N^o D'ORDRE : 8809

UNIVERSITÉ DE PARIS SUD
U.F.R. SCIENTIFIQUE D'ORSAY

THÈSE

présentée pour obtenir le titre de

DOCTEUR EN SCIENCES DE L'UNIVERSITÉ PARIS XI

Spécialité : Mathématiques

par

Étienne ROQUAIN

Sujet de la thèse :

Motifs exceptionnels dans des séquences hétérogènes. Contributions à la théorie et à la méthodologie des tests multiples.

Rapporteurs :	Mme	Gesine	REINERT
	M.	Joseph P.	ROMANO

Soutenue le 25 octobre 2007 devant le jury composé de :

M.	Gilles	BLANCHARD	Examinateur
M.	Stéphane	BOUCHERON	Examinateur
M.	Pascal	MASSART	Président du jury
Mme	Marie-Agnès	PETIT	Examinatrice
M.	Stéphane	ROBIN	Examinateur
Mme	Sophie	SCHBATH	Directrice de thèse

Remerciements

Je tiens à remercier Sophie Schbath pour m'avoir encadré pendant toute la durée de cette thèse et pour m'avoir initié aux joies des approximations poissonniennes. Ses qualités scientifiques et humaines m'ont été très précieuses. De même, je remercie Gilles Blanchard qui m'a accueilli chaleureusement à Berlin. Ce fut un grand privilège de pouvoir travailler avec lui sur l'épineux problème des tests multiples. Bien sûr, je n'oublie pas Pascal Massart qui est à l'origine de ces deux rencontres et qui n'a pas hésité à prendre de son temps pour m'orienter dans mes choix scientifiques tout au long de cette thèse.

Je remercie Gesine Reinert et Joseph P. Romano pour m'avoir fait l'honneur de rapporter mon travail, ainsi que pour leur remarques pertinentes qui ont contribué à améliorer ce manuscrit. Merci à Stéphane Boucheron, Marie-Agnès Petit et Stéphane Robin d'avoir accepté de faire partie du jury, cela témoigne de l'intérêt qu'ils portent à mon travail, et je leur en suis très reconnaissant.

Merci à toutes les personnes avec qui j'ai eu la chance de discuter de science. Je pense à Antoine Chambaz qui m'a amicalement accueilli à plusieurs reprises dans son laboratoire, mais aussi aux riches échanges avec Sylvie Huet, François Rodolphe, Grégory Nuel, Jean-Jacques Daudin, Magalie Fromont et aux précieuses rencontres avec Yoav Benjamini, Helmut Finner et Sanat Sarkar dans des conférences internationales. Je remercie aussi Meriem El Karoui et Sylvain Baillet pour leur patience lors de nos discussions à propos de biologie et Mark Hoebeke pour toute son aide en informatique. Je dis également un grand merci à tous mes courageux relecteurs qui se reconnaîtront.

Je remercie ceux qui m'ont éveillé à la rigueur mathématique : mon professeur de classes préparatoires Denis Choimet ainsi que les autres excellents professeurs de l'Université de Rennes 1 et de l'antenne de Bretagne de l'ENS Cachan : Nicolas Lerner, Grégory Vial, Hubert Hennion et Michel Pierre. J'exprime aussi toute ma gratitude à Philippe Berthet pour m'avoir initié aux statistiques.

Je tiens également à remercier Mark van de Wiel et Aad van der Vaart de m'avoir spontanément offert une situation de post-doctorant dans leur laboratoire, ce qui m'a permis d'achever ma thèse avec plus de sérénité.

Je n'oublie pas Jean-François avec qui j'ai partagé mon bureau dans la plus grande harmonie, ainsi que tous les membres du laboratoire MIG et tous les autres doctorants que j'ai pu côtoyer. En particulier, merci à mes compagnons de promo Fanny et Sylvain, ainsi bien sûr qu'à Bobby qui reste “tranquille le chat” en toutes circonstances.

Pour finir, je remercie tendrement toute ma famille ainsi que Sabine pour leur soutien inconditionnel.

Table des matières

Présentation générale

Cette thèse possède deux parties pouvant être lues indépendamment. Chaque partie possède un chapitre de présentation détaillé : chapitre 1 pour la partie I, chapitre 9 pour la partie II. Nous présentons ici comment s'articulent les différents thèmes de ces deux parties et notamment de quelle façon certains enjeux de la partie I m'ont conduit à explorer la thématique de la partie II. Le chapitre 8 est à l'intersection des thématiques des parties I et II.

Partie I. Motifs exceptionnels dans des séquences hétérogènes

L'objectif est d'extraire d'une séquence d'ADN des motifs qui ont potentiellement une fonction biologique particulière. Pour cela, une démarche statistique consiste à rechercher les motifs de fréquence exceptionnelle. L'exceptionnalité d'un motif $\mathbf{w}$ est mesurée avec une probabilité critique (appelée *p-value* de $\mathbf{w}$), définie comme la probabilité que le nombre d'occurrences $N(\mathbf{w})$ du motif $\mathbf{w}$ dans une séquence aléatoire (modèle de référence) dépasse le nombre d'occurrences $N^{obs}(\mathbf{w})$ du motif $\mathbf{w}$ dans la séquence observée, c'est-à-dire :

$$\mathbb{P}(N(\mathbf{w}) \geq N^{obs}(\mathbf{w})).$$

La p-value d'un motif dépend bien entendu de la loi de la variable aléatoire $N(\mathbf{w})$, qui dépend elle-même du modèle probabiliste choisi pour la séquence aléatoire.

Classiquement, on ajuste un *modèle de Markov homogène stationnaire d'ordre m* sur la séquence. Cependant, ce modèle est critiquable car il suppose une homogénéité tout au long de la séquence et donc ne reflète pas l'hétérogénéité pouvant exister entre les différentes régions de la séquence.

Dans ce travail, nous cherchons à prendre en compte l'hétérogénéité d'une séquence dans le calcul de la p-value d'un motif. Pour cela, nous attachons à chaque position de la séquence un *état* (encore appelé *couleur*) pouvant représenter différents types d'information concernant une région de la séquence d'ADN (codant/non codant, variable/conservé, etc). La succession des états de la séquence est appelée la *segmentation*. Nous considérons deux types de modèles de séquences prenant en compte cette segmentation :

- Le modèle de Markov hétérogène par morceaux (noté PM : "**P**iece-wise heterogeneous **M**arkov") où la segmentation est déterministe, connue a priori.
- Le modèle de Markov caché à paramètres connus (noté HMM ou M1-Mm) où la segmentation est aléatoire (markovienne) et de loi connue.

Afin de calculer la p-value d'un motif dans les deux modèles ci-dessus, nous proposons d'approcher la loi du comptage $N(\mathbf{w})$. Nous nous focalisons sur des *approximations poissonniennes*, valables lorque le motif $\mathbf{w}$ est rare, i.e. lorsque $\mathbf{w}$ a un comptage attendu qui reste borné quand la longueur de la séquence tend vers l'infini. Nous proposons plusieurs approximations de Poisson composée, chacune correspondant au comptage d'une certaine partie des occurrences du motif **w** dans la séquence. Les approximations qui tiennent compte du plus grand nombre d'occurrences de **w** sont les plus précises mais les plus difficiles à calculer. Pour chacune de ces approximations, l'erreur est mesurée en distance en variation totale par la méthode de Chen-Stein.

Dans le cas où la segmentation est aléatoire (modèle M1-Mm), les approximations ont des termes d'erreurs explicites qui dépendent des paramètres de la loi de la segmentation. Par ailleurs, ces approximations nécessitent la mise en place d'une approximation pour le comptage d'une famille de mots rares recouvrante dans une séquence markovienne homogène. Ce problème, qui a également un intérêt propre, a donné lieu à l'article Roquain and Schbath (2007) qui constitue le chapitre 5.

Lorsque la segmentation est déterministe et connue a priori, les approximations sont valides sous certaines conditions sur la segmentation (nombre de régions suffisamment petites ou longueurs des régions suffisamment grandes). Ces approximations ont été implémentées dans une extension du logiciel R'MES[1] dédié à la **R**echerche de **M**otifs **E**xceptionnels dans les **S**équences. Sur plusieurs exemples de séquences (simulées ou réelles), nous montrons que le score d'exceptionnalité dans un modèle hétérogène diffère du score d'exceptionnalité dans un modèle homogène dès que la séquence sous-jacente est suffisamment hétérogène.

Un lien entre les thèmes des parties I et II : tester simultanément l'exceptionnalité de plusieurs motifs

Lorsque nous recherchons les motifs de fréquence exceptionnelle dans une séquence d'ADN, un problème de multiplicité se pose : parmi un grand nombre de p-values (par exemple les 16384 p-values correspondant aux motifs de longueur 7), il est probable que certaines p-values soient proches de 0 simplement par chance (*faux positifs*). Le problème est donc d'extraire à partir des p-values les motifs qui sont "réellement exceptionnels". Nous proposons une solution utilisant la *théorie des tests multiples*.

En contrôlant la probabilité d'avoir au moins k faux positifs (k-FWER), nous proposons dans un premier temps d'utiliser la méthode appelée ici "k-Bonferroni", qui est rapide mais réputée assez conservative. Dans un second temps, après avoir remarqué que la loi du k-ième minimum des p-values est facilement simulable dans une séquence markovienne, nous utilisons la méthode "k-min" qui est plus longue à calculer mais plus puissante (*i.e.* elle sélectionne davantage de motifs pour le même contrôle du k-FWER).

Partie II. Contributions à la théorie et à la méthodologie des tests multiples

Les procédures de tests multiples sont des outils statistiques indispensables pour analyser les données issues de nombreux domaines biologiques : puces à ADN, imagerie cérébrale, recherche de motifs exceptionnels dans l'ADN, etc. Les praticiens ont par conséquent besoin de procédures efficaces qui satisfont des critères théoriques de validité.

Dans ce travail, nous donnons des contributions à la théorie générale des tests multiples. Nous considérons un ensemble d'*hypothèses nulles* dont seulement une partie sont vraies. Nous

[1]http ://genome.jouy.inra.fr/ssb/rmes

supposons qu'il existe un ensemble de p-values permettant de tester chacune de ces hypothèses nulles de manière individuelle. Par suite, une *procédure de tests multiples* est définie comme une fonction qui, à partir d'un ensemble de p-values, retourne un certain ensemble d'hypothèses nulles, correspondant aux hypothèses nulles rejetées par la procédure. Une telle procédure peut ainsi faire deux types d'erreurs : une *erreur de type I* correspond à une hypothèse nulle rejetée à tort (*faux positifs*) ; une *erreur de type II* correspond à une hypothèse nulle non-rejetée à tort (*faux négatifs*). Il existe plusieurs façons de mesurer les erreurs de type I :

- La probabilité d'avoir au moins une erreur de type I (FWER) est une mesure assez stricte ; une procédure avec un FWER plus petit qu'un niveau α ne rejette jamais à tort avec probabilité plus grande que $1 - \alpha$.
- Une quantité plus permissive, et souvent préférée en pratique, est le taux de fausses découvertes (FDR), défini comme la proportion attendue de rejets à tort parmi l'ensemble des rejets. Ainsi, une procédure avec un FDR plus petit que α est autorisée à faire des erreurs parmi ses rejets mais en proportion plus petite que α (en moyenne).

