sous-mots de longueur $m+1$ n'a de comptage nul dans $x = x_1 \dots x_1$) et on veut montrer qu'il existe une succession de transitions non nulles menant du premier élément au second.

Les deux mots $y_{h-m} \dots y_h$ et $z_1 \dots z_{m+1}$ sont de longueur $m+1$ et sont donc, à ce titre, présents au moins une fois dans la séquence x. Soit i_1 (resp. i_2) une position où commence le premier mot (resp. le second).

On suppose sans perte de généralité que $x_{n+1} = x_n$ ce qui nous permet de déplacer l'origine de la numérotation des lettres de x si cela s'avère nécessaire.

Si $i_1 = i_2$ c'est que les deux mots sont identiques et il n'y a rien à montrer. Grâce à la structure circulaire de x évoquée ci-avant, on peut supposer que $i_1 < i_2$ (dans le cas inverse on peut en effet renuméroter x à partir de i_2 pour se ramener au premier cas).

Dans le cas où $i_2 - i_1 \leq m$, cela signifie que la fin du premier mot chevauche la seconde dans la séquence ; la transition de l'un à l'autre est évidente.

Si les deux mots ne se chevauchent pas, on se trouve dans la situation décrite par ce schéma :

$$\begin{array}{cccccccccccccc}
 & y_{h-m-1} & y_{h-m} & \cdots & y_{h-1} & & & & z_1 & \cdots & z_m & z_{m+1} & \\
\cdots & x_{i_1} & x_{i_1+1} & \cdots & x_{i_1+m} & x_{i_1+m+1} & \cdots & x_{i_2-1} & x_{i_2} & \cdots & x_{i_2+m-1} & x_{i_2+m} & \cdots
\end{array}$$

il convient alors d'utiliser les transitions présentent dans la séquence. En effet, ces transitions sont définies par

$$\begin{aligned}
P = \ & \mathbb{P}(y_{h-1} | y_1, \dots, y_{h-2}) \times \mathbb{P}(x_{i_1+m+1} | y_2, \dots, y_{h-1}) \dots \\
& \dots \times \mathbb{P}(z_m | x_{i_2-}, \dots, x_{i_2-1}, z_1 \dots, z_{m-1}) \times \mathbb{P}(z_{m+1} | x_{i_2-}, \dots, x_{i_2-1}, z_1 \dots, z_m).
\end{aligned}$$

et il clair que P est non nul puisque le produit le définissant fait intervenir des transitions non nulles dès que les comptages des mots de longueur $m+1$ terminant chacune de ces transitions sont eux-même non nuls. Or, ces mots sont présents dans la séquence x par construction ce qui achève la preuve. ■

6.2 Un motif

Comme l'a vu en section 1.2 (page 21), il existe des familles de mots participant conjointement à un même phénomène biologique. Dans le cas du *chi* d'*Hemophilus influenzæ* par exemple, les quatre mots du *motif* `g.tggtgg` (le symbole "`.`" en deuxième position du motif signifie que l'on peut le remplacer par n'importe quelle lettre de l'alphabet) ont une activité biologique.

Le traitement séparé des différents mots composant un motif peut donner des résultat intéressants, mais l'étude simultanée de l'ensemble de ces mots devrait être beaucoup plus riche d'enseignements.

L'objet de cette partie est de comprendre comment adapter les méthodes évoquées dans les sections précédentes afin de pouvoir examiner les nombres d'occurrences des motifs.

6.2.1 Définitions

On commence par introduire quelques définitions :

Définition 6.6 (Motif)
On appelle motif *une famille finie de mot de la forme :*

$$\mathcal{W} = \{W_1, \dots, W_r\}$$

et si, pour $i \in \{1, \dots, r\}$, on note h_i la longueur du mot W_i alors la longueur h du motif $\mathcal{W}$ est définie par

$$h = \max_{i \in \{1,\dots,r\}} h_i.$$

On appelle motif complété *d'un motif $\mathcal{W}$, et on note $\overline{\mathcal{W}}$, le motif qui contient exactement les mots de longueur h obtenu en ajoutant aux mots de longueur inférieure à h du motif initial tous les suffixes possibles pour en faire des mots de longueur h.*

Voici un exemple de motif de longueur 4 dans l'alphabet $\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$:

$$\mathcal{W} = \{\mathtt{caat}, \mathtt{gta}, \mathtt{gtat}\} \tag{6.3}$$

et voici son motif complété :

$$\overline{\mathcal{W}} = \{\mathtt{caat}, \mathtt{gtaa}, \mathtt{gtac}, \mathtt{gtag}, \mathtt{gtat}\} \tag{6.4}$$

Définition 6.7 (Comptage pondéré)
Soit $\mathcal{W} = \{W_1, \dots, W_r\}$ un motif et $p = \{p_1, \dots, p_r\} \in \mathbb{R}^r$ un ensemble de pondérations, on désigne alors les comptages pondérés *du motif (repectivement pour la variable aléatoire et pour l'observation) par :*

$$N_p(\mathcal{W}) = \sum_{i=1}^{r} p_i N(W_i) \text{ et } n_p(\mathcal{W}) = \sum_{i=1}^{r} p_i n(W_i)$$

6.2.2 Résultats

Proposition 6.8 *Si $\mathcal{W}$ est un motif de taille h donné, alors, si $\overline{\mathcal{W}} = \{W_1, \ldots, W_r\}$ est son motif complété et que $p = \{p_1, \ldots, p_r\} \in \mathbb{R}^r$ est une pondération, on peut calculer une approximation de*

$$\mathbb{P}\big(N_p(\overline{\mathcal{W}}) = n_p(\overline{\mathcal{W}})\big)$$

comme dans le théorème 6.2 en posant $\forall x, y \in \mathcal{A}^{h-1}$

$$f(y,z) = \sum_{i=1}^{r} p_i \times \mathbb{I}_{y=\left[w_1^i,\ldots,w_{h-1}^i\right]} \times \mathbb{I}_{z=\left[w_2^i,\ldots,w_h^i\right]}$$

où $\forall i \in \{1,\ldots,r\}$, $W_i = w_1^i \ldots w_h^i$.

Toute la difficulté consiste à choisir les valeurs des pondérations en fonction de ce que l'on souhaite observer.

On considère le motif $\mathcal{W}$ défini en (6.3). Son motif complété est défini en (6.4) et si on se donne la pondération

$\overline{\mathcal{W}}$	`caat`	`gtaa`	`gtac`	`gtag`	`gtat`
p	1	2	1	1	1

on définit alors f par $f(y,z) = 0$ pour tous $y, z \in \mathcal{A}^3$ sauf

$$\begin{cases} f([\texttt{cca}],[\texttt{cat}]) = 1 \\ f([\texttt{gta}],[\texttt{taa}]) = 2 \\ f([\texttt{gta}],[\texttt{tac}]) = 1 \\ f([\texttt{gta}],[\texttt{tag}]) = 1 \\ f([\texttt{gta}],[\texttt{tat}]) = 1 \end{cases}$$

et, comme

$$N(\texttt{gta}) \sim N(\texttt{gtaa}) + N(\texttt{gtac}) + N(\texttt{gtag}) + N(\texttt{gtat}).$$

on obtient

$$N_p(\overline{\mathcal{W}}) \sim N(\texttt{caat}) + N(\texttt{gta}) + N(\texttt{gtaa})$$

ce qui correspond à la somme des occurences des différents mots de $\mathcal{W}$.

Si, en revanche, on choisit au départ la pondération

$\overline{\mathcal{W}}$	`caat`	`gtaa`	`gtac`	`gtag`	`gtat`
p	1	1	1	1	1

alors cette fois-ci

$$N_p(\overline{\mathcal{W}}) \sim N(\texttt{caat}) + N(\texttt{gta})$$

ce qui constitue le nombre de positions où au moins un élément du motif $\mathcal{W}$ apparaît.

On le voit bien ici, on dispose d'une grande souplesse dans les comptages traitables du fait de la liberté absolue dans le choix des pondérations p. Cependant, dans tous les cas, on ne considère qu'une information résumée concernant les occurrences du motif. En effet, les événements considérés sont de la forme

$$\{N(W_1) + \ldots + N(W_r) = n(W_1) + \ldots + n(W_r)\}$$

alors qu'il serait plus naturel de pouvoir s'intéresser aux événements de la forme

$$\{N(W_1) = n(W_1), \ldots, N(W_r) = n(W_r)\}\,.$$

On touche là aux limites de l'approche du problème par les grandes déviations de niveau 1 et on se reportera à la partie III (page 107) pour comprendre comment l'on peut traiter ce dernier type d'événements.

6.3 Changement de probabilité

Comme dans le cas indépendant (voir section 4.4 page 72), il est possible d'isoler le changement de probabilité mis en œuvre dans la démonstration du théorème de *Cramér-Chernov.*

On rappelle que, $\forall \theta \in \mathbb{R}$, $\widetilde{\Pi}_\theta$ définie par

$$\widetilde{\Pi}_\theta(x,y) = e^{-\Lambda(\theta)} \frac{v_\theta(y)}{v_\theta(x)} \Pi(x,y) e^{\theta f(x,y)} \tag{6.5}$$

est stochastique (voir proposition 5.2 page 85).

Proposition 6.9 *Si $(X_i)_i$ est une chaîne de* Markov *(supposée circulaire ; $X_{n+1} = X_1$) de matrice de transition Π alors la probabilité d'un événement A quelconque peut s'écrire :*

$$\mathbb{P}(A) = \rho(\Pi_\theta)^n \mathbb{E}\left[\mathbb{I}_A \times e^{-\theta \widehat{S}_n}\right]$$

où $\widehat{S}_n = \sum_{i=1}^n f(\widehat{X}_i, \widehat{X}_n)$ avec $(\widehat{X}_i)_i$ chaîne de Markov *de transition $\widehat{\Pi} = \widetilde{\Pi}_\theta$.*

Preuve. En effet on peut écrire

$$\mathbb{P}(A) = \sum_{x_1,\ldots,x_n \in \mathcal{A}} \mathbb{I}_A \times \Pi(x_1,x_2) \times \ldots \times \Pi(x_n,x_1)$$

donc, en utilisant (6.5), on a

$$\mathbb{P}(A) = \sum_{x_1,\ldots,x_n\in\mathcal{A}} \mathbb{I}_A \times \rho(\Pi_\theta)^n \times e^{-\theta s_n} \times \widetilde{\Pi}_\theta(x_1,x_2) \times \ldots \times \widetilde{\Pi}_\theta(x_n,x_1)$$

avec $s_n = f(x_1,x_2) + \ldots + f(x_n,x_{n+1})$, et la preuve de la proposition est achevée. ■

De tels changements de probabilités pourront révéler tout leur intérêt dans le cadre d'études d'événements par le biais de simulations. En effet, les approximations obtenues par de telles approches seront d'autant meilleures que les probabilités des événements considérés seront grandes sous les lois utilisées pour simuler les séquences.

Or le théorème 6.2 montre qu'il existe τ tel que

$$\Lambda^*(a) = \tau a - \Lambda(\tau)$$

et comme on a également $\Lambda'(\tau) = a$, la proposition 5.3 (page 86) nous permet d'affirmer que

$$\mathbb{E}_{q_\tau}^{\widetilde{\Pi}_\tau}[f(X_1,X_2)] = a.$$

ce qui explicite la nature "centrale" de l'événement considéré sous la nouvelle loi $\widetilde{\Pi}_\tau$.

6.4 Algorithmes

On résume ici la marche à suivre pour traiter les occurrences d'un mot ou un motif donné de longueur h dans une séquence x.

6.4.1 Choix du modèle

Il faut commencer par choisir le modèle dans lequel on souhaite calculer la significativité du comptage du mot ou motif examiné. Pour cela, soit on se donne un modèle markovien d'ordre inférieur ou égal à $h-1$ (typiquement lorsque l'on utilise les paramètres issus d'une utilisation précédente de la méthode), soit on choisit l'ordre du modèle markovien dont les paramètres sont estimés sur la séquence x et on se limite dans ce cas aux modèles d'ordre au plus égal à $h-2$.

6.4.2 Construction de $\Pi^{(h-1)}$

On utilise les paramètres de la section précédente pour construire la matrice de transition $\Pi^{(h-1)}$ de la chaîne de *Markov* d'ordre 1 dans l'alphabet agrandi $\mathcal{A}^{h-1}$. Si la matrice résultante n'est pas irréductible, on supprime de l'alphabet $\mathcal{A}^{h-1}$ les "lettres" à l'origine de cette non irréductibilité. Avant de continuer, on s'assure que les mots ou motifs que l'on souhaite traiter ne font intervenir aucune de ces lettres. Si certaines de ces lettres apparaissent dans le mot ou motif en question, on sait qu'il n'apparaît pas dans la séquence, mais on ne peut pas dire à quel point cette absence est exceptionnelle.

6.4.3 Calcul de $\Lambda^*(a)$

On cherche à calculer le minimum de la fonction

$$\begin{array}{rrcl} f: & \mathbb{R} & \to & \mathbb{R} \\ & \theta & \mapsto & \Lambda(\theta) - \theta a \end{array}.$$

Chaque évaluation de f nécessite le calcul de $\rho(\Pi_\theta^{(h-1)})$ la plus grande valeur propre de la matrice creuse $\Pi^{(h-1)}$ d'ordre k^{h-1}. On effectue ce calcul en utilisant la version modifiée de l'algorithme d'*Arnoldi* présentée en section C.1 (page 201).

Pour effectuer la minimisation proprement dite, on va utiliser la méthode de *Brent* (voir section C.2 page 205) qui présente l'avantage d'optimiser le nombre d'évaluations de la fonction f nécessaires.

6.4.4 Résultats

Une fois le calcul de $\Lambda^*(a)$ effectué, il ne reste plus qu'à calculer l'approximation recherchée et, le cas échéant, les paramètres $\widetilde{\Pi}_\theta$ précédemment évoqués, pour effectuer des simulations.

Troisième partie

Grandes déviations de niveau 2

FIG. 7.1 – Représentation graphique des mots de longueur $h = 9$ chez *Caenorhabditis elegans* dans le modèle $M1$. On consultera la section 11.1 page 160 pour plus de détails.

Chapitre 7

Théorie

Pré-requis :

Chapitre 5

Description :

Résultats théoriques des grandes déviations de niveau 2 du nombre d'occurrences de mots sur une chaîne de *Markov*.

Résumé :

Il est possible d'utiliser directement les résultats de grandes déviations de niveau 1 pour établir un PGD pour la distribution empirique des paires de lettres.

On peut retrouver ce résultat de manière plus directe en utilisant le lemme de *Varadhan* qui constitue une généralisation de la technique de changement de probabilité.

Il est possible d'étendre ces résultats à la distribution empirique des mots de h lettres.

Contenu du chapitre

7.1 Application du théorème de *Gärtner-Ellis*

Dans toute cette partie on va s'intéresser aux lois empiriques des lettres ou des paires de lettres. On reprend pour cela les notions et notations de la section 4.5 (page 75).

7.1.1 Les singletons

On suppose que $(X_i)_i$ est une chaîne de *Markov* d'ordre 1 sur l'alphabet $\mathcal{A}$ de cardinal k et de matrice de transition Π irréductible. On rappelle la définition de la distribution empirique des lettres $L_n = (L_n(1), \dots, L_n(k)) \in \mathcal{M}_1(\mathcal{A})$ avec

$$L_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{X_i=x} \quad \forall x \in \mathcal{A}.$$

et on a alors le

Théorème 7.1 *La suite $(L_n)_n$ de v.a. à valeurs dans $\mathcal{M}_1(\mathcal{A})$, suit un principe de grandes déviations de bonne fonction de taux*

$$I(x) = \sup_{\theta \in \mathbb{R}^k} \left\{ \langle \theta, x \rangle - \Lambda(\theta) \right\}$$

où Λ est définie comme au chapitre 5 (page 83).

Preuve. Pour obtenir la preuve, on va se contenter d'appliquer le résultat de la section 5.3 (page 88) avec une fonction f particulière :

$$\begin{array}{rccl} f: & \mathcal{A} \times \mathcal{A} & \to & \mathbb{R}^k \\ & (x, y) & \mapsto & f(x, y) = f(x) = (\mathbb{I}_{x=1}, \dots, \mathbb{I}_{x=k}) \end{array}$$

et il est alors clair que

$$\begin{aligned} L_n &= \frac{1}{n}\sum_{i=1}^{n} f(X_i, X_{i+1}) \\ &= \frac{1}{n}\sum_{i=1}^{n} f(X_i) \end{aligned}$$

ce qui achève la démonstration. ■

Ce résultat est tout à fait valable, mais la forme de la fonction de taux proposée n'est pas très utilisable d'un point de vue numérique. En effet, une simple évaluation de I va requérir la maximisation dans un espace à d dimensions d'une fonction dont chaque évaluation nécessite le calcul de la plus grande valeur propre d'une matrice d'ordre d. De plus, par le mécanisme d'augmentation d'alphabet introduit en section 6 (page 91), d va en fait prendre des valeurs de type k^{h-1} où h est la taille du motif considéré, ce qui va augmenter de manière très sensible la complexité des algorithmes mis en jeu.

Aussi, avant d'aller plus loin, on va s'efforcer ici de trouver une forme plus simple à la fonction de taux, ce qui est l'objet de la

Proposition 7.2 $\forall \nu \in \mathbb{R}^k$ *on a*

$$I(\nu) = \begin{cases} \sup_{u \gg 0} \left\{ \sum_{x\in\mathcal{A}} \nu(x) \log \frac{u(x)}{(u\Pi)(x)} \right\} & si\ \nu \in \mathcal{M}_1(\mathcal{A}) \\ +\infty & sinon \end{cases}$$

Preuve. On trouvera la preuve complète de la proposition en annexe (voir section D.3.2 page 232) dont voici un résumé : on commence par montrer que I est infinie en dehors de $\mathcal{M}_1(\mathcal{A})$ en utilisant le principe de grandes déviations établi en théorème 7.1, puis on montre successivement les deux inégalités du résultat en ce qui concerne les $\nu \in \mathcal{M}_1(\mathcal{A})$. ■

7.1.2 Les paires

On se place dans les mêmes hypothèses que précédemment et on va utiliser le résultat précédent pour en établir un nouveau concernant cette fois-ci la distribution empirique des paires de lettres ou des mots de deux lettres.

On se place dans $\mathcal{A}^2 = \mathcal{A} \times \mathcal{A}$ et on s'intéresse aux déviations de $L_{n,2} \in \mathcal{M}_1(\mathcal{A}^2)$ avec

$$L_{n,2}(x,y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{X_i=x} \times \mathbb{I}_{X_{i+1}=y}.$$

On introduit la notion suivante

Définition 7.3 (mesure invariante par translation)
Si $\nu \in \mathcal{M}_1(\mathcal{A}^d$ *avec* $d \geqslant 2$ *alors on dit que* ν *est une* mesure invariante par translation *si,* $\forall x_1, \ldots, x_{d-1} \in \mathcal{A}$*, on a*

$$\overline{\nu}(x_1, \ldots, x_{d-1}) = \sum_{y \in \mathcal{A}} \nu(x_1, \ldots, x_{d-1}, y) = \sum_{y \in \mathcal{A}} \nu(y, x_1, \ldots, x_{d-1})$$

et on note $\mathcal{S}\left(\mathcal{A}^d\right) \subset \mathbb{R}^d$ *l'ensemble des mesures invariantes par translation. Notons que cette notion signifie simplement qu'il y a autant de mot de longueur* d *possédant comme préfixe ou comme suffixe un même mot de longueur* $d-1$.

On peut alors établir le

Théorème 7.4 *La suite* $(L_{n,2})_n$ *de v.a. à valeurs dans* $\mathcal{M}_1(\mathcal{A})$*, suit un principe de grandes déviations de bonne fonction de taux*

$$I_2(\nu) = \begin{cases} \sum_{x,y \in \mathcal{A}} h_\nu(x,y) & \text{si } \nu \in \mathcal{S} \\ +\infty & \text{sinon} \end{cases}$$

avec

$$h_\nu(x,y) = \begin{cases} 0 \quad \text{si } \overline{\nu}(x) = 0 \text{ ou } \Pi(x,y) = 0 \\ \nu(x,y) \log \dfrac{\nu(x,y)}{\overline{\nu}(x)\Pi(x,y)} \quad \text{sinon} \end{cases}$$

(avec la convention $0 \log 0 = 0$*).*

Preuve. Pour démontrer ce résultat, on va tout d'abord appliquer le théorème 7.1 à la chaîne de *Markov* d'ordre 1 dans l'alphabet $\mathcal{A}^2$ de matrice de transition $\Pi^{(2)}$ (technique d'augmentation de la taille de l'alphabet comme on l'a déjà fait au chapitre 6 page 91). La suite de la preuve est assez calculatoire, on se reportera à la section D.3.2 (page 232) pour en obtenir une version détaillée. ■

Remarque 7.5 *si on note* $\overline{\nu} \otimes \Pi$ *la loi sur* $\mathcal{A}^2$ *définie* $\forall x, y \in \mathcal{A}$ *par* $(\overline{\nu} \otimes \Pi)(x,y) = \overline{\nu}(x)\Pi(x,y)$*, notons alors que pour* $\nu \in \mathcal{S}$ *on a*

$$I_2(\nu) = H(\nu | \overline{\nu} \otimes \Pi)$$

résultat à mettre en relation avec celui de la remarque 4.15 (page 79).

7.2 Application du lemme de *Varadhan*

On peut aussi obtenir directement le résultat du théorème 7.4 à partir du théorème concernant les déviations de la distribution empirique des paires dans le cas i.i.d (voir théorème 4.14 page 78) en utilisant le théorème de *Varadhan* (section A.5 page 185).

On commence par comparer les lois $\mathbb{P}_n$ $\widehat{\mathbb{P}}_n$ de la séquence $X = (X_1, \ldots, X_n)$ respectivement dans les modèles $M0$ (de loi μ) et $M1$ (de matrice de transition Π), et on obtient ainsi la

Proposition 7.6 (Formule de *Radon-Nikodym*)
On a

$$\frac{d\widehat{\mathbb{P}}_n}{d\mathbb{P}_n}(x_1, \ldots, x_n) = e^{nF(L_{n,2}[x_1,\ldots,x_n])}$$

avec $\forall \nu \in \mathcal{M}_1(\mathcal{A}^2)$,

$$F(\nu) = \sum_{i,j \in \mathcal{A}} \nu(i,j) \log \frac{\Pi(i,j)}{\mu(j)}.$$

Preuve. On a

$$\begin{aligned}
\widehat{\mathbb{P}}_n(x_1, \ldots, x_n) &= \Pi(x_1, x_2) \times \ldots \times \Pi(x_{n-1}, x_n) \times \Pi(x_n, x_1) \\
&= \exp\left(\sum_{i=1}^{n} \log \Pi(x_i, x_{i+1})\right) \\
&= \exp\left(n \times \sum_{i,j \in \mathcal{A}} L_{n,2}[x_1, \ldots, x_n](i,j) \log \Pi(i,j)\right)
\end{aligned}$$

et

$$\begin{aligned}
\mathbb{P}_n(x_1, \ldots, x_n) &= \mu(x_1) \times \ldots \times \mu(x_n) \\
&= \exp\left(\sum_{i=1}^{n} \log \mu(x_i)\right) \\
&= \exp\left(n \times \sum_{i,j \in \mathcal{A}} L_{n,2}[x_1, \ldots, x_n](i,j) \log \mu(i)\right)
\end{aligned}$$

si bien que le résultat est immédiat. ■

La fonction F est continue et le théorème 4.14 (page 78) montre que la suite $(\mathbb{P}_n)$ satisfait un PGD de fonction de taux I définie par

$$I(\nu) = \begin{cases} \sum_{i,j\in\mathcal{A}} \nu(i,j) \log \dfrac{\nu(i,j)}{\overline{\nu}(i)\mu(j)} & \text{si } \nu \in \mathcal{S} \\ +\infty & \text{sinon} \end{cases}$$

si bien que le théorème A.12 (page 186) nous permet d'affimer que la suite $(\widehat{\mathbb{P}}_n)$ satisfait un PGD de fonction de taux

$$\widehat{I}(\nu) = \sup_{\nu'\in\mathcal{M}_1(\mathcal{A}^2)} [F(\nu') - I(\nu')] - [F(\nu) - I(\nu)]$$

or il se trouve que

$$I(\nu) - F(\nu) = \begin{cases} H(\nu|\overline{\nu}\otimes\Pi) & \text{si } \nu \in \mathcal{S} \\ +\infty & \text{sinon} \end{cases}$$

et on obtient donc le même résultat qu'au théorème 7.4.

En effet, une entropie étant positive, on a

$$\sup_{\nu'\in\mathcal{M}_1(\mathcal{A}^2)} [F(\nu') - I(\nu')] \leqslant 0$$

de plus, si μ désigne la loi stationnaire de la chaîne de *Markov*, on a ν' définie par

$$\nu'(i,j) = \mu(i)\Pi(i,j) \quad \forall i,j \in \mathcal{A}$$

qui vérifie $\overline{\nu}' = \mu$ si bien que $\overline{\nu}' \otimes \Pi = \nu$ et donc $F(\nu') - I(\nu') = 0$ ce qui montre que le sup est nul et achève la démonstration.

7.3 Cas général

On s'intéresse ici à la distribution empirique des mots de h lettres

$$L_{n,h}(x_1,\ldots,x_h) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{X_i=x_1} \times \ldots \times \mathbb{I}_{X_{i+h-1}=x_h}$$

sur une chaîne de *Markov* d'ordre m avec $m \leqslant h-1$.

On a alors le

Théorème 7.7 *Si, $\forall x_1,\ldots,x_h \in \mathcal{A}$, on note*

$$\Pi(x_1,\ldots,x_h) = \mathbb{P}(X_h = x_h | X_1 = x_1, \ldots, X_{h-1} = x_{h-1})$$

alors la suite $(L_{n,h})_n$ de v.a. à valeurs dans $\mathcal{M}_1(\mathcal{A}^h)$ suit un principe de grandes déviations de bonne fonction de taux

$$I_h(\nu) = \begin{cases} \sum_{x_1,\dots,x_h \in \mathcal{A}} h_\nu(x_1,\dots,x_h) & \textit{si } \nu \in \mathcal{S} \\ +\infty & \textit{sinon} \end{cases}$$

avec

$$h_\nu(x_1,\dots,x_h) = \begin{cases} 0 \quad \textit{si } \overline{\nu}(x_1,\dots,x_{h-1}) = 0 \textit{ ou } \Pi(x_1,\dots,x_h) = 0 \\ \nu(x_1,\dots,x_h) \log \dfrac{\nu(x_1,\dots,x_h)}{\overline{\nu}(x_1,\dots,x_{h-1})\Pi(x_1,\dots,x_h)} & \textit{sinon} \end{cases}$$

(par convention $0 \log 0 = 0$*).*

Preuve. Pour montrer ce résultat, on va simplement utiliser la technique d'augmentation d'alphabet déjà utilisée au chapitre 6 (page 91). On se ramène ainsi du problème concernant X, une chaîne de *Markov* d'ordre m de paramètres Π dans l'alphabet $\mathcal{A}$ à $X^{(h-1)}$, chaîne de *Markov* d'ordre 1 dans l'alphabet $\mathcal{A}^{(h-1)}$ et de matrice de transition $\Pi^{(h-1)}$ définie par

$$\Pi^{(h-1)}\big([x_1 \dots x_{h-1}], [y_1 \dots y_{h-1}]\big) = \begin{cases} 0 \quad \text{si } [x_2 \dots x_{h-1}] \neq [y_1 \dots y_{h-2}] \\ \Pi(x_1,\dots,x_{h-1},y_h) \quad \text{sinon} \end{cases}$$

Le théorème 7.4 nous assure de l'existence d'un principe de grandes déviations pour la suite $(L_{n,h})$ dont la fontion de taux est définie pour tout $\nu^{(h-1)} \in \mathcal{S}\left(\mathcal{A}^{(h-1)} \times \mathcal{A}^{(h-1)}\right)$ par

$$I_h(\nu) = \begin{cases} \sum_{x,y \in \mathcal{A}^{(h-1)}} h_\nu(x,y) & \text{si } \nu \in \mathcal{S}\left(\mathcal{A}^{(h-1)} \times \mathcal{A}^{(h-1)}\right) \\ +\infty & \text{sinon} \end{cases}$$

où $h_\nu(x,y)$ est nul si $\Pi^{(h-1)}(x,y) = 0$ ou si $\overline{\nu}^{(h-1)}(x) = 0$ et

$$h_\nu(x,y) = \nu^{(h-1)}(x,y) \log \frac{\nu^{(h-1)}(x,y)}{\overline{\nu}^{(h-1)}(x)\Pi^{(h-1)}(x,y)}$$

sinon.

Or comme $\Pi^{(h-1)}(x,y) = 0$ si $[x_2 \dots x_{h-1}] \neq [y_1 \dots y_{h-2}]$, les valeurs de $\nu^{(h-1)}$ pour les couples $(x,y) \in \mathcal{A}^{(h-1)} \times \mathcal{A}^{(h-1)}$ ne vérifiant pas cette propriété n'interviennent pas dans l'expression de la fonction de taux. On peut ainsi considérer I_h comme une fonction de $\mathcal{M}_1(\mathcal{A}^h)$ dont on montre aisément qu'elle s'écrit ainsi que le théorème l'affirme, ce qui achève la démonstration. ∎

Remarque 7.8 *De même qu'en remarque 7.5, on constate que la fonction de taux prend la forme d'une entropie relative : si $\nu \in \mathcal{S}$ alors*

$$I_h(\nu) = H(\nu | \overline{\nu} \otimes \Pi)$$

ou l'on note $\overline{\nu} \otimes \Pi$ *la loi sur* $\mathcal{A}^h$ *définie par*

$$(\overline{\nu} \otimes \Pi)(x_1, \ldots, x_h) = \overline{\nu}(x_1, \ldots, x_{h-1})\Pi(x_1, \ldots, x_h)$$

pour tout $x_1, \ldots, x_h \in \mathcal{A}$.

Références

[Buc90] J. A. Bucklew. *Large deviation techniques in decision, simulation, and estimation.* Wiley, 1990.

[dH00] F. den Hollender. *Large Deviations.* American Mathematical Society, 2000.

[DZ98] A. Dembo and O. Zeitouni. *Large deviations techniques and applications.* Springer-Verlag, 1998.

FIG. 8.1 – Représentation graphique des mots de longueur $h = 9$ chez *Lactococcus lactis* dans le modèle $M1$. On consultera la section 11.1 page 160 pour plus de détails.

Chapitre 8

Mise en oeuvre pratique

Pré-requis :

Chapitre 7

Description :

Utilisation pratique des résultats théoriques de niveau 2 pour le calcul de la significativité de comptages de mots et de motifs.

Résumé :

Pour effectuer le calcul pratique des quantités intervenant dans les résultats théoriques du niveau 2, il faut minimiser sous des contraintes dépendant de l'événement considéré, une fonction multivariée dans un espace à k^h dimensions.

En compliquant la fonction à minimiser, on se ramène à un problème sans contrainte dont on propose de déterminer la solution par des méthodes de descente du gradient dont les valeurs en tous points sont, par conséquent, calculées.

Le calcul des paramètres de la projection orthogonale intervenant dans la fonction résultante requérant rapidement un espace mémoire trop important, une méthode de calcul heuristique négligeant cette projection est enfin évoquée.

Contenu du chapitre

8.1 Méthode

Le théorème 7.7 (page 112) nous assure de l'existence d'un principe de grandes déviations pour la suite $(L_{n,h})_n$ c'est-à-dire que si $\Gamma \subset \mathcal{M}_1(\mathcal{A}^h)$ alors

$$-\inf_{\overset{\circ}{\Gamma}} I_h \leqslant \underline{\lim}_{n \to +\infty} \frac{1}{n} \log \mathbb{P}\left(L_{n,h} \in \Gamma\right) \leqslant \overline{\lim}_{n \to +\infty} \frac{1}{n} \log \mathbb{P}\left(L_{n,h} \in \Gamma\right) \leqslant -\inf_{\overline{\Gamma}} I_h$$

de sorte qu'on peut considérer l'approximation suivante à distance finie :

$$\mathbb{P}\left(L_{n,h} \in \Gamma\right) \sim \exp(-n \inf_{\Gamma} I_h).$$

Toute la difficulté se résume donc aux deux points suivants :
- choisir Γ en fonction de l'événement que l'on souhaite étudier ;
- calculer le minimum de I_h sur Γ.

8.1.1 Traduction de l'événement

On va s'intéresser ici à une classe générale d'événements à laquelle on pourra toujours se rapporter par la suite, événements qui vont nous permettre de travailler sur les lois jointes de comptages pondérés.

Soit $\{\mathcal{W}_1, \ldots, \mathcal{W}_r\}$ un ensemble de familles de mots de longueurs h dont on suppose qu'elles vérifient $\mathcal{W}_i \cap \mathcal{W}_j = \emptyset$ si $i \neq j$. On munit chacune de ces familles d'une pondération $\mathcal{P}_i$ et on s'intéresse à l'événement

$$A = \left\{ N_{\mathcal{P}_1}(\mathcal{W}_1) = n_{\mathcal{P}_1}(\mathcal{W}_1), \ldots, N_{\mathcal{P}_r}(\mathcal{W}_r) = n_{\mathcal{P}_r}(\mathcal{W}_r) \right\}.$$

Si, pour tout élément $\nu \in \mathcal{M}_1(\mathcal{A})$ on note $\nu = [\nu(W)]_{W \in \mathcal{A}^h}$ alors on peut poser

$$\Gamma = \left\{ \nu \in \mathcal{M}_1(\mathcal{A}^h), \sum_{W \in \mathcal{W}_i} \nu(W) = \alpha_i \ \forall i, 1 \leqslant i \leqslant r \right\}$$

avec

$$\alpha_i = \frac{n_{\mathcal{P}_i}(W_i)}{n-h+1}$$

et on a ainsi

$$\mathbb{P}(1) = \mathbb{P}(L_{n,h} \in \Gamma).$$

On plonge $\mathcal{M}_1(\mathcal{A}^h)$ dans $(\mathbb{R}_+)^{k^h}$ et, par analogie avec les notations en vigueur dans $\mathcal{M}_1(\mathcal{A}^h)$, on adopte la notation $q = [q(W)]_{W\in\mathcal{A}^h}$ pour tout élément de $(\mathbb{R}_+)^{k^h}$. Ceci étant fait, on pose

$$\begin{array}{rccc} \gamma: & (\mathbb{R}_+)^{k^h} & \to & (\mathbb{R}_+)^{k^h} \\ & q & \mapsto & \gamma(q) \end{array}$$

avec, $\forall W \in \mathcal{A}^h$,

$$\gamma(q)(W) = \begin{cases} \alpha_i \frac{q(W)}{Q_i} & \text{si } \exists i \text{ tel que } W \in \mathcal{W}_i \\ \left(1 - \sum_{i=1}^r \alpha_i\right) \frac{q(W)}{Q} & \text{sinon} \end{cases}$$

où

$$\forall i,\ Q_i = \sum_{W\in\mathcal{W}_i} q(W) \quad \text{et} \quad Q = \sum_{W\notin\bigcup_{i=1}^r \mathcal{W}_i} q(W)$$

et on a

Proposition 8.1

$$\gamma\left([0,1]^{k^h}\right) = \Gamma$$

Preuve. En effet, il est clair que $\gamma(q) \in \Gamma$ pour tout $q \in (\mathbb{R}_+)^{k^h}$ car

$$\begin{aligned} \sum_{W\in\mathcal{A}^h} \gamma(q)(W) &= \sum_{i=1}^r \sum_{W\in\mathcal{W}_i} \gamma(q)(W) + \sum_{W\notin\bigcup_{i=1}^r \mathcal{W}_i} \gamma(q)(W) \\ &= \sum_{i=1}^r \frac{\alpha_i}{Q_i} \sum_{W\in\mathcal{W}_i} q(W) + \frac{1}{Q}\left(1 - \sum_{i=1}^r \alpha_i\right) \sum_{W\notin\bigcup_{i=1}^r \mathcal{W}_i} q(W) \\ &= \frac{Q_i}{Q_i} \sum_{i=1}^r \alpha_i + \frac{Q}{Q}\left(1 - \sum_{i=1}^r \alpha_i\right) \\ &= 1 \end{aligned}$$

(les autres conditions s'obtenant facilement par construction).

Réciproquement, si $\nu \in \Gamma \subset (\mathbb{R}_+)^{k^h}$, il est clair que $\gamma(\nu) = \nu$ car $Q = 1 - \sum_{i=1}^r \alpha_i$ et $Q_i = \alpha_i$ pour tout i. ■

Dans la suite, on utilisera cette proposition pour exprimer l'appartenance à Γ sous la forme de *contraintes* linéaires.

On présente enfin un cas particulier de l'événement général précédemment évoqué dans l'exemple suivant :

Exemple 8.2 (loi jointe simple)
Si on considère une famille simple $\{W_1, \ldots, W_r\}$ *de mots de longueur* h *et que l'on souhaite s'intéresser à l'événement*

$$A = \{N(W_1) = n(W_1), \ldots, N(W_r) = n(W_r)\}$$

il nous suffit de considérer les familles pondérées $(\mathcal{W}_1, \ldots, \mathcal{W}_r)$ *où* $\mathcal{W}_i = \{W_i\}$ *pour tout* i *et que l'on munit des pondérations* $\mathcal{P}_i = (1)$.

Ainsi on obtient

$$\Gamma = \{\nu \in \mathcal{M}_1(\mathcal{A}^h), \nu(W_i) = \alpha_i \ \forall i, 1 \leqslant i \leqslant r\}$$

avec, $\forall i$

$$\alpha_i = \frac{n(W_i)}{n - h + 1}$$

et on a, $\forall W \in \mathcal{A}^h$,

$$\gamma(q)(W) = \begin{cases} \alpha_i & \text{si } \exists i \text{ tel que } W = W_i \\ (1 - \sum_{i=1}^r \alpha_i) \frac{q(W)}{Q} & \text{sinon} \end{cases}$$

où

$$Q = \sum_{W \notin \{W_1, \ldots, W_r\}} q(W)$$

8.1.2 Minimisation de la fonction de taux

On désigne par $\mathcal{S}$l'ensemble des mesures de probabilités qui sont invariantes par translations *shift-invariantes* ;

$$\mathcal{S} = \left\{ \nu \in \mathcal{M}_1(\mathcal{A}^h) \text{ tq } \sum_{x \in \mathcal{A}} \nu(Wx) = \sum_{x \in \mathcal{A}} \nu(xW) \ \forall W \in \mathcal{A}^{h-1} \right\}$$

et comme I_h prend des valeurs infinies en dehors de $\mathcal{S}$ on est ramené au calcul de

$$\inf_{\nu \in \Gamma \cap \mathcal{S}} F(\nu) \tag{8.1}$$

avec

$$F(\nu) = \sum_{W \in \mathcal{A}^h} \nu(W) \log \frac{\nu(W)}{\sum_{k \in \mathcal{A}} \nu(W_- k) \Pi(W)}$$

où $W_- = w_1 \ldots w_{h-1}$ si $W = w_1 \ldots w_h$ est un mot de longueur h, où $\Pi(W) = \Pi(w_1, \ldots, w_h)$ et où on utilise la convention $0 \log 0 = 0$ et $x \log \frac{x}{0} = 0$ pour tout $x \in \mathbb{R}$.

Dans le calcul de (8.1), toute la difficulté provient de la contrainte $\nu \in \Gamma \cap \mathcal{S}$. Comme $\mathcal{S}$ est un sous-espace vectoriel de $\mathcal{M}_1(\mathcal{A})$ on peut considérer la projection orthogonale

$$P: \begin{array}{ccc} \mathcal{M}_1(\mathcal{A}) & \to & \mathcal{S} \\ \nu & \to & P(\nu) \end{array}$$

qui est décrite par une matrice d'ordre k^h dont les coefficients sont obtenus à partir des k^{h-1} contraintes linéaires définissant $\mathcal{S}$.

En utilisant cette projection ont peut alors réécrire (8.1) sous la forme

$$\inf_{\nu \in \Gamma} F \circ P(\nu) \tag{8.2}$$

et la proposition 8.1 permet alors enfin de ramener (8.2) à

$$\inf_{q \in [0,1]^{k^h}} F \circ P \circ \gamma(q) \tag{8.3}$$

La minimisation à effectuer dans le calcul de (8.3) étant dépourvue de contraintes, on peut utiliser des algorithmes classiques comme celui de la descente du gradient (voir C.2.5 page 208) pour déterminer la valeur du minimum.

8.2 Approche optimale

8.2.1 Dérivées partielles

Pour pouvoir mettre en oeuvre la méthode proposée, il faut être en mesure de calculer en tout point le gradient de $F \circ P \circ \gamma$ et c'est l'objet de cette section.

Soit $W_0 \in \mathcal{A}^h$ alors on a pour tout $q \in [0,1]^{k^h}$

$$\frac{\partial F \circ P \circ \gamma}{\partial q(W_0)}(q) = \sum_W \underbrace{\frac{\partial F}{\partial \nu(W)}[P \circ \gamma(q)]}_{(A)} \times \sum_{W'} P(W, W') \underbrace{\frac{\partial}{\partial q(W_0)}[\gamma(q)(W')]}_{(B)}$$

où $P = [P(W, W')]_{W, W' \in \mathcal{A}^h}$ désigne la matrice correspondant à la projection P.

Pour le calcul de (A) on commence par poser $R_\nu(W) = \sum_{x \in \mathcal{A}} \nu(Wx)$ pour tout $W \in \mathcal{A}^{h-1}$ et on peut alors écrire :

$$F(\nu) = \sum_{W' \in \mathcal{A}^h} \nu(W') \log \frac{\nu(W')}{R_\nu(W'_-)\Pi(W')}$$

(toujours avec les mêmes conventions) et on a

Proposition 8.3

$$\frac{\partial F}{\partial \nu(W)}(\nu) = \log \frac{\nu(W)}{R_\nu(W_-)\Pi(W)}$$

Preuve. Il est facile de voir que

$$\frac{\partial}{\partial \nu(W)}[R_\nu(W)] = \begin{cases} 1 & si \quad W'_- = W_- \\ 0 & si \quad W'_- \neq W_- \end{cases}$$

de sorte que

$$\begin{aligned}
\frac{\partial F}{\partial \nu(W)}(\nu) &= \sum_{y\in\mathcal{A}} \frac{\partial}{\partial \nu(W)} \left[\nu(W_-y) \log \frac{\nu(W_-y)}{R_\nu(W_-)\Pi(W_-y)}\right] \\
&= \sum_{y\neq w_h} -\nu(W_-y) \times \frac{R_\nu(W_-)\Pi(W_-y)}{\nu(W_-y)} \times \frac{\nu(W_-y)\Pi(W_-y)}{R_\nu(W_-)^2\Pi(W_-y)^2} \\
&\quad + \log \frac{\nu(W)}{R_\nu(W_-)\Pi(W)} \\
&\quad + \nu(W) \times \frac{R_\nu(W)\Pi(W)}{\nu(W)} \times \frac{R_\nu(W_-)\Pi(W) - \nu(W)\Pi(W)}{R_\nu(W_-)^2\Pi(W)^2} \\
&= \log \frac{\nu(W)}{R_\nu(W_-)\Pi(W)} + \underbrace{\frac{R_\nu(W_-) - \nu(W)}{R_\nu(W_-)} - \sum_{y\neq w_h} \frac{\nu(W_-y)}{R_\nu(W_-)}}_{0}
\end{aligned}$$

ce qui achève la preuve. ■

Pour le calcul de (B), on a la

Proposition 8.4 *On distingue trois cas dans lesquels on a les résultats suivants :*

(1) $W' = W_0$

$$\frac{\partial}{\partial q(W_0)}[\gamma(q)(W')] = \begin{cases} \alpha_i \dfrac{Q_i - q(W_0)}{Q_i^2} & si\ \exists i\ tq\ W_0 \in \mathcal{W}_i \\ \left(1 - \sum_i \alpha_i\right) \dfrac{Q - q(W_0)}{Q^2} & sinon \end{cases}$$

(2) $W' \neq W_0$ *et* $\exists i_0,\ W_0 \in \mathcal{W}_{i_0}$

$$\frac{\partial}{\partial q(W_0)}[\gamma(q)(W')] = \begin{cases} -\alpha_{i_0} \dfrac{q(W')}{Q_{i_0}^2} & si\ W' \in \mathcal{W}_{i_0} \\ 0 \quad sinon \end{cases}$$

(3) $W' \neq W_0$ *et* $\forall i,\ W_0 \notin \mathcal{W}_i$

$$\frac{\partial}{\partial q(W_0)}[\gamma(q)(W')] = \begin{cases} 0 & si\ W' \in \mathcal{W}_i \\ -(1-\sum_i \alpha_i)\dfrac{q(W')}{Q^2} & sinon \end{cases}$$

Preuve. Evident d'après la définition de γ. ∎

8.2.2 Algorithme

Si on suppose que Π et P sont connues l'algorithme permettant de calculer (8.3) se présente alors sous la forme suivante :

- on choisit q_0 un point quelconque de $[0,1]^{k^h}$;
- on calcule G_0, le gradient de $F \circ P \circ \gamma$ pris en q_0. Pour cela on va d'abord calculer les $Q_i(q_0) = Q_i^{q_0}$, $Q(q_0) = Q^{q_0}$ et $R(q_0) = R^{q_0}$, utiliser ces valeurs dans le calcul de G_0 et les conserver pour des calculs ultérieurs ;
- on calcule $\lambda_{\min}$ et $\lambda_{\max}$ de sorte que $\mathcal{D} = \{q_0 + \lambda G_0, \lambda \in [\lambda_{\min}, \lambda_{\max}]\}$ vérifie $\mathcal{D} \cap [0,1]^{k^h} = \mathcal{D}$;
- on minimise f sur $[\lambda_{\min}, \lambda_{\max}]$ où $f(\lambda) = (F \circ P \circ \gamma)(q_0 + \lambda G_0)$. Pour cela, on calcule $Q_i(G_0) = Q_i^{G_0}$, $Q(G_0) = Q^{G_0}$ et $R(G_0) = R^{G_0}$. On pourra dès lors utiliser ces valeurs et la linéarité des fonctions Q_i, Q et R pour effectuer plus rapidement les évaluations de f sur $\mathcal{D}$. Pour la minimisation, on est ramené à un problème à une dimension et on peut donc utiliser l'algorithme de *Brent* (voir section C.2 page 205). On note q_1 l'argument du minimum ainsi obtenu.
- On réitère l'étape précédente en remplaçant q_0 par q_1 et on obtient ainsi une suite $(q_i)_i$ d'approximations du minimum. On stoppe l'algorithme quand une certaine précision a été obtenue :

$$|F \circ P \circ \gamma(q_{i+1}) - F \circ P \circ \gamma(q_i)| < \varepsilon.$$

On préfère ici une condition d'arrêt portant sur la valeur du minimum plutôt que sur l'argument de ce minimum car c'est ce minimum qui va être utilisée dans l'approximation de la probabilité recherchée.

8.3 Heuristique

Si l'algorithme proposé est fonctionnel, la matrice de la projection P pose plusieurs problèmes numériques. Tout d'abord, sa taille. En effet cette matrice nécessite pour être stockée un espace mémoire en $O(k^{2h})$ ce qui représente le carré de l'espace par ailleurs nécessaire pour le reste de l'algorithme.

La quantité de mémoire disponible étant limitée, la capacité de traitement de l'algorithme sera grandement amoindrie par l'usage de cette projection. Par ailleurs, le calcul même de la matrice P à partir des contraintes linéaires définissant $\mathcal{S}$ présente des difficultés numériques.

Pour ces deux raisons, on propose une alternative à l'algorithme initial en se contentant du calcul de

$$\inf_{q \in \Gamma} F(q) = \inf_{q \in [0,1]^{k^h}} F \circ \gamma(q)$$

Dans cette approche, l'algorithme reste très similaire au précédent. Seules les formules des dérivées partielles se simplifient notablement :

$$\frac{\partial F \circ \gamma}{\partial q_{W_0}}(q) = \sum_W \frac{\partial F}{\partial \nu(W)}[\gamma(q)] \times \frac{\partial}{\partial q_{W_0}}[\gamma(q)_W]$$

en utilisant toujours les résultats intermédiaires des propositions 8.3 et 8.4.

Quatrième partie

Applications

FIG. 9.1 – Représentation graphique des mots de longueur $h = 9$ chez *Neisseria meningitidis* dans le modèle $M1$. On consultera la section 11.1 page 160 pour plus de détails.

Chapitre 9

Validation

Pré-requis :

Chapitre 3.

Description :

Validation des calculs effectués par les grandes déviations en comparant les résultats à ceux des méthodes existantes.

Résumé :

REGEXPCOUNT est un *package* MAPLE permettant d'effectuer des calculs exacts de significativité de comptages d'expressions régulières (en particulier de mots ou de motifs) par le biais des automates.

R'MES constitue l'implémentation des méthodes asymptotiques présentées au chapitre 3, tandis que GDon est le programme permettant le calcul des mêmes quantités par les grandes déviations (de niveau 1).

On utilise GDon pour examiner les mots exceptionnels d'une séquence tirée selon un modèle markovien donné, et on constate que la distribution des résultats est assez cohérente avec ce que l'on attendait.

Lorsque l'on s'intéresse à de véritable séquences biologiques, des comparaisons avec R'MES puis avec REGEXPCOUNT montrent la médiocre qualité des résultats de GDon dans le cas de mots peu exceptionnels mais aussi la grande précision de ces mêmes résultats dans le cas des mots très exceptionnels.

Contenu du chapitre

9.1 Outils et notations

On a vu au chapitre 3 (page 47) qu'il existait différentes approches du problème de la détection des mots et motifs exceptionnels. Les notations concernant les programmes REGEXPCOUNT et R'MES sont décrites dans la suite. Le programme DEMOS qui s'intéresse aux distances entre occurrences, en plus des simples nombres d'occurrences, est volontairement omis ici dans la mesure où il se trouve encore dans une phase de développement.

Dans tous les cas, on va utiliser ces programmes pour calculer les quantités

$$\mathcal{P}(W) = \begin{cases} \mathbb{P}\big(N(W) \geqslant n(W)\big) & \text{si le mot est sur-représenté} \\ \mathbb{P}\big(N(W) \leqslant n(W)\big) & \text{si le mot est sous-représenté} \\ \end{cases}$$

dans le cas d'un mot W et

$$\mathcal{P}_p(\mathcal{W}) = \begin{cases} \mathbb{P}\big(N_p(\mathcal{W}) \geqslant n_p(\mathcal{W})\big) & \text{si le motif est sur-représenté} \\ \mathbb{P}\big(N_p(\mathcal{W}) \leqslant n_p(\mathcal{W})\big) & \text{si le motif est sous-représenté} \end{cases}$$

dans le cas d'une famille pondérée $(\mathcal{W}, p)$ (voir la section 6.2.1 page 100 pour les notations concernant les familles pondérées).

Afin de faciliter la lecture de ces probabilités, qui peuvent parfois être très petites, on va adopter la représentation gaussienne déjà évoquée en section 4.2 (page 66) : on utilise l'inverse de la fonction de répartition (la fonction quantile) d'une variable aléatoire gaussienne centrée réduite pour calculer les quantités $\mathcal{X}(W)$ et $\mathcal{X}_p(\mathcal{W})$ telles que

$$\mathbb{P}\big(\mathcal{N}(0,1) \leqslant \mathcal{X}(W)\big) = \mathcal{P}(W) \quad \text{et} \quad \mathbb{P}\big(\mathcal{N}(0,1) \leqslant \mathcal{X}_p(\mathcal{W})\big) = \mathcal{P}_p(\mathcal{W})$$

x	-1	-2	-5	-10	-25	-50
$\mathbb{P}(\mathcal{N} \leqslant x)$	$1.6\ 10^{-1}$	$2.3\ 10^{-2}$	$2.9\ 10^{-7}$	$7.6\ 10^{-24}$	$3.1\ 10^{-138}$	$1.1\ 10^{-545}$

TAB. 9.1 – Quelques valeurs de la fonction de répartition d'une variable aléatoire gaussienne centrée réduite. $\mathcal{N} \sim \mathcal{N}(0,1)$.

(voir table 9.1 pour quelques exemples).

On en prend alors les valeurs absolues que l'on affecte du signe + pour les mots ou motifs sur-représentés et du signe − dans le cas d'une sous-représentation pour obtenir le résultat final $\mathcal{R}(W)$ (cas d'un mot) ou $\mathcal{R}_p(\mathcal{W})$ (cas d'un motif).

Exemple 9.1 *Les résultats obtenus s'interprètent donc ainsi :*

- $\mathcal{R}(\texttt{gctggtgg}) = +32.9$ *signifie*

$$\mathbb{P}\big(N(\texttt{gctggtgg}) \geqslant n(\texttt{gctggtgg})\big) = 1.1\ 10^{-237}$$

- $\mathcal{R}(\texttt{cgcg}) = -9.5$ *signifie*

$$\mathbb{P}\big(N(\texttt{cgcg}) \leqslant n(\texttt{cgcg})\big) = 1.0\ 10^{-21}$$

- $\mathcal{R}_{1,1}(\texttt{ccgct}, \texttt{agcgg}) = +72.2$ *signifie*

$$\mathbb{P}\big(N_{1,1}(\texttt{ccgct}, \texttt{agcgg}) \geqslant n_{1,1}(\texttt{ccgct}, \texttt{agcgg})\big) = 6.1\ 10^{-1135}$$

9.1.1 REGEXPCOUNT

Le package REGEXPCOUNT permet la mise en pratique des méthodes présentées en section 3.2.3 (page 56). Il s'agit d'un ensemble de fonctions utilisant le moteur de calcul formel MAPLE (dans sa version 5 et bientôt également dans sa version 6) qui est donc nécessaire au package.

On peut regretter qu'un programme non libre, et très cher de surcroît, soit indispensable à l'utilisation de ce package, mais il faut bien comprendre que la ré-implémentation des procédures de calculs formels que MAPLE possède déjà est une tâche très importante qui n'a jusqu'à présent pas été envisagée pour ce projet.

On pourra se procurer ce package ainsi qu'une documentation complète à l'adresse web suivante :

`http://algo.inria.fr/libraries/`

REGEXPCOUNT permet d'obtenir pour les moments d'ordre 1 et 2 des comptages de mots ou motifs par l'approche exacte aussi bien que par l'approche asymptotique donnant en cela des résultats comparables à ceux de R'MES dans son approche gaussienne (voir ci-après).

Il permet également de calculer de manière exacte les queues des distributions mais se heurte là à des difficultés numériques lorsque les nombres d'occurrences des mots ou motifs considérés sont importants (il faut effectuer un développement de *Taylor* formel à un ordre égal au nombre d'occurrences). Dans la suite, on notera $\mathcal{R}_x!$ la quantité $\mathcal{R}$ calculée par cette approche avec REGEXPCOUNT.

Dans tous les cas, les fonctions de ce package sont numériquement limitées à des modèles d'ordre assez faibles ($M2$ au mieux) et demandent une bonne connaissance de MAPLE afin de pouvoir être utilisées.

Remarquons également que l'utilisateur devra spécifier les paramètres des modèles utilisés ainsi que les comptages des mots ou motifs qu'il souhaite examiner, ce qui signifie que l'utilisation d'une méthode auxiliaire pour estimer des paramètres à partir d'une séquence et compter les occurrences est par ailleurs nécessaire.

9.1.2 R'MES

Le programme R'MES constitue l'implémentation des méthodes asymptotiques présentée en section 3.1 (page 48). Le "cœur" de R'MES effectue les calculs des quantités $\mathcal{R}$ précédemment évoquées par l'approche gaussienne ($\mathcal{R}_g$) ou *Poisson* composée ($\mathcal{R}_{pc}$) et est implémenté en C++ (et de façon anecdotique en C). Il existe également des fonctionnalités permettant une visualisation graphique des résultats en utilisant le programme SPLUS dont le portage vers un "clone" libre comme R n'est malheureusement pas envisagé dans un proche avenir.

Il est à noter que ce programme a été initialement développé sur plateforme SUN et que, malgré de nombreux efforts d'adaptation, il subsiste toujours des problèmes de compatibilité avec l'architecture x86 (notamment en ce qui concerne les calculs pour les familles de mots). Les sources du package ainsi que les binaires pour SUN et une documentation complète sont disponibles en téléchargement sur le web à l'adresse :

`http://www.inra.fr/bia/J/AB/genome/`

Signalons enfin que ce package ne propose pas le traitement de modèle $M00$ et d'un manière générale, ne permet pas de spécifier les paramètres des modèles considérés (ceux-ci sont systématiquement estimés à partir d'une

h	$M0$	$M1$	$M2$	$M3$	$M4$	$M5$	$M6$	$M7$
3	$4.2\,10^{-4}$	$6.9\,10^{-4}$	–	–	–	–	–	–
4	$1.9\,10^{-4}$	$2.5\,10^{-4}$	$3.8\,10^{-4}$	–	–	–	–	–
5	$1.1\,10^{-4}$	$1.2\,10^{-4}$	$1.6\,10^{-4}$	$1.9\,10^{-4}$	–	–	–	–
6	$5.2\,10^{-5}$	$5.5\,10^{-5}$	$6.4\,10^{-5}$	$8.3\,10^{-5}$	$1.6\,10^{-4}$	–	–	–
7	$3.5\,10^{-5}$	$3.4\,10^{-5}$	$3.9\,10^{-5}$	$5.5\,10^{-5}$	$1.2\,10^{-4}$	$3.8\,10^{-4}$	–	–
8	$2.9\,10^{-5}$	$2.8\,10^{-5}$	$3.1\,10^{-5}$	$4.7\,10^{-5}$	$1.1\,10^{-4}$	$3.7\,10^{-4}$	$1.3\,10^{-3}$	–
9	$2.8\,10^{-5}$	$2.7\,10^{-5}$	$3.0\,10^{-5}$	$4.5\,10^{-5}$	$1.1\,10^{-4}$	$3.6\,10^{-4}$	$1.3\,10^{-3}$	$5.3\,10^{-3}$

TAB. 9.2 – Temps calculs moyen (en secondes) pour un mot de longueur h dans le modèle Mm du programme `rmes.gaussien`. Les calculs ont été effectués pour tous les mots de longueur h du génome de *Mycoplasma genitallium* et les temps obtenus en divisant le temps d'exécution du programme (en faisant abstraction du temps de lecture de la séquence) par le nombre de mots traités. La configuration choisie est un bi-PIII 600 MHz équipé de 512 Mo de mémoire vive fonctionnant sous le système d'exploitation LINUX 2.2.

séquence car la méthode théorique l'éxige), et que l'on est limité à l'alphabet nucléique (limitation de l'implémentation).

`rmes.gaussien`

Le programme `rmes.gaussien` permet d'effectuer les calculs pour l'approche gaussienne (voir la section 3.1.2 page 49) pour tous les mots d'une longueur donnée ou bien pour une famille de mots (cette dernière fonctionnalité n'est pas opérationnelle sur plateforme x86).

La table 9.2 présente les temps de calcul du programme pour différentes longueurs de mots et différents ordres de modèles. Comme il n'est pas possible de distinguer le temps de pré-traitement (lecture de la séquence et estimation des paramètres des modèles) et celui des calculs du temps d'exécution global, on estime le premier par le temps global dans le cas de l'examen des mots de longueur deux (et on le soustrait par la suite au temps global pour les mots de longueur $h > 2$).

Dans ce programme le modèle $M0$ est traité comme un cas particulier du modèle $M1$ ce qui explique les différences mineures entre ces deux modèles. Comme le temps de calcul nécessaire à l'estimation des paramètres du modèle ne peut être réellement déduit du temps de calcul total, on constate une diminution du temps de calcul pour un mot lorsque h augmente (le surplus de calcul est divisé en un plus grand nombre).

h	$M0$	$M1$	$M2$	$M3$	$M4$	$M5$	$M6$	$M7$
3	24	15	–	–	–	–	–	–
4	1.6	1.2	$5.0\,10^{-1}$	–	–	–	–	–
5	$1.2\,10^{-1}$	$9.8\,10^{-2}$	$3.8\,10^{-2}$	$1.9\,10^{-2}$	–	–	–	–
6	$8.1\,10^{-3}$	$6.2\,10^{-3}$	$2.7\,10^{-3}$	$1.0\,10^{-3}$	$5.9\,10^{-4}$	–	–	–
7	$5.6\,10^{-4}$	$4.9\,10^{-4}$	$2.2\,10^{-4}$	$9.8\,10^{-5}$	$6.5\,10^{-5}$	$5.6\,10^{-4}$	–	–
8	$7.6\,10^{-5}$	$6.7\,10^{-5}$	$4.6\,10^{-5}$	$3.6\,10^{-5}$	$3.3\,10^{-5}$	$3.3\,10^{-5}$	$4.1\,10^{-5}$	–
9	$3.8\,10^{-5}$	$3.6\,10^{-5}$	$3.1\,10^{-5}$	$2.9\,10^{-5}$	$2.9\,10^{-5}$	$2.9\,10^{-5}$	$4.1\,10^{-5}$	$6.3\,10^{-5}$

TAB. 9.3 – Temps calculs moyen (en secondes) pour un mot de longueur h dans le modèle Mm du programme `rmes.poisson.composee`. Les calculs ont été effectués pour tous les mots de longueur h du génome de *Mycoplasma genitallium* et les temps obtenus en divisant le temps d'exécution du programme (en faisant abstraction du temps de lecture de la séquence) par le nombre de mots traités. La configuration choisie est un bi-PIII 600 MHz équipé de 512 Mo de mémoire vive fonctionnant sous le système d'exploitation LINUX 2.2.

On observe également une augmentation géométrique du temps de calcul avec l'ordre des modèles considérés, ce qui n'est guère étonnant. La raison de cette progression géométrique est, comme on pouvait s'y attendre, proche de quatre (le cardinal de l'alphabet nucléique).

`rmes.poisson.composee`

Le programme `rmes.poisson.composee` permet, pour sa part, d'effectuer les calculs de l'approche *Poisson* composée (voir la section 3.1.3 page 50) pour tous les mots d'une longueur donnée (mais pas pour des familles de mots).

On peut voir sur la table 9.3 que le temps de calcul pour un mot est croissante selon deux axes : l'ordre du modèle considéré tout d'abord (ce qui est naturel) mais aussi la longueur du mot. L'évolution sur ce dernier axe est assez contre intuitive puisque le temps de calcul augmente lorsque la longueur du mot diminue. En fait, plus le nombre d'occurrences du mot est important, plus le temps de calcul est long (il y a, en moyenne, plus d'occurrences pour les mots courts que pour les mots longs).

Plusieurs problèmes numériques interdisent par ailleurs au programme de donner des résultats valables pour un nombre important de mots selon que ceux ci :

- ont un trop grand nombre d'occurrences ;
- sont "trop" exceptionnels (probabilités inférieures à 10^{-300}).

h	$M0$	$M1$	$M2$	$M3$	$M4$	$M5$	$M6$	$M7$
3	39%	20%	–	–	–	–	–	–
4	21%	26%	5.5%	–	–	–	–	–
5	23%	25%	6.2%	1.4%	–	–	–	–
6	23%	21%	5.2%	1.3%	0.22%	–	–	–
7	80%	9%	2.5%	0.72%	0.21%	0.68%	–	–
8	18%	1.3%	0.54%	0.19%	0.064%	0.99%	8.4%	–
9	0.029%	0.05%	0.038%	0.018%	0.0065%	1.3%	11%	31%

TAB. 9.4 – Pourcentages de mots de longueur h pour lesquels les calculs de `rmes.poisson.composee` sur le génome de *Mycoplasma genitallium* parviennent à une incohérence numérique dans le modèle Mm ; il s'agit de mots très exceptionnels ou ayant un grand nombre d'occurrences. La configuration choisie est un bi-PIII 600 MHz équipé de 512 Mo de mémoire vive fonctionnant sous le système d'exploitation LINUX 2.2.

La table 9.4 donne pour chaque longueur de mot et chaque modèle considéré la proportion de mots que le programme n'a pas été en mesure de calculer. Bien que le génome considéré, celui de *Mycoplasme genitallium*, soit le plus petit génome complet connu (580 000 bases seulement), on compte jusqu'à 39% d'erreurs pour les mots les plus courts. De plus, même en ne considérant que des mots longs ($h = 9$ en l'occurrence) on trouve jusqu'à 31% d'erreurs ce qui montre certaines limitations au programme.

9.1.3 GDon

Le programme `GDon` (Prononcer "GéDéon") est l'implémentation en C des méthodes développées dans cette thèse (voir chapitre 6 page 91 notamment), il est à noter que ce programme utilise également deux fonctions en fortran 77 pour le calcul des valeurs propres et vecteur propres d'une matrice qui ont été tirées de LAPACK (la bibliothèque libre en fortran pour l'algèbre linéaire) mais reste compilable avec les outils standard du GNU (compilateurs `gcc` et `g77` en l'occurrence).

La `glib` (bibliothèque C for utile) ainsi que d'autres standards du GNU comme `automake` (pour faciliter la compilation sur différentes plateformes) ou encore `getopt` (lecture des arguments d'un programme) sont largement utilisés par le programme et sont nécessaires à sa compilation.

Néanmoins, le choix de ces outils, directement issus du GNU, assure a priori une bonne portabilité de GDon qui a déjà été compilé et utilisé avec

h	$M00$	$M0$	$M1$	$M2$	$M3$	$M4$	$M5$	$M6$	$M7$
2	$4.2\,10^{-4}$	$4.3\,10^{-4}$	–	–	–	–	–	–	–
3	$1.1\,10^{-3}$	$1.1\,10^{-3}$	$3.1\,10^{-3}$	–	–	–	–	–	–
4	$2.9\,10^{-3}$	$2.6\,10^{-3}$	$4.8\,10^{-3}$	$9.8\,10^{-3}$	–	–	–	–	–
5	$8.6\,10^{-3}$	$6.9\,10^{-3}$	$1.2\,10^{-2}$	$1.6\,10^{-2}$	$1.8\,10^{-2}$	–	–	–	–
6	$2.9\,10^{-2}$	$2.0\,10^{-2}$	$3.0\,10^{-2}$	$3.7\,10^{-2}$	$4.1\,10^{-2}$	$5.6\,10^{-2}$	–	–	–
7	$2.2\,10^{-1}$	$1.9\,10^{-1}$	$1.9\,10^{-1}$	$2.2\,10^{-1}$	$2.4\,10^{-1}$	$3.1\,10^{-1}$	$3.5\,10^{-1}$	–	–
8	1.7	$9.2\,10^{-1}$	1.0	1.1	1.2	1.5	1.7	1.9	–
9	6.4	4.7	5.5	5.8	6.5	7.0	7.1	7.3	6.7

TAB. 9.5 – Temps calculs moyen (en secondes) pour un mot de longueur h dans le modèle Mm du programme `GDon` en utilisant l'alphabet $\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$. La configuration choisie est un bi-PIII 600 MHz équipé de 512 Mo de mémoire vive fonctionnant sous le système d'exploitation LINUX 2.2.

succès sur plateforme x86 sous Linux ainsi que sur SUN muni de Solaris.

GDon permet d'effectuer les calculs des quantités proposées dans la partie II pour des mots ou des familles de mots dans divers alphabets ; on se reportera à l'annexe E page 239 pour plus de détails concernant les possibilités de ce programme.

On pourra se procurer GDon à l'adresse suivante :

`http://www.genopole.cnrs.fr/GDon`

ou bien en me contactant à l'adresse :

`nuel@genopole.cnrs.fr`

GDon permet d'effectuer les calculs pour les comptages sur une séquence donnée de un ou plusieurs mots ou motifs de longueur h dans les modèles $M00$ à Mm, où $m \leqslant h-1$ dont les paramètres sont estimés à partir de la séquence ou sont spécifiés par l'utilisateur.

Il est possible de spécifier différents types d'alphabets ce qui permet de travailler sur des séquences nucléiques ou protéiques en diminuant la taille de l'alphabet en regroupant les lettres selon leurs propriétés. Typiquement on pourra regrouper en cinq familles (celles qui sont suggérées par la table 1.2 page 19 en séparant le tryptophane qui, par sa forme et sa masse, diffère notablement des autres acides aminés) en utilisant l'alphabet $\mathcal{A} = \{\mathtt{0}, \mathtt{1}, \mathtt{2}, \mathtt{3}, \mathtt{4}\}$

où

$$\begin{cases} \mathtt{0} = \{\mathtt{A}, \mathtt{V}, \mathtt{L}, \mathtt{I}, \mathtt{P}, \mathtt{F}, \mathtt{M}\} & \text{hydrophobes sauf tryptophane} \\ \mathtt{1} = \{\mathtt{W}\} & \text{tryptophane} \\ \mathtt{2} = \{\mathtt{G}, \mathtt{S}, \mathtt{T}, \mathtt{Y}, \mathtt{C}, \mathtt{N}, \mathtt{Q}\} & \text{polaires} \\ \mathtt{3} = \{\mathtt{K}, \mathtt{R}, \mathtt{H}\} & \text{basiques} \\ \mathtt{4} = \{\mathtt{D}, \mathtt{E}\} & \text{acides} \end{cases}$$

De même, il est possible de réduire le cardinal de l'alphabet nucléique en regroupant les bases en *purines* (`a` et `g`) et *pyrimidines* (`c` et `t`) ou en tout autre combinaison. Pour plus de précision concernant ce type de manipulations, on se reportera à la section 11.2 (page 162).

On pourra utiliser cette "astuce" pour traiter des mots et motifs dont les calculs sont trop longs ou bien requièrent une trop grande quantité de mémoire dans l'alphabet initial pour pouvoir être traités.

GDon est également capable de calculer les paramètres de changements de probabilités évoqués en section 6.3 (page 102) et d'utiliser ces paramètres (ou d'autres spécifiés par l'utilisateur) pour effectuer des simulations sous cette nouvelle probabilité. On désignera désormais par $\mathcal{R}_{gd}$ la quantité $\mathcal{R}$ calculée par GDon.

Comme le montre la table 9.5, les performances de GDon en termes de rapidité des calculs dépend essentiellement de la longueur h du mot considéré. Lorsque que cette longueur est fixée, on peut observer des variations du temps de calcul moyen selon le modèle considéré mais ces variations étant essentiellement aléatoires (même si la complexité de la matrice de transition influe légèrement sur ce résultat).

Cela est tout à fait naturel dans la mesure où, comme on l'a vu au chapitre 6 (page 91), le calcul de $\mathcal{R}$ fait intervenir (à plusieurs reprises) le calcul de la valeur propre de *Perron-Frobénius* d'une matrice d'ordre k^{h-1} (où, on le rappelle, k désigne le cardinal de l'alphabet considéré) dont, fort heureusement, seuls k^h éléments sont non nuls.

En termes de ressources mémoire on est néanmoins très rapidement limité. En comptant 32 bits pour stocker un réel (précision double en C sur plateforme x86) et dans le cas d'un alphabet à quatre lettres ($k = 4$), on trouve les quantités de mémoire suivante pour le stockage de la matrice de transition seule :

longueur h	2	3	4	5	6	7	8
mémoire	64 o	256 o	1 Ko	4 Ko	16 Ko	64 Ko	256 Ko

9	10	11	12	13	14
1 Mo	4 Mo	16 Mo	64 Mo	256 Mo	1 Go

Bien évidemment les algorithmes mis en œuvre dans les calculs vont être amenés à utiliser plusieurs fois cette quantité de mémoire si bien qu'en se limitant à 100 Mo pour la mémoire allouée à la matrice de transition, on trouve les limites suivantes pour la longueur des mots ou motifs considérés selon la valeurs de k :

k	2	4	5	20
h max	24	12	10	5

ce qui permet de mieux comprendre l'intérêt des alphabets réduits précédemment évoqués (cas des acides aminés notamment).

Remarquons enfin que la complexité d'une famille de mots (son nombre d'éléments par exemple) n'intervient pas de manière significative dans le programme. On peut donc effectuer les calculs pour des familles quelconques de longueur h pour le même "coût" que dans le cas d'un mot de longueur h.

On pourra se reporter à l'annexe E (page 239) pour une documentation complète de GDon.

9.2 Séquences aléatoires

Les événements que nous sommes amenés à considérer ont des probabilités bien trop faibles pour pouvoir raisonnablement espérer les évaluer par des simulations directes. Nous allons donc nous contenter ici de simuler des séquences selon un modèle donné avant d'y rechercher des mots exceptionnels.

Expliquons notre démarche à travers un exemple : on considère le génome complet de *Mycoplasma genitalium* (séquence `mgenitalium`) dont la longueur totale est $n = 580\,074$ (il s'agit du plus court des génomes bactériens séquencés). Voici les paramètres du modèle $M0$ estimés sur ce génome :

$$\widehat{\mu} = \begin{pmatrix} \mathtt{a} & \mathtt{c} & \mathtt{g} & \mathtt{t} \\ 34.6\% & 15.8\% & 15.9\% & 33.7\% \end{pmatrix}$$

et voici ceux de $M1$:

$$\widehat{\Pi} = \begin{pmatrix} & \mathtt{a} & \mathtt{c} & \mathtt{g} & \mathtt{t} \\ \mathtt{a} & 42.2\% & 15.2\% & 16.8\% & 25.8\% \\ \mathtt{c} & 39.9\% & 18.1\% & 6.2\% & 35.8\% \\ \mathtt{g} & 31.2\% & 18.8\% & 17.8\% & 32.2\% \\ \mathtt{t} & 25.9\% & 13.9\% & 18.7\% & 41.5\% \end{pmatrix}$$

On effectue le tirage dans le modèle $M1$ ci dessus d'une séquence `random` de même longueur n et de même premier caractère `t` que `mgenitalium`. On va maintenant étudier tous les mots exceptionnels de longueur $h = 6$ de cette

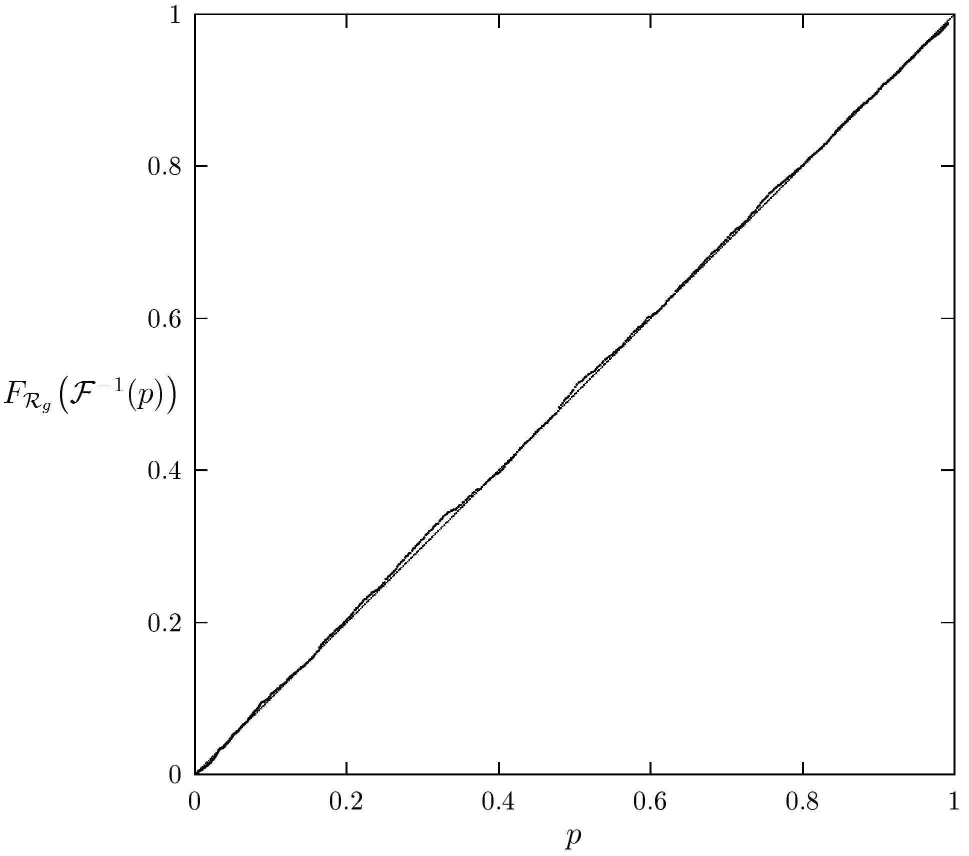


FIG. 9.2 – *qqplot* des 4 096 éléments de $\big(\mathcal{R}_g(W)\big)_{W\in\mathcal{A}^6}$ sur la séquence `random` avec $h = 6$ et dans le modèle $M1$. $\mathcal{F}_{\mathcal{R}_g}$ désigne la fonction de répartition empirique des $\big(\mathcal{R}_g(W)\big)_W$.

séquence et vérifier que la distribution des $\mathcal{R}$ trouvés correspond bien à la distribution attendue.

En effet, en effectuant les calculs de $\mathcal{R}$ sous le modèle $M1$ on est en droit de s'attendre à observer une fonction de répartition empirique "proche" de la fonction de répartition d'une gaussienne centrée réduite (puisque c'est le mode de représentation qui a été choisi) : $\forall r \in \mathbb{R}$ alors

$$F(r) = \frac{\text{nombre de } \mathcal{R} \leqslant r}{\text{nombre total de } \mathcal{R}}$$

désigne la fonction de répartition empirique des résultats et on s'attend à ce que

$$F(\mathcal{F}^{-1}(p)) \text{ soit proche de } p \text{ pour tout } p \in [0,1]$$

où $\mathcal{F}$ est la fonction de répartition d'une variable aléatoire gaussienne centrée réduite.

On peut voir en figure 9.2 que les résultats du programme `rmes.gaussien` correspondent bien notre attente ce qui nous rassure à la fois sur la méthode d'évaluation des résultats et sur la validité des résultats proposés par `rmes.gaussien`.

Sur la figure 9.3, c'est la répartition des résultats de GDon qui est examinée. Même si le graphe observé est loin d'être totalement aberrant, il y a néanmoins de grosses variations par rapport à l'attendu, en particulier il est clair que GDon sous-estime le caractère exceptionnel des mots considérés (et ceci que les mots soient sur- ou sous-représentés).

Fort heureusement, cette imprécision de GDon ne doit pas nous inquiéter outre mesure. En effet les approximations obtenues par les grandes déviations sont d'autant meilleures que les événements considérés sont rares ; or, les comptages observés ici n'ont, bien évidemment, rien d'exceptionnel puisque la séquence à précisément été générée selon le modèle de référence.

On a néanmoins ainsi l'illustration de la faiblesse des approches par les grandes déviations pour des événements trop fréquents.

Voici les résultats obtenus pour les mots les plus exceptionnels par les deux méthodes :

mot	$\mathcal{R}_g$	$\mathcal{R}_{gd}$
`tatcga`	$+3.27$	$+2.30$
`attcac`	$+3.25$	$+2.41$
`tccata`	-3.30	-2.80
`ggaaaa`	-3.36	-2.71

compte tenu des remarques précédentes et même si les différences sont relativement importantes, on peut se satisfaire de constater que les résultats de `GDon` restent du même ordre que ceux de `rmes.gaussien`. On ne s'étendra

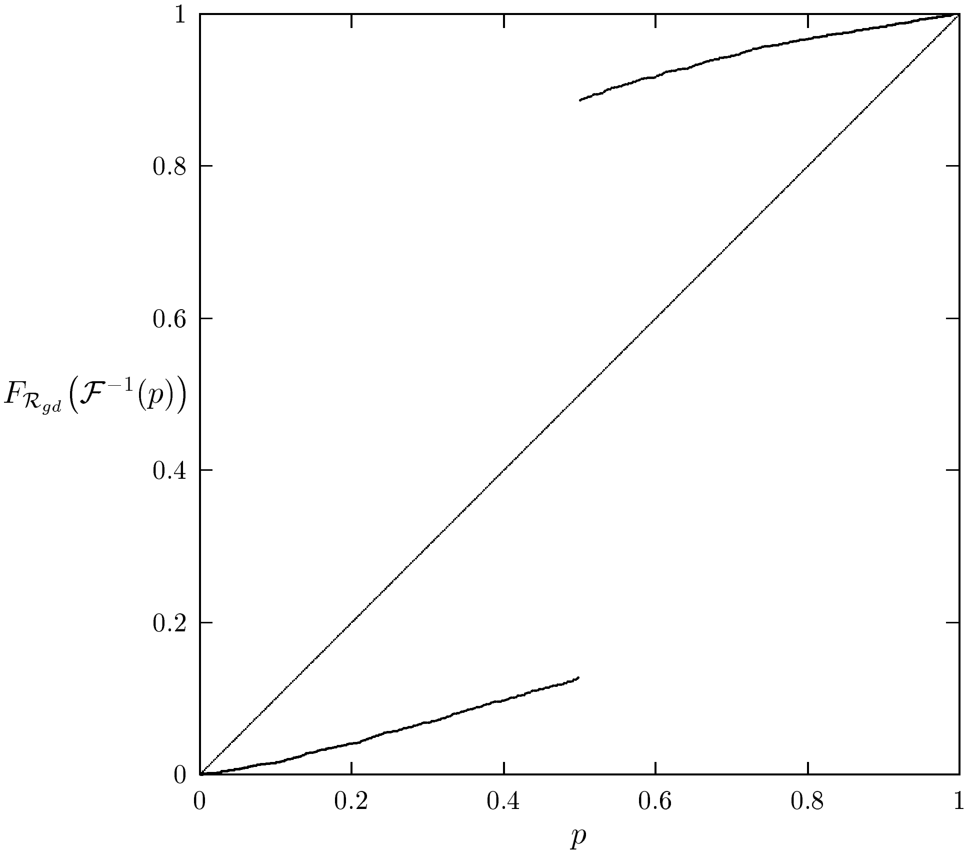


FIG. 9.3 – *qqplot* des 4 096 éléments de $\big(\mathcal{R}_g(W)\big)_{W\in\mathcal{A}^6}$ sur la séquence `random` avec $h = 6$ et dans le modèle $M1$. $\mathcal{F}_{\mathcal{R}_{gd}}$ désigne la fonction de répartition empirique des $\big(\mathcal{R}_{gd}(W)\big)_W$.

cependant pas davantage sur ce point, et on consultera plutôt la section 9.3 pour plus de détails concernant la comparaison des résultats de R'MES et de GDon.