Un objectif de la théorie des tests multiples est de construire des procédures garantissant un contrôle des taux d'erreurs de type I comme le FWER ou le FDR.

La partie II de cette thèse propose :

- Un nouvel éclairage sur les mathématiques mises en jeu dans les résultats les plus classiques du contrôle du FDR en donnant des preuves plus concises, basées sur des lemmes probabilistes explicites (ce qui autorise parfois à généraliser la forme des procédures classiques ou à affaiblir légèrement les hypothèses des théorèmes classiques).
- Des nouvelles procédures qui améliorent ou sont compétitives avec les procédures existantes ; notamment dans les problèmes d'*adaptivité* à π_0 (la proportion d'hypothèses nulles vraies) pour le contrôle du FDR, à l'aide de *procédures à plusieurs étapes*, et dans les problèmes d'adaptivité à la structure de dépendance entre les p-values pour le contrôle du FWER, à l'aide de procédures par *rééchantillonnage*. Ce dernier travail a fait l'objet d'un article, publié sous la référence Arlot *et al.* (2007).

Première partie

Motifs exceptionnels dans des séquences hétérogènes

Notations de la partie I

$\mathbf{1}\{E\}$, $\lvert E\rvert$	indicatrice, cardinal d'un ensemble E
$\mathcal{L}(X)$, $\mathbb{E}X$, d_{vt}	loi, espérance de X, distance en variation totale
$\mathcal{P}(\lambda)$	loi de Poisson de paramètre λ
$\mathcal{CP}(\lambda_k, k \geq 1)$	loi de Poisson composée de paramètres λ_k, $k \geq 1$
$\mathcal{A}$, $\mathbf{X} = (X_i)_{i\in\mathbb{Z}}$	alphabet, séquence (infinie)
π, Π, μ	probabilité de transition, matrice de transition, loi stationnaire de la séquence
$\mathbf{w} = w_1 \cdots w_h$, $\mathcal{W}$	mot de longueur h, famille de mots
$\mathcal{P}(\mathbf{w})$, $\mathcal{P}'(\mathbf{w})$	ensembles des périodes, périodes principales de $\mathbf{w}$
$\mathbf{w}^{(p)}$, $\mathbf{w}_{(p)}$	préfixe, suffixe d'ordre p de $\mathbf{w}$
$\mathbf{bcf} \in \mathcal{C}'_k$	motif maximal caractérisant l'occurrence d'un k-train
$Y_i(\mathbf{w})$, $\widetilde{Y}_i(\mathbf{w})$, $\widetilde{Y}_{i,k}(\mathbf{w})$	indicatrices d'occurrence – de $\mathbf{w}$, d'un train de $\mathbf{w}$, d'un k-train de $\mathbf{w}$ – à la position i
$N(\mathbf{w})$, $\widetilde{N}(\mathbf{w})$, $\widetilde{N}_k(\mathbf{w})$	nombres d'occurrences – de $\mathbf{w}$, de trains de $\mathbf{w}$, de k-trains de $\mathbf{w}$ – dans $X_1 \cdots X_n$
$N^\infty(\mathbf{w})$, $\widetilde{N}^\infty(\mathbf{w})$, $\widetilde{N}_k^\infty(\mathbf{w})$	nombres d'occurrences – de $\mathbf{w}$, de trains de $\mathbf{w}$, de k-trains de $\mathbf{w}$ – calculés dans la séquence $\mathbf{X}$
$a(\mathbf{w})$, $A(\mathcal{W})$	probabilité d'auto-recouvrement de $\mathbf{w}$, matrice d'auto-recouvrement de $\mathcal{W}$
$\mu(\mathbf{w})$, $\pi(\mathbf{w})$	probabilité d'occurrence de $\mathbf{w}$, probabilité d'occurrence de $\mathbf{w}$ sachant w_1
$\mathcal{S}$, $\mathbf{t} = t_1 \cdots t_h$, $\mathbf{s} = s_1 \cdots s_n$	ensemble des états, coloriage, segmentation (fixée)
ρ, τ_i	nombre de ruptures, i-ième instant de rupture
$\mathbf{s}_j$, $L_{\min}$	j-ième segment de $\mathbf{s}$, longueur minimale des $\mathbf{s}_j$
$\mathbf{X}_j$	j-ième segment de $\mathbf{X}$ selon la segmentation $\mathbf{s}$
$N_j(\mathbf{w})$	nombre d'occurences de $\mathbf{w}$ dans le segment $\mathbf{X}_j$
$N(\mathbf{w}, \mathbf{t})$, $N(\mathbf{w}, s)$	nombre d'occurences de $\mathbf{w}$ dans le coloriage $\mathbf{t}$, nombre d'occurences de $\mathbf{w}$ dans l'état s
$N_{\text{uni}}(\mathbf{w})$, $N_{\text{bic}}(\mathbf{w})$, $N'_{\text{bic}}(\mathbf{w})$	nombres d'occurrences de $\mathbf{w}$ unicolores, bicolores, dans les trains bicolores
π_s, Π_s, μ_s	probabilité de transition, matrice de transition, loi stationnaire de la séquence dans l'état s
$\mu_s(\mathbf{w})$, $\pi_s(\mathbf{w})$	probabilité d'occurrence de $\mathbf{w}$ dans l'état s, probabilité d'occurrence de $\mathbf{w}$ sachant w_1 dans l'état s
$a_s(\mathbf{w})$	probabilité d'auto-recouvrement de $\mathbf{w}$ dans l'état s
$\mathbf{S} = S_1 \cdots S_n$	segmentation (aléatoire)
$\pi_{\mathbf{S}}$, $\mu_{\mathbf{S}}$	probabilité de transition, loi invariante de $\mathbf{S}$
Mm, PMm, PSMm, HMMm	modèles de Markov – homogène stationnaire, hétérogène par morceaux, hétérogène stationnaire par morceaux, caché – d'ordre m

Chapitre 1

Présentation de la partie I

1.1 Motivation

L'objectif est de mettre en évidence dans une séquence d'ADN des motifs qui ont une fonction biologique particulière. Pour cela, une méthode consiste à rechercher les motifs sur- ou sous-représentés dans cette séquence, c'est-à-dire des motifs avec un comptage significativement trop grand ou trop petit par rapport à un comptage attendu a priori. Alors, sous la pression de sélection, un motif serait sur-représenté s'il est "bon" pour l'organisme étudié ; par exemple, il peut s'agir d'un site de fixation pour une protéine qui permet la transcription de gènes, ou encore d'un site bloquant la dégradation des brins d'ADN par une enzyme, etc. Inversement, un motif serait sous-représenté s'il est néfaste pour l'organisme ; par exemple, ce peut être un site de fixation d'enzymes de restriction qui coupent les deux brins d'ADN.

1.2 Mesure de l'exceptionnalité d'un motif

D'un point de vue mathématique, une séquence d'ADN est une succession $x_1 \cdots x_n$ de n lettres[1] prises dans l'alphabet $\mathcal{A} = \{\texttt{a},\texttt{g},\texttt{c},\texttt{t}\}$, et un motif $\mathbf{w}$ est une succession de h lettres $w_1 \cdots w_h$ prises dans le même alphabet (étant entendu que h est moralement bien plus petit que n). L'approche statistique classique consiste à supposer que la séquence observée $x_1 \cdots x_n$ est une réalisation d'une variable aléatoire $X_1 \dots X_n$ avec une certaine loi (ou famille de lois) donnée(s) a priori. Pour mesurer l'exceptionnalité de la fréquence du mot $\mathbf{w}$, on compare alors le comptage observé $N^{obs}(\mathbf{w})$ du motif $\mathbf{w}$ dans la séquence observée au comptage aléatoire $N(\mathbf{w})$ du mot $\mathbf{w}$ dans la séquence aléatoire "modèle", en calculant la p-value :

$$\mathbb{P}(N(\mathbf{w}) \geq N^{obs}(\mathbf{w})), \tag{1.1}$$

qui est la probabilité que le comptage de référence soit plus grand que celui observé. Fixons un niveau $\alpha \in]0,1[$ (par exemple $\alpha = 0.05$). Lorsque la p-value du motif est plus petite que α, cela signifie qu'avec probabilité plus grande que $1-\alpha$, on a $N(\mathbf{w}) < N^{obs}(\mathbf{w})$; le motif $\mathbf{w}$ est alors dit *significativement sur-représenté (au niveau α)*. Pour la sous-représentation, la p-value est plutôt définie par $\mathbb{P}(N(\mathbf{w}) \leq N^{obs}(\mathbf{w}))$, de sorte que cette probabilité soit plus petite que α lorsque qu'avec probabilité plus grande que $1-\alpha$, on ait $N(\mathbf{w}) > N^{obs}(\mathbf{w})$; le motif $\mathbf{w}$ est alors

[1]Ces lettres sont parfois appelées "bases" ou "nucléotides" (termes biologiques)

dit *significativement sous-représenté (au niveau α)*. Dans la suite, nous nous concentrons sur le cas sur-représenté (le cas sous-représenté étant analogue).

La p-value $\mathbb{P}(N(\mathbf{w}) \geq N^{obs}(\mathbf{w})) = \sum_{k \geq N^{obs}(\mathbf{w})} \mathbb{P}(N(\mathbf{w}) = k) = 1 - \sum_{k < N^{obs}(\mathbf{w})} \mathbb{P}(N(\mathbf{w}) = k)$ se calcule donc en fonction de la loi de $N(\mathbf{w})$, qui est bien entendu fonction du modèle qu'on a choisi pour la séquence $X_1 \cdots X_n$. Le choix du modèle est discuté dans la section suivante. Par la suite, nous calculerons la loi de $N(\mathbf{w})$, ce qui sera la principale difficulté mathématique du sujet.

Remarque 1.1 (Lien avec les tests) *La p-value (1.1) permet de tester l'hypothèse nulle H_0 : "la loi du comptage de $\mathbf{w}$ est la distribution nulle $\mathcal{L}(N(\mathbf{w}))$" à partir de l'observation $N^{obs}(\mathbf{w})$. L'hypothèse nulle H_0 est rejetée au niveau α si et seulement si la p-value est plus petite que α, c'est-à-dire lorsque $\mathbf{w}$ est significativement sur-représenté au niveau α. Ce test est donc défini individuellement pour chaque motif $\mathbf{w}$; le problème de construire un test simultanément pour plusieurs motifs à la fois sera examiné au chapitre 8.*

1.3 Choix du modèle

Le choix du modèle dépend de l'a priori que nous souhaitons nous donner sur la séquence. Les mots exceptionnels sont alors exceptionnels **par rapport** au modèle choisi. Par exemple,

- Si les lettres de la séquence d'ADN sont supposées indépendantes et valent `a,g,c,t` avec des probabilités égales, le comptage $N(\mathbf{w})$ a la même loi quelque soit $\mathbf{w}$. Ceci signifie que si l'on considère toutes les bases indépendantes et équiprobables, les motifs les plus sur-représentés sont simplement les plus fréquents.
- Si on suppose que les lettres de la séquence d'ADN sont toujours indépendantes mais que le modèle est "ajusté" à la fréquence de chacune des lettres, c'est-à-dire que la probabilité de voir chaque lettre $x \in \{\mathtt{a,g,c,t}\}$ est égale à la fréquence de x dans la séquence observée, on cherche alors les motifs exceptionnels en supposant la composition de la séquence en nucléotides connue.