9.3 Comparaison avec R'MES

9.3.1 Approche gaussienne

On peut voir en fig 9.4 la comparaison des résultats donnés par `rmes.gaussien` et par `GDon` pour les 4 096 mots de longueur 6 du génome de *Mycoplasma genitalium* lorsque l'on se place dans le modèle $M1$ (les résultats sont similaires dans les autres modèles).

On constate tout d'abord une bonne cohérence des valeurs de $\mathcal{R}$ des deux programmes qui nous rassure sur la validité de l'approche GDon, mais il existe néanmoins des différences importantes.

En premier lieu, on observe un "plateau" important pour les valeurs de $\mathcal{R}_{gd}$ proche de 0, c'est à dire pour les mots pas ou peu exceptionnels. Comme on l'a déjà vu, cela s'explique simplement par le fait que les événements considérés dans ce cas sortent largement du cadre d'application des grandes déviations. Bien évidemment ce "défaut" de l'approche a une incidence quasi nulle en ce qui concerne la recherche de mots exceptionnels puisque un mot ne peut être qualifié de la sorte qu'à partir du moment où sa valeur de $\mathcal{R}$ est assez grande en module.

A gauche et à droite du graphe, pour les mots les plus exceptionnels, on observe également des différences importantes qu'il faut très certainement imputer à l'imprécision d'une méthode "centrale" comme l'approximation gaussienne lorsque des événements très rares sont examinés. Même l'ordre dans lequel les mots sont classés par les deux méthodes peut varier ce qui peut avoir une importance capitale dans la détection des mots exceptionnels.

On illustre bien ainsi tout l'intérêt que peut avoir le recours à des méthodes spécialement conçues pour l'étude des événements rares comme les grandes déviations par rapport à une approche "centrale" comme l'approximation gaussienne qui semble totalement inadaptée au problème considéré.

9.3.2 Approche *Poisson* composée

La figure 9.5 propose la comparaison des résultats de l'approche *Poisson* composée de R'MES et de ceux de GDon.

Pour les mots peu exceptionnels (avec des valeurs de $\mathcal{R}$ proches de 0), on remarque une fois de plus des différences expliquées par le manque de

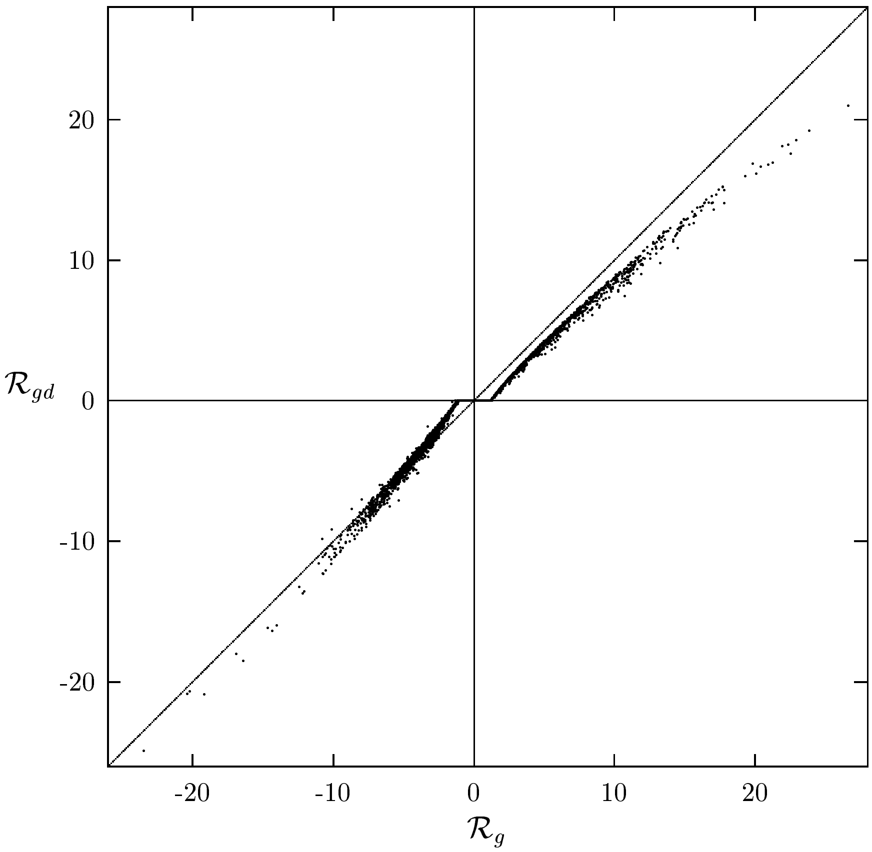


FIG. 9.4 – Comparaison des résultats de R'MES (gaussien) et de GDon pour les 4 096 mots de longueur 6 de *Mycoplasma genitalium* dans $M1$.

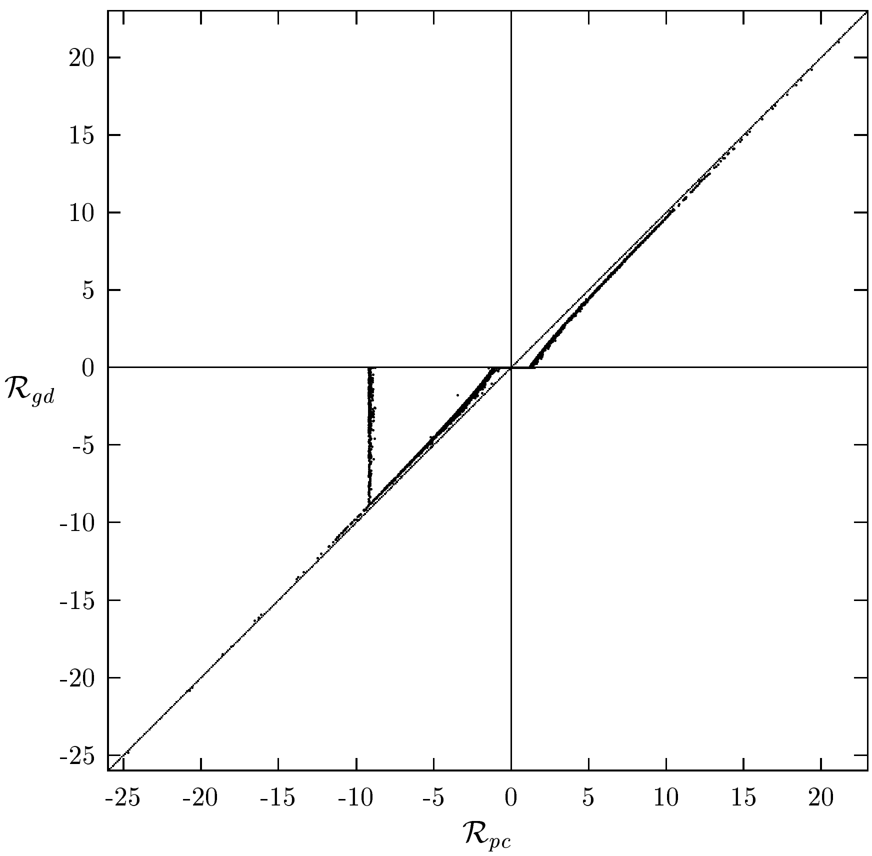


FIG. 9.5 – Comparaison des résultats de R'MES (poisson composée) et de GDon pour les 4 096 mots de longueur 6 de *Mycoplasma genitalium* dans $M1$.

précision de l'approche grandes déviations pour ces événements.

Dans le cas des mots exceptionnels, on constate à l'inverse que les résultats des deux méthodes sont très proches ce qui constitue une validation indéniable du travail effectué sur R'MES et met en évidence la qualité des approximations par *Poisson* composée qui sont utilisées dans ce programme.

Il ne faut cependant pas perdre de vue les problèmes auxquels cette dernière approche est confrontée :

- traitement des familles de mots non disponible ;
- calculs impossibles pour de nombreux mots, 9% des mots de longueur 6 dans notre cas, parmi les plus exceptionnels (on a par exemple $\mathcal{R}_{gd}(\texttt{aacaac}) = +18.1$) alors que $\mathcal{R}_{pc}(\texttt{aacaac})$ n'est pas disponible).

Enfin, on remarque en dernier lieu une étrange accumulation de points dans le graphe de la figure 9.5 : les valeurs de $\mathcal{R}_{pc}$ sont proches de -9 pour un grand nombre de mots tandis que les valeurs de $\mathcal{R}_{gd}$ pour ces mêmes mots s'étalent entre -9 et 0.

La nature pour le moins singulière de ce résultat ainsi que sa persistance sur d'autres exemples ainsi que lorsque l'on compare $\mathcal{R}_{pc}$ à $\mathcal{R}_{g}$ (voir fig 9.6) permettent de supposer que ce problème est issu d'une erreur d'implémentation dans R'MES qui n'a, pour l'instant, malheureusement pas été corrigée.

Dans le cas présent, plus de 400 mots sont concernés par ce *bug* ce qui porte à 20% la proportion de mots de longueur 6 que le programme `poisson.composee` se trouve finalement incapable de traiter correctement.

Ainsi, même si l'approche *Poisson* composée rejoint la précision des résultats par les grandes déviations en ce qui concerne les mots exceptionnels, son incapacité à traiter de nombreux mots rend (dans cette implémentation) son usage hasardeux dans les cas courants.

9.4 Comparaison avec REGEXPCOUNT

La complexité des calculs permettant d'examiner la queue de distribution d'un mot (ou d'un motif) étant directement liée à son nombre d'occurrences on se contente ici d'examiner les mots de longueur 6 et 7 dont les comptages sont nuls ou égaux à un dans `mgenitalium`.

On obtient ainsi un total de 1818 ainsi répartis selon leur longueur h et leur nombre d'occurrences N :

$h\backslash N$	0	1
6	14	35
7	851	918

Le calcul de l'ensemble des queues de distributions de ces mots a requis environ 5h30 de calculs (soit plus de 10s par mot) sur un calculateur Alpha.

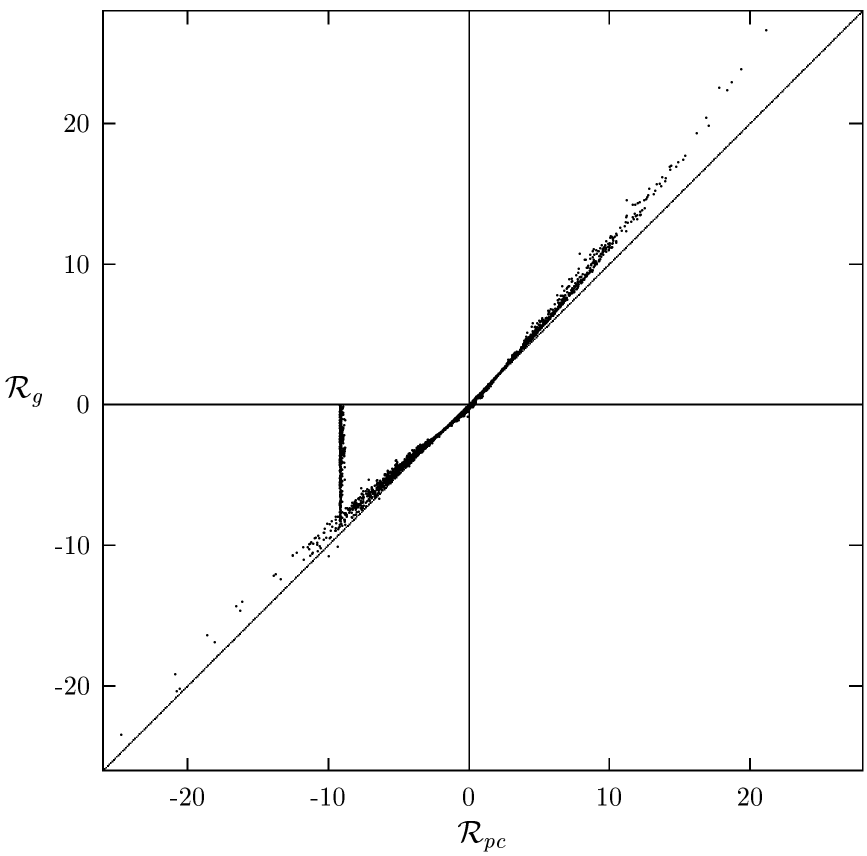


FIG. 9.6 – Comparaison des résultats de R'MES (poisson composée) et de R'MES (gaussien) pour les 4 096 mots de longueur 6 de *Mycoplasma genitalium* dans $M1$.

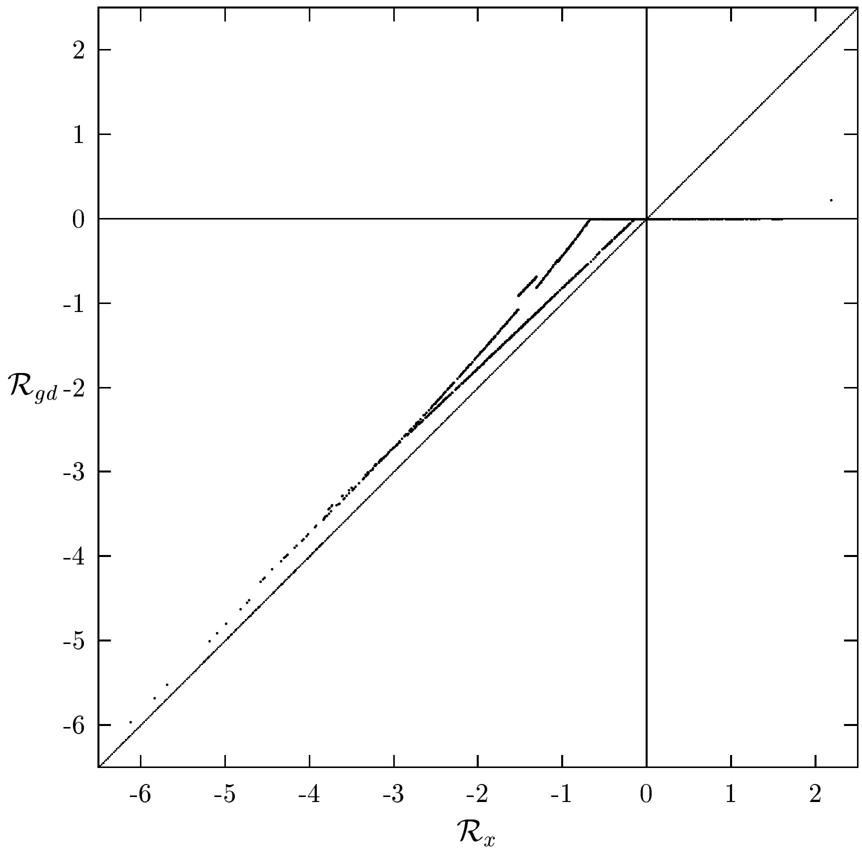


FIG. 9.7 – Comparaison des résultats de REGEXPCOUNT et de GDon pour les 1 818 mots de comptage inférieur ou égal à un et de longueur 6 ou 7 de *Mycoplasma genitalium* dans $M1$.

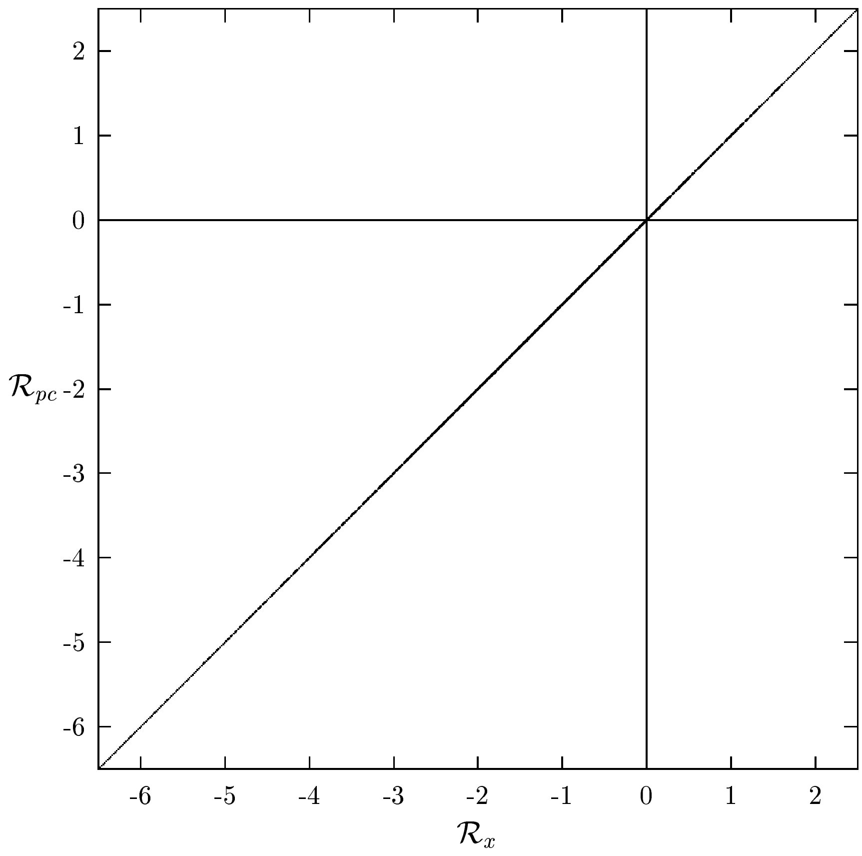


FIG. 9.8 – Comparaison des résultats de REGEXPCOUNT et de R'MES (poisson composée) pour les 1 818 mots de comptage inférieur ou égal à un et de longueur 6 ou 7 de *Mycoplasma genitalium* dans $M1$.

Les résultats et leur comparaison avec ceux de GDon sont donnés sur la figure 9.7.

A titre de vérification la figure 9.8 montre que les résultats obtenus par la méthode exacte sont bien identiques à ceux obtenus avec l'approximation par lois de *Poisson* composée.

Une fois de plus, on constate le manque de qualité prévisible des résultats de GDon concernant les mots peu exceptionnels. Mais on constate par ailleurs l'augmentation de la qualité des approximations par les grandes déviations lorsque les événements considérés se font plus rares, ce qui constitue une nouvelle validation de la méthode étudiée.

Références

[NSF99] P. Nicodème, B. Salvy, and P. Flajolet. Motif statistics. *Theoretical Computer Science*, 1999. to appear.

[RD99] S. Robin and J.J. Daudin. Exact distribution of word occurrences in a random sequence of letters. *J. App. Prob.*, 36 :179–193, 1999.

[RD00] S. Robin and J.J. Daudin. Exact distribution of the distances between any occurrences of a set of words. *Ann. Inst. Statist. Math.*, 2000.

[RS01] S. Robin and S. Schbath. Numerical comparison of several approximations of the word count distribution in random sequences. *J. Comp. Biol.*, 2001. To appear.

[Sch97] S. Schbath. An efficient statistic to detect over- and under- represented words in dna sequences. *J. Comp. Biol.*, 4 :189–192, 1997.

Fig. 10.1 – Représentation graphique des mots de longueur $h = 9$ chez *Mycobacterium leprae* dans le modèle $M1$. On consultera la section 11.1 page 160 pour plus de détails.

Chapitre 10

Premiers résultats

Pré-requis :
Chapitre 1 et chapitre 9

Description :
Résultats des grandes déviations pour l'étude des mots exceptionnels du chapitre 1.

Résumé :
Les sites de restrictions sont statistiquement sous-représentés ainsi qu'on l'attendait et, tout comme l'annonçaient nos connaissances biologiques, les motifs *chi* et *uptake* sont pour leurs parts très sur-représentés.

Contenu du chapitre

10.1 Introduction

On va considérer les exemples de mot exceptionnels présentés au chapitre 1 (page 15) afin de vérifier si leur caractère exceptionnel d'un point de vue biologique se traduit également par un caractère exceptionnel du point de vue des statistiques.

Pour pouvoir tenir compte de structure double brin de l'ADN, on va s'efforcer d'étudier les mots et motifs de chaque exemple aussi bien sur un brin que sur son complémentaire. Pour cela, on utilise la capacité de GDon à traiter des familles de mots pour s'intéresser aux familles formées des mots de chacun des exemples ainsi que de leurs complémentaires.

10.2 Sites de restrictions

Comme on a vu en section 1.2.1 (page 21), la pression de la sélection doit limiter le nombre d'occurrences des sites de restrictions dans les génomes des organismes synthétisant les enzymes auxquels ils correspondent. Il est vraisemblable qu'un tel comportement aura pour conséquence la sous-représentation de ces motifs.

10.2.1 *Bacillus subtilis*

Chez *Bacillus subtilis*, l'enzyme de restriction `Bsu6633I` est associée au site `cgcg` (qui est un *palindrome*, c'est à dire que son complémentaire n'est autre que lui-même). Les résultats de l'étude de ce site sont donnés par le tableau suivant (quantiles gaussiens obtenus par GDon) :

modèle \ motif	`cgcg`
$M00$	-66.0
$M0$	-9.5
$M1$	-29.5
$M2$	-20.4

De même pour l'enzyme `Bsu8565I` qui est associée au palindrome `ggatcc` on trouve :

modèle \ motif	`ggatcc`
$M00$	-29.5
$M0$	-21.6
$M1$	-23.5
$M2$	-29.2
$M3$	-23.1
$M4$	-18.0

Dans les deux cas, les résultats sont parfaitement cohérents avec nos attentes. Ces mots sont extrêmement sous-représentés et ceci quel que soit l'ordre du modèle markovien choisi.

10.2.2 *Escherichia coli*

Chez *Escherichia coli*, les résultats sont eux aussi en accord avec l'intuition qu'on en a. Voici les résultats concernant l'enzyme `Eco92I` qui est associée au palindrome `ccgcgg` :

modèle \ motif	`ccgcgg`
$M00$	-15.1
$M0$	-18.1
$M1$	-30.0
$M2$	-38.2
$M3$	-29.5
$M4$	-22.1

L'enzyme `Eco120I` est pour sa part associée au site de restriction `ggtctc`, qui n'est pas un palindrome. C'est pourquoi on fait intervenir dans les calculs son complémentaire inverse :

modèle \ motif	ggtctc	gagacc	ggtctc gagacc
$M00$	-38.2	-37.5	-53.7
$M0$	-39.0	-38.3	-54.8
$M1$	-25.0	-24.5	-35.1
$M2$	-23.3	-22.8	-32.7
$M3$	-21.3	-21.1	-30.1
$M4$	-12.9	-13.7	-18.9

Remarquons que si la sous-représentation de site de restriction est clairement établie sur chacun des brins, l'étude simultanée des deux brins (3$^{\text{ème}}$ colonne du tableau) met en évidence une sous-représentation encore plus marquée.

10.3 Chi

Comme la section 1.2.2 (page 22) il va de la survie des organismes utilisant des nucléases de disposer dans leurs génomes des motifs *chi* en nombre suffisant pour inhiber l'action de ces nucléases avant que des dégâts irréversibles ne soient occasionnés. Il est donc naturel de s'attendre un trouver un grand nombre de ces motifs *chi* dans leurs génomes respectifs.

10.3.1 *Bacillus subtilis*

Chez *Bacillus subtilis* c'est le motif très court `ccgct` dont l'activité de *chi* a pu être mesurée en laboratoire. Voici les résultats :

modèle \ motif	ccgct	agcgg	ccgct agcgg
$M00$	$+23.1$	$+23.0$	$+32.9$
$M0$	$+50.5$	$+51.1$	$+72.2$
$M1$	$+41.1$	$+41.9$	$+59.0$
$M2$	$+18.2$	$+16.3$	$+24.7$
$M3$	$+4.0$	$+2.5$	$+5.1$

On observe une forte sur-représentation des comptages de ces motifs qui est en parfaite adéquation avec le comportement que l'on attendait.

10.3.2 *Escherichia coli*

De même, chez *Escherichia coli*, le motif `gctggtgg` est sur-représenté comme le montrent les résultats :

modèle \ motif	`gctggtgg`	`ccaccagc`	`gctggtgg` `ccaccagc`
$M00$	$+32.9$	$+33.6$	$+47.1$
$M0$	$+32.2$	$+32.7$	$+46.0$
$M1$	$+33.0$	$+33.2$	$+47.0$
$M2$	$+19.9$	$+19.8$	$+28.1$
$M3$	$+10.8$	$+10.9$	$+15.5$
$M4$	$+11.4$	$+11.5$	$+16.2$
$M5$	$+9.7$	$+9.0$	$+13.5$
$M6$	$+3.1$	$+1.7$	$+3.8$

10.3.3 *Haemophilus influenzae*

Chez *Haemophilus influenzae*, le motif *chi* est dégénéré ; quelle que soit la deuxième lettre de ce motif, on a pu mesurer en laboratoire une action inhibitrice sur les nucléases. Les résultats statistiques sont les suivants :

modèle \ motif	`g.tggtgg`	`ccacca.c`	`g.tggtgg` `ccacca.c`
$M00$	$+3.5$	$+7.0$	$+7.8$
$M0$	$+13.1$	$+16.3$	$+21.0$
$M1$	$+12.6$	$+15.7$	$+20.3$
$M2$	$+5.0$	$+8.2$	$+9.7$
$M3$	$+2.3$	$+5.6$	$+6.0$
$M4$	$+0.0$	$+2.9$	$+2.6$
$M5$	-0.0	$+0.7$	$+0.0$
$M6$	-0.0	$+0.0$	$+0.0$

Si la sur-représentation de la famille entière est acquise dans les modèles markoviens d'ordres faibles, elle est beaucoup moins évidente pour les modèles d'ordres 5 et 6.

Afin de mieux comprendre ce qui se passe on étudie séparément chacune des versions du *chi* dans le tableau suivant :

modèle \ motif	gatggtgg ccaccatc	gctggtgg ccaccagc	ggtggtgg ccaccacc	gttggtgg ccaccaac
$M00$	-0.0	$+4.6$	$+6.0$	$+4.3$
$M0$	$+3.8$	$+13.3$	$+14.7$	$+9.9$
$M1$	$+4.4$	$+11.7$	$+15.1$	$+8.9$
$M2$	$+0.0$	$+6.2$	$+7.9$	$+3.7$
$M3$	-1.9	$+6.2$	$+6.1$	$+2.0$
$M4$	-0.8	$+3.4$	$+3.8$	-0.0
$M5$	-2.3	$+0.8$	$+2.5$	-0.0
$M6$	-1.6	$+0.0$	$+0.7$	$+0.0$

Dans certains modèles, on peut observer la très faible sur-représentation voir même la sous-représentation des versions `gatggtgg` et `gttggtgg` du *chi* tandis que la sur-représentation du `gctggtgg` et surtout de `ggtggtgg` est confirmée.

Au vu de ces résultats, on est en droit de ce demander si, pour des raisons qui nous échappent, l'évolution n'a pas privilégié l'usage des versions `gctggtgg` et `ggtggtgg` du *chi* au détriment des deux autres chez *Haemophilus influenzae*. Et même si l'activité de *chi* des versions les moins usitées du motifs peut encore être mesurée en laboratoire, il est possible que celle-ci finisse par disparaître peu à peu.

10.4 Uptake

Ainsi qu'on l'a vu en section 1.2.3 (page 23), les motifs uptake doivent, tout comme les motifs *chi*, être présents en nombre important dans leur génomes respectifs.

10.4.1 *Neisseria menigitidis*

Chez *neisseria menigitidis*, on observe en effet (exepté dans le modèle $M8$) l'extrême sur-représentation du mot `gccgtgtgaa` comme le montrent les résultats suivants :

modèle \ motif	`gccgtgtgaa`	`ttcagacggc`	`gccgtgtgaa` `ttcagacggc`
$M00$	+99.3	+99.5	+140.6
$M0$	+98.7	+98.9	+139.8
$M1$	+97.5	+98.0	+138.2
$M2$	+89.4	+90.0	+126.9
$M3$	+81.0	+81.7	+115.2
$M4$	+65.7	+65.4	+92.8
$M5$	+40.1	+39.7	+56.6
$M6$	+13.7	+13.9	+19.7
$M7$	+3.1	+2.4	+4.4
$M8$	+0.0	+0.0	+0.0

10.4.2 *Haemophilus influenzae*

De même, chez *Haemophilus influenzae*, le motif `aagtgcggt` est également, tel que l'on s'y attendait, extrêmement sur-représenté (à part dans le modèle $M8$) :

modèle \ motif	`aagtgcggt`	`accgcactt`	`aagtgcggt` `accgcactt`
$M00$	+73.5	+73.3	+103.9
$M0$	+78.5	+78.0	+110.7
$M1$	+75.9	+75.5	+107.1
$M2$	+73.3	+72.5	+103.2
$M3$	+64.5	+63.8	+90.9
$M4$	+50.8	+49.6	+71.1
$M5$	+27.2	+27.3	+38.7
$M6$	+6.1	+7.0	+9.6
$M7$	+0.0	+0.0	+0.0

Références

[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.

[RM01] R.J. Roberts and D. Macelis. Rebase - restriction enzymes and methylases. *Nucleic Acids Research*, 29 :268–269, 2001.

[SGS99] H.O. Smith, M.L. Gwinn, and Salzberg S.L. Dna uptake signal sequences in naturally transformable bacteria. *Res. Microbiol.*, 150(9-10) :603–616, Nov-Dec 1999.

FIG. 11.1 – Représentation graphique des mots de longueur $h = 9$ chez *Homo sapiens* dans le modèle $M1$. On consultera la section 11.1 page 160 pour plus de détails.

Chapitre 11

Exploitation avancée

Pré-requis :
Chapitre 9.

Description :
Représentation et exploitation des résultats des grandes déviations.

Résumé :
Il est possible de représenter sur une seule image les résultats concernant la sur- ou sous-représentation des mots d'une longueur donnée en utilisant la *CGR* (*Chaos Game Representation*).

L'utilisation d'alphabets réduits pour le traitement de mots étant pour le moins hasardeuse, il faut s'efforcer de l'éviter.

L'alignement peut être employé comme outils pour former des *clusters* de mots exceptionnels de façons à éliminer le *bruit* produit par l'évolution.

Contenu du chapitre

11.1 Représentations graphique

Il peut être intéressant de représenter graphiquement les résultats concernant le caractère exceptionnel de l'ensemble des mots d'une longueur donnée. L'article [DGV⁺99] propose d'adopter la *Chaos Game Representation* (ou *CGR*) pour représenter les fréquences des mots d'une longueur donnée, et relie l'image obtenue à la notion de *signature génomique* ; de caractérisation de la spécificité d'un génome.

Le principe de la *CGR* est très simple : on commence par choisir une disposition arbitraire des quatre lettres de l'alphabet $\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$

$$CGR_1 = \begin{array}{|c|c|} \hline \mathtt{a} & \mathtt{c} \\ \hline \mathtt{g} & \mathtt{t} \\ \hline \end{array}$$

par exemple. Notons que ce choix va influer sur la résultat final (pour obtenir une représentation indépendante du choix initial, on pourra préférer une disposition spatiale sur les sommets d'un tétraèdre mais nous ne nous intéresserons pas ici à ce point).

Ce choix étant fait, il est désormais possible de faire correspondre tous les mots d'une longueur h donnée à une "case" précise d'une image carrée possédant 2^h lignes et colonnes. L'algorithme permettant de trouver cette correspondance pour un mot donné est de nature récursive :

- On initialise l'algorithme en désignant par I l'image entière et par W le mot entier ;
- Tant que W n'est pas vide :
  - On divise I en quatre partie égale nommées $I_\mathtt{a}$, $I_\mathtt{c}$, $I_\mathtt{g}$ ou $I_\mathtt{t}$ selon leur position (en accord avec la répartition initialement choisie) ;
  - On pose $I = I_{\text{première lettre de } W}$ et on élimine la première lettre de W ;
  - On itère.

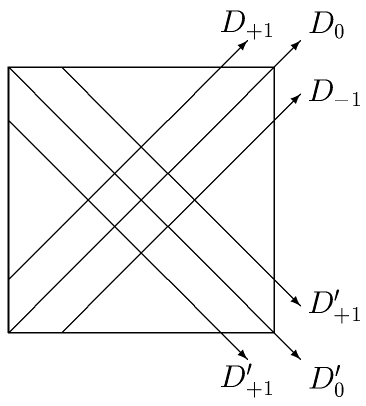


FIG. 11.2 – Interprétation de la disposition des mots dans la *CGR*. La diagonale D_0 (resp. D'_0) contient des mots ne comportant que des `g` ou `c` (resp. `a` ou `t`) tandis que les diagonales D_{-1} et D_{+1} (resp. D'_{-1} et D'_{+1}) comportent des mots dont toutes les lettres sauf une sont des `g` ou des `c` (resp. des `a` ou des `t`).

Ainsi, on aura, pour les mots de deux lettres, la répartition suivante :

$$CGR_1 \otimes CGR_1 = CGR_2 =$$

aa	ac	ca	cc
ag	at	cg	ct
ga	gc	ta	tc
gg	gt	tg	tt

et, pour ceux de trois lettres :

$$CGR_1 \otimes CGR_2 = CRG_3 =$$

aaa	aac	aca	acc	caa	cac	cca	ccc
aag	aat	acg	act	cag	cat	ccg	cct
aga	agc	ata	atc	cga	cgc	cta	ctc
agg	agt	atg	att	cgg	cgt	ctg	ctt
gaa	gac	gca	gcc	taa	tac	tca	tcc
gag	gat	gcg	gct	tag	tat	tcg	tct
gga	ggc	gta	gtc	tga	tgc	tta	ttc
ggg	ggt	gtg	gtt	tgg	tgt	ttg	ttt

Comme le montre la figure 11.2, le choix qui a été effectué positionne sur chaque diagonale les mots respectant dans leur composition certaines proportions en `g-c` ou `a-t`, tandis que les colonnes correspondent aux diverses proportions et dispositions des purines et pyrimidines dans chacun des mots ; la troisième colonne du tableau correspond par exemple aux mots s'écrivant purine-pyrimidine-purine.

Cette disposition étant acquise, les auteurs de [DGV$^+$99] proposent alors de donner à chacun des points de l'image une couleur d'autant plus foncée

génome	**figure**	**page**
Haemophilus influenzae	1.1	14
Mycoplasma genitalium	2.1	26
Methanococcus jannaschii	3.1	46
Saccharomyces cerevisiae	4.1	62
Escherichia coli	5.1	82
Bacillus subtilis	6.1	90
Caenorhabditis elegans	7.1	106
Lactococcus lactis	8.1	116
Neisseria meningitidis	9.1	126
Mycobacterium leprae	10.1	148
Homo sapiens	11.1	158

TAB. 11.1 – table des *CGR* disponibles.

que le mot correspondant est fréquent. Puisque nous disposons pour notre part d'une statistique $\mathcal{R}$ rendant compte de la sur- ou sous-représentation de chacun de ce mots, nous allons colorier notre image d'une manière légèrement différente : un point de l'image sera rouge s'il correspond à un mot sur-représenté et bleu s'il correspond à un mot sous-représenté, La nuance du point étant d'autant plus foncée que le mot est exceptionnel.

La table 11.1 contient la liste des *CGR* ainsi obtenues pour différent de génomes et leurs position dans cet ouvrage. Dans tous les cas, on peut remarquer que la structure récursive de la représentation se traduit par une "allure fractale" des images obtenues lorsque l'on considère des modèles markoviens d'ordres faibles. En effet, si un mot (court) est significativement sur-représenté, il en ira très certainement de même pour ces sur-mots dont la disposition dans l'image se verra répétée à différents niveaux.

Cette "allure fractale" disparaît cependant complètement lorsque l'on travaille dans le modèle markovien maximal ($Mh-2$ pour un mot de longueur h). On obtient alors des images sans caractéristiques visuelles évidentes (c'est pourquoi elle n'ont pas été représentées ici), mais qui constituent peut-être davantage une représentation non biaisée de la signature génomique des organismes étudiés.

11.2 Alphabets réduits

Ainsi qu'il l'a été suggéré en section 9.1 (page 128), il est parfois utile, d'un point de vu numérique, de réduire la taille de l'alphabet que l'on considère. Il est par exemple possible de considérer un alphabet protéique à 5 lettres au

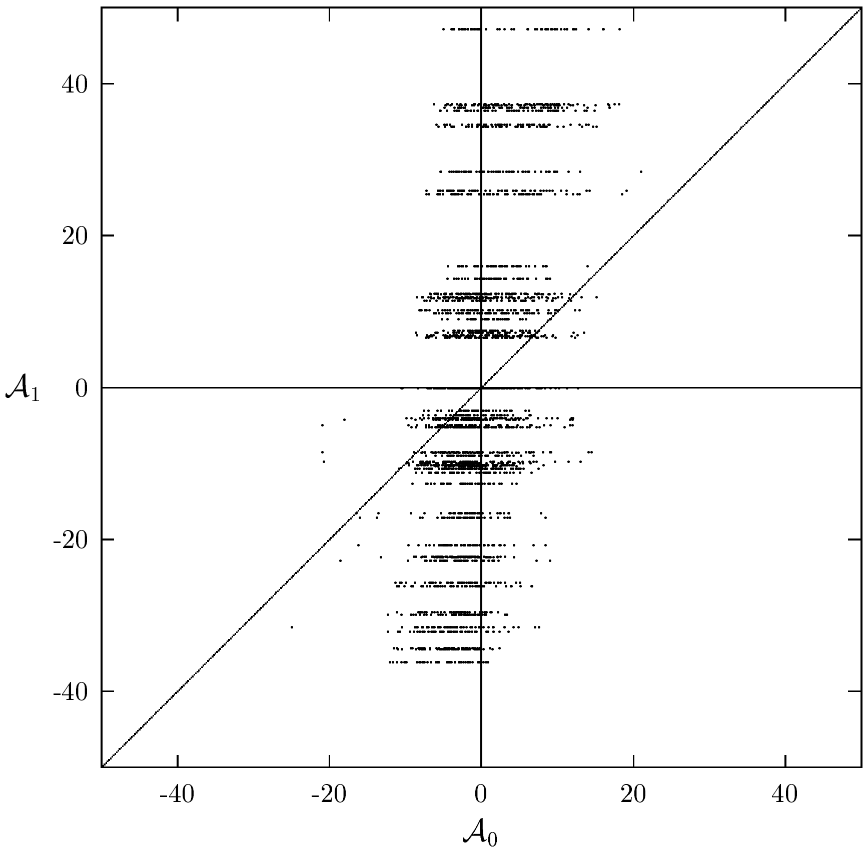


FIG. 11.3 – Comparaison des résultats de GDon avec les alphabets $\mathcal{A}_0 = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$ et $\mathcal{A}_1 = \{\mathtt{ag}, \mathtt{ct}\}$ pour les mots de longueur $h = 6$ de *Mycoplasma genitalium* dans $M1$.

lieu de 20 ou encore des alphabets nucléiques à 2 lettres au lieu de 4.

C'est ce dernier point que l'on va ici considérer en détail. On pose

$$\mathcal{A}_0 = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$$

et on définit l'alphabet purine-pyrimidine d'une part avec

$$\mathcal{A}_1 = \{\mathtt{ag}, \mathtt{ct}\}$$

et l'alphabet regroupant les liaison fortes et faibles d'autre part avec

$$\mathcal{A}_2 = \{\mathtt{at}, \mathtt{cg}\}.$$

Pour pouvoir évaluer l'erreur commise en travaillant avec ces informations résumées, on va comparer les résultats du programme GDon pour tous les mots de longueur $h = 6$ en se plaçant dans le modèle $M1$ successivement dans chacun de ces trois alphabets.

Sur la figure 11.3, on remarque en tout premier lieu une répartition horizontale des points qui ne doit cependant pas nous étonner outre mesure puisque cette répartition s'explique par le fait qu'il existe $2^6 = 64$ mots dans $\mathcal{A}_0$ pour chaque mot dans $\mathcal{A}_1$.

Ceci étant posé, il faut bien constater que la correspondance entre les résultats obtenus dans $\mathcal{A}_0$ et ceux obtenus dans $\mathcal{A}_1$ sont très loin d'être les mêmes. Néanmoins, on peut remarquer une certaine "tendance" à la sous-représentation des mots dans $\mathcal{A}_0$ d'un mot sous-représenté dans $\mathcal{A}_1$ (et on peut faire la même remarque pour les mots sur-représentés).

On peut effectuer les mêmes observations sur la figure 11.4 qui compare les résultats obtenus dans $\mathcal{A}_0$ à ceux obtenus dans $\mathcal{A}_2$. La question que l'on est naturellement en droit de se poser est : qu'en est-il de la relation entre les résultats obtenus dans les deux alphabets réduits ? La figure 11.5 effectue cette comparaison. On retrouve ici une répartition des points en ligne et en colonnes due aux répétitions déjà évoquées pour les figures précédentes qui sont, cette fois-ci, croisées.

On constate que les deux approches peuvent donner des résultats en totale contradiction : un mot peut être extrêmement sous-représenté dans $\mathcal{A}_1$ tout en étant sur-représenté dans $\mathcal{A}_2$. Si un tel comportement est caractéristique d'un phénomène particulier il nous est malheureusement inconnu. On peut donc conclure qu'il apparaît sage d'utiliser ces résultats avec la plus grande prudence et qu'il convient certainement de réserver ces approches aux mots et motifs dont le traitement n'est pas numériquement viable autrement.

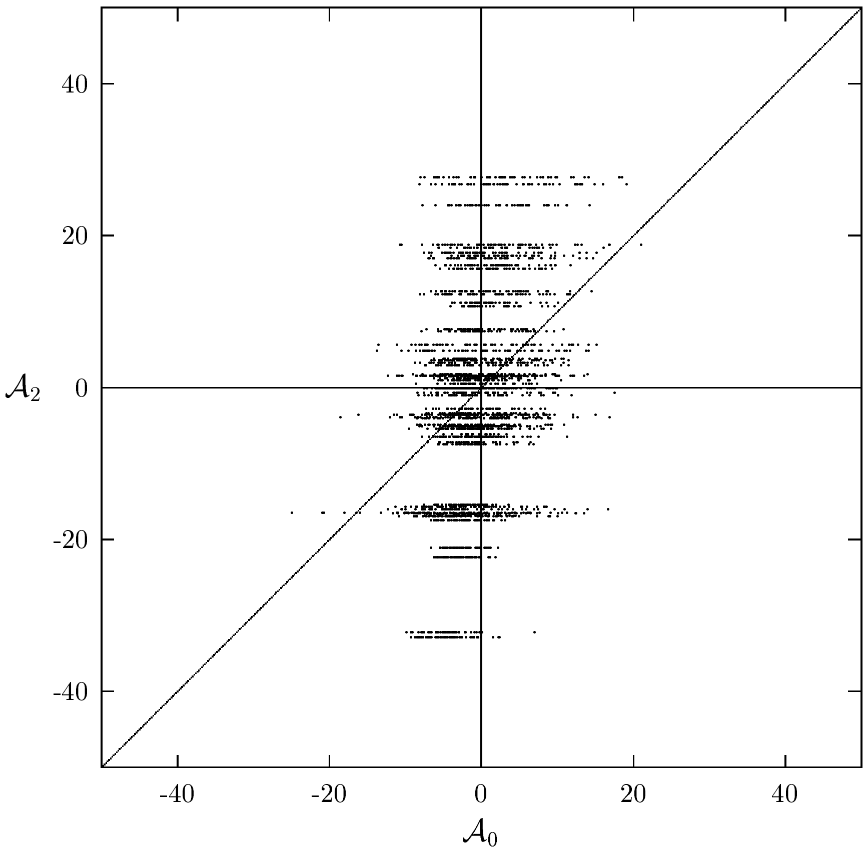


FIG. 11.4 – Comparaison des résultats de GDon avec les alphabets $\mathcal{A}_0 = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$ et $\mathcal{A}_2 = \{\mathtt{at}, \mathtt{cg}\}$ pour les mots de longueur $h = 6$ de *Mycoplasma genitalium* dans $M1$.

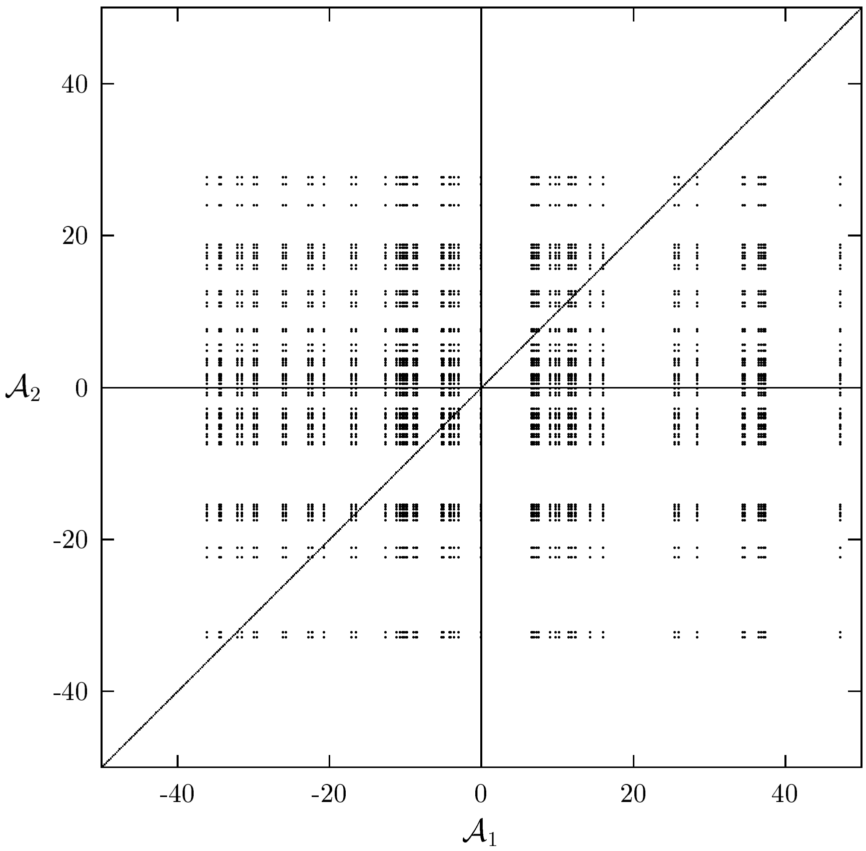


FIG. 11.5 – Comparaison des résultats de GDon avec les alphabets $\mathcal{A}_1 = \{\mathtt{ag}, \mathtt{ct}\}$ $\mathcal{A}_2 = \{\mathtt{at}, \mathtt{cg}\}$ pour les mots de longueur $h = 6$ de *Mycoplasma genitalium* dans $M1$.

mot	complémentaire	$\mathcal{R}_{gd}$
ggcgctgg	ccagcgcc	+52.4
cgccagcg	cgctggcg	+50.7
gctggcgg	ccgccagc	+48.5
gctggtgg	ccaccagc	+47.0
tgctggcg	cgccagca	+45.9
cttccagc	gctggaag	+45.4
ctggctgg	ccagccag	+44.3
tggcgctg	cagcgcca	+40.9
caccagcg	cgctggtg	+40.8
tcttccag	ctggaaga	+40.2
tcgccagc	gctggcga	+39.8
aactggcg	cgccagtt	+39.5
ctggcgat	atcgccag	+39.0
ctggcggc	gccgccag	+38.8
ctgccagc	gctggcag	+38.3
ctgctggc	gccagcag	+38.0
agcgccag	ctggcgct	+37.3
gcgctggc	gccagcgc	+36.9
gctggcgc	gcgccagc	+36.8
ctggcggt	accgccag	+36.3
ctggctga	tcagccag	+36.2
ctggcgaa	ttcgccag	+36.0
ttccagca	tgctggaa	+35.1
ccagcagc	gctgctgg	+35.1
tccgccag	ctggcgga	+34.7
tgctggtg	caccagca	+34.6
accaccag	ctggtggt	+34.6
ctggcagg	cctgccag	+34.1
tttccagc	gctggaaa	+33.9
tccagcgc	gcgctgga	+33.7

FIG. 11.6 – Les 30 mots de longueur $h = 8$ les plus sur-représentés chez *Escherichia coli* dans le modèle $M1$. Pour chaque mot, on étudie la famille formée par ce mot et son complémentaire inverse afin de traiter le caractère exceptionnel de ce mot sur les deux brins d'ADN en même temps. Le motif *chi* de l'organisme est souligné.

11.3 *Clusters* de mots

On a pu constater au chapitre 10 que les méthodes statistiques qui ont été mises au point donnent des résultats en accord avec ce que l'on attend lorsque l'on examine des mots exceptionnels du point de vue biologique.

Est-il possible d'utiliser ces résultats statistiques pour mettre en évidence des phénomènes biologiques jusqu'alors inconnus ? Peut-on utiliser ces approches pour créer de l'information ?

Dans le but de répondre affirmativement à ces questions, [Sch95] propose d'examiner tous les mots d'une longueur donnée et de classer les résultats en séparant les mots sur-représenté des mots sous-représentés et en les ordonnant par significativité décroissante. Il ne reste plus alors qu'à donner ces listes au biologistes pour qu'ils puissent y déceler les mots intéressants par expérience ou par homologie.

La figure 11.6 présente la liste des 30 mots de longueur $h = 8$ les plus sous-représentés dans le modèle $M1$ chez *Escherichia coli*. Le motif *chi* (`gctggtgg`) apparaît en quatrième position ce qui permet d'espérer qu'un traitement expérimental des premiers résultats aurait pu ici mettre en evidence l'activité du bon motif. Néanmoins, même le dernier mot de la liste de la figure 11.6 est désigné comme étant très intéressant par l'approche statistique et il est bien évident que les recherches expérimentales sont limitées dans le nombre d'essais que l'on peut y mettre en œuvre. L'idéal serait donc de limiter le champ d'investigation des biologistes en ne mettant en avant que les motifs ayant un fort potentiel.

On peut remarquer sur la figure 11.6 qu'un grand nombre de mots de la liste présentent une certaine ressemblance avec le motif *chi*. Il peut s'agir de mots recouvrant en partie ce motif (`ctggctgg` ou `tggcgctg`) ou encore de versions dégénérées (`gctgctgg` ou `gctggcgg`) de celui-ci.

Si un mot est nécessaire en grand nombre dans un génome pour assurer la survie de l'organisme auquel il appartient (c'est le cas du motif *chi* chez *Escherichia coli*), on peut s'attendre à ce que les mutations, insertions et délétions survenant lors de l'évolution de ce génome fasse apparaitre un nombre tout aussi important de versions dégénérées de ce motif. Bien que ces motifs dégénérés ne jouent pas de rôle particulier pour l'organisme, il est probable qu'une méthode statistique les identifie comme exceptionnels alors qu'il ne sont que les conséquences du phénomène que l'ont souhaite mettre en évidence.

De même, dans le cas de motifs dont la trop grande fréquence limite les chances de survie de l'organisme (cas des sites de restriction par exemple), les mutations auront également tendance à générer un nombre plus faible de versions dégénérées de ce motif que ce que l'on serait en droit d'attendre.

Dans les deux cas, un moyen d'éliminer ce "bruit" peut consister à grouper les mots désignés par la méthode statistique en différents *clusters* où l'on regroupe les mots qui s'alignent bien les uns avec les autres. Pour tenir compte de l'ordre dans lequel les mots sont sélectionnés par la méthode statistique, on propose ici l'utilisation d'un algorithme itératif.

Un cluster $\mathcal{C}$ donné contient r mots alignés les uns avec les autres (sans *gaps*) et en chaque position k on note $n_k(a)$ le nombre de a à cette position pour tout $a \in \mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$ et on désigne par $a_k \in \mathcal{A}$ le caractère majoritaire de chaque positions (on peut effectuer un choix aléatoire ou arbitraire en cas d'égalité car cela n'a pas d'incidence sur les calculs qui suivent).

Si on se donne `r` le score obtenu pour un *match* et `p` la pénalisation pour un *mismatch*, on peut alors définir le score $s(\mathcal{C})$ du *cluster* $\mathcal{C}$ de la façon suivante :

$$s(\mathcal{C}) = \sum_k \mathtt{r} \times n_k(a_k) - \mathtt{p} \times \left(\sum_{a \neq a_k} n_k(a) \right).$$

On considère une suite ordonnée de n mots et de leurs complémentaires $(W_i, \overline{W}_i)_{1 \leqslant i \leqslant n}$ que l'on cherche à répartir en *clusters* $(\mathcal{C}_j)_j$ dont c désigne le nombre à tout instant de l'algorithme alors :

- $i = 1$: On intialise l'algorithme avec $c = 1$ et $\mathcal{C}_1 = W_1$;
- $i \Rightarrow i + 1$:
  - Pour j variant de 1 à c, on aligne W_{i+1} dans $\mathcal{C}_j$ en maximisant le score de ce *cluster* si le *cluster* est compatible avec le mot ; c'est à dire si au moins un des mots de $\mathcal{C}_j$ possède au moins `min` *matchs* avec W_{i+1} et au plus `max` *mismatch*.
  - On note $s(W_{i+1})$ le meilleur score ainsi obtenu et $c(W_{i+1})$ l'indice du *cluster* correspondant (ces deux quantités sont mises à 0 si le mot ne s'aligne dans aucun des *clusters* existant) ;
  - On procède de la même façon pour $\overline{W}_{i+1}$ pour lequel on calcule $s(\overline{W}_{i+1})$ $c(\overline{W}_{i+1})$;
  - Si $c(W_{i+1}) \neq 0$ ou $c(\overline{W}_{i+1}) \neq 0$ alors on place le mot dont le score est le plus fort dans le *cluster* correspondant (en cas d'égalité, on privilégie le cluster comportant le plus de mot, puis le plus ancien et enfin l'odre alphabétique des mots) ;
  - Si $c(W_{i+1}) = c(\overline{W}_{i+1}) = 0$ alors on créé un nouveau *cluster* $\mathcal{C}_{c+1} = \{W_{i+1}\}$;
- $i = n$: L'algorithme est terminé.

La figure 11.7 représente le résultat de l'algorithme obtenu avec les paramètres `min` = 4, `max` = 2, `r` = 2 et `p` = 1. Voici la table des fréquences

```
      ggcgctgg
  cgctggcg
   gctggcgg
   gctggtgg
  tgctggcg
   gctggaag
ctggctgg
     tggcgctg
  cgctggtg
    ctggaaga
   gctggcga
  aactggcg
    ctggcgat
    ctggcggc
   gctggcag
 ctgctggc
    ctggcgct
 gcgctggc
   gctggcgc
    ctggcggt
    ctggctga
    ctggcgaa
  tgctggaa
gctgctgg
    ctggcgga
  tgctggtg
    ctggtggt
    ctggcagg
   gctggaaa
 gcgctgga
```

FIG. 11.7 – Seul *Cluster* formés à partir des 30 mots de longueur $h = 8$ les plus sur-représentés chez *Escherichia coli* dans le modèle $M1$. Les paramètres de l'algorithmes sont : $\texttt{min} = 4$, $\texttt{max} = 2$, $\texttt{r} = 2$ et $\texttt{p} = 1$.

d'apparition des lettres aux différentes positions du *cluster* :

`a`	0	0	1	1	0	0	0	0	5	6	4	4	0	0
`c`	1	2	4	0	28	0	0	0	19	0	4	1	0	0
`g`	1	2	1	17	0	0	30	30	0	18	11	1	2	1
`t`	0	1	5	0	0	29	0	0	4	1	0	6	0	0

On peut alors utiliser ce profil pour définir un motif consensus pour le *cluster* en éliminant les positions sans préférence marquée pour une lettre sur les autres et on obtient ainsi : `gctggcggt` dont la partie soulignée ressemble beaucoup au motif *chi* (une seule différence en sixième position).

Dans le cas de *Haemophilus influenzae* on s'intéresse aux 250 mots de longueur $h = 8$ les plus sur-représentés dans le modèle $M1$ et, avec les mêmes paramètres que ci-dessus, on trouve uniquement deux *clusters* contenant plusieurs mots dont voici les consensus :

`gaaagtgaggta[ag]a` et `ttcttcttc`

(`[ag]` désignant "`a` ou `g`"). Ici encore, on peut constater l'efficacité de cette approche puisque, à une lettre près, on retrouve dans le premier consensus le motif *uptake* (`aagtgcggt`) qui est ainsi bien mis en évidence.

Il est remarquable de constater qu'on a pu isoler des mots de longueurs bien supérieures (jusqu'à $h = 14$) à celles de ceux initialement étudiés ($h = 8$) ce qui constitue un avantage important à la méthode. En effet, il est important de noter que :

- la complexités des calculs effectués par GDon augmente géométriquement avec la longueur des motifs ;
- le nombre de mots d'une longueur donnée augmente de la même façon ce qui risque d'interdire l'étude systématique de chacun de ces mots ;

et on comprend donc immédiatement l'intérêt de limiter la longueur des mots à étudier.

Même si l'algorithme assez "naïf" qui a été mis en place ici n'est pas nécessairement le plus adapté au problème qui nous préoccupe, il permet néanmoins d'illustrer la possibilité d'utiliser de l'alignement pour éliminer le "bruit" que l'on observe usuellement dans l'étude des mots exceptionnels et constitue très certainement une "voie" à ne pas négliger si l'on désire véritablement créer de l'information.

Références

[DGV+99] P.J. Deschavanne, A. Giron, J. Vilain, G. Fagot, and B. Fertil. Genomic signature : characterization and classification of species assessed by chaos game representation of sequences. *Mol Biol Evol.*, 16(10) :1391–1399, 1999.

[Sch95] S. Schbath. *Etude asymptotique du nombre d'occurrences d'un mot dans une chaîne de Markov et application à la recherche de mots de fréquence exceptionnelle dans les séquences d'ADN*. PhD thesis, PARIS V, 1995.

Cinquième partie

Conclusion

Ce travail a permis d'effectuer la synthèse de résultats théoriques pour les grandes déviations de niveaux 1 et 2 du nombre d'occurrences d'un mot ou d'un motif sur une chaîne de *Markov* d'ordre quelconque. Les moyens pratiques de mise en œuvre de ces résultats ont été développés et, dans le cas du niveau 1, ont conduit à la création d'un outils informatique opérationnel : GDon.

La validation de l'ensemble de la démarche a pu être obtenue en comparant les résultats de ce programme avec ceux des programmes existants. Ainsi que l'on pouvait s'y attendre, il ressort de cette validation que les grandes déviations constituent un outils particulièrement bien adapté à l'étude des événements rares tandis que la qualité des approximations décroît lorsque la fréquences des événements augmente. Des méthodes comme les approximations gaussienne proposées par R'MES, montrent le comportement inverse et font ainsi la preuve de leur manque total de précision pour l'étude des mots exceptionnels.

Les versions *Poisson* composée de R'MES ainsi que les approches exactes de REGEXPCOUNT s'avèrent, pour leurs parts, d'une grande précision en toutes circonstances mais comportent de nombreuses limitations numériques dans leurs implémentations présentes :

- pas de traitement de motifs pour R'MES *Poisson* composée ;
- ordres des modèles limités pour REGEXPCOUNT ;
- nombre d'occurrences des mots considérés influant sur le temps de calculs des deux programmes ;
- impossibilités de traiter certains motifs trop complexes avec REGEXPCOUNT et certains mots avec R'MES.

GDon constitue donc une alternative intéressante à ces méthodes. Avec un espace mémoire en $O(k^h)$ (k cardinal de l'alphabet et h taille du motif), cette approche est cependant, elle aussi, confrontée à des limitations (mots de longueur 12 dans un alphabet à 4 lettres par exemple). En contre-partie, la complexité numérique des calculs reste indépendante de la longueur des séquences, de l'ordre du modèle et de la complexité des motifs considérés.

Par manque de temps, les résultats de grandes déviations de niveau 2 n'ont pu être implémentés, mais le fort potentiel de cette approche ne doit pas rester inexploité. Les lois jointes des comptages de mots étant accessibles par cette approche, elles représentent en effet le moyen d'effectuer des calculs conditionnels dont l'intérêt sur le plan biologique est évident. De plus, il est possible que les calculs mis en œuvre dans le niveau 2 soient numériquement plus viables (du moins dans la version heuristique de ces calculs, c'est à dire en négligeant la projection orthogonale) que ceux du niveau 1.

Le changement de probabilités qui intervient dans les grandes déviations et auquel on a explicitement accès au cours des calculs du niveau 1 présente un grand intérêt pour le domaine des simulations. Avec cette nouvelle probabilité on rend en effet possible l'observation d'événements de faibles probabilités qui ne pourraient être observés autrement. Pour l'étude de certains motifs complexes particulièrement long et dégénérés, le recours aux simulations reste bien souvent le seul outil valable et, pour ces motifs, l'usage systématique des changements de probabilités pourrait permettre de limiter de façon très importante le nombre des simulations à effectuer pour un coût numérique quasi nul.

Enfin, on a pu montrer la difficulté que représente l'interprétation des résultats lorsque l'on s'efforce de créer de l'information (et non plus simplement de valider des résultat biologiques connus). Le recours à l'alignement des mots exceptionnels dans le but de former des motifs pertinents et d'éliminer le "bruit" de l'évolution semble très prometteur et il est certain que cette approche assez novatrice mérite d'être explorée plus avant.

Sixième partie

Annexes

Annexe A

Théorie générale des grandes déviations

Contenu du chapitre

A.1 Duale de *Legendre*

On désigne par $\langle \cdot, \cdot \rangle$ un produit scalaire du $\mathbb{R}^d$ et on considère

$$\begin{array}{rccc} f: & \mathbb{R}^d & \to & \overline{\mathbb{R}} \\ & \theta & \mapsto & f(\theta) \end{array}$$

dont on peut alors définir la *duale de Legendre* :

$$\begin{array}{rccc} f^*: & \mathbb{R}^d & \to & \overline{\mathbb{R}} \\ & x & \mapsto & f^*(x) \end{array}$$

avec

$$f^*(x) = \sup_{\theta \in \mathbb{R}^d} \left\{ \langle \theta, x \rangle - f(\theta) \right\}. \tag{A.1}$$

Avant d'aller plus loin, on introduit les deux définitions suivantes :

Définition A.1 *On considère $\mathcal{X}$ un* espace polonais *(espace topologique séparé) et $f : \mathcal{X} \to \overline{\mathbb{R}}$ une fonction quelconque alors on dit :*

(i) f *est* propre *si* $\exists x \in \mathcal{X}$ *tel que* $|f(x)| \neq +\infty$ *;*

(ii) f *est* semi-continue inférieurement *(*s.c.i*)* *si ses ensembles de niveaux sont fermés :*

$$f^{-1}(]-\infty, c[) = \{x \in \mathcal{X}, f(x) \leq c\} \text{ est fermé } \forall c \in \mathbb{R}.$$

On peut maintenant énoncer la

Proposition A.2 *On a les propriétés suivantes :*

(i) *si f est propre alors f^* est convexe et s.c.i. ;*

(ii) *si f est convexe et s.c.i alors $f^{**} = f$.*

En prenant des hypothèses plus restrictives sur la nature de f, on peut obtenir des résultats intéressants concernant sa duale de *Legendre* :

Proposition A.3 *On suppose que $f : \mathbb{R} \to \overline{\mathbb{R}}$ est convexe et que $f \in \mathcal{C}^1$ alors $\mathcal{D}_f \triangleq \{t \in \mathbb{R}, f(t) < +\infty\} =]a, b[$ et $\mathcal{D}_{f'} \triangleq \{t \in \mathbb{R}, f'(t) < +\infty\} =]\alpha, \beta[$ avec $a, b, \alpha, \beta \in \overline{\mathbb{R}}$ et si $x \in \mathbb{R}$ alors*

(i) *si $x \leq \alpha$ (en particulier $\alpha \in \mathbb{R}$ nécessairement) alors*

$$f^*(x) = \left\{ \begin{array}{ll} +\infty & \text{si } a = -\infty \text{ et } x < \alpha \\ \lim\limits_{t \to a} \{tx - f(t)\} & \text{sinon} \end{array} \right. ;$$

(ii) *si $x \in]\alpha, \beta[$, $\exists \tau \in \mathbb{R}$ tel que $f'(\tau) = x$ et on a*

$$f^*(x) = \tau x - f(\tau);$$

(iii) *si $x \geq \beta$ (en particulier $\beta \in \mathbb{R}$ nécessairement) alors*

$$f^*(x) = \left\{ \begin{array}{ll} +\infty & \text{si } b = +\infty \text{ et } x > \beta \\ \lim\limits_{t \to b} \{tx - f(t)\} & \text{sinon} \end{array} \right. .$$

Preuve. f' est continue et monotone sur $]a, b[$ donc son image est clairement un intervalle ouvert de $\mathbb{R}$ que l'on choisit ici de désigner par $]\alpha, \beta[$.

Soit $x \in \mathbb{R}$, pour effectuer le calcul de $f^*(x)$ il convient d'étudier $f_x : t \mapsto tx - f(t)$ pour en déterminer le sup. On étudie simplement les variations de f_x et on distingue trois cas :

cas 1 Si $x \leq \alpha$ alors f_x est monotone décroissante sur $]a, b[$ et donc

$$f^*(x) = \lim_{t \to a} tx - f(t)$$

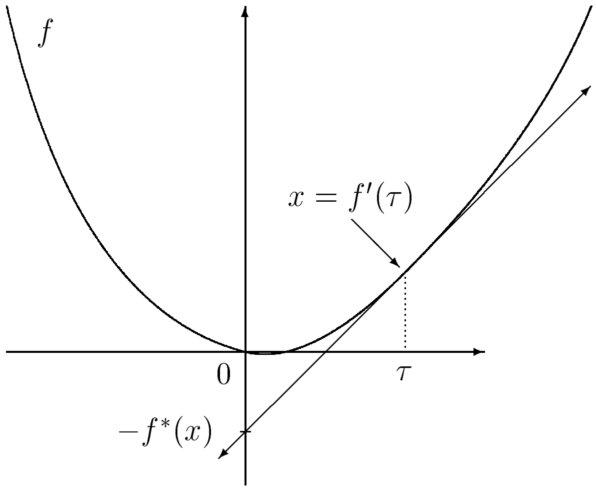


FIG. A.1 – Interprétation géométrique de la duale de *Legendre*

si, de plus, $a = -\infty$ et que que $x < \alpha$ alors

$$\begin{aligned} f^*(x) &= \lim_{t \to -\infty} t \left[(x - \alpha) + \left(\alpha - \frac{f(t)}{t} \right) \right] \\ &= +\infty \end{aligned}$$

car $x - \alpha < 0$ et $\frac{f(t)}{t} \xrightarrow[t \to -\infty]{} \alpha$. On a ainsi achevé la preuve de **(i)**.

cas 2 si $\alpha < x < \beta$, f_x est croissante puis décroissante et atteint son maximum en un $\tau \in \mathbb{R}$ tel que $f'(\tau) = x$ ce qui prouve **(ii)** (voir figure A.1).

cas 3 Enfin si $\beta \leq x$ alors on travaille de manière symétrique au premier cas et on montre ainsi facilement **(iii)**.

■

A.2 Principe de grandes déviations

On considère $\mathcal{X}$, un espace polonais quelconque, que l'on dote d'une structure d'espace probabiliste. On pose d'abord la définition suivante :

Définition A.4 (Fonction de taux)
On dit que $I : \mathcal{X} \to [0, +\infty]$, *non identiquement égale à* $+\infty$, *est une* fonction de taux *si elle est semi-continue inférieurement (voir définition A.1). Une fonction de taux est dite* bonne *si ses ensembles de niveaux sont compacts (en plus d'être fermés).*

Après quoi on peut énoncer la

Définition A.5 (Principe de Grandes Déviations)
Une suite $(\mathbb{P}_n)$ *de mesures sur* $\mathcal{X}$ *satisfait un* Principe de Grandes Déviations *(*PGD*) de vitesse* n *et de fonction de taux* I *si on a*

(MAJ) $\overline{\lim\limits_{n\to+\infty}} \frac{1}{n} \log \mathbb{P}_n(F) \leq -\inf\limits_F I \quad \forall F$ *fermé* $\subset \mathcal{X}$ *;*

(MIN) $\underline{\lim\limits_{n\to+\infty}} \frac{1}{n} \log \mathbb{P}_n(O) \geq -\inf\limits_O I \quad \forall O$ *ouvert* $\subset \mathcal{X}$*.*

Par extension de cette définition on dira qu'une suite (X_n) *de variables aléatoire sur* $\mathcal{X}$ *satisfait un* PGD *de fonction de taux* I *si* $(\mathbb{P}(X_n \in \cdot))$ *le satisfait, c'est à dire si*

(MAJ) $\overline{\lim\limits_{n\to+\infty}} \frac{1}{n} \log \mathbb{P}(X_n \in F) \leq -\inf\limits_F I \quad \forall F$ *fermé* $\subset \mathcal{X}$ *;*

(MIN) $\underline{\lim\limits_{n\to+\infty}} \frac{1}{n} \log \mathbb{P}(X_n \in O) \geq -\inf\limits_O I \quad \forall O$ *ouvert* $\subset \mathcal{X}$*.*

Cette définition correspond de façon intuitivement à la propriété suivante : si $\Gamma \subset \mathcal{X}$ alors on a

$$\mathbb{P}(X_n \in \Gamma) \sim e^{-n \inf\limits_{\Gamma} I}.$$

Remarquons de plus que cette information ne présente d'intérêt que lorsque la vitesse de décroissance exponentielle $(\inf_\Gamma I)$ est non nulle (car sinon on apprend que la probabilité d'un événement est de l'ordre de 1 ce qui n'est pas très utile).

Cette définition formelle, si elle nous éloigne de l'intuition de la notion de grandes déviations telle qu'elle était introduite au chapitre 4 (page 63), présente l'avantage de permettre la construction d'une théorie unifiant les différentes techniques provenant *a priori* de sources très variées.

A.3 Théorème de *Gärtner-Ellis*

Il existe un résultat général permettant l'établissement de principes de grandes déviations pour des suites de variables aléatoires dépendantes, il s'agit du théorème de *Gärtner-Ellis*.

On considère une suite (Z_n) de variables aléatoires dans $\mathbb{R}^d$. On désigne par $\langle \cdot, \cdot \rangle$ un produit scalaire sur $\mathbb{R}^d$ et on note φ la transformée de *Laplace* de Z_n définie par

$$\varphi_n(\theta) = \mathbb{E}[e^{\langle \theta, Z_n \rangle}] \quad \forall \theta \in \mathbb{R}^d$$

On fait alors les hypothèses suivantes :

$$\begin{aligned} &\textbf{(i)} \quad \lim_{n\to+\infty} \log \varphi_n(n \cdot \theta) = \Lambda(\theta) \in [-\infty, +\infty]; \\ &\textbf{(ii)} \quad 0 \in \overset{\circ}{\mathcal{D}_\Lambda} \ \text{ avec } \mathcal{D}_\Lambda = \{\theta \in \mathbb{R}^d, \Lambda(\theta) < +\infty\}; \end{aligned} \tag{A.2}$$

où $\overset{\circ}{\mathcal{D}}_\Lambda$ désigne l'*intérieur* de $\mathcal{D}_\Lambda$; on ajoute des conditions de régularité sur Λ : Λ est différentiable sur $\mathcal{D}_\Lambda = \mathbb{R}^d$ ou, à défaut,

$$\begin{array}{ll} \textbf{(i)} & \Lambda \text{ est s.c.i. sur } \mathbb{R}^d \text{ (voir définition A.4);} \\ \textbf{(ii)} & \Lambda \text{ est différentiable sur } \overset{\circ}{\mathcal{D}}_\Lambda; \\ \textbf{(iii)} & \Lambda \text{ est } \textit{escarpée} \left(\lim\limits_{\theta \to \theta_0} |\nabla \Lambda(\theta)| = +\infty, \forall \theta_0 \in \partial \mathcal{D}_\Lambda \right); \end{array} \tag{A.3}$$

où $\partial\mathcal{D}_\Lambda$ désigne le *bord* de $\mathcal{D}_\Lambda$.

Théorème A.6 *Si les hypothèses (A.2) et (A.3) sont vérifiées alors* (Z_n) *suit un PGD de bonne fonction de taux* Λ^* *où* Λ^* *désigne la duale de* Legendre *de* Λ *(voir section A.1).*

Cette version du théorème de *Gärtner-Ellis* est moins générale que la version “usuelle” qui se contente de l'hypothèse (A.2) pour donner un résultat très proche : seule la minoration sur les ouverts est légèrement modifiée (le minimum de la fonction de taux est pris sur un sous-ensemble de l'ouvert). La version simplifiée proposée ici est cependant suffisante pour les différents cas auxquels on souhaite appliquer le résultat.

Afin d'illustrer le théorème, on considère l'exemple suivant : soit (X_n), un échantillon de la variable aléatoire X à valeurs dans $\mathbb{R}$ (pour simplifier). On a alors le

Corollaire A.7 (***Cramér-Chernov***)
On suppose que

$$\varphi(t) = \mathbb{E}[e^{tX}] < +\infty \quad \forall t \in \mathbb{R} \tag{A.4}$$

alors $\left(\frac{X_1+\ldots+X_n}{n}\right)_n$ *suit un PGD de bonne fonction de taux* I*, où* I *est la transformée de* Cramér *(duale de* Legendre *de la* log-Laplace*) de* X*.*

Preuve. On pose

$$Z_n = \frac{S_n}{n} \quad \text{avec} \quad S_n = \sum_{i=1}^{n} X_i$$

et on calcule pour tout $t \in \mathbb{R}$

$$\begin{aligned} \varphi_n(n \cdot t) &= \mathbb{E}[e^{ntZ_n}] \\ &= \mathbb{E}[e^{tS_n}] \\ &= \varphi(t)^n \end{aligned}$$

et donc

$$\lim_{n\to+\infty} \frac{1}{n} \log \varphi_n(n \cdot t) = \log \varphi(t) = \Lambda(t) < +\infty \quad \forall t \in \mathbb{R}.$$

La condition (A.4) nous assurant que φ , donc Λ, est $\mathcal{C}^\infty$ sur $\mathbb{R}$ (voir la section D.1.1 page 211 pour s'en convaincre), les hypothèses (A.2) et (A.3) sont clairement vérifiées et le théorème A.6 s'applique pour obtenir le corollaire. ■

Le résultat ainsi obtenu est cependant différent du résultat dit de *Cramér-Chernov* que donne le théorème 4.5 (page 69). En fait, le résultat présent se trouve être plus général que celui du théorème 4.5 comme le prouve le

Corollaire A.8 *Si $a > \mathbb{E}[X]$ alors*

$$\lim_{n\to+\infty} \frac{1}{n}\mathbb{P}(S_n \geq na) = -\Lambda^*(a).$$

Preuve. On va successivement appliquer les relations **(MAJ)** et **(MIN)** de la définition A.5 à $[a,+\infty[$ et à $]a,+\infty[$. Comme Λ^* est positive, continue et strictement convexe et que $\Lambda^*(m) = 0$ (avec $m = \Lambda'(0) = \mathbb{E}[X]$), il est clair que

$$\inf_{[a,+\infty[} \Lambda^* = \inf_{]a,+\infty[} \Lambda^* = \Lambda^*(a)$$

ce qui achève la preuve. ■

A.4 Entropie

On introduit ici la notion d'*entropie relative* qui apparaît bien souvent dans les grandes déviations.

Définition A.9 (Entropie relative)
Si ν et μ sont deux lois sur l'alphabet fini $\mathcal{A}$ et que $\mu \gg 0$ alors, avec la convention $0\log 0 = 0$*, on définit $H(\nu|\mu)$ l'entropie de ν relativement à μ par*

$$H(\nu|\mu) = \sum_{x\in\mathcal{A}} \nu(x)\log\frac{\nu(x)}{\mu(x)}.$$

Proposition A.10 *Si $\mu \in \mathcal{M}_1(\mathcal{A})$ vérifie $\mu \gg 0$ alors*

(i) *$\nu \mapsto H(\nu|\mu)$ est une fonction convexe ;*

(ii) *$H(\nu|\mu) \geqslant 0$ avec égalité si et seulement si $\nu = \mu$.*

Preuve. On commence par poser $h(t) = t\log t$ pour tout $t \in \mathbb{R}$ et on remarque que

$$H(\nu|\mu) = \sum_{x\in\mathcal{A}} h\left(\frac{\nu(x)}{\mu(x)}\right) \times \mu(x).$$

Montrons d'abord **(i)**. Soient $\mu, \mu' \in \mathcal{M}_1(\mathcal{A})$ et soit $\lambda \in [0,1]$, on a

$$\begin{aligned} H\big(\lambda\nu + (1-\lambda)\nu'|\mu\big) &= \sum_{x\in\mathcal{A}} h\left(\lambda\frac{\nu(x)}{\mu(x)}(1-\lambda)\frac{\nu'(x)}{\mu(x)}\right) \times \mu(x) \\ &\leqslant \sum_{x\in\mathcal{A}} \left[\lambda h\left(\frac{\nu(x)}{\mu(x)}\right) + (1-\lambda)h\left(\frac{\nu'(x)}{\mu(x)}\right)\right] \times \mu(x) \\ &\leqslant \lambda H(\nu|\mu) + (1-\lambda)H(\nu'|\mu) \end{aligned}$$

grâce à la convexité de $t \mapsto h(t)$.

Pour le point **(ii)**, on remarque que $h(t) \geqslant t-1$ pour tout $t \geqslant 0$ avec égalité si et seulement si $t=1$. Ainsi on a $\forall x \in \mathcal{A}$

$$h\left(\frac{\nu(x)}{\mu(x)}\right) \geqslant \frac{\nu(x)}{\mu(x)} - 1$$

si bien qu'en sommant sur x chacun de ces termes multiplié par $\mu(x)$ on trouve :

$$H(\nu|\mu) = \sum_{x\in\mathcal{A}} \nu(x) - \mu(x) = 1 - 1 = 0.$$

Si $\exists x \in \mathcal{A}$ tel que

$$h\left(\frac{\nu(x)}{\mu(x)}\right) > \frac{\nu(x)}{\mu(x)} - 1$$

il est de la même façon clair que $H(\nu|\mu) > 0$, de sorte que la deuxième partie de **(ii)** est également démontrée. ■

Remarque A.11 *Remarquons que les résultats de la proposition A.10 tiennent toujours si on généralise la définition A.9 à*

$$H(\nu|\mu) = \sum_{x\in\mathcal{A},\mu(x)>0} \nu(x)\log\frac{\nu(x)}{\mu(x)}.$$

et on utilisera bien souvent cette définition en lieu et place de la première.

A.5 Lemme de *Varadhan*

Dans les grandes déviations, la technique clé consiste souvent à effectuer un changement de probabilité (voir preuve section D.1.1 page 211). Le résultat suivant, donne un cadre formel à cette technique en permettant l'établissement d'un PGD à partir d'un PGD existant et d'un changement de mesure.