Ainsi, prendre un modèle qui ne ressemble pas à la séquence observée est peu informatif, puisque cette méthode détectera des motifs exceptionnels par rapport à quelque chose de faux. Il faut donc choisir un modèle qui prend en compte une partie relativement importante de l'information contenue dans la séquence observée pour pouvoir interpréter les motifs détectés comme exceptionnels par rapport à cette information (par exemple par rapport à la composition de la séquence en nucléotides). Les modèles de Markov vont être très utiles en ce sens.

Remarque 1.2 *La p-value (1.1) a toujours un sens même lorsque le modèle est autorisé à dépendre des données. Pourtant, pour rester cohérent avec la démarche statistique, il convient de montrer que lorsque le modèle est ajusté à partir des données, la valeur de la p-value "est proche" de celle calculée avec la vraie loi du modèle. Pour des modèles markoviens, ceci se justifie asymptotiquement.*

1.3.1 Modèle de Markov homogène

On dit que la séquence suit un modèle de Markov homogène d'ordre $m \geq 1$ si pour tout i la loi de la lettre X_i conditionnellement aux lettres $X_j, j < i$ vaut z avec la probabilité

$\pi(X_{i-m} \cdots X_{i-1}, z)$, où $\pi : \mathcal{A}^m \times \mathcal{A} \to [0,1]$ est un paramètre du modèle (indépendant de la position i) vérifiant $\forall y_1 \cdots y_m \in \mathcal{A}^m, \sum_{z \in \mathcal{A}} \pi(y_1 \cdots y_m, z) = 1$. En particulier, la loi de X_i conditionnellement aux lettres $X_j, j < i$ ne dépend que de la valeur des m lettres précédant X_i (mémoire d'ordre m). Souvent, on choisit comme loi pour $X_1 \cdots X_m$ (loi initiale) la loi stationnaire de la chaîne, c'est-à-dire l'unique mesure μ sur $\mathcal{A}^m$ vérifiant $\sum_{y_1 \cdots y_m} \mu(y_1 \cdots y_m)\pi(y_1 \cdots y_m, z) = \mu(y_2 \cdots y_m z)$. L'existence et l'unicité d'une telle mesure sont assurées dès que $\forall y_1 \cdots y_m \in \mathcal{A}^m$, $\forall z \in \mathcal{A}$, $\pi(y_1 \cdots y_m, z) > 0$ (ou alternativement sous des hypothèses d'ergodicité). Ainsi défini, ce modèle sur $X_1 \cdots X_n$ est stationnaire et est donc communément appelé le modèle de Markov stationnaire d'ordre m.

Les paramètres π et μ sont classiquement ajustés sur le modèle en fonction de la séquence observée en choisissant les estimateurs :

$$\hat{\pi}(y_1 \cdots y_m, z) = \frac{N^{obs}(y_1 \cdots y_m z)}{N^{obs}(y_1 \cdots y_m +)}$$
$$\hat{\mu}(y_1 \cdots y_m) = \frac{N^{obs}(y_1 \cdots y_m)}{n},$$

où $N^{obs}(y_1 \cdots y_m z)$ est le nombre d'occurrences de $y_1 \cdots y_m z$ dans la séquence observée, où $N^{obs}(y_1 \cdots y_m)$ est le nombre d'occurrences de $y_1 \cdots y_m$ dans la séquence observée, et où on a noté $N^{obs}(y_1 \cdots y_m +) := \sum_{z \in \mathcal{A}} N^{obs}(y_1 \cdots y_m z)$. L'estimateur $\hat{\pi}$ n'a bien sûr de sens que si $N^{obs}(y_1 \cdots y_m +) > 0$, mais on peut montrer que c'est le cas lorsque n est suffisamment grand. Lorsque n tend vers l'infini, la loi forte des grands nombres pour les chaînes de Markov (cf. Dacunha-Castelle and Duflo (1983)) nous assure que ces estimateurs sont consistants, c'est-à-dire qu'ils convergent vers les vrais paramètres π et μ.

Comme la connaissance de $\hat{\pi}$ et $\hat{\mu}$ est équivalente à la connaissance des $(m+1)$-mots[2] dans la séquence observée (à la première lettre de la séquence près), choisir pour la séquence le modèle de Markov stationnaire d'ordre m avec les paramètres $\hat{\pi}$ et $\hat{\mu}$ revient à choisir comme a priori la composition de la séquence en mots de longueur $m+1$. Par conséquent, un mot de longueur $h \geq m+2$ sera exceptionnel dans ce modèle si son comptage ne peut s'expliquer à partir du comptage de ses sous-mots de longueur $m+1$. L'intérêt des modèles markoviens n'est donc pas dans le fait de modéliser une séquence biologique, mais plutôt dans le fait de bien contrôler l'a priori que nous nous donnons sur le modèle et donc de savoir quelle exceptionnalité nous regardons. Ainsi, calculer la p-value d'un mot de longueur h dans tous les modèles de Markov d'ordre m, $m \leq h-2$, peut avoir un intérêt pour déceler si l'exceptionnalité de ce mot est due à lui-même ou à un (ou plusieurs) de ses sous-mots[3].

Cependant, les séquences d'ADN étudiées présentent souvent une **hétérogénéité** de composition (codant/non-codant, variable/conservée), c'est-à-dire que la loi de génération des lettres n'est pas la même tout au long de la séquence. Par conséquent, ajuster un modèle homogène sur une séquence hétérogène peut être aberrant et il faut dans ce cas utiliser un modèle de Markov hétérogène.

[2]C'est-à-dire des mots de longueur $m+1$.

[3]Un sous-mot de $\mathbf{w} = w_1 \cdots w_h$ est défini comme un mot de ℓ lettres *consécutives* de $\mathbf{w}$, avec $\ell < h$.

1.3.2 Modèle de Markov hétérogène

L'hétérogénéité d'un modèle est décrite par une segmentation attachée à la séquence observée. La segmentation est définie comme une succession d'états pris dans un ensemble fini $\mathcal{S}$. On considèrera deux cas : le cas où la segmentation est **fixée et connue a priori** — on la notera alors $\mathbf{s} = s_1 \cdots s_n$ avec $s_i \in \mathcal{S}$ — et le cas où la segmentation est **aléatoire et de loi markovienne connue a priori**.

Cas d'une segmentation fixée

Lorsque la segmentation de la séquence est connue par le biologiste, nous pouvons l'intégrer dans un modèle de Markov hétérogène à segmentation fixée ("fixée" signifiant "déterministe" i.e. non aléatoire). Pour tout i, la loi de la lettre X_i conditionnellement aux lettres $X_j, j < i$ vaut z avec la probabilité $\pi_{s_i}(X_{i-m} \cdots X_{i-1}, z)$ où les $\pi_s, s \in \mathcal{S}$ sont des paramètres propres à des chaînes de Markov homogènes. Une version simplifiée de ce modèle est celle où la séquence globale n'est qu'une concaténation de modèles de Markov stationnaires indépendants (la concaténation se faisant selon la segmentation). L'estimation des transitions π_s et des mesures μ_s dans un état s se fait alors avec les estimateurs :

$$\begin{aligned}
\hat{\pi}_s(y_1 \cdots y_m, z) &= \frac{N^{obs}(y_1 \cdots y_m z, s)}{N^{obs}(y_1 \cdots y_m +, s)} \\
\hat{\mu}_s(y_1 \cdots y_m) &= \frac{N^{obs}(y_1 \cdots y_m, s)}{n_{\mathbf{s}}(s)},
\end{aligned}$$

où $N^{obs}(y_1 \cdots y_m z, s)$ est le nombre d'occurrences de $y_1 \cdots y_m z$ dans la séquence observée et entièrement dans l'état s, $N^{obs}(y_1 \cdots y_m +, s) := \sum_{z \in \mathcal{A}} N^{obs}(y_1 \cdots y_m z, s)$ et où $n_{\mathbf{s}}(s)$ désigne le nombre de fois où l'état s apparaît dans la segmentation $\mathbf{s}$. Nous remarquons que d'après le cas homogène, la convergence de ces estimateurs est garantie sur chaque segment si leur longueur tend vers l'infini.

Remarque 1.3 (Etat, couleur, coloriage) *Chaque état $s \in \mathcal{S}$ est aussi appelé "couleur". Une succession d'états est ainsi appelé "coloriage" et on parlera de mot "colorié". Ainsi, un mot $(y_1 \cdots y_m, s)$ représente un mot unicolore (colorié avec un coloriage unicolore) de couleur s.*

L'a priori d'un tel modèle réside donc dans le comptage des sous-mots d'ordre $m+1$ de $\mathbf{w}$ coloriés dans chaque état $s \in \mathcal{S}$. Un mot exceptionnel dans ce modèle aura donc un comptage qui ne s'explique pas à partir de la composition en $(m+1)$-mots unicolores.

Cas d'une segmentation aléatoire markovienne (HMM)

Lorsque la segmentation est inconnue par le biologiste, il est commode de modéliser la séquence par un modèle de Markov caché (HMM) ; la segmentation est alors un processus caché $\mathcal{S}$ (non observable), et la séquence $\mathbf{X}$ suit une chaîne de Markov (hétérogène) conditionnellement à la segmentation. Les paramètres du modèle sont consitués à la fois de la probabilité de transition $\pi_{\mathcal{S}}$ de la segmentation et des probabilités de transitions hétérogènes $\pi_s, s \in \mathcal{S}$ de la séquence $\mathbf{X}$ conditionnellement aux états de $\mathcal{S}$. Ces paramètres sont estimés classiquement avec l'algorithme EM (cf. par exemple Muri (1997)). Ici, on négligera cette étape d'estimation, en supposant ces paramètres connus a priori.

Comme il n'existe pas de statistique qui résume l'information que prend en compte un tel modèle, l'interprétation de l'exceptionnalité d'un motif dans un HMM est un peu moins intuitive : les motifs sont exceptionnels par rapport au modèle considéré (avec les paramètres connus ou estimés par une autre méthode statistique spécifique).

1.4 Approximations de la loi du comptage d'un mot : rappel du cas homogène

Il existe de nombreuses approches pour évaluer la loi du comptage d'un mot dans une séquence markovienne homogène stationnaire, et le lecteur pourra pour cela consulter le chapitre 6 de Lothaire (2005). Dans cette première partie de thèse, nous nous focaliserons principalement sur les approximations de type poissonniennes, valables pour des mots rares, c'est-à-dire pour des mots avec un comptage attendu borné avec la longueur de la séquence. Nous rappelons ainsi l'approximation de Poisson composée de Schbath (1995a). Cette démarche sera expliquée en détail dans le chapitre 2.