Théorème A.12 (***Varadhan***)
On suppose que $(\mathbb{P}_n)$ *satisfait un PGD de fonction de taux* I. *On considère* $F : \mathcal{X} \to \mathbb{R}$ *une fonction continue majorée alors la suite* $(\widehat{\mathbb{P}}_n)$ *définie par*

$$\widehat{\mathbb{P}}_n(S) = \frac{1}{C}\int_S e^{nF(x)}\mathbb{P}_n(dx) \quad \forall S \subset \mathcal{X} \textit{ mesurable}$$

avec

$$C = \int_{\mathcal{X}} e^{nF(x)}\mathbb{P}_n(dx),$$

satisfait un PGD de fonction de taux $\widehat{I}$ *définie par*

$$\widehat{I}(x) = \sup_{y\in\mathcal{X}}[F(y) - I(y)] - [F(x) - I(x)].$$

En exemple d'application, considérons le cas suivant : (X_n) est un échantillon de taille n de X variable aléatoire réelle de densité f. On suppose que $\mathbb{E}[e^{tX}] < +\infty$ pour tout $t \in \mathbb{R}$ et on considère $a > m = \mathbb{E}[X]$ alors les conditions d'application du corollaire 4.10 (page 73) sont réunies. Si on a $\tau \in \mathbb{R}$ tel que $\Lambda'(\tau) = a$ on considère la densité

$$\widehat{f}(x) = \frac{1}{\varphi(\tau)}e^{\tau x}f(x)$$

où

$$\varphi(\tau) = \mathbb{E}[e^{\tau X}] = \int_{\mathbb{R}} e^{\tau x}f(x)dx$$

et, si $\widehat{X}$ est une variable aléatoire de densité f, on a $\mathbb{E}[\widehat{X}] = a$. On considère $(\widehat{X}_n)$ un échantillon de $\widehat{X}$ et on s'intéresse au comportement de $\widehat{S}_n = \widehat{X}_1 + \ldots + \widehat{X}_n$. On note $\mathbb{P}_n(\cdot) = \mathbb{P}(\frac{S_n}{n} \in \cdot)$, $\widehat{\mathbb{P}}_n(\cdot) = \mathbb{P}(\frac{\widehat{S}_n}{n} \in \cdot)$ et on a, pour toute partie $S \subset \mathbb{R}$ mesurable

$$\begin{aligned}
\widehat{\mathbb{P}}_n(S) &= \mathbb{P}(\frac{\widehat{S}_n}{n} \in S) \\
&= \int_{\mathbb{R}^n} \mathbb{I}_{\frac{x_1+\ldots+x_n}{n}\in S}\widehat{f}(x_1)\ldots\widehat{f}(x_n)dx_1\ldots dx_n \\
&= \frac{1}{\varphi(\tau)^n}\int_{\mathbb{R}^n} \mathbb{I}_{\frac{s_n}{n}\in S}e^{\tau s_n}f(x_1)\ldots f(x_n)dx_1\ldots dx_n \\
&= \frac{\int_S e^{nF(x)}\mathbb{P}_n(dx)}{\int_{\mathbb{R}} e^{nF(x)}\mathbb{P}_n(dx)}
\end{aligned}$$

avec $F(x) = \tau x$.

Le corollaire A.7 montre que $(\mathbb{P}_n)$ suit un PGD de fonction de taux Λ^* duale de Legendre de Λ définie par

$$\Lambda(t) = \log \mathbb{E}(e^{tX}) \quad \forall t \in \mathbb{R}$$

et le théorème A.12 nous permet alors d'affirmer que $(\widehat{\mathbb{P}}_n)$ suit un PGD de fonction de taux

$$\widehat{I}(x) = \sup_{y \in \mathbb{R}} [\tau y - \Lambda^*(y)] - [\tau x - \Lambda^*(x)].$$

En particulier,

$$\lim_{n \to +\infty} \frac{1}{n} \log \mathbb{P}(\widehat{S}_n \geq na) = 0 \tag{A.5}$$

car

$$\begin{aligned} \widehat{I}(a) &= \sup_{y \in \mathbb{R}} [\tau y - \Lambda^*(y)] - [\tau a - \Lambda^*(a)] \\ &= \Lambda^{**}(\tau) - \tau a - [\tau a - \Lambda(\tau)] \\ &= \Lambda^{**}(\tau) - \Lambda(\tau) \end{aligned}$$

qui est bien nul dès lors que Λ est "assez régulière" ce qui est le cas ici (voir [RW98]).

Notons que le résultat (A.5) n'est guère étonnant. En effet, $\mathbb{E}[\widehat{X}] = a$ et par conséquent, la loi des grands nombres nous assure que

$$\frac{\widehat{S}_n}{n} \xrightarrow[n \to +\infty]{} a \quad \text{p.s.}$$

Ainsi on retrouve via *Varadhan* un résultat attendu mais on comprend bien l'utilisation que l'on pourra faire d'un tel résultat en général : utiliser le théorème pour établir des PGD dans des cas complexes à partir de cas plus simples. On se reportera à la section 7.2 (page 111) pour découvrir une illustration de cette idée.

Références

[Buc90] J. A. Bucklew. *Large deviation techniques in decision, simulation, and estimation.* Wiley, 1990.

[DZ98] A. Dembo and O. Zeitouni. *Large deviations techniques and applications.* Springer-Verlag, 1998.

[RW98] R. T. Rockafellar and R. J-B. Wets. *Variational analysis.* Princeton University Press, 1998.

[Tor98] N. Torrent. *Application des grandes déviations et de la loi d'Erdös-Réyni pour des variables indépendantes ou de dépendance markovienne.* PhD thesis, Université PARIS VII, 1998.

Annexe B

Compléments

Contenu du chapitre

B.1 Théorème de *Perron-Frobénius*

Cette travail constitue une synthèse de différents ouvrages (notamment [Gan90], [Kar69], [LT85] et [Sen81]) portant sur l'étude du spectre des matrices positives.

On commence par poser les notations qui vont revenir tout au long de cette section :

Notation B.1 *Soit* $\mathbf{A} = (a_{ij}) \in \mathcal{M}_{p,q}(\mathbb{R})$ *une matrice réelle. On note :*
- $\mathbf{A} \geq \mathbf{0}$ *si* $a_{ij} \geq 0$ *pour tout* i, j *;*
- $\mathbf{A} > \mathbf{0}$ *si* $\mathbf{A} \geq \mathbf{0}$ *et* $\mathbf{A} \neq \mathbf{0}$ *;*
- $\mathbf{A} \gg \mathbf{0}$ *si* $a_{ij} > 0$ *pour tout* i, j*.*

Si $\mathbf{B} = (b_{ij}) \in \mathcal{M}_{p,q}(\mathbb{R})$ *est une autre matrice on note* $\mathbf{A} \geq \mathbf{B}$ *(resp.* $>$ *et* $\gg$*) si* $\mathbf{A} - \mathbf{B} \geq \mathbf{0}$ *(resp.* $>$ *et* $\gg$*).*

On peut ensuite énoncer le premier résultat :

Théorème B.2 (*Perron-Frobénius* fort)
Si $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$ *est une matrice* $A > 0$ *telle qu'il existe* $m \geq 1$ *telle que* $\mathbf{A}^m \gg \mathbf{0}$ *(c'est à dire que* $\mathbf{A}$ *est* primitive*) alors*

(i) $\mathbf{A}$ *admet une valeur propre* $\rho > 0$ *dont un vecteur propre associé est à coordonnées toutes strictement positives ;*

(ii) *ρ est une valeur propre de multiplicité* 1 *;*

(iii) *si $\lambda \neq \rho$ est une autre valeur propre de* $\mathbf{A}$ *alors* $|\lambda| < \rho$.

On appelle ρ la valeur propre de Frobénius *de* $\mathbf{A}$.

Pour prouver ce théorème on introduit l'ensemble

$$\mathcal{S} = \left\{ \mathbf{X} = [x_i] \in \mathbb{R}^n,\ \mathbf{X} \geq \mathbf{0} \text{ et } \sum_i x_i = 1 \right\}.$$

Cet ensemble est clairement un fermé borné de $\mathbb{R}^n$ et par conséquent un compact, de plus il est également clair qu'un élément $\mathbf{X}$ de $\mathcal{S}$ vérifie $\mathbf{X} > \mathbf{0}$.

On pose également

$$\Lambda = \{\lambda \in \mathbb{R}, \exists \mathbf{X} \in \mathcal{S} \text{ tel que } \mathbf{AX} \geq \lambda \mathbf{X}\}.$$

qui est un ensemble non vide puisque $0 \in \Lambda$.

Preuve de (i). On va montrer que Λ admet une borne supérieure et qu'il s'agit de ρ. Pour cela, on va commencer par montrer l'existence d'un majorant à l'ensemble Λ, puis on vérifiera que le plus petit majorant se trouve précisément être le ρ du théorème.

Soit $\lambda \in \Lambda$ et soit $\mathbf{X} \in \mathcal{S}$ tel que $\mathbf{AX} \geq \lambda \mathbf{X}$. Cette inégalité se traduit par les relations suivantes :

$$\sum_{j=1}^{n} a_{ij} x_j \geq \lambda x_i \quad \forall i$$

qui peuvent se sommer pour obtenir

$$\lambda \leq \sum_{i,j} a_{ij} x_j.$$

Comme par ailleurs $\mathbf{X} \in \mathcal{S} \Rightarrow x_j \leq 1$ pour tout j, on obtient

$$\lambda \leq \sum_{i,j} a_{ij}$$

si bien que l'existence d'un majorant pour Λ se trouve démontrée.

On pose

$$\rho = \sup \Lambda$$

et on va montrer qu'il s'agit bien là du ρ du théorème.

Par définition de ρ il existe une suite (ρ_n) de Λ et une suite $(\mathbf{X}_n)$ de $\mathcal{S}$ telles que

$$\lim_{n\to\infty} \rho_n = \rho \text{ et } \mathbf{AX}_n \geq \rho_n \mathbf{X}_n.$$

Comme $\mathcal{S}$ est compact il existe une sous-suite de $(\mathbf{X}_n)$ qui converge. On la note $\left(\mathbf{X}_{\phi(n)}\right)$ et on note $\mathbf{X}$ sa limite. En faisant tendre n vers ∞ on trouve alors :

$$\mathbf{AX} \geq \rho\mathbf{X}.$$

Montrons qu'il s'agit là d'une égalité en utilisant un raisonnement par l'absurde.

On suppose que $\mathbf{AX} \neq \rho\mathbf{X}$ alors $\mathbf{AX} > \rho\mathbf{X}$, en composant cette inégalité par $\mathbf{A}^m$ on obtient $\mathbf{A}^{m+1}\mathbf{X} \gg \rho\mathbf{A}^m\mathbf{X}$ (car $\mathbf{B} \gg \mathbf{0}$ et $\mathbf{Z} > \mathbf{0} \Rightarrow \mathbf{BZ} \gg \mathbf{0}$).
On pose $\mathbf{Y} = \mathbf{A}^m\mathbf{X} \gg \mathbf{0}$ et on a $\mathbf{AY} \gg \rho\mathbf{Y}$. Pour un ε assez petit on a donc encore $\mathbf{AY} \gg (\rho + \varepsilon)\,\mathbf{Y}$.
On a donc $\mathbf{AY} \geq (\rho + \varepsilon)\,\mathbf{Y}$ et quitte à multiplier $\mathbf{Y}$ par une constante positive on peut supposer que $\mathbf{Y} \in \mathcal{S}$.
Cela implique que $(\rho + \varepsilon) \in \Lambda$ ce qui est contraire à l'hypothèse initiale.

Par conséquent on a

$$\mathbf{AX} = \rho\mathbf{X}$$

avec $\mathbf{X} \in \mathbf{S}$ donc $\mathbf{X} > \mathbf{0}$. De plus, $\mathbf{A}^m\mathbf{X} \gg \mathbf{0} \Rightarrow \rho^m\mathbf{X} \gg \mathbf{0} \Rightarrow \rho > 0$ et $\mathbf{X} \gg \mathbf{0}$ si bien que la preuve se trouve achevée. ■

Preuve de (ii). On désigne par $E_\rho = \mathrm{Ker}(\mathbf{A} - \rho\mathbf{I})$ le sous-espace propre associé à la valeur propre ρ et on va raisonner par l'absurde.

On suppose que $\dim(E_\rho) \geq 2$, on peut alors considérer $\mathbf{Y} \neq \mathbf{0}$ un élément de E_ρ tel que $(\mathbf{X}, \mathbf{Y})$ forme une famille libre.
Soit $\mu \in \mathbb{R}$ tel que $\mathbf{X} - \mu\mathbf{Y} \geq \mathbf{0}$ et que l'une au moins des composantes de $\mathbf{X} - \mu\mathbf{Y}$ soit nulle (on peut prendre par exemple $\mu = \inf\left(\frac{x_i}{|y_i|}, y_i \neq 0\right)$).
On a alors $\mathbf{X} - \mu\mathbf{Y} > \mathbf{0}$ car $(\mathbf{X}, \mathbf{Y})$ est libre et $\mathbf{A}^m \gg \mathbf{0}$ donc $\mathbf{A}^m(\mathbf{X} - \mu\mathbf{Y}) \gg \mathbf{0}$ ce qui implique que $\rho^m(\mathbf{X} - \mu\mathbf{Y}) \gg \mathbf{0}$ c'est à dire $\mathbf{X} - \mu\mathbf{Y} \gg \mathbf{0}$.

On a bien là une contradiction avec les hypothèses. ■

Preuve de (3). Soit $\lambda \neq \rho$ une autre valeur propre de $\mathbf{A}$. Soit $\mathbf{Z} > \mathbf{0}$ un vecteur propre associé à cette valeur propre. On a

$$\begin{aligned}
\mathbf{AZ} = \lambda\mathbf{Z} \quad &\Leftrightarrow \quad \sum_{j=1}^{n} a_{ij} z_j = \lambda z_i \quad \forall i \\
&\Leftrightarrow \quad \sum_{j=1}^{n} a_{ij}\,|z_j| \geq \lambda\,|z_i| \quad \forall i \\
&\Leftrightarrow \quad \mathbf{A}\,|\mathbf{Z}| \geq |\lambda|\,|\mathbf{Z}|\,.
\end{aligned}$$

Quitte à multiplier $|\mathbf{Z}|$ par une constante, on peut supposer que $|\mathbf{Z}| \in \mathcal{S}$ et donc on obtient $|\lambda| \in \Lambda$ ce qui implique que $|\lambda| \leq \rho$.

Il s'agit maintenant de montrer que $|\lambda| \neq \rho$, pour cela on va, une fois de plus, effectuer un raisonnement par l'absurde.

On suppose que $|\lambda| = \rho$, et, dans un premier temps, on se limite au cas où $m = 1$.
Comme $\mathbf{A} \gg \mathbf{0}$, il existe $\delta > 0$ tel que $\mathbf{A}_\delta = \mathbf{A} - \delta \mathbf{I}_n \gg \mathbf{0}$. Le spectre de $\mathbf{A}_\delta$ étant clairement celui de $\mathbf{A}$ modulo la soustraction de δ à chaque valeur propre, $\rho - \delta$ est la plus grande valeur propre réelle positive de $\mathbf{A}_\delta$ et $\lambda - \delta$ est une valeur propre de $\mathbf{A}_\delta$. Le même raisonnement que ci-dessus permet de montrer que $|\lambda - \delta| \leq \rho - \delta$ on obtient donc

$$\rho = |\lambda| = |\lambda - \delta + \delta| \leq |\lambda - \delta| + \delta \leq \rho$$

ce qui implique que

$$|\lambda| = |\lambda - \delta| + \delta$$

c'est à dire $\lambda > 0$ et donc $\lambda = \rho$ ce qui est contraire à l'hypothèse.

Dans le cas où m est quelconque, le point crucial consiste à montrer que ρ^m est la plus grande valeur propre réelle positive de $\mathbf{A}^m$. En effet, ceci étant démontré et comme λ^m est également une valeur propre de $\mathbf{A}^m$ on obtient, en appliquant le théorème à la matrice $\mathbf{A}^m$, que $|\lambda|^m \neq \rho^m$ et donc que $|\lambda| \neq \rho$.
On suppose que $\mathbf{A}^m$ n'admet pas ρ comme plus grande valeur propre réelle positive on peut donc supposer que $\mathbf{A}^m$ admet deux valeurs propres réelles positives $\lambda_1 < \lambda_2$; notons $\mathbf{X}_1$ et $\mathbf{X}_2$, les vecteurs propres associés ; ils sont $\gg \mathbf{0}$. Soit alors $\mu > 0$ tel que $\mathbf{X}_2 - \mu \mathbf{X}_1 > \mathbf{0}$ mais non $\gg \mathbf{0}$ (On prend $\mu = \inf \left(\frac{x_{2i}}{|x_{1i}|}, x_{1i} \neq 0 \right)$). On a alors $\mathbf{A}^m (\mathbf{X}_2 - \mu \mathbf{X}_1) \gg \mathbf{0}$ mais

$$\mathbf{A}^m (\mathbf{X}_2 - \mu \mathbf{X}_1) = \lambda_2 \mathbf{X}_2 - \mu \lambda_1 \mathbf{X}_1 = \lambda_2 (\mathbf{X}_2 - \mu \mathbf{X}_1) - (\lambda_1 - \lambda_2) \mu \mathbf{X}_1$$

ce qui implique que

$$\lambda_2 (\mathbf{X}_2 - \mu \mathbf{X}_1) = \mathbf{A}^m (\mathbf{X}_2 - \mu \mathbf{X}_1) + (\lambda_1 - \lambda_2) \mu \mathbf{X}_1$$

et donc $(\mathbf{X}_2 - \mu \mathbf{X}_1) \gg \mathbf{0}$ ce qui est impossible. ∎

Pour pouvoir généraliser le théorème B.2 au cas des matrices non plus primitives mais irréductibles, on introduit tout d'abord le lemme et la proposition suivante (tirés de [LT85]) :

Lemme B.3 *Si* $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$ *vérifie* $\mathbf{A} \geq \mathbf{0}$ *et peut se mettre sous la forme*

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{A}_{22} \end{bmatrix}$$

où $\mathbf{A}_{11} \in \mathcal{M}_k(\mathbb{R})$ *et* $1 \leq k \leq n-1$ *alors* A *n'est pas irréductible.*

Preuve. En effet, si on part d'un état $i > k$ il n'est possible de joindre qu'un état $j > k$ et il est donc impossible de revenir en un état compris entre 1 et k. ∎

Proposition B.4 *Si* $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$ *vérifiant* $\mathbf{A} \geq \mathbf{0}$ *est irréductible (*$\forall i, j$ $\exists m(i,j)$ *tel que* $\mathbf{A}^m(i,j) > 0$*) alors*

$$(\mathbf{I} + \mathbf{A})^{n-1} \gg \mathbf{0}.$$

Preuve. Soit $\mathbf{X} \in \mathbb{R}^n$, $\mathbf{X} > \mathbf{0}$. On pose $\mathbf{Y} = (\mathbf{I} + \mathbf{A})\mathbf{X} = \mathbf{X} + \mathbf{AX}$. Comme $\mathbf{A} \geq \mathbf{0}$ on a $\mathbf{AX} \geq \mathbf{0}$ et donc $\mathbf{Y}$ a au moins autant de termes non nul que $\mathbf{X}$. On va montrer que si $\mathbf{X}$ est non $\gg \mathbf{0}$ alors $\mathbf{Y}$ a au moins un terme non nuls de plus que $\mathbf{X}$.

On considère Σ une matrice de permutation telle que

$$\mathbf{\Sigma X} = \begin{bmatrix} \mathbf{U} \\ \mathbf{0} \end{bmatrix}$$

avec $\mathbf{U} \gg \mathbf{0}$. Alors on a $\mathbf{\Sigma Y} = \mathbf{\Sigma X} + \mathbf{\Sigma AX} = \mathbf{\Sigma X} + \mathbf{\Sigma A \Sigma' \Sigma X}$ car $\mathbf{\Sigma'\Sigma} = \mathbf{\Sigma\Sigma'} = \mathbf{I}$ et par conséquent, on a

$$\mathbf{\Sigma Y} = \begin{bmatrix} \mathbf{U} \\ \mathbf{0} \end{bmatrix} + \mathbf{\Sigma A \Sigma'} \begin{bmatrix} \mathbf{U} \\ \mathbf{0} \end{bmatrix}.$$

Si on note

$$\mathbf{\Sigma Y} = \begin{bmatrix} \mathbf{V} \\ \mathbf{W} \end{bmatrix} \quad \text{et} \quad \mathbf{\Sigma A \Sigma'} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix},$$

on obtient $\mathbf{V} = \mathbf{U} + \mathbf{A}_{11}\mathbf{U}$ et $\mathbf{W} = \mathbf{A}_{21}\mathbf{U}$. Il est clair que $\mathbf{\Sigma A \Sigma'}$ est positive et irréductible, puisque $\mathbf{A}$ l'est, et on a donc $\mathbf{A}_{21} > \mathbf{0}$ en utilisant le lemme précédant (si $\mathbf{A}_{21} = \mathbf{0}$ alors $\mathbf{\Sigma A \Sigma'}$ n'est pas irréductible).

Par conséquent $\mathbf{W} > \mathbf{0}$ et donc $\mathbf{\Sigma Y}$, donc $\mathbf{Y}$, a au moins un terme non nul de plus que $\mathbf{X}$. ■

On peut alors énoncer le

Théorème B.5 (*Perron-Frobénius* faible)
Si $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$ *est une matrice* $A > 0$ *telle que,* $\forall i, j$ *il existe* $m(i,j) \geq 1$ *telle que* $\mathbf{A}^m(i,j) > 0$ *(c'est à dire que* $\mathbf{A}$ *est* irréductible*) alors*

(i) $\mathbf{A}$ *admet une valeur propre* $\rho > 0$ *dont un vecteur propre associé est à coordonnées toutes strictement positives ;*

(ii) ρ *est une valeur propre de multiplicité* 1 *;*

(iii)' *si* $\lambda \neq \rho$ *est une autre valeur propre de* $\mathbf{A}$ *alors* $|\lambda| \leq \rho$.

On appelle ρ *la* ***valeur propre de Frobénius*** *de* $\mathbf{A}$ *.*

Notons que les résultats sont quasiment identiques à ceux du théorème B.2, seul le "$<$" de **(iii)** est remplacé par un "$\leq$" dans le **(iii)'** du nouveau théorème.

Avant de s'attaquer à la preuve du théorème B.5 résultat, on effectue la remarque suivante :

Remarque B.6 *Si* $\mathbf{A} \geq 0$ *est irréductible alors pour tout* $\mathbf{X} \gg \mathbf{0}$ *on a* $\mathbf{AX} \gg \mathbf{0}$ *(en effet, si* $\exists i$ *tel que* $\sum_{j=1}^{n} A_{ij} X_j = 0$ *alors on a* $A_{ij} = 0$ *pour tout* j *et comme, a une permutation près on peut supposer que* $i = n$ *le lemme permet de conclure que* $\mathbf{A}$ *n'est pas irréductible).*

Preuve de (i). De la même façon que dans la première preuve, on montre qu'il existe $\mathbf{X} \in \mathcal{S}$ tel que $\mathbf{AX} \geq \rho\mathbf{X}$. Il s'agit encore une fois de montrer que l'on a l'égalité.

Pour cela, on raisonne encore une fois par l'absurde. Si $\mathbf{AX} \neq \rho\mathbf{X}$ alors $\mathbf{AX}-\rho\mathbf{X} > \mathbf{0}$, de plus on sait (proposition B.4) que $(\mathbf{I}+\mathbf{A})^{n-1} \gg \mathbf{0}$ donc $(\mathbf{I}+\mathbf{A})^{n-1}(\mathbf{AX}-\rho\mathbf{X}) \gg \mathbf{0}$.

On remarque ensuite que $(\mathbf{I}+\mathbf{A})^{n-1}$ et $\mathbf{A}$ commutent et on pose $\mathbf{Y} = (\mathbf{I}+\mathbf{A})^{n-1}\mathbf{X}$.

On a alors $\mathbf{AY}-\rho\mathbf{Y} \gg \mathbf{0}$ avec $\mathbf{Y} \gg \mathbf{0}$ et le même raisonnement que dans la première preuve permet alors de conclure à une absurdité ; on a donc $\mathbf{AX} = \rho\mathbf{X}$.

On remarque ensuite que $(\mathbf{I}+\mathbf{A})^{n-1}\mathbf{X} \gg \mathbf{0}$ et que $(\mathbf{I}+\mathbf{A})^{n-1}\mathbf{X} = (1+\rho)^{n-1}\mathbf{X}$ on obtient donc $\mathbf{X} \gg \mathbf{0}$.

En utilisant la remarque B.6, on a $\mathbf{AX} = \rho\mathbf{X} \gg \mathbf{0}$ et donc $\rho > 0$.

Il est également possible de montrer que $\rho > 0$ en raisonnant une nouvelle fois par l'absurde. On suppose que $\rho \leq 0$ et comme on a déjà vu que $\rho \geq 0$ car $0 \in \Lambda$ on a $\rho = 0$.

On utilise alors le **[(iii)']** pour remarquer que $\forall\lambda$ valeur propre de $\mathbf{A}$ on a $|\lambda| \leq 0$ c'est à dire $\lambda = 0$.

On obtient ainsi finalement $\mathbf{A} = \mathbf{0}$ ce qui est absurde (il suffit pour cela de ce placer dans $\mathcal{M}_n(\mathbb{C})$). ■

Preuve de (ii). On applique simplement le théorème B.2 à la matrice primitive $\mathbf{I}+\mathbf{A}$ ce qui permet de conclure que la valeur propre $\rho+1$ est de multiplicité 1. On remarque ensuite que les valeurs propres et les espaces propres de $\mathbf{A}$ s'obtient trivialement à partir de ceux de $\mathbf{I}+\mathbf{A}$, et la preuve est achevée. ■

Preuve de (iii)'. On répète simplement le même raisonnement que dans la preuve de **(iii)**. ■

On peut maintenant établir l'important corollaire suivant dont la preuve est tirée de [DZ98] :

Corollaire B.7 *Soit* $\mathbf{A} \in \mathbb{R}^n$ *une matrice positive irréductible et soit* ρ *sa valeur propre de Frobénius alors,* $\forall\mathbf{\Phi} \gg \mathbf{0}$ *et* $\forall i$ *on a*

$$\lim_{k\to+\infty} \frac{1}{k} \log\left(\sum_{j=1}^{n} \mathbf{A}^k(i,j)\,\Phi_j\right) = \log\rho.$$

Preuve. Soit $\mathbf{X}$ un vecteur propre à droite à coordonnées strictement positives associé à ρ. On pose $\alpha = \sup X_j$, $\beta = \inf X_j > 0$, $\gamma = \sup \Phi_j$ et $\delta = \inf \Phi_j$.

On a donc pour tout j

$$\delta \leq \Phi_j \leq \gamma \quad \text{et} \quad 0 < \beta \leq X_j \leq \alpha$$

donc

$$\frac{\delta}{\alpha} X_j \leq \Phi_j \leq \frac{\gamma}{\beta} X_j$$

si bien que

$$\begin{aligned} \lim_{k \to +\infty} \frac{1}{k} \log \left(\sum_{j=1}^{n} \mathbf{A}^k (i,j) \, \Phi_j \right) &= \lim_{k \to +\infty} \frac{1}{k} \log \left(\sum_{j=1}^{n} \mathbf{A}^k (i,j) \, X_j \right) \\ &= \lim_{k \to +\infty} \frac{1}{k} \log \left(\rho^n X_j \right) \\ &= \log \rho. \end{aligned}$$

■

Ce corollaire nous permet alors d'effectuer une remarque bien utile :

Remarque B.8 *Si* $\mathbf{\Phi} \gg \mathbf{0}$ *est un vecteur propre associé à* λ *alors* $\lambda = \rho$.

B.2 Formule de *Whittle*

On s'intéresse ici à un résultat de combinatoire dû à *Whittle* ([Whi55]). Soit $\mathcal{A}$ un alphabet fini de cardinal k et soit $(n_{xy})_{x,y \in \mathcal{A}}$ un ensemble de comptage ($n_{xy} \geq 0, \ \forall x, y \in \mathcal{A}$).

Soit

$$\mathcal{S} = \{(X_i)_{i \in 1,\ldots,n} \text{ tq } N(xy) = n_{xy} \ \forall x, y \in \mathcal{A}\}$$

où n désigne un entier et où, $\forall x, y \in \mathcal{A}$,

$$N(xy) = \sum_{i=1}^{n} \mathbb{I}_{X_i X_{i+1} = xy}$$

avec la convention $X_{n+1} = X_1$.

On pose $\forall x \in \mathcal{A}$

$$N(x+) = \sum_{y \in \mathcal{A}} N(xy) \text{ et } N(+x) = \sum_{y \in \mathcal{A}} N(xy)$$

(on peut en effet supposer que $N(x+) > 0 \ \forall x \in \mathcal{A}$ car si ce n'est pas le cas, il suffit alors d'exclure de $\mathcal{A}$ les lettres ne vérifiant pas la condition) et on a alors le résultat suivant :

Théorème B.9 (*Whittle*)
On note $|\mathcal{S}|$ *désigne le cardinal de* $\mathcal{S}$.

Si

$$N(x+) = N(+x) \ \forall x \in \mathcal{A} \tag{B.1}$$

alors $|\mathcal{S}| > 0$

$$|\mathcal{S}| = \frac{\prod_{x\in\mathcal{A}} N(x+)!}{\prod_{x,y\in\mathcal{A}} N(xy)!} \times H \tag{B.2}$$

avec H *mineur du terme* $(1,1)$ *de la matrice* $I - \widehat{\Pi}$ *où*

$$\widehat{\Pi}(x,y) = \frac{N(xy)}{N(x+)}.$$

et si (B.1) n'est pas vérifiée alors $|\mathcal{S}| = 0$.

Si la fraction contenue dans la formule (B.2) est assez naturelle, l'intervention du terme H l'est beaucoup moins ; on se reportera à l'article de *Whittle* pour plus de précisions.

Afin de pouvoir manipuler plus facilement cette formule on établit l'encadrement suivant :

Proposition B.10 *On a*

$$\frac{1}{n^{k-1}} \leq H \leq (k-1)!$$

Preuve. H est un déterminant d'ordre $k-1$ dont tous les termes (non diaganaux) sont de la forme

$$\frac{N(xy)}{N(x+)} \quad x, y \neq 1.$$

tandis que les termes de la diagonale s'écrivent sous la forme

$$1 - \frac{N(xy)}{N(x+)} \quad x, y \neq 1.$$

En effectuant un développement récursif de ce déterminant, on obtient une somme de $(k-1)!$ termes de la forme

$$\frac{d}{\prod_{x=2}^{k} N(x+)} \quad d \in \mathbb{Z} \tag{B.3}$$

et appartenant à $[0,1]$ comme produit de termes de $[0,1]$.

Il est donc évident que $H \leq (k-1)!$ et de plus, H étant également de la forme (B.3) et le théorème B.9 assurant que $H > 0$ on obtient aisément la minoration

$$H \geq \frac{1}{\prod_{x=2}^{k} N(x+)} \geq \frac{1}{n^{k-1}}.$$

■

B.3 Théorème de la limite centrale

Théorème B.11 *Soient $\mathcal{A}$ un espace d'état fini, $(X_i)_i$ une chaîne de* Markov *irréductible sur $\mathcal{A}$ de matrice de transition Π et de loi stationnaire μ. Soit enfin $f : \mathcal{A}^2 \to \mathbb{R}$ une fonction déterministe. On note $S_n = \sum_{i=1}^n f(x_i, X_{i+1})$ et, $\forall s \in \mathcal{A}$, on note $\mathbb{P}_s$ la loi sur la chaîne de* Markov *en partant de l'état initial s alors on a :*

$$\frac{S_n}{n} \xrightarrow[n\to\infty]{} m = \mathbb{E}_\mu^\Pi [f(X_1, X_2)] \ \mathbb{P}_s \ p.s. \ \forall s \in \mathcal{A} \tag{B.4}$$

et

$$Y_n = \frac{1}{\sqrt{n}}(S_n - nm) \xrightarrow{\mathcal{L}(\mathbb{P}_s)} \mathcal{N}(0, \sigma^2) \tag{B.5}$$

avec

$$\sigma^2 = \sum_{i\in\mathbb{Z}} \mathbb{C}ov_\mu^\Pi \left[f(X_1, X_2), f(X_i, X_{i+1})\right]$$

Preuve. Les hypothèses du théorème impliquant en particulier l'ergodicité de la chaîne de *Markov*, le premier point du théorème est trivial. Le second point demande pour sa part un peu de travail : on va montrer que $\phi(t) = \mathbb{E}[e^{tY_n}]$ est asymptotiquement de la forme $e^{\frac{t^2}{2}\sigma^2}$.

On a

$$\phi(t) = \mathbb{E}[e^{\frac{t}{\sqrt{n}}S_n} \cdot e^{-tm\sqrt{n}}]$$

donc

$$\frac{1}{n}\log\phi(t) = \Lambda\left(\frac{t}{\sqrt{n}}\right) + D_n - \frac{\sqrt{n}tm}{n}$$

avec

$$D_n = \frac{1}{n}\log\mathbb{E}[e^{\frac{t}{\sqrt{n}}S_n}] - \Lambda\left(\frac{t}{\sqrt{n}}\right).$$

et $\Lambda = \log\phi$.

Il est clair que D_n tend vers 0 quand n tend vers l'infini, mais qu'en est-il de nD_n ? On note $\rho(\theta)$ la plus grande valeur propre de Π_θ ($\Pi_\theta(x,y) = \Pi(x,y)e^{\theta f(x,y)}$ pour tout $x, y \in \mathcal{A}$ et pour $\theta \in \mathbb{R}$) et $v_\theta \gg 0$ un vecteur propre associé à cette valeur propre. On considère que la chaîne part de l'état initial $s \in \mathcal{A}$ et on rappelle qu'alors

$$\mathbb{E}[e^{\theta S_n}] = \sum_{x_{n+1}\in\mathcal{A}} \Pi_\theta^n(s, x_{n+1}).$$

Soit $\alpha = \sup v_\theta(i)$ et $\beta = \inf v_\theta(i)$ alors

$$\frac{v_\theta(i)}{\alpha} \le 1 \le \frac{v_\theta(i)}{\beta} \ \forall i \in \mathcal{A}$$

si bien que

$$\begin{aligned}
\frac{1}{\alpha}\sum_{x_{n+1}\in\mathcal{A}}\Pi_\theta^n(s,x_{n+1})v_\theta(x_{n+1}) &\leq \mathbb{E}[e^{\theta S_n}] \leq & \frac{1}{\beta}\sum_{x_{n+1}\in\mathcal{A}}\Pi_\theta^n(s,x_{n+1})v_\theta(x_{n+1})\\
\frac{1}{\alpha}(\Pi_\theta^n\times v_\theta)(s) &\leq \mathbb{E}[e^{\theta S_n}] \leq & \frac{1}{\beta}(\Pi_\theta^n\times v_\theta)(s)\\
n\log\rho(\theta)+\log v_\theta(s)-\log\alpha &\leq \log\mathbb{E}[e^{\theta S_n}] \leq & n\log\rho(\theta)+\log v_\theta(s)-\log\beta
\end{aligned}$$

et par conséquent

$$\log v_\theta(s)-\log\alpha \leq \log\mathbb{E}[e^{\theta S_n}]-n\log\rho(\theta)\leq \log v_\theta(s)-\log\beta.$$

et donc

$$\log v_{\frac{t}{\sqrt{n}}}(s)-\log\alpha \leq \log\mathbb{E}[e^{\frac{t}{\sqrt{n}}S_n}]-n\log\rho(\frac{t}{\sqrt{n}})\leq \log v_{\frac{t}{\sqrt{n}}}(s)-\log\beta.$$

Comme $v_0=(1,\ldots,1)$ il est clair que

$$\lim_{n\to\infty}\log v_{\frac{t}{\sqrt{n}}}(s)-\log\alpha=\lim_{n\to\infty}\log v_{\frac{t}{\sqrt{n}}}(s)-\log\beta=0$$

et donc

$$\lim_{n\to\infty} nD_n=0.$$

On va maintenant utiliser un développement limité au voisinage de l'infini de

$$n\Lambda\left(\frac{t}{\sqrt{n}}\right)=n\Lambda(0)+\sqrt{n}t\Lambda'(0)+\frac{t^2}{2}\Lambda''(0)+o(1)$$

et comme on a (voir propositions 5.3 et 5.4 page 86)

$$\begin{aligned}
\Lambda(0) &= 0\\
\Lambda'(0) &= m\\
\Lambda''(0) &= \sigma^2
\end{aligned}$$

la preuve est achevée. ■

Références

[DZ98] A. Dembo and O. Zeitouni. *Large deviations techniques and applications.* Springer-Verlag, 1998.

[Gan90] F. R. Gantmakher. *Théorie des matrices.* Editions Jacques Gabay, 1990.

[Kar69] S. Karlin. *Initiation aux processus aléatoires.* Dunod, Paris, 1969.

[LT85] A Lancaster and M. Tismenetsky. *The theory of matrices.* Academic press, Orlando, 1985.

[Sen81] E. Seneta. *Non-negative matrices and Markov chains.* Springer-Verlag, New York, 1981.

[Whi55] P. Whittle. Some distribution and moment fomulæ for the markov chain. *J. R. Statist. Soc. B.*, 17 :235–242, 1955.

Annexe C

Algorithmes

Contenu du chapitre

C.1 *Arnoldi*

C.1.1 Procédure de *Rayleigh-Ritz*

Soit A une matrice réelle d'ordre n et soit $\mathcal{S} \subset \mathbb{R}^n$ un sous-espace vectoriel, alors il existe une procédure permettant d'extraire des approximations des valeurs et vecteurs propres de A à partir du sous-espace $\mathcal{S}$; la procédure de *Rayleigh-Ritz*.

La méthode est la suivante :

- on commence par calculer $Q = (q_1, \ldots, q_m)$ une base orthogonale de $\mathcal{S}$.
- On calcule $B = Q'AQ$ le quotient de *Rayleigh* (remarquons que la matrice B est d'ordre m et non n) ;

– soient (λ_i, x_i) les valeurs et vecteurs propres de B ;
– l'approximation de *Ritz* est alors constituée par les couples (λ_i, y_i) avec $y_i = Qx_i$.

De plus, cette approximation est optimale dans le sens où $|||AY - \Lambda Y|||$ avec $Y = (y_1, \ldots, y_m)$ et $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_m)$ (norme d'application linéaire mesurant la qualité globale de l'approximation) minore la même quantité pour tout autre choix de Y (avec $y_i \in \mathcal{S}$) et Λ.

Remarque C.1 *L'idée consiste évidemment à ramener un problème de dimension n à une dimension (souvent beaucoup) plus petite m et on conçoit sans peine que toute la difficulté consiste à bien choisir le sous-espace $\mathcal{S}$; en prenant un exemple extrême, si $\mathcal{S} = vect(y)$ où y est un vecteur propre de A, la méthode donne immédiatement y et sa valeur propre associée comme approximation.*

C.1.2 Projection pour le problème des valeurs propres

On s'intéresse désormais au problème :

$$Ax = \lambda x$$

dont on va chercher une solution approchée de sorte que $x \in \mathcal{K}$ et que $Ax - \lambda x$ soit orthogonal à $\mathcal{L}$, $\mathcal{K}$ et $\mathcal{L}$ étant deux sous-espaces vectoriels de dimension m donnés dont on notera V et W les bases respectives.

Soit y tel que $x = Vy$, on a alors $W'(AVy - \lambda Vy) = 0$ ce qui donne :

$$(W'AV)y = \lambda W'Vy$$

équation correspondant à un problème de valeurs propres généralisées de dimension m dont les solutions (λ_i, y_i) fournissent les approximations (λ_i, x_i) avec $x_i = Vy_i$.

Dans la suite on utilisera cette méthode avec $\mathcal{K} = \mathcal{L}$ et on prendra pour V une base orthogonale de sorte que $V'V = Id$; on sera ainsi ramené à un problème de valeurs propres non généralisées de dimension m.

C.1.3 Méthode d'*Arnoldi*

On va utiliser la méthode ci-dessous en choisissant pour $\mathcal{K}$ le sous-espace de *Kyrov* de dimension m :

$$\mathcal{K}_m = \text{vect}(v_1, Av_1, \ldots, A^{m-1}v_1)$$

où v_1 est un vecteur initial de norme 1. On va de plus supposer que $\mathcal{K}_m$ est de dimension m, hypothèse qui sera toujours vérifiée dans nos applications numériques.

On construit $V_m = (v_1, \dots, v_m)$ une base orthogonale de $\mathcal{K}_m$ et on pose $H_m = V_m' A V_m$ qui n'est autre que la restriction de A à $\mathcal{K}_m$ écrite dans la base V_m.

On calcule (λ_i, y_i) les valeurs propres et vecteurs propres de H_m et (λ_i, x_i) avec $x_i = V_m y_i$ constitue une approximation des solutions du problème initial (valeurs propres et vecteurs propres de A).

On calcule V_m en utilisant le procédé d'orthogonalisation de *Gram-Schmidt* :

- on pose $v_2 = \frac{w_2}{\|w_2\|}$ et $w_2 = Av_1 + c_1 v_1$ et on va déterminer c_1 ;
  - $v1'w2 = 0 \Leftrightarrow v_1' A v_1 + c_1 v_1' v1 = 0 \Leftrightarrow c_1 = -v_1' A v_1 = h_{11}$
  - $w_2' w_2 = \|w_2\| \Leftrightarrow w_2' A v_1 + c_1 w_2' v1 = \|w_2\| \Leftrightarrow \|w_2\| = w_2' A v_1$
- on reprend la même chose au rang $j+1$, avec $(v_1, \dots, v_j)$ déjà construits : $v_{j+1} = \frac{w_{j+1}}{\|w_{j+1}\|}$ et $w_{j+1} = Av_j + \sum_{i=1}^n c_i v_i$ et on va déterminer les c_i ;
  - $vi' w_{j+1} = 0 \Leftrightarrow v_i' A v_j + \sum_{k=1}^n c_k v_i' vk = 0 \Leftrightarrow c_i = -v_i' A v_j = h_{ij}$
  - $w_{j+1}' w_{j+1} = \|w_{j+1}\| \Leftrightarrow w_{j+1}' A v_j + \sum_{k=1}^n c_k w_{j+1}' v_k = \|w_{j+1}\| \Leftrightarrow \|w_{j+1}\| = w_{j+1}' A v_j$

On obtient ainsi un algorithme n'utilisant la matrice A que dans le cadre de produits avec des vecteurs, ce qui permet de tirer parti de la nature creuse de la matrice lorsque cela est le cas.

Algorithme : pour $j = 1, 2, \dots, m$

- $w = Av_j$
- pour $i = 1, 2, \dots, j$
  - $h_{ij} = v_i' w (= v_i' A v_j)$
  - $w = w - h_{ij} v_i$
- $h_{j+1,j} = \|w\|_2$
- $v_{j+1} = \frac{w}{h_{j+1,j}} (= \frac{w}{\|w\|_2})$

Remarque C.2 *Remarquons au passage la forme particulière de la matrice H_m dont on doit calculer valeurs et vecteurs propres : $h_{ij} = 0$ si $i > j+1$ (ceci est directement lié à l'algorithme d'orthogonalisation de* Gram-Schmidt*). Lorsqu'une matrice vérifie cette propriété on dit qu'elle est sous la forme de* Hessenberg *et il existe pour ce type de matrice des algorithmes performants en ce qui concerne le problème des valeurs et vecteurs propres (algorithme HQR en particulier).*

C.1.4 *Arnoldi* pour le problème de *Perron-Frobénius*

Dans le cas particulier du problème de *Perron-Frobénius*, on ne s'intéresse qu'à la plus grande valeur propre ρ de la matrice (et éventuellement du vecteur propre x associé). On va utiliser pour résoudre ce problème une version légèrement modifiée et itérative d'*Arnoldi*.

On commence par se donner un entier m_{max} qui constituera la dimension maximum du sous-espace de *Kyrov* utilisé. On va ensuite simplement reprendre l'algorithme décrit ci-dessus mais on va évaluer une approximation de ρ à chaque itération ($j = j + 1$). Lorsque $j = m_{max}$ est atteint, on évalue également x et on reprend l'algorithme avec $v_1 = x$ (après avoir normalisé le vecteur). On continue ainsi jusqu'à ce que la valeur de ρ n'évolue plus.

Algorithme : Tant que $|\rho_{new} - \rho_{old}| > \epsilon$
- $v_1 = x$ (quelconque la première fois)
- pour $j = 1, 2, \ldots, m_{max}$
  - $\rho_{old} = \rho_{new}$
  - $w = Av_j$
  - pour $i = 1, 2, \ldots, j$
    - $h_{ij} = v_i' w (= v_i' A v_j)$
    - $w = w - h_{ij} v_i$
  - $h_{j+1,j} = \|w\|_2$
  - $v_{j+1} = \frac{w}{h_{j+1,j}} (= \frac{w}{\|w\|_2})$
  - on calcule ρ_{new} la plus grande valeur propre de H_j (algorithme HQR)
  - si $|\rho_{new} - \rho_{old}| \leq \epsilon$ l'algorithme est terminé.
- on calcule ρ_{new} plus grande valeur propre de $H_{m_{max}}$ et y son vecteur propre associé (algorithme HQR2) qui nous donne une approximation $x = V_{m_{max}} y$.

Remarque C.3 *En toute rigueur, il possible d'éviter la dernière évaluation de x si l'algorithme se termine sur $j = m_{max}$ et il est nécessaire de rajouter une évaluation de x lorsque l'algorithme ne se termine pas en $j = m_{max}$ et que l'on souhaite également obtenir x.*

La question naturelle lorsque l'on voit un tel algorithme est : quelle valeur donner à m_{max} ? Lorsque ce paramètre augmente, l'espace mémoire requis pour l'exécution d'une boucle ainsi que le temps de calcul nécessaire augmente sensiblement (mémoire en $n \times m_{max}$ avec n désignant l'ordre de A et temps en m_{max}^2), mais il est également clair que le nombre total de boucles effectuées va décroître sensiblement.

En l'absence de moyen thérorique pour fixer cette valeur, la seule méthode restant alors est l'optimisation empirique.

C.2 *Brent*

C.2.1 Généralités

On ramène le problème consistant à trouver le minimum local d'une fonction régulière à celui consistant à trouver le minimum global d'une fonction régulière et convexe. Dans la suite, f désigne une fonction $\mathcal{C}^1$ convexe dont on cherche à déterminer $x_{\min}$ l'argument où est atteint $f_{\min} = \min f$.

Proposition C.4 *Si $a < b < c$ et que $f_b < f_a$ et $f_b < f_c$ alors $x_{min} \in [a, c]$.*

Preuve. Si $x_{\min} \notin [a, c]$, soit $x_{\min} < a$, soit $x_{\min} > c$. Dans un cas comme dans l'autre, la régularité et la convexité de f (f' croissante) montre que f doit être monotone sur $[a, c]$ ce qui est contraire à l'hypothèse effectuée. ■

Corollaire C.5 *On suppose que les conditions de la proposition C.4 sont vérifiées et on considère x tel que $a < x < b$ (resp. $b < x < c$) alors, si $f_x < f_b$ on a $x_{min} \in [a, b]$ (resp. $\in [b, c]$) et si $f_x > f_b$, $x_{min} \in [x, c]$ (resp. $\in [a, x]$).*

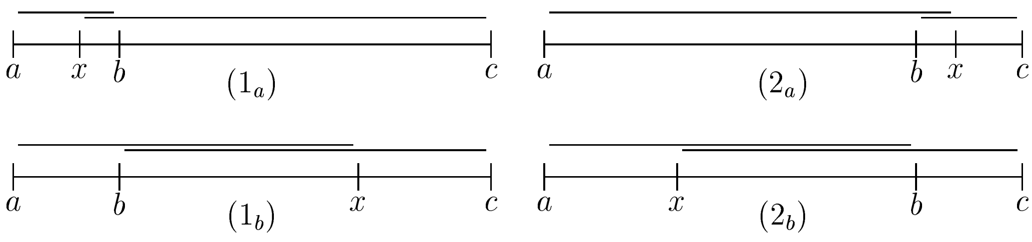
C.2.2 Nombre d'or

Il est ainsi possible de se ramener d'un triplet de point (a, b, c) dont les extrémités encadrent $x_{\min}$ à un nouveau triplet à partir duquel on peut réitérer la procédure. Pour cela, il suffit de choisir un nouveau point à chaque itération et de s'arrêter lorsque $|c - a|$ (taille de l'encadrement) atteint une valeur convenable.

L'algorithme du nombre d'or va être la simple optimisation de cette algorithme quant aux positions des différents points à chaque itération.

La première difficulté consiste à choisir le segment ou positionner x. Si on veut que la taille du nouvel encadrement soit peu (ou pas) sensible à la position de $x_{\min}$ il faut simplement choisir x dans le plus long des segments $[a, b]$ et $[b, c]$.

En effet, comme le montre l'illustration ci-après, le choix d'un x dans le segment le plus court (1_a et 2_a) donne des encadrements beaucoup plus déséquilibrés que dans l'autre cas (1_b et 2_b).

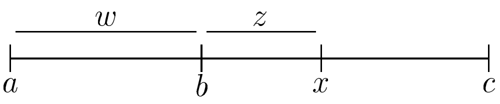


Pour simplifier le raisonnement, on supposera désormais que nous nous trouvons dans le cas 1_b.

A partir des proportions

$$\frac{b-a}{c-a} = w \text{ et } \frac{x-b}{c-a} = z$$

qui sont représentées ici :



il est possible d'exprimer le fait que dans, tous les cas, les encadrements résultants ($[a, x]$ ou $[b, c]$) doivent être de même taille :

$$w + z = 1 - w \iff z = 1 - 2w \tag{C.1}$$

Il semble également naturel d'imposer que le positionnement de b ait lui même respecté les proportions que l'on souhaite imposer à x. On obtient ainsi :

$$w = \frac{z}{1-w} \iff w - w^2 = z \tag{C.2}$$

En combinant (C.1) et (C.2) on obtient

$$w^2 - 3w + 1 = 0 \iff w = \frac{3-\sqrt{5}}{2} \quad (\text{car } w \in]0, 1[)$$

En effet, si on suppose par exemple que $f_x < f_b$ alors on est ramené de (a, b, c) à (b, x, c). Il alors faut comparer la proportion w à $\frac{z}{1-w}$ si $z < 0.5$ et à $\frac{1-z}{1-w}$ sinon, mais dans ce dernier cas (obtenu pour $w < 0.25$) on trouve alors l'équation $w^2 - w = 0$ dont les solutions ne sont pas acceptables.

C.2.3 Approximation parabolique

Si la fonction est assez régulière, on peut tabler sur sa ressemblance avec une parabole au voisinage de son minimum. L'abscisse du minimum d'une parabole définie par $g(x) = \alpha x^2 + \beta x + \gamma$ ($\alpha > 0$) étant immédiatement donnée par $x_{\min} = -\frac{\beta}{2\alpha}$, on comprend l'intérêt d'utiliser une approximation parabolique de la fonction a minimiser.

Proposition C.6 *Soit (a, b, c) un triplet réalisant un encadrement de x_{min} ($a < b < c$, $f_b < f_a$ et $f_b < f_c$) alors*

$$x_{min} = \frac{1}{2}\frac{(f_b - f_c)(b-a)(b+a) - (f_b - f_a)(b-c)(b+c)}{(f_b - f_c)(b-a) - (f_b - f_a)(b-c)} \tag{C.3}$$

Preuve. Déterminons α et β telle que la parabole d'équation $g(x) = \alpha x^2 + \beta x + \gamma$ passe par (a, f_a), (b, f_b) et (c, f_c) :

$$\left\{\begin{array}{l} f_a = \alpha a^2 + \beta a + \gamma \\ f_b = \alpha b^2 + \beta b + \gamma \\ f_c = \alpha c^2 + \beta c + \gamma \end{array}\right. \Rightarrow \left\{\begin{array}{ll} \frac{f_b - f_c}{b-c} = \alpha(b+c) + \beta & (L1) \\ \frac{f_b - f_a}{b-a} = \alpha(b+a) + \beta & (L2) \end{array}\right.$$

En effectuant $(L1) - (L2)$ on obtient

$$\begin{aligned} \alpha(c-a) &= \frac{f_b - f_c}{b-c} - \frac{f_b - f_a}{b-a} \\ \alpha &= \frac{(f_b - f_c)(b-a) - (f_b - f_a)(b-c)}{(b-a)(b-c)(c-a)} \end{aligned} \tag{C.4}$$

En effectuant $(L1) + (L2)$ on obtient également

$$\begin{aligned} 2\beta &= \frac{f_b - f_c}{b-c} + \frac{f_b - f_a}{b-a} - \alpha(a + 2b + c) \\ \beta &= \frac{(f_b - f_c)(b-a) + (f_b - f_a)(b-c)}{2(b-a)(b-c)} - \frac{\alpha}{2}(a + 2b + c) \end{aligned} \tag{C.5}$$

En combinant (C.4) et (C.5) et en réduisant au même dénominateur on obtient

$$-\frac{\beta}{2\alpha} = \frac{1}{2}\frac{A(f_b - f_c) + B(f_b - f_a)}{(f_b - f_c)(b-a) - (f_b - f_a)(b-c)}$$

avec

$$\begin{aligned} A &= (b-a)\frac{1}{2}\left[(a + 2b + c) - (c - a)\right] \\ &= (b-a)(b+a) \end{aligned}$$

et

$$\begin{aligned} B &= (b-c)\frac{1}{2}\left[-(a + 2b + c) - (c - a)\right] \\ &= -(b-c)(b+c) \end{aligned}$$

on obtient bien ainsi l'équation (C.3). ■

C.2.4 Algorithme de *Brent*

L'algorithme de *Brent* pour la minimisation des fonctions combine les deux approches vues précédemment alternant itérations du nombre d'or et approximations paraboliques.

Voila en quoi consiste l'algorithme :

- itération n : on possède un triplet (a_n, b_n, c_n) qui réalise un encadrement du minimum qui est alors approché par b_n.
- itération $n+1$:
  - on choisit un nouveau point x selon la méthode du nombre d'or ou selon l'approximation parabolique.
  - on évalue f_x la valeurs de f en x.
  - on obtient alors

$$(a_{n+1}, b_{n+1}, c_{n+1}) = \left\{ \begin{array}{c} \left.\begin{array}{c} (x, b_n, c_n) \text{ si } f_x > f_{b_n} \\ (a_n, x, b_n) \text{ si } f_x < f_{b_n} \end{array}\right\} \text{si } x \in]a_n, b_n[\\ \left.\begin{array}{c} (a_n, b_n, x) \text{ si } f_x > f_{b_n} \\ (b_n, x, c_n) \text{ si } f_x < f_{b_n} \end{array}\right\} \text{si } x \in]b_n, c_n[\end{array} \right.$$

Au départ on ne possède qu'un triplet (a_0, b_0, c_0) et une seule évaluation de f au point b_0. L'algorithme commence par deux itérations par la méthode du nombre d'or (en effet, l'approche parabolique nécessite les valeurs de f en chacun des points du triplet). Par la suite, le choix entre nombre d'or et approximation parabolique est fonction des "mouvements" de l'approximation du minimum : si l'approche parabolique permet un déplacement du minimum inférieur à la moitié du déplacement précédent (il faut donc conserver à cet effet les deux dernières approximations du minimum) c'est cette approche qui est choisie. L'algorithme se termine lorsque l'encadrement du minimum est suffisamment précis par exemple.

C.2.5 Application : descente du gradient

Il est également possible d'utiliser l'algorithme de *Brent* pour effectuer des minimisations multi-dimensionnelles. On considère une fonction

$$F : \mathbb{R}^d \to \mathbb{R}$$

dont on cherche à calculer le minimum. On ramène la dimension du problème à un en choisissant une droite $\mathcal{D} = \{\lambda a + b, \lambda \in \mathbb{R}\}$ où $a, b \in \mathbb{R}^d$, et en cherchant à calculer

$$\inf_{\lambda \in \mathbb{R}} F(\lambda a + b)$$

ce qui peut se faire à l'aide de l'algorithme de *Brent*.

Toute la difficulté ensuite, consiste à choisir la droite $\mathcal{D}$ sur laquelle effectuer cette minimisation. Dans l'algorithme de descente du gradient on choisit en chaque point la direction correspondant à la plus grande variabilité de F ; la direction indiquée par le gradient de F en ce point.

L'algorithme fonctionne alors de la façon suivante :

- itération n : on possède une approximation $x_n \in \mathbb{R}^d$ du minimum recherché ;
- itération $n+1$:
  - On pose
    $$a = \nabla F(q_n)$$
    et on calcule $b \in \mathcal{R}^d$ tel que $x_n \in \mathcal{D}$ avec $\mathcal{D} = \{\lambda a + b, \lambda \in \mathbb{R}\}$.
  - On calcule x_{n+1} le minimum de
    $$\begin{array}{rccl} f: & \mathbb{R} & \to & \mathbb{R} \\ & \lambda & \mapsto & f(\lambda) = F(\lambda a + b) \end{array}$$

On initialise l'algorithme avec une valeur x_0 quelconque, mais il est clair que plus cette valeur initiale sera proche du minimum recherché plus l'algorithme aura des chances de converger rapidement vers la bonne valeur.

Remarquons également ici que l'algorithme ne peut faire la distinction entre un minimum local et global, on pourra donc être amené à effectuer l'algorithme à plusieurs reprises avec des initialisations différentes dans le cas de suspicion d'existence de mimima locaux.

Références

[Arn51] W. E. Arnoldi. The principle of minimised itérations in the solution of the matrix eigenvalue problem. *Quart. Appl. Math.*, 9 :17–29, 1951.

[Bre73] R. P. Brent. *Algorithms for minimisation without derivatives.* Englewood Cliffs, NJ : Prentice-Hall, 1973.

[PTVF97] W. H. Press, S. A. Teukolsky, W. T. Vettering, and B. P. Flannery. *Numerical recipes in C.* Cambridge University Press, 1997.

[Saa92] Y. Saad. *Numerical methods for large eigenvalue problems.* Manchester University Press, 1992.

[Ste94] W. J. Stewart. *Introduction to numerical solution to Markov chains.* Princeton University Press, 1994.

[VTPF99] W. T. Vettering, S. A. Teukolsky, W. H. Press, and B. P. Flannery. *Numerical recipes : example book (C) second edition.* Cambridge University Press, 1999.

Annexe D

Démonstrations

Contenu du chapitre

D.1 Variables aléatoires *i.i.d.*

D.1.1 Théorème de *Cramér-Chernov*

Voici la preuve complète du théorème 4.5 (page 69). Cette preuve, directement tirée de [dH00], est néanmoins rappelée pour son analogie avec la preuve concernant l'analogue de ce théorème dans le cas des chaînes de *Markov*.

On rappelle le résultat à démontrer :

Théorème D.1 *Soit $(X_i)_{1\leqslant i\leqslant n}$ un échantillon de X est une variable aléatoire à valeurs dans $\mathcal{A} \subset \mathbb{R}$ de cardinal fini qui vérifie :*

$$\varphi(t) = \mathbb{E}[e^{tX}] < \infty, \forall t \in \mathbb{R}.$$

On pose $S_n = \sum_{i=1}^n X_i$ *et on considère* $a > \mathbb{E}[X]$ *alors :*

$$\lim_{n->+\infty} \frac{1}{n} \log \mathbb{P}(S_n \geqslant an) = -I(a)$$

avec

$$I(z) = \sup_{t \in R} [zt - \log \varphi(t)].$$

On commence par considérer la version simplifiée suivante :

Théorème D.2 *Soit* (X_i) *un échantillon de* X*, v.a.r vérifiant :*

$$\varphi(t) = \mathbb{E}[e^{tX}], \forall t \in \mathbb{R}$$

et telle que $\mathbb{E}[e^{tX}] < 0$ *alors*

$$\lim_{n->+\infty} \frac{1}{n} \log \mathbb{P}(S_n \geqslant 0) = -I(0)$$

il suffira de prouver ce second théorème puisqu'on constate que

Lemme D.3 *Le théorème D.1 est un corollaire du théorème D.2.*

Preuve. En effet, si X et a vérifient les hypothèses du théorème D.1, on pose $\bar{X} = X - a$ et on a

$$\varphi_X(t) = e^{ta} \varphi_{\bar{X}}(t)$$

donc

$$\log \varphi_X(t) = ta + \log \varphi_{\bar{X}}(t)$$

si bien que

$$I_X(a) = \sup_{t \in \mathbb{R}} (-\log \varphi_{\bar{X}}(t)) = I_{\bar{X}}(0)$$

et comme

$$\mathbb{P}(S_n^X \geqslant na) = \mathbb{P}(S_n^{\bar{X}} \geqslant 0)$$

on obtient bien le résultat souhaité. ∎

On va donc désormais s'atteler à la démonstration du théorème D.2. On note f la densité de X.

Etant donné que

$$\varphi(t) = \int_{-\infty}^{+\infty} e^{tx} f(x)\, dx < \infty \;\; \forall t \in \mathbb{R}$$

on en déduit que φ est $\mathcal{C}^\infty$ et qu'on a

$$\varphi'(t) = \int_{-\infty}^{+\infty} x e^{tx} f(x)\, dx \Rightarrow \varphi'(0) = \mathbb{E}[X] < 0$$

et
$$\varphi''(t) = \int_{-\infty}^{+\infty} x^2 e^{tx} f(x)\, dx \Rightarrow \varphi'' > 0 \text{ (car } f \equiv 0 \text{ impossible).}$$

Il est en effet facile de montrer que $y^n e^{tx} f(x)$ est intégrable pour toutes valeurs de $n \in \mathbb{N}$ car $\exists T > 0$ tel que

$$\begin{cases} |x^n| \leqslant e^{Tx} & \text{pour } x \geqslant 0 \\ |x^n| \leqslant e^{-Tx} & \text{pour } x < 0 \end{cases}$$

donc
$$|x^n e^{tx} f(x)| \leqslant e^{t-T} f(x) \mathbb{I}_{x<0} + e^{t+T} f(x) \mathbb{I}_{x \geqslant 0}$$
qui est bien intégrable.

Puisque $\mathbb{E}[X] < 0$ on a nécessairement $\mathbb{P}(X < 0) > 0$ et montre facilement que cela implique
$$\lim_{t \to -\infty} \varphi(t) = +\infty. \tag{D.1}$$

On distingue ensuite deux cas :

(i) si $\mathbb{P}(X > 0) = 0$ alors il est aisé de montrer que

$$\lim_{t \to +\infty} \varphi(t) = \mathbb{P}(X = 0)$$

Si on pose
$$\rho = \inf_{t \in \mathbb{R}} \varphi(t) \tag{D.2}$$
la décroissance de φ permet alors de montrer que $\rho = \mathbb{P}(X = 0)$ et on obtient donc clairement

$$\mathbb{P}(S_n \geqslant 0) = \mathbb{P}(S_n = 0) = \rho^n$$

ce qui achève la preuve.

(ii) Si, au contraire, $\mathbb{P}(X > 0) > 0$ alors a, de manière similaire à (D.1), la limite suivante :
$$\lim_{t \to +\infty} \varphi(t) = +\infty. \tag{D.3}$$

Dans le cas **(ii)**, on utilise conjointement (D.1) et (D.3) ainsi que la régularité et convexité de φ pour montrer qu'il existe un unique $\tau > 0$ (strictement positif car $\varphi'(0) < 0$) vérifiant

$$\varphi(\tau) = \rho \text{ et } \varphi'(\tau) = 0.$$

où ρ est toujours défini par (D.2).

On a alors

$$\mathbb{P}(S_n \geqslant 0) = \mathbb{P}(e^{\tau S_n} \geqslant 1) \leqslant \mathbb{E}[e^{\tau S_n}] = \varphi(\tau)^n = \rho^n$$

et donc

$$\overline{\lim_{n\to+\infty}} \frac{1}{n} \log \mathbb{P}(S_n \geqslant 0) \leqslant \log \rho \qquad \text{(D.4)}$$

On va maintenant s'attacher à démontrer l'autre inégalité et on va pour cela utiliser une technique de changement de probabilité. On considère $(\widehat{X}_i)$ un échantillon de la variable aléatoire $\widehat{X}$ de densité

$$\widehat{f}(x) = \frac{1}{\rho} e^{\tau x} f(x)$$

verifiant bien

$$\begin{aligned}\int_{-\infty}^{+\infty} \widehat{f}(x)dx &= \frac{1}{\rho}\int_{-\infty}^{+\infty} e^{\tau x} f(x)dx \\ &= \frac{1}{\rho}\varphi(\tau) = 1.\end{aligned}$$

On pose $\widehat{S}_n = \sum_{i=1}^n \widehat{X}_i$ et on a :

$$\begin{aligned}\mathbb{P}(S_n \geqslant 0) &= \int_{x_1+\ldots+x_n \geqslant 0} f(x_1)\ldots f(x_n)dx_1\ldots dx_n \\ &= \int_{x_1+\ldots+x_n \geqslant 0} [\rho e^{-\tau x_1}\widehat{f}(x_1)]\ldots[\rho e^{-\tau x_n}\widehat{f}(x_n)]dx_1\ldots dx_n \\ &= \rho^n \mathbb{E}[e^{-\tau \widehat{S}_n} \mathbb{I}_{\widehat{S}_n \geqslant 0}].\end{aligned}$$

On a donc

$$\frac{1}{n}\log \mathbb{P}(S_n \geqslant 0) = \log \rho + \frac{1}{n}\log \mathbb{E}[e^{-\tau \widehat{S}_n} \mathbb{I}_{\widehat{S}_n \geqslant 0}] \qquad \text{(D.5)}$$

dès lors que $\mathbb{E}[e^{-\tau \widehat{S}_n} \mathbb{I}_{\widehat{S}_n \geqslant 0}] > 0$

Pour montrer ce dernier point, on va utiliser le théorème central limite et pour cela, on va commencer par considérer les moments d'ordre 1 et 2 de $\widehat{X}$ par le biais de sa transformée de *Laplace* :

$$\begin{aligned}\widehat{\varphi}(t) = \mathbb{E}[e^{t\widehat{X}}] &= \int_{-\infty}^{+\infty} e^{tx}\widehat{f}(x)dx \\ &= \frac{1}{\rho}\int_{-\infty}^{+\infty} e^{tx}e^{\tau x} f(x)dx \\ &= \frac{1}{\rho}\varphi(t+\tau).\end{aligned}$$

Par conséquent $\widehat{\varphi}$ est finie partout et

$$\begin{aligned}\mathbb{E}[\widehat{X}] &= \widehat{\varphi}'(0) = \frac{1}{\rho}\varphi'(\tau) = 0\\ \mathbb{V}ar[\widehat{X}] &= \widehat{\varphi}''(0) = \frac{1}{\rho}\varphi''(\tau) = \sigma^2 > 0\end{aligned}$$

si bien que les conditions d'application du théorème central limite sont réunies et on obtient donc

$$\frac{1}{\sigma\sqrt{n}}\widehat{S}_n \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

et il existe donc $C > 0$ tel que, pour n assez grand,

$$\mathbb{P}\left(\frac{1}{\sigma\sqrt{n}}\widehat{S}_n \in [0,C]\right) > \frac{1}{4}$$

et donc, pour ces mêmes n,

$$\begin{aligned}\mathbb{E}[e^{-\tau\widehat{S}_n}\mathbb{I}_{\widehat{S}_n\geqslant 0}] &\geqslant e^{-\tau C\widehat{\sigma}\sqrt{n}}\mathbb{P}\left(\frac{1}{\sigma\sqrt{n}}S_n \in [0,C]\right)\\ &> 0\end{aligned}$$

si bien qu'en passant à la limite inférieure en n dans (D.5) on obtient

$$\varliminf_{n\to+\infty}\frac{1}{n}\log\mathbb{P}(S_n \geqslant 0) \geqslant \log\rho$$

qui, combiné à (D.4), achève la preuve.

D.1.2 Théorème de *Sanov*

Voici la preuve complète du théorème 4.12 (page 76). Cette preuve, est en partie tirée de [dH00]. Il est néanmoins intéressant de l'examiner en détails dans la mesure où la preuve du théorème concernant la distribution empirique des paires s'en inspire largement.

Théorème D.4 *Si L_n désigne la distribution empirique d'un échantillon de taille n de X variable aléatoire de loi μ sur $\mathcal{A}$ un alphabet fini, et si on suppose (sans perte de généralité) que $\mu(x) > 0$ pour tout $x \in \mathcal{A}$, alors on a les deux résultats suivants :*

(MAJ) *pour tout $F \subset \mathcal{M}_1(\mathcal{A})$ fermé on a*

$$\varlimsup_{n\to+\infty}\frac{1}{n}\log\mathbb{P}\left(L_n \in F\right) \leq -\inf_{\nu\in F} I(\nu);$$

(MIN) *et pour tout $O \subset \mathcal{M}_1(\mathcal{A})$ ouvert on a*

$$\varliminf_{n\to+\infty} \frac{1}{n} \log \mathbb{P}\left(L_n \in O\right) \geqslant - \inf_{\nu\in O} I(\nu)$$

où

$$I(\nu) = H(\nu|\mu) = \sum_{x\in\mathcal{A}} \nu(x) \log \frac{\nu(x)}{\mu(x)}$$

*désigne l'*entropie relative *de ν par rapport à μ.*

On va en fait se contenter de montrer une version simplifiée du théorème D.4 :

Théorème D.5 *Sous les mêmes hypothèses on a :*

$$\lim_{n\to+\infty} \frac{1}{n} \log \mathbb{P}\left(L_n \in B\right) = - \inf_{\nu\in B} I(\nu)$$

où B désigne une boule ouverte quelconque pour la distance de variation totale sur $\mathcal{M}_1(\mathcal{A})$.

car on peut montrer que

Lemme D.6 *Le théorème D.4 est un corollaire du théorème D.5*

Preuve.

(MIN) Soit $O \subset \mathcal{M}_1(\mathcal{A})$ un ouvert et soit $x \in O$. Comme O est ouvert, $\exists \rho_x > 0$ tel que, si $B(x,\rho_x) = \{y \in \mathcal{M}_1(\mathcal{A}), d(x,y) < \rho_x\}$ désigne la boule ouverte de centre x et de rayon ρ_x, alors on a $B(x,\rho_x) \subset O$.
Cette relation d'inclusion assure que

$$\mathbb{P}(L_n \in O) \leqslant \mathbb{P}\left(L_n \in B(x,\rho_x)\right)$$

et en passant à la limite inférieure on trouve

$$\begin{aligned} \varliminf_{n\to+\infty} \frac{1}{n} \log \mathbb{P}(L_n \in O) &\leq \varliminf_{n\to+\infty} \frac{1}{n} \log \mathbb{P}\left(L_n \in B(x,\rho_x)\right) \\ &\leq -\inf_{B(x,\rho_x)} \end{aligned} \tag{D.6}$$

en utilisant le théorème D.5.
Comme

$$\inf_{B(x,\rho_x)} I \leqslant I(y) \quad \forall y \in B(x,\rho_x)$$

on obtient

$$\varliminf_{n\to+\infty} \frac{1}{n} \log \mathbb{P}(L_n \in O) \leqslant -I(x) \tag{D.7}$$

en appliquant ce résultat à $y = x$ et en utilisant (D.6).
$x \in O$ ayant été choisi arbitrairement il ne reste plus qu'à prendre le sup en x sur O de (D.7) pour achever la preuve de ce point.

(MAJ) Soit $F \subset \mathcal{M}_1(\mathcal{A})$ un fermé. Comme $\mathcal{A}$ est fini, il est clair que $\mathcal{M}_1(\mathcal{A})$ est compact et donc que F est compact.

Soit $\varepsilon > 0$. Par continuité de I on sait que $\forall x \in F$, $\exists \eta_x > 0$ tel que $\forall y \in \mathcal{M}_1(\mathcal{A})$ on ait

$$d(y,x) < \eta_x \Rightarrow |I(y) - I(x)| \leqslant \varepsilon. \tag{D.8}$$

L'ensemble $(B_x)_{x \in F}$, où $B_x = B(x, \eta_x)$ désigne la boule ouverte de centre x et de rayon η, réalise clairement un recouvrement ouvert de F dont on peut, par compacité, extraire un sous-recouvrement fini.

On considère donc une suite finie $x_1, \ldots, x_N$ réalisant

$$F \subset \bigcup_{i=1}^{N} B_{x_i}$$

et on a alors la majoration naturelle

$$\mathbb{P}(L_n \in F) \leqslant \sum_{i=1}^{N} \mathbb{P}(L_n \in B_{x_i}) \leqslant N \times \max_i \mathbb{P}(L_n \in B_{x_i})$$

En passant à la limite supérieure dans cette dernière équation on obtient

$$\begin{aligned} \overline{\lim_{n\to+\infty}} \frac{1}{n} \log \mathbb{P}(L_n \in F) &\leq \overline{\lim_{n\to+\infty}} \frac{1}{n} \log \max_i \mathbb{P}(L_n \in B_{x_i}) \\ &\leq \max_i - \inf_{B_{x_i}} I. \end{aligned} \tag{D.9}$$

Or si $y \in B_{x_i}$, (D.8) nous assure que

$$I(x_i) - \varepsilon \leqslant I(y) \leqslant I(x_i) + \varepsilon$$

et donc on a

$$\inf_{B_{x_i}} I \geqslant I^{\varepsilon}(x_i) \tag{D.10}$$

avec I^{ε} définie $\forall y \in \mathcal{M}_1(\mathcal{A})$ par $I^{\varepsilon}(y) = I(y) - \varepsilon$.

Comme $x_i \in F$, il est par ailleurs clair que

$$I^{\varepsilon}(x_i) \geqslant \inf_F I^{\varepsilon}. \tag{D.11}$$

Le résultat s'obtient en combinant (D.9), (D.10) et (D.11).

■

On note k le cardinal de $\mathcal{A}$ et on pose

$$\mathcal{S}_n = \left\{ s = (s_x)_{x \in \mathcal{A}} \in (\mathbb{N}^*)^k, \sum_{x \in \mathcal{A}} s_x = n \right\}.$$

Il est alors bien clair que $\frac{1}{n}\mathcal{S}_n \subset \mathcal{M}_1(\mathcal{A})$.

Soit $s \in \mathcal{S}_n$ alors

$$\begin{aligned} \mathbb{P}\left(L_n = \nu_n(s)\right) &= \mathbb{P}\left(L_n(x) = \nu_n(s_x),\ \forall x \in \mathcal{A}\right) \\ &= \mathbb{P}\left(\sum_{i=1}^{n} \mathbb{I}_{X_i = x} = s_x,\ \forall x \in \mathcal{A}\right) \\ &= n! \prod_{x \in \mathcal{A}} \frac{\mu_x^{s_x}}{s_x!} \end{aligned} \tag{D.12}$$

où $\mu = (\mu_x)_{x \in \mathcal{A}}$ désigne la loi des $(X_i)_i$ et $\nu_n(s) = \frac{1}{n}s$ et $\nu_n(s_x) = \frac{1}{n}s_x$.

Grâce à la formule de *Stirling* qui donne un équivalent de factoriel n en l'infini

$$n! \sim \sqrt{\frac{2\pi}{n}} e^{-n} n^n,$$

on peut établir que

$$\begin{aligned} n! \prod_{x \in \mathcal{A}} \frac{\mu_x^{s_x}}{s_x!} &\sim \left(\sqrt{\frac{n}{2\pi}}\right)^{k-1} e^{-n} e^{n \log n} \frac{\prod_{x \in \mathcal{A}} e^{s_x \log \mu_x}}{\prod_{x \in \mathcal{A}} e^{-s_x} e^{s_x \log s_x}} \\ &\sim \left(\sqrt{\frac{n}{2\pi}}\right)^{k-1} \prod_{x \in \mathcal{A}} e^{s_x \log \frac{\mu_x}{\nu_n(s_x)}}. \end{aligned} \tag{D.13}$$

En combinant (D.12) et (D.13) on obtient :

$$\frac{1}{n} \log \mathbb{P}\left(L_n = \nu_n(s)\right) = O\left(\frac{\log n}{n}\right) - I(\nu_n(s)) \tag{D.14}$$

uniformément en $s \in \mathcal{S}_n$.

Si on pose par ailleurs

$$M_n = \max_{s \in \mathcal{S}_n \text{ tq } \nu_n(s) \in B} \mathbb{P}\left(L_n = \nu_n(s)\right)$$

alors on a clairement

$$M_n \leqslant \mathbb{P}(L_n \in B \cap \frac{1}{n}\mathcal{S}_n) \leqslant |\mathcal{S}_n| M_n. \tag{D.15}$$

Comme $|\mathcal{S}_n| \sim n^{k-1}$ on a également $\frac{1}{n}\log|\mathcal{S}_n| = O\left(\frac{\log n}{n}\right)$ si bien qu'avec l'aide de (D.13) et (D.15) on trouve

$$\frac{1}{n}\log\mathbb{P}(L_n \in B \cap \frac{1}{n}\mathcal{S}_n) = O\left(\frac{\log n}{n}\right) - \inf_{B\cap\frac{1}{n}\mathcal{S}_n} I. \tag{D.16}$$

Enfin, comme

$$\bigcup_{n>0}\frac{1}{n}\mathcal{S}_n$$

est dense dans $\mathcal{M}_1(\mathcal{A})$ et que I est continue, on achève la démonstration en passant simplement à la limite en n dans (D.16).

D.1.3 Théorème de *Sanov* pour les paires

On présente ici la preuve complète du théorème 4.14 (page 78). Cette preuve s'inspire de [dH00] et fonctionne sur le même principe que la preuve de la section précédente.

Théorème D.7 *On a les deux résultats suivants :*

(MAJ) *pour tout $F \subset \mathcal{M}_1(\mathcal{A}^2)$ fermé on a*

$$\overline{\lim_{n\to+\infty}}\frac{1}{n}\log\mathbb{P}\left(L_n^2 \in F\right) \leq -\inf_{\nu\in F} I_2(\nu);$$

(MIN) *et pour tout $O \subset \mathcal{M}_1(\mathcal{A}^2)$ ouvert on a*

$$\overline{\lim_{n\to+\infty}}\frac{1}{n}\log\mathbb{P}\left(L_n^2 \in O\right) \geqslant -\inf_{\nu\in O} I_2(\nu)$$

avec

$$I_2(\nu) = \begin{cases} \sum_{x,y\in\mathcal{A}} \nu(x,y)\log\frac{\nu(x,y)}{\overline{\nu}(x)\mu(y)} & \text{si } \nu \in \mathcal{S} \\ +\infty & \text{sinon} \end{cases}$$

où $\mathcal{S}$ désigne l'ensemble des éléments ν de $\mathcal{M}_1(\mathcal{A}^2)$ qui sont invariants par translations *c'est à dire qui vérifient :*

$$\overline{\nu}(x) = \sum_{y\in\mathcal{A}}\nu(x,y) = \sum_{y\in\mathcal{A}}\nu(y,x). \tag{D.17}$$

Grâce à un argument identique à celui du lemme D.6, on se contente de montrer le

Théorème D.8 *Si B désigne une boule ouverte quelconque pour la distance de variation totale sur $\mathcal{M}_1(\mathcal{A}^2)$ alors*

$$\lim_{n\to+\infty} \frac{1}{n} \log \mathbb{P}\left(L_n \in B\right) = -\inf_{\nu\in B} I_2(\nu)$$

avec les mêmes hypothèses et définitions qu'au théorème D.7.

On note k le cardinal de $\mathcal{A}$ et on pose

$$\mathcal{S}_n = \left\{ s = (s_{xy})_{x,y\in\mathcal{A}} \in (\mathbb{N}^*)^{2k}, \sum_{x,y\in\mathcal{A}} s_{xy} = n \right\}$$

et il est alors bien clair que $\frac{1}{n}\mathcal{S}_n \subset \mathcal{M}_1(\mathcal{A}^2)$.

Pour tout $s \in \mathcal{S}_n$ on pose $\nu_n(s) = \frac{1}{n}s$ et on a alors

$$\mathbb{P}\left(L_n^2 = \nu_n(s)\right) = K(s) \times \prod_{x\in\mathcal{A}} \mu_x^{\overline{s}_x} \tag{D.18}$$

où $K(s)$ désigne le nombre de séquences de longueur n dont les comptages vérifient

$$N_{xy} = s_{xy} \quad \forall x, y \in \mathcal{A}$$

et où

$$\overline{s}_x = \frac{1}{2}\left(\sum_{y\in\mathcal{A}} s_{xy} + \sum_{y\in\mathcal{A}} s_{yx}\right).$$

Si on suppose que $X_{n+1} = X_n$ (hypothèse non restrictive), il se trouve que la valeur de $K(s)$ est donnée par *Wittle* ([Whi55]) voir section B.2 (page 195) :

$$K(s) = \frac{\prod_{x\in\mathcal{A}} \overline{s}_x!}{\prod_{x,y\in\mathcal{A}} s_{xy}!} \times H(s)$$

avec

$$\frac{1}{n^{k-1}} \leqslant H(s) \leqslant (k-1)! \tag{D.19}$$

si

$$\overline{s}_x = \sum_{y\in\mathcal{A}} s_{xy} = \sum_{y\in\mathcal{A}} s_{yx}$$

et $K(s) = 0$ sinon.

Par conséquent, on s'intéresse désormais uniquement aux cas des $s \in \mathcal{S}'_n \subset \mathcal{S}_n$ avec

$$\mathcal{S}'_n = \left\{ s = (s_{xy})_{x,y\in\mathcal{A}} \in (\mathbb{N}^*)^{2k}, \sum_{x,y\in\mathcal{A}} s_{xy} = n \text{ et } \sum_{x\in\mathcal{A}} s_{xy} = \sum_{x\in\mathcal{A}} s_{yx} \right\}$$

pour lesquels la formule D.18 devient alors

$$\mathbb{P}\left(L_n^2 = \nu_n(s)\right) = H(s) \times \frac{\prod_{x\in\mathcal{A}} \overline{s}_x!}{\prod_{x,y\in\mathcal{A}} s_{xy}!} \times \prod_{x\in\mathcal{A}} \mu_x^{\overline{s}_x}. \tag{D.20}$$

La formule de *Stirling* (voir page 218) donne :

$$\begin{aligned}\frac{\prod_{x\in\mathcal{A}} \overline{s}_x!}{\prod_{x,y\in\mathcal{A}} s_{xy}!} \times \prod_{x\in\mathcal{A}} \mu_x^{\overline{s}_x} &\sim \left(\sqrt{\frac{n}{2\pi}}\right)^k \times \frac{\prod_{x\in\mathcal{A}} e^{-\overline{s}_x}\overline{s}_x^{\overline{s}_x}}{\prod_{x,y\in\mathcal{A}} e^{-s_{xy}} s_{xy}^{s_{xy}}} \times \prod_{x\in\mathcal{A}} \mu_x^{\overline{s}_x} \\ &\sim \left(\sqrt{\frac{n}{2\pi}}\right)^k \times \exp\left(\sum_{x,y\in\mathcal{A}} s_{xy} \log \frac{\overline{s}_x \mu_x}{s_{xy}}\right) \\ &\sim \left(\sqrt{\frac{n}{2\pi}}\right)^k \times e^{-nI_2(\nu_n(s))} \end{aligned} \tag{D.21}$$

En utilisant l'encadrement (D.19) dans la formule (D.20) et l'équivalent trouvé en (D.21) on obtient finalement

$$\mathbb{P}\left(L_n^2 = \nu_n(s)\right) = O\left(\frac{\log n}{n}\right) - I_2(\nu_n(s)).$$

On conclut alors de la même façon qu'en section D.1.2 en utilisant la densité de

$$\bigcup_{n>0} \frac{1}{n}\mathcal{S}'_n$$

dans l'ensemble $\mathcal{S}$ des éléments $\mathcal{M}_1(\mathcal{A})$ qui sont invariants par translation.