Dans une séquence markovienne homogène stationnaire et lorsque le mot **w** est rare, c'est-à-dire que son comptage attendu $\mathbb{E}N(\mathbf{w})$ est borné avec n, Schbath (1995a) a proposé d'approcher la loi de $N(\mathbf{w})$ par une approximation de type poissonnienne. Lorsque le mot ne peut pas se recouvrir, on peut voir asymptotiquement le processus des occurrences de **w** dans la séquence comme un processus de Poisson, et son comptage peut s'approcher par une loi de Poisson de paramètre $\mathbb{E}N(\mathbf{w})$. Lorsque le mot peut se recouvrir, le processus des occurrences de **w** n'a plus rien d'un processus de Poisson, car les occurrences de **w** arrivent par "paquets", que l'on appelle **train de w** (cf. FIG. 1.1). Ainsi, si on appelle **taille** d'un train de **w** le nombre d'occurrences de **w** dans le train et si on note $\widetilde{N}_k(\mathbf{w})$ le nombre de trains de **w** de taille k, le comptage du mot **w** s'écrit :

$$N(\mathbf{w}) = \sum_{k \geq 1} k \widetilde{N}_k(\mathbf{w}). \tag{1.2}$$

Avec le théorème de Chen-Stein (cf. Barbour *et al.* (1992)), on peut montrer que le processus des occurrences des k-trains (i.e. trains de taille k) de **w** est asymptotiquement un processus de Poisson, et donc que les $(\widetilde{N}_k(\mathbf{w}), k \geq 1)$ suivent asymptotiquement des lois de Poisson indépendantes de paramètres respectifs $\mathbb{E}\widetilde{N}_k(\mathbf{w})$, pour $k \geq 1$. Par définition, (1.2) permet de montrer que la loi de $N(\mathbf{w})$ suit asymptotiquement une loi de Poisson composée de paramètres $(\mathbb{E}\widetilde{N}_k(\mathbf{w}), k \geq 1)$.

`a c a c a c a c a g c a c t a c a c a`

FIG. 1.1 – Cette séquence contient 2 trains du mot `acaca` : l'un commençant à la position 1 (taille 3) et l'autre commençant à la position 15 (taille 1). On doit faire attention de ne pas oublier l'occurrence de `acaca` commençant à la position 3.

De plus, d'après Schbath (1995a), les paramètres $\mathbb{E}\widetilde{N}_k(\mathbf{w})$, $k \geq 1$ s'expriment de la façon suivante : pour tout $k \geq 1$,

$$\mathbb{E}\widetilde{N}_k(\mathbf{w}) = (a(\mathbf{w}))^{k-1}(1 - a(\mathbf{w}))^2 \mathbb{E}N(\mathbf{w}),$$

$a(\mathbf{w})$ désignant la probabilité d'auto-recouvrement du mot $\mathbf{w}$ (lorsqu'il n'y aura pas d'ambiguïté, on notera simplement a au lieu de $a(\mathbf{w})$). Ceci s'obtient en montrant que la probabilité d'occurrence d'un train de taille k de $\mathbf{w}$ à une position donnée est $a^{k-1}(1-a)^2\mu(\mathbf{w})$, où $\mu(\mathbf{w})$ désigne la probabilité d'occurrence de $\mathbf{w}$. Cette dernière formule peut se déduire de l'heuristique suivante (cf. FIG. 1.2) : s'il y a un train à une position donnée, il ne doit pas être précédé d'une occurrence de $\mathbf{w}$ (facteur en $1-a$), il est constitué de $k-1$ chevauchements successifs du mot $\mathbf{w}$ (facteur en a^{k-1}) et d'une occurrence de $\mathbf{w}$ (facteur en $\mu(\mathbf{w})$), et finalement il ne doit pas être suivi d'une autre occurrence de $\mathbf{w}$ (facteur en $1-a$). On obtient donc la probabilité finale en faisant le produit des différents facteurs.

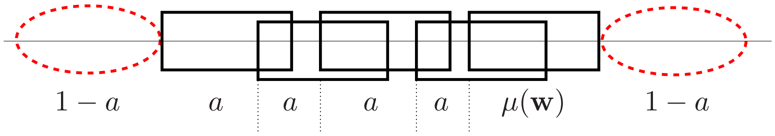


FIG. 1.2 – La probabilité d'occurrence d'un train de taille k dans le cas stationnaire est $a^{k-1}(1-a)^2\mu(\mathbf{w})$. Chaque rectangle représente une occurrence de $\mathbf{w}$, les ellipses représentent l'absence de recouvrement.

Remarque 1.4 *La stationnarité du modèle est essentielle dans le calcul des paramètres.*

L'asymptotique considérée ici est celle où $\mathbb{E}N(\mathbf{w}) = O(1)$, c'est-à-dire $(n-h+1)\mu(\mathbf{w}) = O(1)$, où h désigne la longueur du mot. Comme les paramètres du modèle sont fixes, cela suppose que $\mathbf{w}$ a une longueur qui tend vers l'infini. Formellement, $\mathbf{w}$ est en fait une suite de mots $\mathbf{w}_n$ et on considère la suite des lois des comptages des $\mathbf{w}_n$ dans des séquences markoviennes stationnaires de longueur n. Pourtant, cette notation étant un peu lourde, on omettra la dépendance en n dans $\mathbf{w}$ lorsque l'on imposera l'hypothèse de rareté.

1.5 Présentation des nouveaux résultats hétérogènes

Le but de cette première partie de thèse est d'établir des approximations poissonniennes pour la loi du comptage d'un mot dans le cas où la séquence suit un modèle hétérogène.

1.5.1 Différents types d'approximations considérés

Lorsque l'on attache une segmentation (fixée ou aléatoire) à la séquence observée, les occurrences de $\mathbf{w} = w_1 \cdots w_h$ apparaissent selon un certain **coloriage** $\mathbf{t} = t_1 \cdots t_h \in \mathcal{S}^h$, où les t_i sont des états de $\mathcal{S}$, de sorte que si $N(\mathbf{w}, \mathbf{t})$ désigne le nombre d'occurrences de $\mathbf{w}$ dans le coloriage $\mathbf{t}$, on a $N(\mathbf{w}) = \sum_{\mathbf{t}=t_1\cdots t_h} N(\mathbf{w}, \mathbf{t})$, c'est-à-dire que le comptage du mot $\mathbf{w}$ s'écrit comme la somme des comptages de $\mathbf{w}$ coloriés, dans tous les coloriages possibles. Pour simplifier le problème, nous allons considérer trois (sous-) comptages pour $\mathbf{w}$, chacun correspondant à un type de coloriage spécifique :

- le comptage unicolore $N_{\text{uni}}(\mathbf{w})$ qui est le nombre d'occurrences de $\mathbf{w}$ dans les coloriages unicolores $\mathbf{t} = s^h$, $s \in \mathcal{S}$,

– le comptage unicolore ou bicolore à une rupture d'état (noté dans la suite "au plus bicolore à une rupture d'état") $N_{\text{bic}}(\mathbf{w})$ qui est le nombre d'occurrences de $\mathbf{w}$ dans les coloriages $\mathbf{t} = s^\ell t^{h-\ell}$ pour $s \neq t$ et $\ell = 1, \ldots, h$ (le cas $\ell = h$ correspondant au comptage des occurrences unicolores),
– le comptage $N'_{\text{bic}}(\mathbf{w})$ de $\mathbf{w}$ égal à $\sum_{k\geq 1} k\widetilde{N}_{k,\text{bic}}(\mathbf{w})$, où le comptage $\widetilde{N}_{k,\text{bic}}(\mathbf{w})$ est le nombre de k-trains au plus bicolores à une rupture d'état.

Les coloriages "bicolores à une rupture d'état" n'incluent donc pas les coloriages du type $ststst\cdots$ avec $s \neq t$. Cependant, comme il n'y aura pas ambiguïté dans la suite, on s'autorisera parfois à utiliser seulement le terme "bicolore" au lieu de "bicolore à une rupture d'état" (c'est le cas par exemple dans la notation $N_{\text{bic}}(\mathbf{w})$). Pour mieux comprendre les comptages définis ci-dessus, nous proposons de traiter un exemple ; nous considérons le mot $\mathbf{w} = \texttt{aca}$ avec la séquence et la segmentation données comme dans la figure FIG. 1.3. Les comptages valent alors $N(\mathbf{w}) = 6$, $N_{\text{uni}}(\mathbf{w}) = 2$, $N_{\text{bic}}(\mathbf{w}) = 5$ et $N'_{\text{bic}}(\mathbf{w}) = 3$. Pour calculer $N'_{\text{bic}}(\mathbf{w})$, nous remarquons qu'il y a trois trains : le premier contient trois ruptures d'état et donc n'est pas pris en compte, le second a bien une seule rupture d'état et contient une occurrence de $\mathbf{w}$ et le dernier a bien une seule rupture d'état et contient deux occurrences de $\mathbf{w}$. Remarquons que, d'une manière générale, les inégalités suivantes sont vraies :

$$N_{\text{uni}}(\mathbf{w}) \leq N_{\text{bic}}(\mathbf{w}) \leq N(\mathbf{w})$$
$$N'_{\text{bic}}(\mathbf{w}) \leq N_{\text{bic}}(\mathbf{w}).$$

La dernière relation vient du fait qu'un train bicolore (à au plus une rupture d'état) ne contient que des occurrences de $\mathbf{w}$ bicolores (à au plus une rupture d'état). Remarquons que le comptage $N'_{\text{bic}}(\mathbf{w})$ n'est pas toujours plus grand que $N_{\text{uni}}(\mathbf{w})$ (cf. figure FIG. 1.3 en ne considérant que les 7 premières positions). Cependant, c'est le cas dès que tous les segments de la segmentation sont plus grands que la longueur maximale des trains, car dans ce cas tous les trains sont bicolores et $N'_{\text{bic}}(\mathbf{w}) = N(\mathbf{w}) \geq N_{\text{uni}}(\mathbf{w})$. Similairement, lorsque tous les segments de la segmentation sont plus grands que la longueur du mot h, on a $N_{\text{bic}}(\mathbf{w}) = N(\mathbf{w})$.

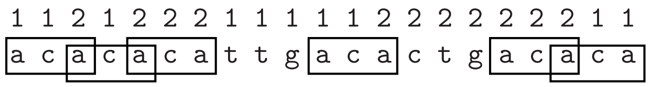


FIG. 1.3 – Réalisation d'une séquence avec sa segmentation attachée et illustration pour les comptages $N_{\text{uni}}(\mathbf{w})$, $N'_{\text{bic}}(\mathbf{w})$ et $N_{\text{bic}}(\mathbf{w})$ avec $\mathbf{w} = \texttt{aca}$.

Pour approcher la loi de $N(\mathbf{w})$ nous allons nous baser sur celle de l'un des comptages $N_{\text{uni}}(\mathbf{w})$, $N'_{\text{bic}}(\mathbf{w})$ ou $N_{\text{bic}}(\mathbf{w})$: une approximation est dite de type "mot unicolore", "train bicolore", "mot bicolore" ou "mot multicolore" si elle est basée sur le comptage de $N_{\text{uni}}(\mathbf{w})$, $N'_{\text{bic}}(\mathbf{w})$, $N_{\text{bic}}(\mathbf{w})$ ou $N(\mathbf{w})$ respectivement. Parmi ces approximations, celles qui tiennent compte du plus grand nombre d'occurrences de $\mathbf{w}$ sont les plus précises mais les plus difficiles à calculer. Les termes d'erreurs liés à ces différentes approximations vont bien entendu dépendre de la longueur des segments de la segmentation.