D.2 Propriétés de Λ

On trouve ici les calculs concernant les preuves des deux résultats concernant les dérivées premières et secondes de la fonction Λ en section 5.1 (page 84). On généralise dans cette partie les résultats données dans [Tor98].

On commence par rappeler quelques définitions et hypothèses : Π est une matrice stochastique irréductible sur l'espace d'état fini $\mathcal{A}$ de cardinal k ; $f : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ est une fonction déterministe quelconque ; $\forall \theta \in \mathbb{R}$ la matrice Π_θ est définie par

$$\Pi_\theta(x,y) = \Pi(x,y)e^{\theta f(x,y)} \quad \forall x,y \in \mathcal{A};$$

on note $\rho(\Pi_\theta)$ la valeur propre de *Perron-Frobénius* de Π_θ ; $\forall \theta \in \mathbb{R}$ on note $\Lambda(\theta) = \log \rho(\Pi_\theta)$.

Si on note v_θ (resp. w_θ) un vecteur propre à droite (resp. à gauche) de Π_θ associé à la valeur propre $\rho(\Pi_\theta)$ on définit alors la matrice stochastique et irréductible $\widetilde{\Pi}_\theta$ par

$$\widetilde{\Pi}_\theta(x,y) = e^{-\Lambda(\theta)}\frac{v_\theta(y)}{v_\theta(x)}\Pi_\theta(x,y) = e^{-\Lambda(\theta)}\frac{v_\theta(y)}{v_\theta(x)}\Pi(x,y)e^{\theta f(x,y)}$$

dont la distribution stationnaire q_θ est définie (à un facteur multiplicatif près) par

$$q_\theta = [v_\theta(1)w_\theta(1)\dots v_\theta(k)w_\theta(k)].$$

On commence alors par établir le lemme suivant :

Lemme D.9 *On a :*

$$\begin{cases} \bullet \sum\limits_{y\in\mathcal{A}} e^{-\Lambda(\theta)}\Pi_\theta(x,y)v_\theta(y) = v_\theta(x) & \forall x \in \mathcal{A} \\ \bullet \sum\limits_{x\in\mathcal{A}} e^{-\Lambda(\theta)}w_\theta(x)\Pi_\theta(x,y) = w_\theta(y) & \forall y \in \mathcal{A} \end{cases} \tag{D.22}$$

et

$$\begin{cases} \bullet\ \Lambda'(\theta)v_\theta(x) + v'_\theta(x) = \sum\limits_{y\in\mathcal{A}} f(x,y)e^{-\Lambda(\theta)}\Pi_\theta(x,y)v_\theta(y) \\ \quad + \sum\limits_{y\in\mathcal{A}} e^{-\Lambda(\theta)}\Pi_\theta(x,y)v'_\theta(y) \quad \forall x \in \mathcal{A} \\ \bullet\ \Lambda'(\theta)w_\theta(y) + w'_\theta(y) = \sum\limits_{x\in\mathcal{A}} f(x,y)w_\theta(x)e^{-\Lambda(\theta)}\Pi_\theta(x,y) \\ \quad + \sum\limits_{x\in\mathcal{A}} w'_\theta(x)e^{-\Lambda(\theta)}\Pi_\theta(x,y) \quad \forall y \in \mathcal{A} \end{cases} \tag{D.23}$$

Preuve. On obtient (D.22) en utilisant les définitions de v_θ et w_θ :

$$\Pi_\theta v_\theta = \rho(\Pi_\theta)v_\theta \text{ et } w_\theta\Pi_\theta = \rho(\Pi_\theta)w_\theta$$

et (D.23) s'obtient par simple dérivation de (D.22). ■

D.2.1 Dérivées premières

On rappelle ici la proposition 5.3 (page 86) :

Proposition D.10 *Pour tout $\theta \in \mathbb{R}$ on a*

$$\Lambda'(\theta) = \mathbb{E}_{q_\theta}^{\widetilde{\Pi}_\theta}[f(X_1, X_2)].$$

Preuve. $\forall x \in \mathcal{A}$ on dérive l'équation suivante :

$$\sum_{y\in\mathcal{A}} \widetilde{\Pi}_\theta(x,y) = \sum_{y\in\mathcal{A}} e^{-\Lambda(\theta)} \frac{v_\theta(y)}{v_\theta(x)} \Pi(x,y)\, e^{\theta f(x,y)} = 1$$

et on obtient

$$\begin{aligned}
\sum_{y\in\mathcal{A}} f(x,y)\,\widetilde{\Pi}_\theta(x,y) &= \sum_{y\in\mathcal{A}} \Lambda'(\theta)\,\widetilde{\Pi}_\theta(x,y) - \sum_{y\in\mathcal{A}} e^{-\Lambda(\theta)} \frac{v'_\theta(y)\,v_\theta(x)}{v^2_\theta(x)} \Pi_\theta(x,y) \\
&\quad + \sum_{y\in\mathcal{A}} e^{-\Lambda(\theta)} \frac{v_\theta(y)\,v'_\theta(x)}{v^2_\theta(x)} \Pi_\theta(x,y) \\
&= \Lambda'(\theta) - \sum_{y\in\mathcal{A}} e^{-\Lambda(\theta)} \frac{v'_\theta(y)}{v_\theta(x)} \Pi_\theta(x,y) + \frac{v'_\theta(x)}{v_\theta(x)}.
\end{aligned}$$

on a donc

$$\begin{aligned}
&\sum_{x,y\in\mathcal{A}} f(x,y)\,q_\theta(x)\,\widetilde{\Pi}_\theta(x,y) \\
=\;& \sum_{x\in\mathcal{A}} q_\theta(x) \sum_{y\in\mathcal{A}} f(x,y)\,\widetilde{\Pi}_\theta(x,y) \\
=\;& \sum_{x\in\mathcal{A}} q_\theta(x)\,\Lambda'(\theta) - \sum_{x\in\mathcal{A}} q_\theta(x) \sum_{y\in\mathcal{A}} e^{-\Lambda(\theta)} \frac{v'_\theta(y)}{v_\theta(x)} \Pi_\theta(x,y) + \sum_{x\in\mathcal{A}} q_\theta(x)\, \frac{v'_\theta(x)}{v_\theta(x)}
\end{aligned}$$

et donc

$$\begin{aligned}
&\sum_{x,y\in\mathcal{A}} f(x,y)\,q_\theta(x)\,\widetilde{\Pi}_\theta(x,y) \\
=\;& \Lambda'(\theta) - \sum_{x,y\in\mathcal{A}} \frac{v_\theta(x)\,w_\theta(x)}{C} e^{-\Lambda(\theta)} \frac{v'_\theta(y)}{v_\theta(x)} \Pi_\theta(x,y) + \sum_{x\in\mathcal{A}} \frac{v_\theta(x)\,w_\theta(x)}{C} \times \frac{v'_\theta(x)}{v_\theta(x)} \\
=\;& \Lambda'(\theta) - \frac{1}{C} \sum_{x,y\in\mathcal{A}} v'_\theta(y)\, e^{-\Lambda(\theta)} w_\theta(x)\,\Pi_\theta(x,y) + \frac{1}{C} \sum_{x\in\mathcal{A}} w_\theta(x)\,v'_\theta(x) \\
=\;& \Lambda'(\theta) - \frac{1}{C} \sum_{y\in\mathcal{A}} v'_\theta(y) \sum_{x\in\mathcal{A}} e^{-\Lambda(\theta)} w_\theta(x)\,\Pi_\theta(x,y) + \frac{1}{C} \sum_{x\in\mathcal{A}} w_\theta(x)\,v'_\theta(x) \\
=\;& \Lambda'(\theta) - \frac{1}{C} \sum_{y\in\mathcal{A}} v'_\theta(y)\,w_\theta(y) + \frac{1}{C} \sum_{x\in\mathcal{A}} w_\theta(x)\,v'_\theta(x) \\
=\;& \Lambda'(\theta).
\end{aligned}$$

■

D.2.2 Dérivées secondes

On rappelle ici la proposition 5.4 (page 86) :

Proposition D.11 *Pour tout* $\theta \in \mathbb{R}$ *on a*

$$\begin{aligned} \Lambda''(\theta) &= \sum_{n\in\mathbb{Z}} \mathbb{C}ov_{q_\theta}^{\widetilde{\Pi}_\theta} \left[f(X_1, X_2), f(X_n, X_{n+1})\right] & \text{(D.24)} \\ &= \lim_{n\to+\infty} \frac{1}{n} \mathbb{V}ar_{q_\theta}^{\widetilde{\Pi}_\theta} \left[\sum_{i=1}^{n} f(X_i, X_{i+1})\right]. & \text{(D.25)} \end{aligned}$$

Preuve. On commence par dériver

$$\Lambda'(\theta) = \sum_{y\in\mathcal{A}} f(x,y)\widetilde{\Pi}_\theta(x,y) + \sum_{y\in\mathcal{A}} \frac{v'_\theta(y)}{v_\theta(y)}\widetilde{\Pi}_\theta(x,y) - \frac{v'_\theta(x)}{v_\theta(x)} \tag{D.26}$$

ce qui donne

$$\begin{aligned} \Lambda''(\theta) &= \sum_{y\in\mathcal{A}} f(x,y)^2\widetilde{\Pi}_\theta(x,y) - \Lambda'(\theta)\sum_{y\in\mathcal{A}} f(x,y)\widetilde{\Pi}_\theta(x,y) \\ &+ \sum_{y\in\mathcal{A}} f(x,y)e^{-\Lambda(\theta)}\frac{v'_\theta(y)v_\theta(x)}{v_\theta(x)^2}\Pi_\theta(x,y) \\ &- \sum_{y\in\mathcal{A}} f(x,y)e^{-\Lambda(\theta)}\frac{v_\theta(y)v'_\theta(x)}{v_\theta(x)^2}\Pi_\theta(x,y) - \Lambda'(\theta)\sum_{y\in\mathcal{A}} e^{-\Lambda(\theta)}\frac{v'_\theta(y)}{v_\theta(x)}\Pi_\theta(x,y) \\ &+ \sum_{y\in\mathcal{A}} e^{-\Lambda(\theta)}\frac{v''_\theta(y)v_\theta(x)}{v_\theta(x)^2}\Pi_\theta(x,y) - \sum_{y\in\mathcal{A}} e^{-\Lambda(\theta)}\frac{v'_\theta(y)v'_\theta(x)}{v_\theta(x)^2}\Pi_\theta(x,y) \\ &+ \sum_{y\in\mathcal{A}} f(x,y)e^{-\Lambda(\theta)}\frac{v'_\theta(y)}{v_\theta(x)}\Pi_\theta(x,y) - \frac{v''_\theta(x)v_\theta(x)}{v_\theta(x)^2} + \frac{v'_\theta(x)^2}{v_\theta(x)^2} \end{aligned}$$

que l'on simplifie en

$$\begin{aligned} \Lambda''(\theta) &= \sum_{y\in\mathcal{A}} f(x,y)^2\widetilde{\Pi}_\theta(x,y) + \sum_{y\in\mathcal{A}} f(x,y)\frac{v'_\theta(y)}{v_\theta(y)}\widetilde{\Pi}_\theta(x,y) \\ &- \Lambda'(\theta)\left(\sum_{y\in\mathcal{A}} f(x,y)\widetilde{\Pi}_\theta(x,y) + \sum_{y\in\mathcal{A}}\frac{v'_\theta(y)}{v_\theta(y)}\widetilde{\Pi}_\theta(x,y)\right) \\ &- \frac{v'_\theta(x)}{v_\theta(x)}\left(\sum_{y\in\mathcal{A}} f(x,y)\widetilde{\Pi}_\theta(x,y) + \sum_{y\in\mathcal{A}}\frac{v'_\theta(y)}{v_\theta(y)}\widetilde{\Pi}_\theta(x,y)\right) \\ &+ \sum_{y\in\mathcal{A}} \frac{v''_\theta(y)}{v_\theta(y)}\widetilde{\Pi}_\theta(x,y) + \sum_{y\in\mathcal{A}} f(x,y)\frac{v'_\theta(y)}{v_\theta(y)}\widetilde{\Pi}_\theta(x,y) - \frac{v''_\theta(x)}{v_\theta(x)} + \frac{v'_\theta(x)^2}{v_\theta(x)^2} \end{aligned}$$

et, en utilisant (D.26), on se ramène ainsi à

$$\begin{aligned}
\Lambda''(\theta) &= \sum_{y\in\mathcal{A}} f(x,y)^2\widetilde{\Pi}_\theta(x,y) - \Lambda'(\theta)^2 - \Lambda'(\theta)\frac{v'_\theta(x)}{v_\theta(x)} \\
&+ \sum_{y\in\mathcal{A}} f(x,y)\frac{v'_\theta(y)}{v_\theta(y)}\widetilde{\Pi}_\theta(x,y) - \Lambda'(\theta)\frac{v'_\theta(x)}{v_\theta(x)} - \frac{v'_\theta(x)^2}{v_\theta(x)^2} \\
&+ \sum_{y\in\mathcal{A}} \frac{v''_\theta(y)}{v_\theta(y)}\widetilde{\Pi}_\theta(x,y) + \sum_{y\in\mathcal{A}} f(x,y)\frac{v'_\theta(y)}{v_\theta(y)}\widetilde{\Pi}_\theta(x,y) - \frac{v''_\theta(x)}{v_\theta(x)} + \frac{v'_\theta(x)^2}{v_\theta(x)^2}
\end{aligned}$$

que l'on peut réécrire

$$\begin{aligned}
\Lambda''(\theta) &= \sum_{y\in\mathcal{A}} f(x,y)^2\widetilde{\Pi}_\theta(x,y) - \Lambda'(\theta)^2 + \sum_{y\in\mathcal{A}} \frac{v''_\theta(y)}{v_\theta(y)}\widetilde{\Pi}_\theta(x,y) - \frac{v''_\theta(x)}{v_\theta(x)} \\
&+ 2\sum_{y\in\mathcal{A}} f(x,y)\frac{v'_\theta(y)}{v_\theta(y)}\widetilde{\Pi}_\theta(x,y) - 2\Lambda'(\theta)\frac{v'_\theta(x)}{v_\theta(x)}.
\end{aligned}$$

On somme cette dernière relation sur x avec $q_\theta(x)$ et on obtient

$$\begin{aligned}
\Lambda''(\theta) &= \mathbb{V}ar_{q_\theta}^{\widetilde{\Pi}_\theta}[f(X_1,X_2)] + \sum_{x,y\in\mathcal{A}} \frac{v''_\theta(y)}{v_\theta(y)}q_\theta(x)\widetilde{\Pi}_\theta(x,y) - \sum_{x\in\mathcal{A}} q_\theta(x)\frac{v''_\theta(x)}{v_\theta(x)} \\
&+ 2\sum_{x,y\in\mathcal{A}} f(x,y)\frac{v'_\theta(y)}{v_\theta(y)}q_\theta(x)\widetilde{\Pi}_\theta(x,y) - 2\Lambda'(\theta)\sum_{x\in\mathcal{A}} q_\theta(x)\frac{v'_\theta(x)}{v_\theta(x)}.
\end{aligned}$$

on remplace $q_\theta(x)$ par $w_\theta(x)v_\theta(x)$ (on suppose que $C=1$ pour simplifier les calculs) et on trouve

$$\begin{aligned}
\Lambda''(\theta) &= \mathbb{V}ar_{q_\theta}^{\widetilde{\Pi}_\theta}[f(X_1,X_2)] + \sum_{x,y\in\mathcal{A}} w_\theta(x)v''_\theta(y)\Pi_\theta(x,y)e^{-\Lambda(\theta)} - \sum_{x\in\mathcal{A}} w_\theta(x)v''_\theta(x) \\
&+ 2\sum_{x,y\in\mathcal{A}} f(x,y)w_\theta(x)v'_\theta(y)\Pi_\theta(x,y)e^{-\Lambda(\theta)} - 2\Lambda'(\theta)\sum_{x\in\mathcal{A}} w_\theta(x)v'_\theta(x).
\end{aligned}$$

En utilisant (D.22) et (D.23) on se ramène finalement a

$$\Lambda''(\theta) = \mathbb{V}ar_{q_\theta}^{\widetilde{\Pi}_\theta}[f(X_1,X_2)] + 2R \tag{D.27}$$

avec

$$R = \sum_{y\in\mathcal{A}} w'_\theta(y)v'_\theta(y) - \sum_{x,y\in\mathcal{A}} w'_\theta(x)v'_\theta(y)\Pi_\theta(x,y)e^{-\Lambda(\theta)}$$

Pour tout $n\in\mathbb{Z}$ on définit

$$C_n = \mathbb{C}ov_{q_\theta}^{\widetilde{\Pi}_\theta}\left[f(X_1,X_2), f(X_n,X_{n+1})\right]$$

et on remarque d'une part, que $C_1 = \mathbb{V}ar_{q_\theta}^{\widetilde{\Pi}_\theta}[f(X_1, X_2)]$ et d'autre part, que $C_n = C_{2-n}$ $\forall n \geqslant 2$ par stationarité de la chaîne de *Markov* ; on va donc désormais s'intéresser aux termes C_n pour $n \geqslant 2$.

Comme

$$\mathbb{P}_{q_\theta}^{\widetilde{\Pi}_\theta}(X_1 = x, X_2 = y, X_3 = z) = q_\theta(x)\widetilde{\Pi}_\theta(x, y)\widetilde{\Pi}_\theta(y, z)$$

on a

$$\begin{aligned} C_2 &= \sum_{x,y,z\in\mathcal{A}} \left(f(x, y) - \Lambda'(\theta)\right) \left(f(y, z) - \Lambda'(\theta)\right) q_\theta(x)\widetilde{\Pi}_\theta(x, y)\widetilde{\Pi}_\theta(y, z) \\ &= \sum_{x,y\in\mathcal{A}} \left(f(x, y) - \Lambda'(\theta)\right) q_\theta(x)\widetilde{\Pi}_\theta(x, y) \sum_{z\in\mathcal{A}} \left(f(y, z) - \Lambda'(\theta)\right) \widetilde{\Pi}_\theta(y, z) \end{aligned}$$

on utilise alors (D.26) pour obtenir

$$\begin{aligned} C_2 &= \sum_{x,y\in\mathcal{A}} \left(f(x, y) - \Lambda'(\theta)\right) q_\theta(x)\widetilde{\Pi}_\theta(x, y)\frac{v_\theta'(y)}{v_\theta(y)} \\ &- \sum_{x,y,z\in\mathcal{A}} \left(f(x, y) - \Lambda'(\theta)\right) q_\theta(x)\widetilde{\Pi}_\theta(x, y)\widetilde{\Pi}_\theta(y, z)\frac{v_\theta'(z)}{v_\theta(z)}. \end{aligned}$$

On considère maintenant $n > 2$ et on s'intéresse à C_n. De même que précédemment, on constate tout d'abord que

$$\mathbb{P}_{q_\theta}^{\widetilde{\Pi}_\theta}(X_1 = x, X_2 = y, X_n = z, X_{n+1} = t) = q_\theta(x)\widetilde{\Pi}_\theta(x, y)\widetilde{\Pi}_\theta^{n-1}(y, z)\widetilde{\Pi}_\theta(z, t)$$

en utilisant (D.26) et la définition d'un produit matriciel on trouve

$$\begin{aligned} A_n(y) &= \sum_{z,t\in\mathcal{A}} \left(f(z, t) - \Lambda'(\theta)\right) \widetilde{\Pi}_\theta^{n-1}(y, z)\widetilde{\Pi}_\theta(z, t) \\ &= \sum_{z\in\mathcal{A}} \widetilde{\Pi}_\theta^{n-1}(y, z) \sum_{t\in\mathcal{A}} \left(f(z, t) - \Lambda'(\theta)\right) \widetilde{\Pi}_\theta(z, t) \\ &= \sum_{z\in\mathcal{A}} \frac{v_\theta'(z)}{v_\theta(z)}\widetilde{\Pi}_\theta^{n-1}(y, z) - \sum_{z,t\in\mathcal{A}} \frac{v_\theta'(t)}{v_\theta(t)}\widetilde{\Pi}_\theta^{n-1}(y, z)\widetilde{\Pi}_\theta(z, t) \\ &= \sum_{z\in\mathcal{A}} \frac{v_\theta'(z)}{v_\theta(z)}\widetilde{\Pi}_\theta^{n-1}(y, z) - \sum_{z\in\mathcal{A}} \frac{v_\theta'(z)}{v_\theta(z)}\widetilde{\Pi}_\theta^{n}(y, z) \end{aligned}$$

si bien que

$$\begin{aligned} C_n &= \sum_{x,y\in\mathcal{A}} \left(f(x, y) - \Lambda'(\theta)\right) q_\theta(x)\widetilde{\Pi}_\theta(x, y)A_n(y) \\ &= T_{n-1} - T_n \end{aligned}$$

en posant

$$\left\{\begin{array}{l} T_0 = \sum_{x,y\in\mathcal{A}} \left(f(x,y) - \Lambda'(\theta)\right) q_\theta(x)\widetilde{\Pi}_\theta(x,y)\frac{v'_\theta(y)}{v_\theta(y)} \\ \\ T_n = \sum_{x,y,z\in\mathcal{A}} \left(f(x,y) - \Lambda'(\theta)\right) q_\theta(x)\widetilde{\Pi}_\theta(x,y)\widetilde{\Pi}^n_\theta(y,z)\frac{v'_\theta(z)}{v_\theta(z)} \quad \text{pour } n \geqslant 1 \end{array}\right. .$$

Effectuons le calcul de T_0 :

$$\begin{aligned} T_0 &= \sum_{x,y\in\mathcal{A}} f(x,y)q_\theta(x)\widetilde{\Pi}_\theta(x,y)\frac{v'_\theta(y)}{v_\theta(y)} - \Lambda'(\theta)\sum_{x,y\in\mathcal{A}} q_\theta(x)\widetilde{\Pi}_\theta(x,y)\frac{v'_\theta(y)}{v_\theta(y)} \\ &= A - B \end{aligned}$$

avec

$$\begin{aligned} A &= \sum_{x,y\in\mathcal{A}} f(x,y)w_\theta(x)v'_\theta(y)\Pi_\theta(x,y)e^{-\Lambda(\theta)} \\ &= \sum_{y\in\mathcal{A}} v'_\theta(y)\sum_{x\in\mathcal{A}} f(x,y)w_\theta(x)\Pi_\theta(x,y)e^{-\Lambda(\theta)} \\ &= \Lambda'(\theta)\sum_{y\in\mathcal{A}} w_\theta(y)v'_\theta(y) + \sum_{y\in\mathcal{A}} w'_\theta(y)v'_\theta(y) - \sum_{x,y\in\mathcal{A}} w'_\theta(x)v'_\theta(y)\Pi_\theta(x,y)e^{-\Lambda(\theta)} \end{aligned}$$

et

$$\begin{aligned} B &= \Lambda'(\theta)\sum_{x\in\mathcal{A}} q_\theta(x)\sum_{y\in\mathcal{A}}\widetilde{\Pi}_\theta(x,y)\frac{v'_\theta(y)}{v_\theta(y)} \\ &= \Lambda'(\theta)\left(\sum_{x\in\mathcal{A}} q_\theta(x)\Lambda'(\theta) + \sum_{x\in\mathcal{A}} q_\theta(x)\frac{v'_\theta(x)}{v_\theta(x)} - \sum_{x,y\in\mathcal{A}} f(x,y)q_\theta(x)\widetilde{\Pi}_\theta(x,y)\right) \\ &= \Lambda'(\theta)\sum_{x\in\mathcal{A}} w_\theta(x)v'_\theta(x) \end{aligned}$$

si bien que $T_0 = R$.

Par ailleurs, on a

$$\lim_{n\to\infty}\widetilde{\Pi}^n_\theta(x,y) = q_\theta(y) \quad \forall x,y\in\mathcal{A}$$

et donc

$$\begin{aligned} \lim_{n\to\infty} T_n &= \sum_{x,y,z\in\mathcal{A}} \left(f(x,y) - \Lambda'(\theta)\right) q_\theta(x)\widetilde{\Pi}_\theta(x,y)q_\theta(z)\frac{v'_\theta(z)}{v_\theta(z)} \\ &= \left(\sum_{x,y\in\mathcal{A}} \left(f(x,y) - \Lambda'(\theta)\right) q_\theta(x)\widetilde{\Pi}_\theta(x,y)\right)\sum_{z\in\mathcal{A}} q_\theta(z)\frac{v'_\theta(z)}{v_\theta(z)} \\ &= 0. \end{aligned}$$

Ainsi achève-t-on la preuve de (D.24). En ce qui concerne (D.25), la preuve est beaucoup plus directe ; on calcule simplement

$$\begin{aligned}\mathbb{V}ar_{q_\theta}^{\widetilde{\Pi}_\theta}\left[\sum_{i=1}^{n} f(X_i, X_{i+1})\right] &= \sum_{1\leqslant i,j\leqslant n} \mathbb{C}ov_{q_\theta}^{\widetilde{\Pi}_\theta}\left[f(X_i, X_{i+1}), f(X_j, X_{j+1})\right] \\ &= \sum_{1\leqslant i,j\leqslant n} \mathbb{C}ov_{q_\theta}^{\widetilde{\Pi}_\theta}\left[f(X_1, X_2), f(X_{j-i+1}, X_{j-i+2})\right] \\ &= n \sum_{i=-(n-2)}^{n} \mathbb{C}ov_{q_\theta}^{\widetilde{\Pi}_\theta}\left[f(X_1, X_2), f(X_i, X_{i+1})\right]\end{aligned}$$

d'où le résultat. ■

D.3 Chaînes de *Markov*

D.3.1 Théorème de *Cramér-Chernov*

Voici la preuve complète du théorème 5.2 (page 87). Cette preuve s'inspire directement de la preuve dans le cas i.i.d (voir section D.1.1) et on ne saurait trop conseiller à ce titre d'examiner en premier lieu cette preuve modèle avant de commencer la lecture de celle-ci.

On rappelle le résultat à démontrer : soit $(X_i)_{i=1,\ldots,n}$ une chaîne de *Markov* d'ordre 1 sur $\mathcal{A}$ (de cardinal fini k), de matrice de transition Π irréductible et de distribution stationnaire μ. On pose

$$S_n = \sum_{i=1}^{n} f(X_i, X_{i+1})$$

et on a alors

Théorème D.12 *Soit $a > \mathbb{E}_\mu^\Pi[f(X_1, X_2)]$ alors*

$$\lim_{n\to+\infty} \frac{1}{n} \log \mathbb{P}(S_n \geqslant an) = -\Lambda^*(a)$$

où Λ^* désigne la duale de Legendre *de Λ (définie en section 5.1).*

Comme dans le cas i.i.d, on commence par considérer une version simplifiée du théorème :

Théorème D.13 *On suppose que $0 > \mathbb{E}_\mu^\Pi[f(X_1, X_2)]$ alors :*

$$\lim_{n->+\infty} \frac{1}{n} \log \mathbb{P}(S_n \geqslant 0) = -\Lambda^*(0)$$

qu'il suffira désormais de montrer puisque

Lemme D.14 *le théorème D.12 est un corollaire du théorème D.13.*

Preuve. On se place dans les hypothèses du théorème D.12 et on pose $\bar{f} = f - a$. On a alors

$$\mathbb{E}_{\mu}^{\Pi}[\bar{f}(X_1, X_2)] = \mathbb{E}_{\mu}^{\Pi}[f(X_1, X_2)] - a < 0$$

et donc, par le théorème D.13, on obtient

$$\lim_{n->+\infty} \frac{1}{n} \log \mathbb{P}(\bar{S}_n \geqslant 0) = -\bar{\Lambda}^*(0)$$

avec

$$\bar{S}_n = \sum_{i=1}^{n} \bar{f}(X_i, X_{i+1}) = S_n - na$$

et

$$\bar{\Lambda}^*(0) = \Lambda^*(a)$$

car si $v_\theta \gg 0$ est le vecteur propre associé à $\rho(\Pi_\theta)$ et que $\bar{\Pi}_\theta$ est définie par

$$\begin{aligned} \bar{\Pi}_\theta(x, y) &= \Pi(x, y) e^{\theta \bar{f}(x,y)} \\ &= \Pi(x, y) e^{\theta f(x,y) - \theta a} \\ &= \Pi_\theta(x, y) e^{-\theta a} \end{aligned}$$

on a

$$\begin{aligned} \bar{\Pi}_\theta v_\theta &= e^{-\theta a} \Pi_\theta v_\theta \\ &= e^{-\theta a} \rho(\Pi_\theta) v_\theta \end{aligned}$$

et donc (voir remarque B.8 page 195) $\rho(\bar{\Pi}_\theta) = e^{-\theta a} \rho(\Pi_\theta)$ si bien que $\bar{\Lambda}(\theta) = \Lambda(\theta) - \theta a$ ce qui donne le résultat annoncé. ■

On va donc maintenant montrer le théorème D.13. On considère en premier lieu le cas où $0 \in \Lambda'(\mathbb{R})$ alors il existe τ tel que

$$\Lambda^*(0) = -\Lambda(\tau) \text{ et } \Lambda'(\tau) = 0.$$

On pose $\Lambda_n(\theta) = \log \mathbb{E}[e^{\theta S_n}]$ et on a (grâce au lemme 5.7 page 88)

$$\lim_{n \to +\infty} \frac{1}{n} \Lambda_n(\theta) = \Lambda(\theta). \tag{D.28}$$

Comme par l'inégalité de *Markov* on a

$$\begin{aligned} \mathbb{P}(S_n \geqslant 0) &= \mathbb{P}(e^{\tau S_n} \geqslant 1) \\ &\leq \mathbb{E}[e^{\theta S_n}] \end{aligned}$$

on trouve finalement

$$\overline{\lim}_{n\to+\infty} \frac{1}{n} \log \mathbb{P}(S_n \geqslant 0) \leqslant -\Lambda^*(0) = \Lambda(\tau) \tag{D.29}$$

De la même façon que dans le cas i.i.d. il va maintenant falloir utiliser un changement de probabilité pour montrer l'autre inégalité.

On considère $(\widehat{X}_i)$ une chaîne de *Markov* de matrice de transition $\widehat{\Pi} = \widetilde{\Pi}_\tau$ et de loi stationnaire $\widehat{\mu} = q_\tau$.

On suppose pour simplifier que $X_{n+1} = X_1 = s$ avec $s \in \mathcal{A}$ désignant l'état initial et on note $\widehat{S}_n = \sum_{i=1}^n f(\widehat{X}_i, \widehat{X}_{i+1})$. On a alors :

$$\begin{aligned}
\mathbb{P}(S_n \geqslant 0) &= \sum_{x_2,\ldots,x_n} \mathbb{I}_{S_n \geqslant 0} \Pi(s, x_2) \ldots \Pi(x_n, s) \\
&= \sum_{x_2,\ldots,x_n} \mathbb{I}_{S_n \geqslant 0} \left[\rho(\Pi_\tau) \frac{v_\tau(s)}{v_\tau(x_2)} e^{-\tau f(s,x_2)} \widehat{\Pi}(s, x_2)\right] \ldots \\
&\quad \ldots \left[\rho(\Pi_\tau) \frac{v_\tau(s)}{v_\tau(x_2)} e^{-\tau f(s,x_2)} \widehat{\Pi}(s, x_2)\right] \\
&= \rho(\Pi_\tau)^n \sum_{x_2,\ldots,x_n} \mathbb{I}_{\widehat{S}_n \geqslant 0} e^{-\tau \widehat{S}_n} \widehat{\Pi}(s, x_2) \ldots \widehat{\Pi}(s, x_2) \\
&= \rho(\Pi_\tau)^n \mathbb{E}\left[e^{-\tau \widehat{S}_n} \mathbb{I}_{\widehat{S}_n \geqslant 0}\right]
\end{aligned}$$

et donc

$$\frac{1}{n} \log \mathbb{P}(S_n \geqslant 0) = \Lambda(\tau) + \frac{1}{n} \log \mathbb{E}\left[e^{-\tau \widehat{S}_n} \mathbb{I}_{\widehat{S}_n \geqslant 0}\right] \tag{D.30}$$

dès que $\mathbb{E}\left[e^{-\tau \widehat{S}_n} \mathbb{I}_{\widehat{S}_n \geqslant 0}\right] > 0$.

Comme dans le cas i.i.d., on va utiliser un théorème central limite mais il faut pour cela tout d'abord calculer

$$\begin{aligned}
m &= \mathbb{E}_{\widehat{\mu}}^{\widehat{\Pi}}[f(X_1, X_2)] \\
&= \widehat{\Lambda}'(0)
\end{aligned}$$

et

$$\begin{aligned}
\sigma^2 &= \sum_{i \in \mathbb{Z}} \mathbb{C}ov_{\widehat{\mu}}^{\widehat{\Pi}}[f(X_1, X_2), f(X_i, X_{i+1})] \\
&= \widehat{\Lambda}''(0)
\end{aligned}$$

avec $\widehat{\Lambda}$ définie par

$$\widehat{\Lambda}(\theta) = \rho(\widehat{\Pi}_\theta) \text{ pour tout } \theta \in \mathbb{R}.$$

Lemme D.15 *On a*

$$\widehat{\Lambda}(\theta) = \Lambda(\theta + \tau) - \Lambda(\tau) \tag{D.31}$$

Preuve. On considère v_θ et $v_{\theta+\tau}$ les vecteurs propres ($\gg 0$) respectivement associés à $\rho(\Pi_\theta)$ et $\rho(\Pi_{\theta+\tau})$. On pose $\widehat{v}_\theta = \frac{v_{\theta+\tau}}{v_\theta}$ et on calcule :

$$\begin{aligned}
\left(\widehat{\Pi}_\theta \widehat{v}_\theta\right)(x) &= \sum_{y\in\mathcal{A}} \widehat{\Pi}_\theta(x,y)\widehat{v}_\theta(y) \\
&= \sum_{y\in\mathcal{A}} \frac{1}{\rho(\Pi_\tau)} \frac{v_\tau(y)}{v_\tau(x)} e^{(\theta+\tau)f(x,y)} \frac{v_{\theta+\tau}(y)}{v_\tau(y)} \Pi(x,y) \\
&= \frac{1}{\rho(\Pi_\tau)} \frac{1}{v_\tau(x)} \sum_{y\in\mathcal{A}} \Pi_{\theta+\tau}(x,y) v_{\theta+\tau}(y) \\
&= \frac{\rho(\Pi_{\theta+\tau})}{\rho(\Pi_\tau)} \frac{v_{\theta+\tau}(x)}{v_\tau(x)} \\
&= \frac{\rho(\Pi_{\theta+\tau})}{\rho(\Pi_\tau)} \widehat{v}_\theta(x)
\end{aligned}$$

donc $\widehat{v}_\theta \gg 0$ est un vecteur propre associé à $\frac{\rho(\Pi_{\theta+\tau})}{\rho(\Pi_\tau)}$. On sait alors (voir *Perron-Frobénius*) que $\rho(\widehat{\Pi}_\theta) = \frac{\rho(\Pi_{\theta+\tau})}{\rho(\Pi_\tau)}$ ce qui acheve la preuve du lemme. ■

En dérivant (D.31) on obtient donc

$$m = \widehat{\Lambda}'(0) = \Lambda'(\tau) = 0$$

et

$$\sigma^2 = \widehat{\Lambda}''(0) = \Lambda''(\tau) > 0$$

de sorte que, grâce au théorème énoncé en section B.3, (page 197)

$$\frac{\widehat{S}_n}{\sqrt{n}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2)$$

et il existe donc $C > 0$ tel que, pour n assez grand,

$$\mathbb{P}\left(\frac{\widehat{S}_n}{\sqrt{n}} \in [0, C]\right) > \frac{1}{4}$$

et donc, pour ces mêmes n,

$$\begin{aligned}
\mathbb{E}[e^{-\tau \widehat{S}_n} \mathbb{I}_{\widehat{S}_n \geqslant 0}] &\geqslant e^{-\tau C \widehat{\sigma} \sqrt{n}} \mathbb{P}\left(\frac{1}{\sigma\sqrt{n}} S_n \in [0, C]\right) \\
&> 0
\end{aligned}$$

si bien qu'en passant à la limite inférieure en n dans (D.30) on obtient

$$\varliminf_{n\to+\infty} \frac{1}{n} \log \mathbb{P}(S_n \geqslant 0) \geqslant \log \rho$$

qui, combiné à (D.29), achève la preuve.

D.3.2 Fonctions de taux

La section propose ici les démonstration des résultats concernant les différentes expressions des fonctions de taux intervenant dans les résultats concernant les déviations des mesures empiriques ; les grandes déviations de niveau 2.

On commence par rappeler la forme initiale de la fonction de taux I intervenant dans les grandes déviations pour la mesure empirique des singlettons :

$$I(x) = \sup_{\theta\in\mathbb{R}^k} \langle \theta, x\rangle - \Lambda(\theta) \tag{D.32}$$

où $\Lambda(\theta) = \log \rho(\Pi_\theta)$ correspond aux notations habituelles, et dont la proposition 7.2 (page 109) donne une expression simplifiée.

Proposition D.16 *$\forall \nu \in \mathbb{R}^k$ on a*

$$I(\nu) = \begin{cases} \sup\limits_{u\gg 0} \left\{ \sum\limits_{x\in\mathcal{A}} \nu(x) \log \dfrac{u(x)}{(u\Pi)(x)} \right\} & si\ \nu \in \mathcal{M}_1(\mathcal{A}) \\ +\infty & sinon \end{cases}$$

Preuve. On rappelle qu'on dispose ici d'un principe de grandes déviations de bonne fonction de taux I pour la suite des $(L_n^1)_n$ de sorte que, $\forall \Gamma \subset \mathbb{R}^k$ on a :

$$-\inf_{\mathring{\Gamma}} I \leqslant \varliminf_{n\to+\infty} \frac{1}{n} \log \mathbb{P}_n(\Gamma) \leqslant \varlimsup_{n\to+\infty} \frac{1}{n} \log \mathbb{P}_n(\Gamma) \leqslant -\inf_{\overline{\Gamma}} I$$

et donc, en appliquant ce résultat à $\mathcal{M}_1(\mathcal{A})^c = \mathbb{R}^k \backslash \mathcal{M}_1(\mathcal{A})$, on obtient

$$\varliminf \frac{1}{n} \log \mathbb{P}\big(L_n^1 \in \mathcal{M}_1\left(\mathcal{A}\right)^c\big) \geqslant -\inf_{\nu\in\mathcal{M}_1(\mathcal{A})^c} I\left(\nu\right).$$

Comme, par ailleurs, $L_n^1 \in \mathcal{M}_1(\mathcal{A})$, il est clair que

$$\inf_{\nu\in\mathcal{M}_1(\mathcal{A})^c} I\left(\nu\right) \geqslant +\infty$$

si bien que

$$\forall \nu \notin \mathcal{M}_1\left(\mathcal{A}\right) \text{ on a } I\left(\nu\right) = +\infty$$

ce qui prouve la première partie de la proposition.

Pour montrer le résultat concernant les $\nu \in \mathcal{M}_1(\mathcal{A})$, on pose

$$J(\nu) = \sup_{u \gg 0} \left\{ \sum_{x \in \mathcal{A}} \nu(x) \log \frac{u(x)}{(u\Pi)(x)} \right\}$$

et on va établir successivement les deux inégalités suivantes : $I(\nu) \geqslant J(\nu)$ et $I(\nu) \leqslant J(\nu)$.

Soient $\nu \in \mathcal{M}_1(\mathcal{A})$ et $u \gg 0$. Pour tout $x \in \mathcal{A}$ on pose $\lambda(x) = \log \frac{u(x)}{(u\Pi)(x)}$. Comme $u \gg 0$ et Π est irréductible on a $u\Pi \gg 0$.
On a aussi

$$\begin{aligned} \Pi_\lambda(x,y) &= \Pi(x,y)\, e^{\langle \lambda, f(y) \rangle} \\ &= \Pi(x,y) \frac{u_y}{(u\Pi)_y} \end{aligned}$$

car $\langle \lambda, f(y) \rangle = \sum_{x \in \mathcal{A}} \log \frac{u(x)}{(u\Pi)(x)} . \mathbb{I}_{x=y} = \log \frac{u(y)}{(u\Pi)(y)}$.
Par ailleurs

$$\begin{aligned} (u\Pi_\lambda)(y) &= \sum_{x \in \mathcal{A}} u(x) \Pi_\lambda(x,y) = \sum_{x \in \mathcal{A}} \frac{u(x)\Pi(x,y)\, u(y)}{(u\Pi)(y)} \\ &= u(y) \end{aligned}$$

par conséquent on a pour tout $n \in \mathbb{N}$ le résultat suivant $u(\Pi_\lambda)^n = u$; le corollaire B.7 du théorème de *Perron-Frobénius* (page 194) donne donc $\rho(\Pi_\lambda) = 1$.
On obtient ainsi

$$I(\nu) = \sup_{\theta \in \mathbb{R}^{|\mathcal{A}|}} \{ \langle \theta, \nu \rangle - \log \rho(\Pi_\theta) \} \geqslant \langle \lambda, \nu \rangle - \log \rho(\Pi_\lambda) = \sum_{x \in \mathcal{A}} \nu(x) \log \frac{u(x)}{(u\Pi)(x)}$$

et comme $u \gg 0$ est arbitraire on a finalement

$$I(\nu) \geqslant J(\nu) \text{ pour tout } \nu \in \mathcal{M}_1(\mathcal{A}).$$

Pour l'inégalité réciproque on fixe $\lambda \in \mathbb{R}^k$ quelconque et on considère $u^* \gg 0$ un vecteur propre à gauche correspondant à $\rho(\Pi_\lambda)$. On a ainsi $u^*\Pi_\lambda = \rho(\Pi_\lambda) u^*$ ce qui s'écrit

$$\sum_{x \in \mathcal{A}} u(x)^* \Pi_\lambda(x,y) = \rho(\Pi_\lambda)\, u(y)^*.$$

On a

$$\langle \lambda, \nu \rangle + \sum_{x \in \mathcal{A}} \nu(x) \log \frac{(u^*\Pi)(x)}{u^*(x)} = \sum_{x \in \mathcal{A}} \nu(x) \left(\log e^{\lambda(x)} + \log \frac{(u^*\Pi)(x)}{u^*(x)} \right)$$

or

$$\begin{aligned}\log e^{\lambda(x)} + \log \frac{(u^*\Pi)(x)}{u^*(x)} &= \log \frac{(u^*\Pi)(x) e^{\lambda(x)}}{u^*(x)} \\ &= \log \frac{\sum_{y\in\mathcal{A}} u(y)^* \Pi(y,x)\, e^{\lambda(x)}}{u(x)^*}\end{aligned}$$

de plus $e^{\lambda(x)} = e^{\langle \lambda, f(x)\rangle}$ donc

$$\begin{aligned}\langle \lambda, \nu\rangle + \sum_{x\in\mathcal{A}} \nu(x) \log \frac{(u^*\Pi)(x)}{u^*(x)} &= \sum_{x\in\mathcal{A}} \nu(x) \left(\log \frac{(u^*\Pi_\lambda)(x)}{\nu(x)^*} \right) \\ &= \log \rho(\Pi_\lambda).\end{aligned}$$

Cela implique que

$$\begin{aligned}\langle \lambda, \nu\rangle - \log \rho(\Pi_\lambda) &= \sum_{x\in\mathcal{A}} \nu(x) \log \frac{u(x)^*}{(u^*\Pi)(x)} \\ &\leqslant \sup_{u\gg 0} \left(\sum_{x\in\mathcal{A}} \nu(x) \log \frac{u(x)^*}{(u^*\Pi)(x)} \right) = J(\nu)\end{aligned}$$

en passant au sup en λ on obtient donc

$$I(\nu) \leqslant J(\nu)$$

et la preuve est achevée. ■

On utilise ce résultat pour énoncer le théorème 7.4 (page 110) dont voici l'énoncé :

Théorème D.17 *La suite $(L_{n,2})_n$ de v.a. à valeurs dans $\mathcal{M}_1(\mathcal{A})$, suit un principe de grandes déviations de bonne fonction de taux*

$$I_2(\nu) = \begin{cases} \sum\limits_{x,y\in\mathcal{A}} h_\nu(x,y) & si\ \nu \in \mathcal{S} \\ +\infty & sinon \end{cases}$$

avec

$$h_\nu(x,y) = \begin{cases} 0 & si\ \overline{\nu}(x) = 0\ ou\ \Pi(x,y) = 0 \\ \nu(x,y) \log \dfrac{\nu(x,y)}{\overline{\nu}(x)\Pi(x,y)} & sinon \end{cases}$$

(avec la convention $0 \log 0 = 0$).

Preuve. Comme dans le chapitre 6, on va augmenter la taille de l'alphabet. On se place dans l'alphabet $\mathcal{A}^2$ et on considère la chaîne de *Markov* $X^{(2)}$ d'ordre 1 de matrice de transition $\Pi^{(2)}$ définie par

$$\Pi^{(2)}([x_1,x_2],[y_1,y_2]) = \mathbb{I}_{x_2=y_1}\Pi(y_1,y_2)$$

et on applique simplement la proposition D.16 à cette chaîne de *Markov* de sorte que l'on a

$$I_2(\nu) = \sup_{u\gg 0}\left\{\sum_{x,y\in\mathcal{A}}\nu(x,y)\log\frac{u(x,y)}{(u\Pi^{(2)})(x,y)}\right\}$$

si $\nu\in\mathcal{M}_1(\mathcal{A}^2)$.

Comme on a

$$\begin{aligned}\left(u\Pi^{(2)}\right)(x,y) &= \sum_{z,t} u\left(z,t\right)\Pi^{(2)}\left([z,t],[x,y]\right)\\ &= \sum_{z,t} u\left(z,t\right)\mathbb{I}_{t=x}\Pi\left(x,y\right)\\ &= \sum_{z} u\left(z,x\right)\Pi\left(x,y\right)\end{aligned}$$

il est clair que l'on a

$$I_2\left(\nu\right) = \sup_{u\gg 0}\sum_{x,y\in\mathcal{A}}\nu\left(x,y\right)\log\left(\frac{u\left(x,y\right)}{\sum_z u\left(z,x\right)\Pi\left(x,y\right)}\right).$$

Dans un premier temps, on suppose que ν n'est pas invariant par translation, il existe donc un y_0 tel que $\nu_1(y_0)\neq\nu_2(y_0)$ avec

$$\nu_1(y_0) = \sum_{x\in\mathcal{A}}\nu(x,y_0)\quad\text{et}\quad \nu_2(y_0) = \sum_{x\in\mathcal{A}}\nu(y_0,x).$$

Supposons que $\nu_1\left(y_0\right) < \nu_2\left(y_0\right)$ (resp. $>$). Soit u tel que $u\left(\cdot,y\right)=1$ si $y\neq y_0$ et $u\left(\cdot,y_0\right)=e^{\alpha}$ ou $\alpha\in\mathbb{R}$ est quelconque.
On a

$$(*) = \sum_{x,y\in\mathcal{A}}\nu\left(x,y\right)\log\left(\frac{u\left(x,y\right)}{\sum_z u\left(z,x\right)\Pi\left(x,y\right)}\right) = \sum_{x,y\in\mathcal{A}}\nu\left(x,y\right)\log\left(\frac{u\left(1,y\right)}{\sum_k u\left(1,x\right)\Pi\left(x,y\right)}\right)$$

et

$$\sum_z u\left(1,x\right)\Pi\left(x,y\right) = \begin{cases} |\mathcal{A}|\,\Pi\left(x,y\right) & \text{si } x\neq y_0\\ |\mathcal{A}|\,\Pi\left(y_0,y\right)e^{\alpha} & \text{sinon}\end{cases}$$

on a ainsi

$$
\begin{aligned}
(*) \quad = \quad & -\sum_{x\neq y_0}\sum_{y\neq y_0}\nu(x,y)\log\left(|\mathcal{A}|\,\Pi(x,y)\right)+\nu(y_0,y_0)\left(\alpha-\log\left[|\mathcal{A}|\,\Pi(y_0,y_0)\,e^{\alpha}\right]\right)\\
& +\sum_{x\neq y_0}\nu(x,y_0)\left(\alpha-\log\left[|\mathcal{A}|\,\Pi(x,y_0)\right]\right)-\sum_{y\neq y_0}\nu(y_0,y)\left(\log\left[|\mathcal{A}|\,\Pi(y_0,y)\,e^{\alpha}\right]\right)\\
= \quad & -\sum_{x\neq y_0}\sum_{y\neq y_0}\nu(x,y)\log\left(|\mathcal{A}|\,\Pi(x,y)\right)+\nu(y_0,y_0)(\alpha-\alpha)\\
& -\nu(y_0,y_0)\log\left[|\mathcal{A}|\,\Pi(y_0,y_0)\right]+\sum_{x\neq y_0}\nu(x,y_0)\,\alpha-\sum_{y\neq y_0}\nu(y_0,y)\,\alpha\\
& -\sum_{y\neq y_0}\nu(y_0,y)\log\left[|\mathcal{A}|\,\Pi(y_0,y)\right]-\sum_{x\neq y_0}\nu(x,y_0)\log\left[|\mathcal{A}|\,\Pi(x,y_0)\right]\\
= \quad & -\sum_{x,y\in\mathcal{A}}\nu(x,y)\log\left[|\mathcal{A}|\,\Pi(x,y)\right]+\alpha\left(\nu_2(y_0)-\nu_1(y_0)\right).
\end{aligned}
$$

On fait $\alpha\to+\infty$ (resp. $-\infty$) et on trouve $I_2(\nu)=\infty$.

On se place maintenant dans le cas où ν est invariant par translation. On commence par remarquer qu'on a alors

$$
\begin{aligned}
& \sum_{x,y\in\mathcal{A}}\nu(x,y)\log\frac{\sum_z u(z,x)\,\nu_2(y)}{\sum_z u(z,y)\,\nu_1(x)}\\
= \quad & \sum_{x,y\in\mathcal{A}}\nu(x,y)\log\left[\sum_z u(z,x)\right]-\sum_{x,y\in\mathcal{A}}\nu(x,y)\log\left[\sum_z u(z,y)\right]\\
& +\sum_{x,y\in\mathcal{A}}\nu(x,y)\log\nu_2(y)-\sum_{x,y\in\mathcal{A}}\nu(x,y)\log\nu_1(x)\\
= \quad & \sum_{x\in\mathcal{A}}\nu_1(x)\log\left[\sum_z u(z,x)\right]\nu_1(x)-\sum_{y\in\mathcal{A}}\nu_2(y)\log\left[\sum_z u(z,y)\right]\\
& +\sum_{y\in\mathcal{A}}\nu_2(y)\log\nu_2(y)-\sum_{x\in\mathcal{A}}\nu_1(x)\log\nu_1(x)\\
= \quad & 0.
\end{aligned}
$$

$$
\begin{aligned}
& I_2(\nu) - \sum_{x,y\in\mathcal{A}} \nu(x,y)\log\frac{\nu(x,y)}{\overline{\nu}(x)\Pi(x,y)} \\
= & \sup_{u\gg 0}\sum_{x,y\in\mathcal{A}} \nu(x,y)\log\left(\frac{u(x,y)}{\sum_z u(z,x)\,\Pi(x,y)}\right) - \sum_{x,y\in\mathcal{A}} \nu(x,y)\log\frac{\nu(x,y)}{\overline{\nu}(x)\Pi(x,y)} \\
= & \sup_{u\gg 0}\sum_{x,y\in\mathcal{A}} \nu(x,y)\log\left(\frac{u(x,y)\,\overline{\nu}(x)}{\sum_z u(z,x)\,\nu(x,y)}\right) \\
= & \sup_{u\gg 0}\sum_{x,y\in\mathcal{A}} \nu(x,y)\log\frac{\mu_x(y)}{\nu_x(y)} \\
= & \sup_{u\gg 0}\left\{-\sum_{x\in\mathcal{A}}\overline{\nu}(x)H(\nu_x|\mu_x)\right\} \\
= & (**)
\end{aligned}
$$

où ν_x et μ_x sont des mesures de probabilité sur $\mathcal{A}$ définies par

$$\nu_x(y) = \frac{u(x,y)}{\sum_z u(z,x)} \quad \text{et} \quad \mu_x(y) = \frac{\nu(x,y)}{\overline{\nu}(x)}.$$

et où $H(\nu_x|\mu_x)$ désigne l'entropie de la première relativement à la seconde.

Un entropie étant toujours positive on a $(**) \leqslant 0$ si bien que

$$I_2(\nu) \leqslant \sum_{x,y\in\mathcal{A}} \nu(x,y)\log\frac{\nu(x,y)}{\overline{\nu}(x)\Pi(x,y)}.$$

Pour l'autre inégalité, on considère d'abord le cas où $\nu \gg 0$, on pose alors $u=\nu$ dans le sup et on obtient

$$(**) \geqslant \sum_{x,y\in\mathcal{A}} \nu(x,y)\log\left(\frac{\nu(x,y)\,\overline{\nu}(x)}{\overline{\nu}(x)\nu(x,y)}\right) = 0$$

Ainsi, on a montré le résultat si $\nu \gg 0$. si ν est quelconque on considère alors une suite $u_n \gg 0$ telle que $u_n \to \nu$ et telle que $\overline{\nu}(x)H(\nu_x|\mu_x) \to 0$ pour tout x ce qui achève la preuve. ■

Références

[dH00] F. den Hollender. *Large Deviations.* American Mathematical Society, 2000.

[Tor98] N. Torrent. *Application des grandes déviations et de la loi d'Erdös-Réyni pour des variables indépendantes ou de dépendance markovienne.* PhD thesis, Université PARIS VII, 1998.

[Whi55] P. Whittle. Some distribution and moment fomulæ for the markov chain. *J. R. Statist. Soc. B.*, 17 :235–242, 1955.

Annexe E

Documentation de GDon

Contenu du chapitre

E.1 Disponibilité

Comme il l'a déjà été signalé au chapitre 9 (page 127), un certain nombre de programmes et librairies sont nécessaires à GDon mais notons que tous ces outils sont disponibles en licences GPL ou LGPL.

Il s'agit tout d'abord des outils standard du GNU que sont `tar`, `make`, `automake`, la librairie `getopt` et, enfin, un compilateur C (comme `gcc`). On pourra se procurer ces librairies et programmes ainsi que les documentations qui vont avec sur le site

`http://www.gnu.org`

Il faut également se procurer la librairie de programmation `glib` qui est issue du projet *the GIMP* et est diponible sur le site

`http://www.gtk.org`

Enfin il faut disposer d'un archive `.tar` du package GDon dont voici la dernière version :

`GDon − 0.4.58.tar`

Ne disposant pas à l'heure actuelle de site web permettant la distribution de ce package, je prie toute personne intéressée de bien vouloir me contacter directement par e-mail à l'adresse :

`nuel@genopole.cnrs.fr`

E.2 Installation

Voici un extrait de l'arborescence de l'archive `GDon-0.4.58.tar` :

`AUTHORS`	informations sur les auteurs du programme
`COPYING`	licence GPL
`ChangeLog`	descriptif détaillé de l'évolution du programme au fil des versions
`INSTALL`	informations concernant l'installation du programme
`KNOW_BUGS`	problèmes connus
`README`	descriptif succint du programme
`NEWS`	résumé du `ChangeLog`
`TODO`	liste des tâches à effectuer dans les prochaines versions
`configure*`	programme de configuration automatique
`doc/`	répertoire des sources de la documentation
`src/`	répertoire des sources du programme

Pour installer le package il suffit de restaurer l'archive et de ce placer dans le répertoire ainsi créé :

```
tar xvf GDon-0.4.58.tar
cd GDon-0.4.58.tar
```

puis de lancer la configuration automatique :

```
./configure
```

En cas de message d'erreur à ce niveau, vérifiez que vous avez bien installé les programmes et librairies évoqués en section E.1. Une fois la phase de configuration terminée il suffit d'utiliser la commande

```
make
```

pour lancer la compilation du programme et de sa documentation.

Un simple

```
make install
```

copie alors l'exécutable ainsi que les pages de manuel dans les répertoires usuels. On pourra, le cas échéant sauter cette étape en copiant "à la main" les fichiers résultant de la compilation (`src/GDon` et `doc/GDon.*`).

E.3 Syntaxe

La syntaxe minimum de GDon est la suivante :

`GDon genome`

où `genome` est un fichier au format FASTA contenant une ou plusieurs séquences. Lorsque l'on invoque cette commande, la ou les séquences du fichier `genome` sont lues et les mots de l'alphabet par défaut et de longueur $1 \leqslant h \leqslant 8$ (8 est la longueur maximum par défaut) sont comptés. Toutes les informations recueillies par le programme sont alors écrites dans le fichier de sortie par défaut (`out.GDon`).

A cette ligne de commande minimum, il est bien évidemment possible de rajouter de nombreuses options que nous allons détailler.

E.3.1 Options générales

`-verbose` ou `-v` :
Option permettant l'affichage de davantage d'informations lors de l'exécution du programme (par defaut, cette option est désactivée)
`-debug` ou `-d` :
Option permettant l'affichage d'infomations de débuggage lors de l'exécution du programme (par défaut, cette option est désactivée).
`-help` ou `-h` :
Option permettant l'affichage d'un récapitulatif rapide des options du programme. Une ligne de commande a la syntaxe erronée provoquera le même résultat.
`-hmax argument` :
Option permettant de changer la valeur de la longueur maximale des mots examinés à la valeur donnée par l'argument (qui doit être entier). Par défaut, la longueur maximale des mots est fixée à 8.
`-output argument` ou `-o argmument` :
Option permettant de spécifier le fichier de sortie de programme. Si cette option est utilisée, le programme écrira dans le fichier spécifié en argument au lieu du fichier de sortie par défaut (`out.GDon`).
`-alphabet argument` :
Option permettant de spécifier l'alphabet dans lequel on désire travailler. L'argument peut prendre plusieurs valeurs numériques entières :

0 (défaut) : $\{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$ (alphabet nucléique standard) ;

1 : $\{\mathtt{a}, \mathtt{g}, \mathtt{c}, \mathtt{t}\}$ (alphabet nucléique, ordre utilisé par R'MES) ;

2 : $\{\mathtt{a}, \mathtt{b}, \ldots, \mathtt{y}, \mathtt{z}\}$ (alphabet romain)

3 : $\{\mathtt{0}, \mathtt{1}\}$ (binaire)

4 : $\{\mathtt{ac}, \mathtt{gt}\}$ (purines et pyrimidines)

5 : $\{\mathtt{at}, \mathtt{cg}\}$ (liaisons hydrogene fortes et faibles)

Notons qu'il est bien sur possible d'ajouter de nouveaux alphabets mais que cela, pour l'instant, requiert une légère modification des sources du programme et une recompilation.

E.3.2 Options concernant le traitement des mots

On doit choisir un seul mode parmi les possibilités suivantes :

mode

`-word argument` ou `-w argument` :
On spécifie le traitement du mot donné en argument. Il est possible d'utiliser plusieurs fois cette option pour traiter successivement plusieurs mots.
`-wordlist argument` :
Le programme va traiter successivement chacun des mots contenu dans le fichier donné en argument (il suffit de séparer les mots par un retour à la ligne dans le fichier donné en argument).
`-all-words` :
Le programme traite successivement tous les mots de la longueur maximale (par défaut ou spécifiée par l'option `-hmax`).
`-familly argument` ou `-f argument` :
Traitement de la famille de mots spécifiée en argument. Cette option ne peut être utilisée qu'une seule fois par ligne de commande et en voici la syntaxe :

`[{commande1}motif1,score1][{commande2}motif2,score2]` ...

sachant que les commandes sont optionnelles et que l'on peut omettre de spécifier un score (auquel cas un score de 1 est sélectionné par défaut).
Pour les motifs, on utilise la syntaxe suivante :
- '`_`' signifie toutes les lettres de l'alphabet ;
- plusieurs lettres entre crochets signifient l'une quelconque de ces lettres.

Les différentes commandes sont les suivantes :
- `{inv}` inverse l'ordre des lettres dans le motif ;

– `{comp}` utilise le complémentaire du motif (sans effet lorsque cette manipulation n'a pas de sens comme dans le cas de l'alphabet romain par exemple) ;
– `{bar}` combine les deux commandes précédentes ;
– `{#}` où `#` est un entier, génére les mots ayant exactement `#` *mismatchs* avec le motif proposé.

Il faut ensuite spécifier un modèle :

modèle

`-m argument` :
On définit l'ordre du modèle markovien à utiliser (et dont les paramètres seront estimés sur la séquence). Par convention, on utilise la valeur `-1` pour spécifier le modèle $M00$ (puis `0` pour $M0$, `1` pour $M1 \ldots$). Par défaut, c'est l'ordre $h-2$ qui sera choisi pour un mot ou motif de longueur h. Notons qu'il est possible de spécifier cette option à plusieurs reprises, ce qui entraînera autant de traitements successifs dans chacun des modèles spécifiés.
`-all-models` :
Tous les modèles possibles (c'est à dir de $M00$ à $Mh-2$) vont être utilisés les uns après les autres.
`-use argument` ou `-u argument` :
Permet d'utiliser les paramètres spécifiés dans le fichier argument. Pour le format de ce fichier, on pourra examiner les fichiers générés par l'option `-p` (ci-après).

options

`-bar` ou `-b` :
Chaque mot sera traité comme une famille composé de lui-même et de son inverse complémentaire.
`-parameters argument` ou `-p argument` :
Ecrit dans le fichier spécifié en argument les paramètres du changement de probabilité.
`-simulation argument` ou `-s argument` :
Effectue les calculs par simulations sous la probabilité spécifiée par `-u`. L'argument définit le nombre de tirages à effectuer.

E.4 Exemples

E.4.1 Comptage de mots

Comptage des mots de longueur $h = 1$ à $h = 3$ dans le génome de *Escherichia coli* :

```
$ GDon ecoli --hmax 3 -v -o
mode verbose
fichier output "sortie" spécifié
ecoli: 4639221 caractères lus en 0.490532 secondes
ecriture de Occ dans out.GDon en 0.000112 secondes
temps de calcul total (hormis lecture de la séquence): 0.000592 secondes
```

et voici le contenu du fichier `sortie` :

```
# Fichier généré par GDon version développement 0.4.58
# le Fri May 25 09:26:01 2001
#
# paramètres généraux:
# séquence = ecoli
# n = 4639221
# A = {a,c,g,t}
# k = 4
# hmax = 3
# fin des paramètres
#
# h = 1, 4 mots
   1142136    1179433    1176775    1140877
# h = 2, 16 mots
    337835     256658     237851     309792
    325118     271649     346636     236029
    267234     383865     270083     255593
    211948     267261     322205     339463
# h = 3, 64 mots
    108901      82578      63364      82992
     58633      74899      73263      49863
     56618      80848      50611      49774
     63692      86476      76229      83395
     76607      66752     104785      76974
     86442      47764      87031      50412
     70934     115673      86870      73159
     26762      42714     102900      63653
```

```
     83490      54737      42460      86547
     96010      92961     114609      80285
     56199      92123      47470      74291
     52670      54225      66108      82590
     68837      52591      27241      63279
     84033      56025      71733      55469
     83483      95221      85132      58369
     68824      83846      76968     109825
# achevé le Fri May 25 09:26:01 2001
# fin du fichier
```

E.4.2 Traitement d'un mot

Significativité de `gctggtgg` chez *Escherichia coli* dans les modèles $M0$ et $M1$:

```
$ GDon ecoli -w gctggtgg -m 0 -m 1 -v
mode verbose
ecoli: 4639221 caractères lus en 1.36179 secondes
modèle M0: estimation en 0.015516 secondes
gctggtgg a=+1.0756e-04 xmin=+1.8960e+00 fmin=-1.1259e-04 p=[+]1.403e-227
N=+32.185619 en 1.4001 secondes
gctggtgg +32.185619
modèle M1: estimation en 0.015083 secondes
gctggtgg a=+1.0756e-04 xmin=+1.9633e+00 fmin=-1.1869e-04 p=[+]7.379e-240
N=+33.051592 en 1.91508 secondes
gctggtgg +33.051592
temps de calculs total (hormis lecture de la séquence): 3.34727 secondes
```

et voici le contenu du fichier `out.GDon` :

```
# Fichier généré par GDon version developpement 0.4.58
# le Fri May 25 09:30:47 2001
#
# paramètres généraux:
# sequence = ecoli
# n = 4639221
# A = {a,c,g,t}
# k = 4
# hmax = 8
# h = 8
# fin des paramètres
#
```

```
#
# matrice de transition du modèle M0
2.461913e-01 2.542308e-01 2.536579e-01 2.459200e-01
# fin de la matrice de transition
#
# [+] gctggtgg (40634) compté 499 fois
# t a fmin xmin proba
1.4001 +1.0756e-04 -1.1259e-04 +1.8960e+00 1.403e-227
#
# matrice de transition du modèle M1
2.957923e-01 2.247175e-01 2.082510e-01 2.712392e-01
2.756564e-01 2.303219e-01 2.939008e-01 2.001209e-01
2.270901e-01 3.262008e-01 2.295112e-01 2.171979e-01
1.857764e-01 2.342593e-01 2.824187e-01 2.975457e-01
# fin de la matrice de transition
#
# [+] gctggtgg (40634) compté 499 fois
# t a fmin xmin proba
1.91508 +1.0756e-04 -1.1869e-04 +1.9633e+00 7.379e-240
# achevé le Fri May 25 09:30:50 2001
# fin du fichier
```

E.4.3 Traitement d'une famille de mots

```
$ GDon hinfluenzae -f [g_tggtgg] -all-models -v
mode verbose
hinfluenzae: 1830023 caratères lus en 0.519051 secondes
famille [g_tggtgg]
4 mots dans la famille
modèle M-1: estimation en 0.004989 secondes
a=+8.6884e-05 xmin=+3.4016e-01 fmin=-4.6670e-06 p=[+]1.953e-4
N=+3.546294 en 1.01436 secondes
+3.546294
modèle M0: estimation en 0.015024 secondes
a=+8.6884e-05 xmin=+1.2464e+00 fmin=-4.8537e-05 p=[+]2.654e-39
N=+13.063698 en 1.074 secondes
+13.063698
modèle M1: estimation en 0.015232 secondes
a=+8.6884e-05 xmin=+1.2021e+00 fmin=-4.5187e-05 p=[+]1.22e-36
N=+12.588568 en 1.53944 secondes
+12.588568
```

```
modèle M2: estimation en 0.01525 secondes
a=+8.6884e-05 xmin=+4.6042e-01 fmin=-8.2883e-06 p=[+]2.587e-7
N=+5.019781 en 1.76712 secondes
+5.019781
modèle M3: estimation en 0.016126 secondes
a=+8.6884e-05 xmin=+2.4529e-01 fmin=-2.5238e-06 p=[+]9.867e-3
N=+2.331355 en 1.86544 secondes
+2.331355
modèle M4: estimation en 0.019047 secondes
a=+8.6884e-05 xmin=+7.8220e-02 fmin=-2.6636e-07 p=[+]6.142e-1
N=+0.000000 en 2.07866 secondes
+0.000000
modèle M5: estimation en 0.031945 secondes
a=+8.6884e-05 xmin=-2.6125e-02 fmin=-3.3231e-08 p=[-]9.41e-1
N=-0.000000 en 2.48564 secondes
-0.000000
modèle M6: estimation en 0.081569 secondes
a=+8.6884e-05 xmin=-2.7362e-02 fmin=-3.3847e-08 p=[-]9.399e-1
N=-0.000000 en 2.94221 secondes
-0.000000
modèle M7: non traité, 7 > 6 = h-2
temps de calculs total (hormis lecture de la séquence): 14.969 secondes
```

et voici (le début) du fichier `out.GDon` :

```
# Fichier généré par GDon version developpement 0.4.58
# le Fri May 25 09:33:50 2001
#
# paramètres généraux:
# sequence = hinfluenzae
# n = 1830023
# A = {a,c,g,t}
# k = 4
# hmax = 8
# h = 8
# sequence = hinfluenzae
# n = 1830023
# A = {a,c,g,t}
# k = 4
# hmax = 8
# h = 8
# fin des paramètres
```

```
#
# famille [g_tggtgg]
# 4 mots dans la famille
# gatggtgg score 1
# gctggtgg score 1
# ggtggtgg score 1
# gttggtgg score 1
#
# matrice de transition du modèle M-1
2.500000e-01 2.500000e-01 2.500000e-01 2.500000e-01
# fin de la matrice de transition
#
# [+]
# t a fmin xmin proba
1.01436 +8.6884e-05 -4.6670e-06 +3.4016e-01 1.953e-4
#
# matrice de transition du modèle M0
3.101726e-01 1.916495e-01 1.898534e-01 3.083245e-01
# fin de la matrice de transition
#
...
```

Références

[Del00] C. Delannoy. *Programmer en language C.* Eyrolles, 2000.

[PTVF97] W. H. Press, S. A. Teukolsky, W. T. Vettering, and B. P. Flannery. *Numerical recipes in C.* Cambridge University Press, 1997.

[VTPF99] W. T. Vettering, S. A. Teukolsky, W. H. Press, and B. P. Flannery. *Numerical recipes : example book (C) second edition.* Cambridge University Press, 1999.

Bibliographie

[Arn51] W. E. Arnoldi. The principle of minimised itérations in the solution of the matrix eigenvalue problem. *Quart. Appl. Math.*, 9 :17–29, 1951.

[BP66] L. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The Annals of Mathematicals Statistics.*, (37) :1554–1563, 1966.

[Bre73] R. P. Brent. *Algorithms for minimisation without derivatives.* Englewood Cliffs, NJ : Prentice-Hall, 1973.

[Buc90] J. A. Bucklew. *Large deviation techniques in decision, simulation, and estimation.* Wiley, 1990.

[CMU+98] R. Cole, J. Mariani, H. Uszkoreit, G. B. Varile, A. Zaenen, and A. Zampolli. *Survey of the state of the art in human language technology.* Cambridge University Press, 1998.

[Cra38] H. Cramér. Sur un nouveau théorème-limite dans la théorie des probabilités. *Actualités Scientifiques et Industrielles*, pages 5–23, 1938.

[DCD83] D Dacunha-Castelle and M. Duflo. *Probabilités et statistiques : 2. Problèmes à temps mobile.* Masson, 1983.

[Deg00] B. Degraupes. *LaTeXapprentissage, guide et référence.* Vuibert, 2000.

[Del00] C. Delannoy. *Programmer en language C.* Eyrolles, 2000.

[Deu57] J. D. Deuschel. *Large Deviations.* Academic press, Boston, 1957.

[DGV+99] P.J. Deschavanne, A. Giron, J. Vilain, G. Fagot, and B. Fertil. Genomic signature : characterization and classification of species assessed by chaos game representation of sequences. *Mol Biol Evol.*, 16(10) :1391–1399, 1999.

[dH00] F. den Hollender. *Large Deviations.* American Mathematical Society, 2000.

[DZ98] A. Dembo and O. Zeitouni. *Large deviations techniques and applications.* Springer-Verlag, 1998.

[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.

[Ell85] R. S. Ellis. *Entropy, large deviations, and statistical mechanics.* Springer-Verlag, New York, 1985.

[Gan90] F. R. Gantmakher. *Théorie des matrices.* Editions Jacques Gabay, 1990.

[Gou94a] X. Gourdon. *Mathématiques pour M' : Algèbre.* Ellipses, 1994.

[Gou94b] X. Gourdon. *Mathématiques pour M' : Analyse.* Ellipses, 1994.

[JKK92] N. L. Johnson, S. Kotz, and A. W. Kemp. *Univariate discrete distributions.* Wiley : New-York, 1992.

[Kar69] S. Karlin. *Initiation aux processus aléatoires.* Dunod, Paris, 1969.

[KB92] J. Kleffe and M. Borodovsky. First and second moment of counts of words in random text generated by markov chains. *Comp. Applic. Biosci.*, (8) :433–441, 1992.

[LBB+95] H. Lodish, D. Baltimore, A. Berk, S. L. Zipursky, P. Matsudaira, and J. Darnell. *Molecular Cell Biology.* Scientific American Book, Inc, 1995.

[LMS] M.-Y. Leung, G. M. Marsh, and T. P. Speed. Over- and under-representation of short dna words in herpesvirus genomes. *J. Comp. Bio.*, (3) :345–360.

[LT85] A Lancaster and M. Tismenetsky. *The theory of matrices.* Academic press, Orlando, 1985.

[Mur97] F. Muri. *Comparaison d'algorithmes d'identification de chaînes de Markov cachées et applications à la détection de régions homogènes dans les séquences d'ADN.* PhD thesis, Université René Descartes, Paris V, 1997.

[Nic01] P. Nicodème. Fast approximate motif statistics. *J. Comp. Biol.*, 2001. to appear.

[NSF99] P. Nicodème, B. Salvy, and P. Flajolet. Motif statistics. *Theoretical Computer Science*, 1999. to appear.

[PRT95] B. Prum, F. Rodolphe, and E. de Turckheim. Finding words with unexpected frequencies in dna sequences. *J. R. Statist. Soc. B.*, 11 :190–192, 1995.

[PTVF97] W. H. Press, S. A. Teukolsky, W. T. Vettering, and B. P. Flannery. *Numerical recipes in C.* Cambridge University Press, 1997.

[Rab89] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEE.*, (77) :257–286, 1989.

[RD99] S. Robin and J.J. Daudin. Exact distribution of word occurrences in a random sequence of letters. *J. App. Prob.*, 36 :179–193, 1999.

[RD00] S. Robin and J.J. Daudin. Exact distribution of the distances between any occurrences of a set of words. *Ann. Inst. Statist. Math.*, 2000.

[Rég00] M. Régnier. A unified approach to word occurrence probabilities. *Discrete applied mathematics*, 104(1) :259–280, 2000.

[RM01] R.J. Roberts and D. Macelis. Rebase - restriction enzymes and methylases. *Nucleic Acids Research*, 29 :268–269, 2001.

[Ros97] E. Rostand. *Cyrano de Bergerac*. 1897.

[RS98a] G Reinert and S. Schbath. Compound poisson and poisson process approximations for occurrences of multiple words in markov chains. *J. Comp. Biol.*, 5 :223–254, 1998.

[RS98b] M. Régnier and W. Szpankowski. On pattern frequency occurrences in a markovian sequence. *Algorithmica*, 22(4) :631–649, 1998.

[RS99] G. Reinert and S. Schbath. Compound poisson approximations for occurrences of multiple words. *Statistics in Molecular Biology and Genetics*, 33 :257–275, 1999.

[RS01] S. Robin and S. Schbath. Numerical comparison of several approximations of the word count distribution in random sequences. *J. Comp. Biol.*, 2001. To appear.

[RW98] R. T. Rockafellar and R. J-B. Wets. *Variational analysis*. Princeton University Press, 1998.

[Saa92] Y. Saad. *Numerical methods for large eigenvalue problems*. Manchester University Press, 1992.

[San61] I.N. Sanov. On the probability of large deviations of random variables. *Mat. Sb. 42 (en Russe). Traduction en englais dans : Selected translations in Mathematical Statistics and Probability I*, pages 213–244, 1961.

[Sch95] S. Schbath. *Etude asymptotique du nombre d'occurrences d'un mot dans une chaîne de Markov et application à la recherche de mots de fréquence exceptionnelle dans les séquences d'ADN*. PhD thesis, PARIS V, 1995.

[Sch97] S. Schbath. An efficient statistic to detect over- and under-represented words in dna sequences. *J. Comp. Biol.*, 4 :189–192, 1997.

[Sch00] S. Schbath. An overview on the distribution of word counts in markov chains. *J. Comp. Biol.*, (7) :193–202, 2000.

[Sen81] E. Seneta. *Non-negative matrices and Markov chains*. Springer-Verlag, New York, 1981.

[SGS99] H.O. Smith, M.L. Gwinn, and Salzberg S.L. Dna uptake signal sequences in naturally transformable bacteria. *Res. Microbiol.*, 150(9-10) :603–616, Nov-Dec 1999.

[Ste94] W. J. Stewart. *Introduction to numerical solution to Markov chains*. Princeton University Press, 1994.

[SWW75] G. Salton, A. Wong, and C. S. Wang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11) :613–620, 1975.

[Tor98] N. Torrent. *Application des grandes déviations et de la loi d'Erdös-Réyni pour des variables indépendantes ou de dépendance markovienne*. PhD thesis, Université PARIS VII, 1998.

[Ver01] J-Ph. Vert. *Méthodes satitistiques pour la modélisation du langage naturel*. PhD thesis, Université Paris 6, 2001.

[VTPF99] W. T. Vettering, S. A. Teukolsky, W. H. Press, and B. P. Flannery. *Numerical recipes : example book (C) second edition*. Cambridge University Press, 1999.

[Whi55] P. Whittle. Some distribution and moment fomulæ for the markov chain. *J. R. Statist. Soc. B.*, 17 :235–242, 1955.

[WST95] F. M. J. Willems, Y. M. Shtarkov, and Tj. J. Tjalkens. The context-tree weighting method : basic properties. *IEEE Trans. Inform. Theory*, IT-41 :653–664, 1995.

Index

Sujet : Grandes déviations et chaînes de *Markov* pour l'étude des occurrences de mots dans les séquences biologiques.

Mots clés : Grandes déviations, chaîne de *Markov*, nombre d'occurrences de mots, génome, ADN, chi, uptake, site de restriction.

Résumé :

On peut assimiler l'information contenue dans l'ADN d'un organisme à une longue séquence écrite dans un alphabet à quatre lettres : `a`, `c`, `g` et `t`. Certains mots ou motifs que l'on trouve dans ces séquences interviennent directement dans des mécanismes biologiques. Du fait de la pression de la sélection, il est naturel de relier le caractère exceptionnels de ces mots à leurs fréquences d'apparition.

On utilise l'outil statistique des grandes déviations pour mesurer la significativité du comptage d'un mot ou d'un motif dans un texte supposé aléatoire et généré selon une chaîne de *Markov* d'un ordre donné. Grâce à des algorithmes numériques performants (*Brent*, *Arnoldi*, descente du gradient), les résultats théoriques de grandes déviations de niveaux 1 et 2 sont utilisés par le programme GDon pour effectuer les calculs pour motif de taille h en $O(k^h)$ en temps et en espace.

La comparaison des résultats de GDon avec ceux d'autres méthodes asymptotiques (approximations gaussiennes et poissoniennes) ou exactes montre la grande qualité des approximations obtenues en ce qui concerne les événements rares. De plus, divers exemples biologiques concrets sont étudiés par le biais de ce programme et les résultats obtenus sont cohérents avec les connaissances biologiques des mécanismes qui leurs sont liés.

La démarche inverse, c'est à dire la création d'information à partir des résultats statistiques seuls n'est cependant pas si simple. Un méthode de retraitement automatique des résultats par le biais d'alignement est dans ce but envisagée et se fixe pour objectif de distinguer les mots véritablement significatifs du point de vue biologique de ceux dont la nature exceptionnelle est due à l'évolution.

Subject : Large deviations and *Markov* chains for the study of word counts in biological sequences.

key words : Large deviations, *Markov* chain, word count, genome, DNA, chi, uptake, restriction site.

Abstract :

For any organism, the DNA information can be viewed as a long sequence written with the four following letters : `a`, `c`, `g` et `t`. In those sequences, we can find specific patterns involved in biological mecanisms. The natural selection links those exceptional patterns to their frequencies. The aim of this study is to obtain tools for looking at the exceptionality of a given pattern through its number of occurrences.

We use the statistical tool of the large deviations to obtain the significativity of a pattern count in a text randomly generated through a *Markov* chain of a given order. With several numerical algorithm (*Brent*, *Arnoldi*, gradient method), the large deviation results of levels 1 and 2 are used by the software GDon to compute the results for the count of a size h pattern in $O(k^h)$ space and time.

The comparaison between GDon and other asymptotic methods (gaussian or *Poisson* approximation) or exact methods, shows the high quality of the results when we consider rare events. Moreover, several biological examples are studied with GDon and the results we obtain agree with our experimental knowledge.

At the contrary, the use of the statistical results to create new biological informations is not easy. In this aim we developped an automatic processing of the results using some kind of alignments in order to find words with a real biological potential.