Remarque 1.5 *Pour les approximations de type "mot unicolore", "mot bicolore" et "mot multicolore", les approximations que nous proposerons seront directement des approximations pour* *la loi des comptages $N_{\text{uni}}(\mathbf{w})$, $N_{\text{bic}}(\mathbf{w})$ et $N(\mathbf{w})$, respectivemement. Pour l'approximation de type "train bicolore", ce sera dans les paramètres de la loi d'approximation que nous nous restreindrons aux occurrences bicolores des k-trains.*

Pour établir les différentes approximations, l'outil mathématique de base sera toujours la méthode de Chen-Stein (cf. Arratia *et al.* (1989)).

1.5.2 Cas d'une segmentation fixée

Le chapitre 3 propose deux approximations de Poisson composée pour le comptage $N(\mathbf{w})$ d'un mot $\mathbf{w}$ rare dans le cas où la segmentation $\mathbf{s}$ est fixée. Elles sont toutes les deux valables lorsque la séquence est une concaténation de chaînes de Markov homogènes stationnaires indépendantes.

- La première approximation est de type "mot unicolore" et elle s'effectue avec une loi notée $\mathcal{CP}_{\text{uni}}$; les paramètres de cette loi sont donnés par : $\forall k \geq 1$,

$$\sum_{s \in \mathcal{S}} (n_{\mathbf{s}}(s) - h + 1) a_s^{k-1} (1 - a_s)^2 \mu_s(\mathbf{w}),$$

  où a_s est la probabilité d'auto-recouvrement du mot $\mathbf{w}$ dans l'état s, $n_{\mathbf{s}}(s)$ est le nombre de fois où l'état s apparaît dans la segmentation $\mathbf{s}$ et $\mu_s(\mathbf{w})$ est la probabilité d'occurrence de $\mathbf{w}$ dans l'état s. La proposition 3.11 (page 49) établit que sous la condition de rareté, l'erreur en variation totale entre la loi de $N(\mathbf{w})$ et $\mathcal{CP}_{\text{uni}}$ tend vers 0 lorsque n tend vers l'infini dès que le nombre de ruptures ρ dans la segmentation est fixe avec n (ou alternativement si $\rho h = o(n)$).
- La seconde approximation s'effectue avec une loi notée $\mathcal{CP}_{\text{bic}}$:
  - dans le cas où $\mathbf{w}$ n'est pas recouvrant, il s'agit d'une approximation de Poisson de type "mot bicolore" de paramètre $\mathbb{E}N_{\text{bic}}(\mathbf{w})$. Le théorème 3.13 (page 51) montre que si la longueur minimale $L_{\min}$ des segments de $\mathbf{s}$ est plus grande que h et sous la condition de rareté, l'erreur en variation totale entre la loi de $N(\mathbf{w})$ et cette loi de Poisson tend vers 0 lorsque n tend vers l'infini.
  - dans le cas où $\mathbf{w}$ est recouvrant, il s'agit d'une approximation de Poisson composée de type "train bicolore". Les paramètres ont une expression explicite mais complexe (cf. Proposition 3.15 page 52). Le temps de calcul est donc plus long que pour la loi $\mathcal{CP}_{\text{uni}}$. Sous la condition de rareté, l'erreur tend vers 0 dès que $L_{\min}$ est suffisamment grand (précisément $\frac{L_{\min}-3h}{\max(\mathcal{P}'(\mathbf{w}))} \to \infty$, où $\max(\mathcal{P}'(\mathbf{w}))$ est la plus grande des périodes principales[4] de $\mathbf{w}$).

Remarque 1.6 *La condition sur $L_{\min}$ peut paraître contraignante ; en effet, elle n'est pas vérifiée s'il existe un seul segment de $\mathbf{s}$ de longueur petite. Cependant, comme nous nous plaçons sous la condition de rareté, nous pouvons toujours omettre un nombre fini de lettres dans la séquence sans changer la loi du comptage du mot asymptotiquement. Ainsi, dans la condition de validité $\frac{L_{\min}-3h}{\max(\mathcal{P}'(\mathbf{w}))} \to \infty$, $L_{\min}$ peut être remplacé par la longueur minimale des segments de $\mathbf{s}$ dans laquelle on a omis un nombre fini d'états.*

D'après les conditions de validité ci-dessus, l'approximation par $\mathcal{CP}_{\text{bic}}$ sera meilleure que celle par $\mathcal{CP}_{\text{uni}}$ pour des séquences comportant beaucoup de ruptures. Ces deux approximations seront

[4]La définition est donnée dans la section 2.1.2.

comparées précisément sur des simulations dans la section 6.2 du chapitre 4. Par exemple, pour le mot recouvrant $\mathbf{w} = \mathtt{aaaaaaa}$, dans une séquence de taille $n = 100000$ avec une segmentation à deux états (1 et 2) composée de 2000 segments de même longueur, la figure 1.4 représente les deux lois d'approximations $\mathcal{CP}_{\mathrm{uni}}$ et $\mathcal{CP}_{\mathrm{bic}}$, la loi de Poisson ajustée et la loi empirique du comptage (100000 simulations). Le modèle est hétérogène d'ordre 0 ; la probabilité d'émission de `a` et `g` est 0.35 dans l'état 1 et 0.15 dans l'état 2 ; la probabilité d'émission de `c` et `t` est 0.15 dans l'état 1 et 0.35 dans l'état 2. On voit que la loi $\mathcal{CP}_{\mathrm{bic}}$ (en tirets) est beaucoup plus proche de la loi empirique du comptage (en pointillés) que $\mathcal{CP}_{\mathrm{uni}}$ (en tirets-pointillés). Par ailleurs, comme le mot considéré est très recouvrant, l'approximation de Poisson (trait plein) n'est pas non plus valide. Cependant, nous verrons que la loi du comptage est suffisamment bien approchée par $\mathcal{CP}_{\mathrm{uni}}$ lorsque le nombre de segments est suffisamment petit (entre 1 et 100 dans les conditions ci-dessus).

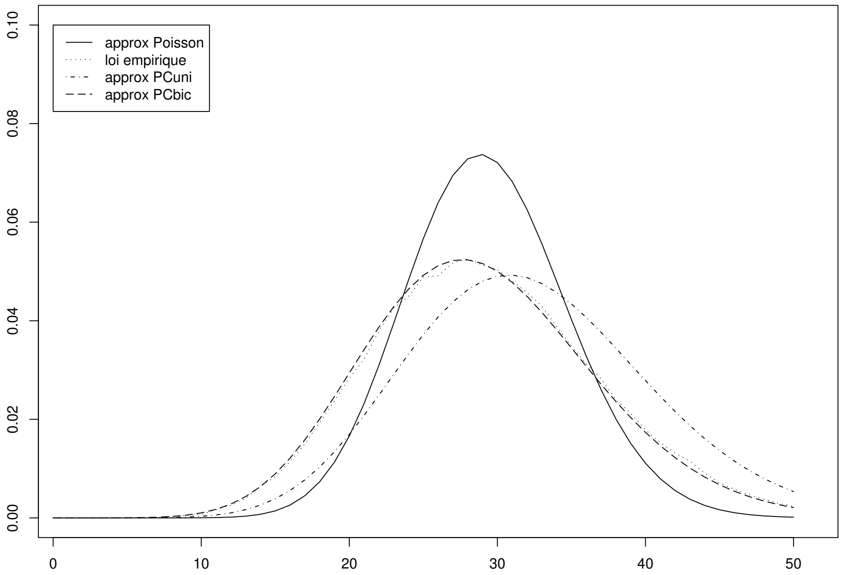


Fig. 1.4 – Densité des deux lois d'approximations $\mathcal{CP}_{\mathrm{uni}}$ ("approx. PCuni") et $\mathcal{CP}_{\mathrm{bic}}$ ("approx. PCbic"), de la loi de Poisson ajustée ("approx. Poisson") et de la loi empirique du comptage ("loi empirique") pour le mot `aaaaaaa` dans une séquence de longueur $n = 100\,000$ sous un modèle hétérogène d'ordre 0.

1.5.3 Cas d'un HMM

Le chapitre 4 examine le cas où la séquence suit un modèle HMM. On peut montrer qu'alors le couple $\mathbf{X}^\star = (\mathbf{X}, \mathbf{S})$ (séquence, segmentation) suit un modèle de Markov homogène stationnaire, quitte à considérer l'alphabet $\mathcal{A}^\star = \{(x, s), x \in \mathcal{A}, s \in \mathcal{S}\}$ des lettres de $\mathcal{A}$ coloriées par les états de $\mathcal{S}$. Ainsi, les nombres d'occurrences $N_{\mathrm{uni}}(\mathbf{w})$, $N_{\mathrm{bic}}(\mathbf{w})$ ou $N(\mathbf{w})$ dans $\mathbf{X}$ sont égaux respectivement aux nombres d'occurrences des familles de mots $\mathcal{W}_{\mathrm{uni}}$, $\mathcal{W}_{\mathrm{bic}}$ et $\mathcal{W}$ dans $\mathbf{X}^\star$, avec :

$$\mathcal{W}_{\text{uni}} := \{(\mathbf{w}, s^h), s \in \mathcal{S}\},$$
$$\mathcal{W}_{\text{bic}} := \{(\mathbf{w}, s^\ell t^{h-\ell}), s, t \in \mathcal{S}, 1 \leq \ell \leq h\},$$
$$\mathcal{W} := \{(\mathbf{w}, \mathbf{t}), \mathbf{t} \in \mathcal{S}^h\}.$$

Pour établir des approximations pour la loi des comptages $N_{\text{uni}}(\mathbf{w})$, $N_{\text{bic}}(\mathbf{w})$ et $N(\mathbf{w})$ dans une séquence HMM, **il suffit donc d'établir une approximation pour la loi d'une famille de mots dans une chaîne de Markov homogène stationnaire**. Une telle approximation a été proposée par Reinert and Schbath (1998), avec un terme d'erreur qui tend vers 0 sous la condition de rareté, pour des familles de mots non recouvrantes (comme par exemple $\mathcal{W}_{\text{uni}}$). Par contre, si la famille de mots est recouvrante (comme par exemple $\mathcal{W}_{\text{bic}}$ et $\mathcal{W}$), il n'est plus garanti que cette erreur tende vers 0. Ainsi, pour avoir accès à une approximation valable pour $N_{\text{bic}}(\mathbf{w})$ ou $N(\mathbf{w})$, nous avons dû dans un premier temps établir une approximation de Poisson composée valable pour les familles de mots rares recouvrantes (dans un modèle homogène stationnaire). Ce nouveau résultat, qui a par ailleurs un intérêt propre, est présenté dans la section 4.2 du chapitre 4, et le chapitre 5 constitue l'article Roquain and Schbath (2007) que nous avons publié.

En appliquant ce nouveau résultat, on déduit donc trois approximations de Poisson composée pour la loi de $N(\mathbf{w})$ dans un HMM, respectivement de type "mot unicolore" ($\mathcal{CP}'_{\text{uni}}$), "mot bicolore"($\mathcal{CP}'_{\text{bic}}$) et "mot multicolore" ($\mathcal{CP}'_{\text{mult}}$). Sous la condition de rareté, l'approximation par $\mathcal{CP}'_{\text{mult}}$ a une erreur qui tend vers 0, mais les paramètres sont complexes à calculer. Les lois $\mathcal{CP}'_{\text{uni}}$ et $\mathcal{CP}'_{\text{bic}}$ sont plus rapides à calculer, mais introduisent un terme d'erreur supplémentaire, qui est d'autant plus petit que la probabilité de quitter un état dans la chaîne $\mathcal{S}$ est petite. Au final, l'approximation par $\mathcal{CP}'_{\text{bic}}$ semble réaliser un bon compromis complexité/précision.

1.5.4 Compléments

Nous présentons dans le chapitre 7 quelques compléments dans le cas où la segmentation est fixée. Tout d'abord, lorsque le mot est fréquent, nous proposons une approximation gaussienne de type "mot unicolore" avec une erreur tendant vers 0, puis une approximation gaussienne de type "mot multicolore" dans le cas indépendant mais sans contrôle de l'erreur. Nous présentons également un algorithme pour calculer la loi exacte du comptage dans un modèle hétérogène, ce qui peut être utile pour des séquences "courtes". On traite finalement le problème de l'estimation des paramètres du modèle hétérogène (à segmentation fixée), ce qui nous permet d'estimer les paramètres de la loi de Poisson composée $\mathcal{CP}_{\text{uni}}$. Par suite, en notant $\widehat{\mathcal{CP}}_{\text{uni}}$ la loi $\mathcal{CP}_{\text{uni}}$ où l'on a remplacé les vrais paramètres par les paramètres estimés, on montre que l'erreur de l'approximation de la loi du comptage d'un mot rare par $\widehat{\mathcal{CP}}_{\text{uni}}$ a une erreur qui tend vers 0 (sous certaines conditions).

1.6 À la recherche de motifs exceptionnels dans des séquences hétérogènes

Dans la section 6.3 du chapitre 6, je présente plusieurs cas concrets de recherche de motifs exceptionnels dans des séquences d'ADN réelles. Plusieurs types d'hétérogénéités biologiques ont ainsi été étudiées.

J'ai implémenté les approximations hétérogènes **à segmentation fixée** par $\mathcal{CP}_{\text{uni}}$ et $\mathcal{CP}_{\text{bic}}$ dans une extension du logiciel R'MES[5]. Pour cela, les p-values sont calculées dans le modèle **hétérogène PSMm avec la segmentation fournie par l'utilisateur**, puis converties en score par une tranformation quantile d'une loi normale centrée réduite. Pour des raisons de temps de calcul, la loi utilisée par défaut pour approcher la loi du comptage est $\mathcal{CP}_{\text{bic}}$ pour les mots non-recouvrants et $\mathcal{CP}_{\text{uni}}$ pour le cas recouvrant (cette dernière étant suffisante lorsque la séquence n'est pas "trop" segmentée i.e. $\rho h/n$ "petit").

En utilisant ce nouveau programme, nous avons recherché les mots de longueur 5 qui sont de fréquence exceptionnelle dans le génome du phage *Lambda* (phage de la bactérie *Escherichia coli*). Son génome est de taille $n = 48502$; il est composé de nombreuses parties codantes (gènes) sur le brin direct ou sur le brin complémentaire. Comme la composition en oligonucléotides (mots) varie en fonction des parties codantes/non-codantes, il est naturel de choisir la segmentation codant/non-codant (précisément codant dans le sens direct/non-codant dans le sens direct). L'emplacement des gènes étant connu, la segmentation est donc connue a priori. Les approximations par défaut sont valides car les mots sont rares et le nombre de ruptures de la segmentation est faible $\rho = 36$. Nous avons donc calculé les **scores hétérogènes** de chaque mot de longueur 5 avec l'extension de R'MES décrite plus haut. Par suite, nous avons comparé cette méthode hétérogène avec la méthode homogène existante, c'est-à-dire avec la version 3 de R'MES (cf. Hoebeke and Schbath (2006)) qui calcule des **scores homogènes** en ne tenant pas compte de la segmentation (selon l'approximation présentée en section 1.4). La figure 1.5 représente les scores hétérogènes en fonction des scores homogènes : on remarque que certains mots (comme `gcaat` par exemple) sont intéressants, car ils sont assez "sur-représentés" et n'ont pas les mêmes scores hétérogène ou homogène. Pour ces mots-là, utiliser un modèle homogène est peu pertinent, et il est davantage conseillé de prendre en compte l'hétérogénéité de la séquence en utilisant la nouvelle méthode hétérogène.

Par ailleurs, cette nouvelle méthode hétérogène s'avère utile pour calculer le score d'exceptionnalité de motifs dans plusieurs séquences simultanément (il suffit pour cela de concaténer ces séquences).

[5]http ://genome.jouy.inra.fr/ssb/rmes

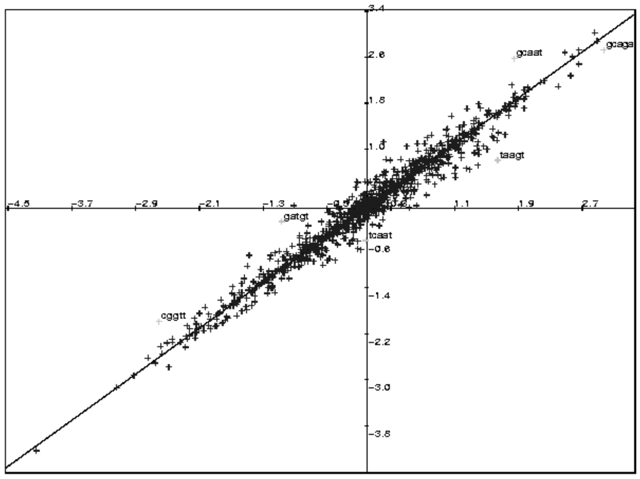


FIG. 1.5 – Scores hétérogènes (en ordonnée) contre les scores homogènes (en abscisse) pour tous les 5-mots dans le phage *Lambda*. L'ordre des modèles de Markov (homogène et hétérogène) est 3.

Chapitre 2

Prérequis : cas homogène

Dans ce chapitre, nous rappelons l'approximation de Schbath (1995a) valable dans le cas homogène, en simplifiant légèrement certaines démarches (notamment la définition de période principale et le calcul du comptage attendu) et en donnant un résultat nouveau (le calcul de la loi de la longueur d'un train de mots).

Nous fixons un ensemble fini $\mathcal{A}$ que l'on appellera **alphabet** (typiquement $\mathcal{A} = \{\mathtt{a}, \mathtt{t}, \mathtt{c}, \mathtt{g}\}$) ; nous appellons ces éléments des **lettres** et toute suite finie de lettres $\mathbf{w} = w_1 \cdots w_h$ des **mots**.

2.1 Comptages de w dans une séquence aléatoire

2.1.1 Définition des comptages $N(\mathbf{w})$ et $N^\infty(\mathbf{w})$

Nous fixons ici $\mathbf{X} = (X_i)_{i \in \mathbb{Z}}$ une séquence infinie de lettres aléatoires de $\mathcal{A}$, et nous ne spécifions pas pour l'instant quelle est la loi de $\mathbf{X}$. Définissons le nombre d'occurrences de $\mathbf{w}$ dans la séquence finie $X_1 \cdots X_n$ par :

$$N(\mathbf{w}) = \sum_{i=1}^{n-h+1} Y_i(\mathbf{w}) \tag{2.1}$$

où

$$Y_i(\mathbf{w}) = \mathbf{1}\{X_i \cdots X_{i+h-1} = w_1 \cdots w_h\}$$

désigne l'indicatrice qui vaut 1 si et seulement si $\mathbf{w} = w_1 \cdots w_h$ a une occurrence à la position i dans $\mathbf{X}$. Comme dans Schbath (1995a), pour tenir compte explicitement des recouvrements éventuels entre différentes occurrences de $\mathbf{w}$, on choisit une autre définition du comptage basée sur la notion de k-train ou encore train de taille k.

Un k-**train** de $\mathbf{w}$ dans une séquence est un ensemble maximal de k recouvrements successifs entre occurrences de $\mathbf{w}$ dans cette séquence ; par suite on dit qu'il y a une occurrence d'un k-train de $\mathbf{w}$ à la position i dans une séquence s'il y a occurrence d'un motif composé d'exactement k recouvrements successifs du mot $\mathbf{w}$ en position i dans la séquence sans que ce motif ne recouvre d'autres occurrences de $\mathbf{w}$ dans la séquence (cf. Exemple 2.1).

Exemple 2.1 *Pour* $\mathbf{w} = \mathtt{aataataa}$*, la séquence*

$$\mathtt{ct}\underline{\mathtt{aataataataataa}}\mathtt{cg}\underline{\mathtt{aataataa}}\mathtt{gca}$$

contient un 3*-train à la position* 3 *et un* 1*-train à la position* 19 *(on doit faire attention à ne pas oublier l'occurrence centrale de* **w** *dans le* 3*-train).*

Remarque 2.2 *Nous insistons sur le fait que la longueur et la taille d'un train de* **w** *désignent deux choses différentes : la longueur d'un train est simplement le nombre de lettres contenues dans le train alors que la taille d'un train est le nombre d'occurrences de* **w** *présentes dans le train. Ainsi, par exemple, la séquence* `ggatatatact` *possède un train* `atatata` *du mot* `atata` *à la position* 3 *de longueur* 7 *et de taille* 2*.*

Avec cette définition, on a donc :

$$N(\mathbf{w}) = \sum_{k\geq 1} k\widetilde{N}_k(\mathbf{w}),$$

où $\widetilde{N}_k(\mathbf{w})$ est le nombre d'occurrences d'un k-train dans la séquence finie $X_1 \dots X_n$. Pour des raisons techniques, nous allons plutôt travailler dans la séquence infinie $\mathbf{X} = (X_i)_{i\in\mathbb{Z}}$; le comptage des occurrences de $\mathbf{w}$ commençant aux positions $\{1, \dots, n-h+1\}$ dans la séquence infinie $\mathbf{X}$ est défini par :

$$N^\infty(\mathbf{w}) = \sum_{k\geq 1} k\widetilde{N}^\infty_k(\mathbf{w}) \quad \text{où} \quad \widetilde{N}^\infty_k(\mathbf{w}) = \sum_{i=1}^{n-h+1} \widetilde{Y}_{i,k}(\mathbf{w}), \tag{2.2}$$

avec $\widetilde{Y}_{i,k}(\mathbf{w})$ désignant l'indicatrice qui vaut 1 si et seulement si il y a une occurrence d'un k-train de $\mathbf{w}$ à la position i dans $\mathbf{X}$. Le nombre de trains de $\mathbf{w}$ commençant à une position de $\{1, \dots, n-h+1\}$ dans la séquence $\mathbf{X}$ s'écrit $\widetilde{N}^\infty(\mathbf{w}) := \sum_{k\geq 1} \widetilde{N}^\infty_k(\mathbf{w})$.

Les deux comptages $N(\mathbf{w})$ et $N^\infty(\mathbf{w})$ peuvent être différents car un train de $\mathbf{w}$ dans $\mathbf{X}$ peut commencer avant la position 1 et finir après la position $h-1$ et/ou commencer avant la position $n-h+2$ et finir après la position n (cf. FIG. 2.1). Cependant, l'événement $\{N(\mathbf{w}) \neq N^\infty(\mathbf{w})\}$ implique qu'il existe une occurrence de $\mathbf{w}$ commençant dans $\{1, \dots, h-1\} \cup \{n-h+2, \dots, n\}$. Ainsi, la distance en variation totale[1] entre la loi de $N(\mathbf{w})$ et celle de $N^\infty(\mathbf{w})$ est majorée par $\sum_{i\in\{1,\dots,h-1\}\cup\{n-h+2,n\}} \mathbb{E}Y_i(\mathbf{w})$. Cette quantité, dans le cas où $\mathbf{X}$ est stationnaire, vaut $2(h-1)\mathbb{E}Y_i(\mathbf{w})$ et tend vers 0 sous la condition de rareté $\mathbb{E}N(\mathbf{w}) = O(1)$ (et $h = o(n)$).

Pour expliciter ce qu'est exactement un recouvrement de k occurrences successives du mot $\mathbf{w}$ à une position, on doit caractériser les distances acceptables entre occurrences **recouvrantes** de $\mathbf{w}$ (resp. entre occurrences recouvrantes **successives** de $\mathbf{w}$) ; ceci nous conduit à la définition de périodes (resp. de périodes principales).

2.1.2 Périodes et périodes principales

L'ensemble des **périodes** de $\mathbf{w}$ est défini par $\mathcal{P}(\mathbf{w}) = \big\{p \in \{1, \dots, h-1\} \mid \forall i \in \{1, \dots, h-p\}, w_{i+p} = w_i\big\}$; les périodes de $\mathbf{w}$ sont les distances acceptables entre occurrences recouvrantes de $\mathbf{w}$. L'ensemble des **périodes principales** de $\mathbf{w}$ est défini par $\mathcal{P}'(\mathbf{w}) = \{p \in \mathcal{P}(\mathbf{w}) \mid \forall i \in \mathcal{P}(\mathbf{w}), p - i \notin \mathcal{P}(\mathbf{w})\}$; les périodes principales de $\mathbf{w}$ représentent donc les distances acceptables entre occurrences recouvrantes *successives* de $\mathbf{w}$. Autrement dit, $p \in \mathcal{P}(\mathbf{w})$ est principale si et seulement si il n'y a que deux occurrences de $\mathbf{w}$ dans le recouvrement $\mathbf{w}^{(p)}\mathbf{w}$.

[1]La distance en variation totale entre deux lois discrètes P et P' sur $\mathbb{N}$ est donnée par $\frac{1}{2}\sum_{x\in\mathbb{N}} |P(x) - P'(x)| = \min \mathbb{P}(X \neq X')$, où le minimum est pris sur l'ensemble des couples (X, X') avec $\mathcal{L}(X) = P$ et $\mathcal{L}(X') = P'$.

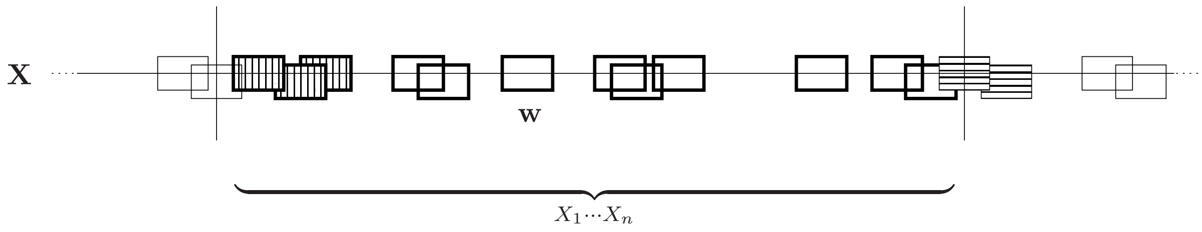


FIG. 2.1 – Différence entre $N(\mathbf{w})$ et $N^\infty(\mathbf{w})$. Les rectangles représentent les occurrences de $\mathbf{w}$ dans $\mathbf{X}$. Les occurrences de $\mathbf{w}$ comptées dans $N(\mathbf{w})$ sont marquées en gras. Les occurrences de $\mathbf{w}$ comptées dans $N(\mathbf{w})$ mais pas dans $N^\infty(\mathbf{w})$ sont remplies avec des traits verticaux. Les occurrences de $\mathbf{w}$ comptées dans $N^\infty(\mathbf{w})$ mais pas dans $N(\mathbf{w})$ sont remplies avec des traits horizontaux. Ici, $N(\mathbf{w}) = 12 = 9+3$, $N^\infty(\mathbf{w}) = 11 = 9+2$, $\widetilde{N}^\infty(\mathbf{w}) = 5$, $\widetilde{N}_1^\infty(\mathbf{w}) = 2$, pour $k = 2,3,4$, $\widetilde{N}_k^\infty(\mathbf{w}) = 1$ et pour $k \geq 5$, $\widetilde{N}_k^\infty(\mathbf{w}) = 0$.

Exemple 2.3 *Pour* $\mathbf{w} = \texttt{aataataa}$, *on a* $\mathcal{P}(\mathbf{w}) = \{3,6,7\}$ *et* $\mathcal{P}'(\mathbf{w}) = \{3,7\}$ *; en effet, la période* 6 *n'est pas principale, car elle correspond au recouvrement* $\texttt{aat}\underline{\texttt{aataataa}}\texttt{taa}$ *dans lequel la première occurrence et la dernière occurrence de* $\mathbf{w}$ *ne sont pas successives (une 3ème occurrence de* $\mathbf{w}$ *apparaît au milieu [occurrence soulignée]).*

Une conséquence directe de la définition des périodes principales est le lemme suivant.

Lemme 2.4 *(i) Une occurrence de* $\mathbf{w}$ *à la position* i *recouvre une occurrence précédente de* $\mathbf{w}$ *dans une séquence si et seulement si il existe une période principale* $p \in \mathcal{P}'(\mathbf{w})$ *telle qu'il y ait occurrence du préfixe* $\mathbf{w}^{(p)} := w_1 \cdots w_p$ *à la position* $i-p$ *dans cette séquence.*
(ii) Dans la précédente assertion, la période principale p *est unique.*

Remarquons que l'on peut obtenir les mêmes résultats en remplaçant "occurrence précédente" par "occurrence suivante" et "préfixe $\mathbf{w}^{(p)} := w_1 \cdots w_p$ à la position $i-p$" par "suffixe $\mathbf{w}_{(p)} := w_{h-p+1} \cdots w_h$ à la position $i+h$".

Remarque 2.5 *Schbath (1995a) a utilisé une définition des périodes principales (équivalente) un peu plus explicite : si* $p_0 = \min \mathcal{P}(\mathbf{w})$, *les périodes principales sont les périodes de* $\mathcal{P}(\mathbf{w})$ *qui ne sont pas de la forme* kp_0 *avec* $k \geq 2$. *La définition que l'on donne ici a l'avantage de se généraliser directement au cas plus complexe d'une famille de mots (cf. Section 5.3 du chapitre 5).*

2.1.3 Caractérisation de l'occurrence d'un k-train

En utilisant le Lemme 2.4, on déduit facilement que l'occurrence d'un k-train à la position i dans $\mathbf{X}$ est équivalente à l'occurrence d'un motif de la forme

$$\mathbf{bcf}$$

à la position $i-h$ avec $\mathbf{b} \in \mathcal{B}$ ("**b**efore"), $\mathbf{c} \in \mathcal{C}_k$ ("**c**omposed motif"), $\mathbf{f} \in \mathcal{F}$ ("**f**ollowing") ; $\mathcal{B}$ étant l'ensemble des h-mots ne finissant pas par un mot $\mathbf{w}^{(p)}$ avec $p \in \mathcal{P}'(\mathbf{w})$, $\mathcal{F}$ étant l'ensemble des h-mots ne commençant pas par un mot $\mathbf{w}_{(p)}$ avec $p \in \mathcal{P}'(\mathbf{w})$, et $\mathcal{C}_k$ étant l'ensemble des motifs de la forme $\mathbf{w}^{(p_1)} \cdots \mathbf{w}^{(p_{k-1})}\mathbf{w}$ avec $p_1, \ldots, p_{k-1} \in \mathcal{P}'(\mathbf{w})$. On note $\mathcal{C}'_k = \{\mathbf{bcf}, \mathbf{b} \in \mathcal{B}, \mathbf{c} \in \mathcal{C}_k, \mathbf{f} \in \mathcal{F}\}$. On a donc la relation explicite :

$$\widetilde{Y}_{i,k}(\mathbf{w}) = \sum_{\mathbf{bcf} \in \mathcal{C}'_k} Y_{i-h}(\mathbf{bcf}). \tag{2.3}$$

En outre, on note que par définition des ensembles $\mathcal{B}$ et $\mathcal{F}$, on a les relations suivantes : pour toute position i,

$$\sum_{\mathbf{b} \in \mathcal{B}} Y_{i-h}(\mathbf{b}w_1) = Y_i(w_1) - \sum_{p \in \mathcal{P}'(\mathbf{w})} Y_{i-p}(\mathbf{w}^{(p+1)}) \tag{2.4}$$

$$\sum_{\mathbf{f} \in \mathcal{F}} Y_i(w_h\mathbf{f}) = Y_i(w_h) - \sum_{p \in \mathcal{P}'(\mathbf{w})} Y_i(\mathbf{w}_{(p+1)}). \tag{2.5}$$

Les relations (2.3), (2.4) et (2.5) sont utiles pour calculer le comptage attendu d'un k-train. Cette quantité va intervenir comme paramètre de la loi aprochée du comptage (cf. Sections 2.2.2 et 3.4.3).

2.2 Approximation de la loi du comptage d'un mot rare lorsque X suit un modèle de Markov homogène

Soit $\mathbf{X} = (X_i)_{i \in \mathbb{Z}}$ une chaîne de Markov homogène sur un espace d'état fini $\mathcal{A}$ et de probabilité de transition π, c'est-à-dire, une suite de variables aléatoires à valeur dans $\mathcal{A}$ vérifiant la propriété markovienne :

$$\forall i \in \mathbb{Z}, \forall y, z \in \mathcal{A}, \mathbb{P}(X_i = z | (X_j)_{j<i-1}, X_{i-1} = y) = \pi(y, z),$$

signifiant que chaque état X_i ne dépend dans son passé que de l'état précédent X_{i-1}. On note Π la matrice (carrée d'ordre $|\mathcal{A}|$) de coefficients $\pi(y, z)$, $y, z \in \mathcal{A}$. On définit la probabilité de transition en $r \geq 2$ sauts par $\pi^r(y, z) := \sum_{y_1, \ldots, y_{r-1} \in \mathcal{A}} \pi(y, y_1) \times \cdots \times \pi(y_{r-1}, z)$. Par convention, $\pi^1(y, z) := \pi(y, z)$ et $\pi^0(y, z) := \mathbf{1}\{y = z\}$.

On suppose que cette chaîne est *irréductible*, i.e. $\forall y, z \in \mathcal{A}, \exists r \geq 0 \mid \pi^r(y, z) > 0$. Classiquement, cela garantit l'existence d'une mesure μ sur $\mathcal{A}$ vérifiant $\forall z \in \mathcal{A}$, $\mu(z) > 0$, et la propriété d'invariance $\mu(z) = \sum_{y \in \mathcal{A}} \mu(y)\pi(y, z)$; cette mesure est ainsi appelée la *mesure invariante* de la chaîne. Nous allons faire l'hypothèse supplémentaire que cette chaîne est *apériodique*, c'est-à-dire que $\forall y \in \mathcal{A}, \mathrm{pgcd}\{r \geq 1 \mid \pi^r(y, y) > 0\} = 1$, ceci implique par le théorème ergodique la propriété de convergence suivante : $\forall y, z \in \mathcal{A}, \pi^r(y, z) \to \mu(z)$, lorsque r tend vers l'infini.
Avec ces hypothèses, et comme la chaîne est indexée par $\mathbb{Z}$, on montre[2] que le processus $\mathbf{X} = (X_i)_{i \in \mathbb{Z}}$ est *stationnaire*. Le modèle résultant est noté classiquement modèle M1 (Markov d'ordre 1).

[2]Pour cela on remarque que pour $i \in \mathbb{Z}$ et $z \in \mathcal{A}$ fixés, on a $\mathbb{P}(X_i = z) = \sum_{y \in \mathcal{A}} \mathbb{P}(X_{i-r} = y)\pi^r(y, z)$ pour tout $r \geq 1$ et on fait tendre r vers l'infini en utilisant l'ergodicité de la chaîne.

2.2.1 Théorème d'approximation

Soit $\mathbf{w} = w_1 \dots w_h$ un mot de longueur $h \geq 2$ vérifiant l'hypothèse

$$\mu(\mathbf{w}) := \mu(w_1)\pi(w_1, w_2) \dots \pi(w_{h-1}, w_h) > 0. \tag{2.6}$$

En utilisant la méthode de Chen-Stein (cf. Barbour *et al.* (1992)), Schbath (1995a) a montré que

$$\begin{aligned} d_{vt}\big(\mathcal{L}(N(\mathbf{w})), \mathcal{CP}(\mathbb{E}\widetilde{N}_k^\infty(\mathbf{w}), k \geq 1)\big) &\leq d_{vt}\big(\mathcal{L}((\widetilde{N}_k^\infty(\mathbf{w}))_k), \otimes_k \mathcal{P}(\mathbb{E}\widetilde{N}_k^\infty(\mathbf{w}))\big) \\ &\leq (n-h+1)\mu(\mathbf{w})[Ch\mu(\mathbf{w}) + C'|\alpha|^h] + 2h\mu(\mathbf{w}), \end{aligned} \tag{2.7}$$

où d_{vt} désigne la distance en variation totale, C et C' sont deux constantes strictement positives qui ne dépendent que de la matrice de transition Π, α désigne la seconde plus grande valeur propre en valeur absolue de Π ($|\alpha| < 1$) et la loi $\mathcal{CP}(\mathbb{E}\widetilde{N}_k^\infty(\mathbf{w}), k \geq 1)$ désigne la loi de Poisson composée[3] de paramètres $(\mathbb{E}\widetilde{N}_k^\infty(\mathbf{w}),\ k \geq 1)$.

L'asymptotique considérée est alors la condition de rareté suivante :

$$\mathbb{E}N(\mathbf{w}) = O(1),$$

c'est-à-dire que le comptage attendu du mot $\mathbf{w}$ est borné lorsque $n \to \infty$. Ici, la matrice de transition Π du modèle est supposée fixée avec n, ce qui impose que $\mathbf{w}$ a une longueur qui tend vers l'infini. Plus précisément, la condition asymptotique $\mathbb{E}N(\mathbf{w}) = O(1)$ et $h = o(n)$ exige que la longueur du mot h tende vers l'infini plus vite que $\log(n)$ i.e. $\log(n)/h = O(1)$. En effet, comme on a par l'hypothèse (2.6) $\forall \ell \in \{1, \dots, h-1\}$, $\pi(w_\ell, w_{\ell+1}) > 0$, si on pose $\delta := \min\big(\{\pi(y,z), y, z \in \mathcal{A}\} \cap]0,1]\big) > 0$, on a $\forall \ell \in \{1, \dots, h-1\}$, $\pi(w_\ell, w_{\ell+1}) \geq \delta$. Ainsi, les conditions $\mathbb{E}N(\mathbf{w}) = O(1)$ et $h = o(n)$ imposent $n\mu(\mathbf{w}) = O(1)$ et donc $n\delta^h = O(1)$ soit $\log(n)/h = O(1)$.

Ainsi, la dernière inégalité de (2.7) montre que l'approximation de Poisson composée de paramètres $(\mathbb{E}\widetilde{N}_k^\infty(\mathbf{w}),\ k \geq 1)$ pour la loi du comptage $N(\mathbf{w})$ a une erreur qui tend vers 0 sous les conditions $\mathbb{E}N(\mathbf{w}) = O(1)$ et $h = o(n)$, c'est-à-dire

$$d_{vt}\big(\mathcal{L}(N(\mathbf{w})), \mathcal{CP}(\mathbb{E}\widetilde{N}_k^\infty(\mathbf{w}), k \geq 1)\big) \xrightarrow[n \to \infty]{} 0. \tag{2.8}$$

Remarque 2.6 *Pour obtenir la convergence (2.8), l'hypothèse (2.6) est superflue car lorsque le mot $\mathbf{w}$ a une probabilité d'occurrence égale à 0, on a $N(\mathbf{w}) = 0$ p.s. et donc trivialement $d_{vt}(\mathcal{L}(N(\mathbf{w})), \delta_0) = 0$, où δ_0 représente la loi Dirac en 0. Cependant, on a choisi ici d'exclure ce cas (un peu marginal) pour garantir que la longueur du mot tende vers l'infini.*

2.2.2 Calcul des paramètres de la loi de Poisson composée limite

Les paramètres de la loi limite $(\mathbb{E}\widetilde{N}_k^\infty(\mathbf{w}),\ k \geq 1)$ peuvent se calculer de la façon suivante : pour tout $k \geq 1$,

$$\mathbb{E}\widetilde{N}_k^\infty(\mathbf{w}) = (n-h+1)(a(\mathbf{w}))^{k-1}(1 - a(\mathbf{w}))^2\mu(\mathbf{w}), \tag{2.9}$$

[3]La loi de Poisson composée de paramètres $(\lambda_k, k \geq 1)$ est définie comme la loi de la variable aléatoire $\sum_{k \geq 1} kZ_k$, où les Z_k sont indépendants et chaque Z_k est de loi de Poisson de paramètre λ_k. Lorsque $\lambda := \sum_{k \geq 1} \lambda_k \in]0, \infty[$, cette loi s'écrit aussi comme la loi de la variable aléatoire $\sum_{j=1}^M K_j$, où M, $K_j, j \geq 1$ sont toutes indépendantes, M suit une loi de Poisson de paramètre λ et les variables aléatoires $K_j, j \geq 1$ sont identiquement distribuées de loi donnée par $\mathbb{P}(K_j = k) = \lambda_k/\lambda$.

où $a(\mathbf{w}) := \sum_{p\in\mathcal{P}'(\mathbf{w})}\prod_{\ell=1}^{p}\pi(w_\ell, w_{\ell+1})$ est la probabilité d'auto-recouvrement du mot $\mathbf{w}$ (lorsqu'il n'y aura pas ambiguïté, on notera simplement a au lieu de $a(\mathbf{w})$).
En effet, la relation (2.3) et la stationnarité de $\mathbf{X}$ donnent $\mathbb{E}\widetilde{N}_k^\infty(\mathbf{w}) = (n-h+1)\sum_{\mathbf{bcf}\in\mathcal{C}'_k}\mathbb{E}Y_{i-h}(\mathbf{bcf})$. Par la propriété de Markov,

$$\mathbb{E}Y_{i-h}(\mathbf{bcf}) = \mathbb{E}Y_{i-h}(\mathbf{b}w_1)\mathbb{E}[Y_i(\mathbf{c}) \mid Y_i(w_1)]\mathbb{E}[Y_{i+|\mathbf{c}|}(\mathbf{f}) \mid Y_{i+|\mathbf{c}|-1}(w_h)].$$

De plus, comme $\mathbf{X}$ est ici stationnaire, $\mathbb{E}Y_{i-h}(\mathbf{bcf}) = \mathbb{E}Y_i(\mathbf{b}w_1)\mathbb{E}[Y_i(\mathbf{c})|Y_i(w_1)]\mathbb{E}[Y_{i+1}(\mathbf{f})|Y_i(w_h)]$. Par suite,

$$\mathbb{E}\widetilde{N}_k^\infty(\mathbf{w}) = (n-h+1)\Bigg(\sum_{\mathbf{b}\in\mathcal{B}}\mathbb{E}Y_i(\mathbf{b}w_1)\Bigg)\Bigg(\sum_{\mathbf{c}\in\mathcal{C}_k}\mathbb{E}[Y_i(\mathbf{c})|Y_i(w_1)]\Bigg)\Bigg(\sum_{\mathbf{f}\in\mathcal{F}}\mathbb{E}[Y_{i+1}(\mathbf{f})|Y_i(w_h)]\Bigg).$$

En notant $\forall p$, $\pi(\mathbf{w}^{(p+1)}) := \prod_{\ell=1}^{p}\pi(w_\ell, w_{\ell+1})$, les expressions (2.4) et (2.5) donnent respectivement $\sum_{\mathbf{b}\in\mathcal{B}}\mathbb{E}Y_i(\mathbf{b}w_1) = \mu(w_1)(1-a)$ et $\sum_{\mathbf{f}\in\mathcal{F}}\mathbb{E}[Y_{i+1}(\mathbf{f})|Y_i(w_h)] = 1-a$. Pour finir, le terme central se traite directement :

$$\begin{aligned}\sum_{\mathbf{c}\in\mathcal{C}_k}\mathbb{E}[Y_i(\mathbf{c})|Y_i(w_1)] &= \sum_{p_1,\dots,p_{k-1}\in\mathcal{P}'(\mathbf{w})}\mathbb{E}[Y_i(\mathbf{w}^{(p_1)}\dots\mathbf{w}^{(p_{k-1})}\mathbf{w})]/\mu(w_1)\\ &= \sum_{p_1,\dots,p_{k-1}\in\mathcal{P}'(\mathbf{w})}\pi(\mathbf{w}^{(p_1+1)})\dots\pi(\mathbf{w}^{(p_{k-1}+1)})\pi(\mathbf{w}) = a^{k-1}\pi(\mathbf{w}).\end{aligned}$$

Exemple 2.7 *Dans $\mathbf{X}$, la probabilité d'auto-recouvrement de $\mathbf{w} = \mathtt{aataataa}$ est*

$$a(\mathbf{w}) = \pi(\mathtt{a},\mathtt{a})\pi(\mathtt{a},\mathtt{t})\pi(\mathtt{t},\mathtt{a})\big(1+\pi(\mathtt{a},\mathtt{a})^2\pi(\mathtt{a},\mathtt{t})\pi(\mathtt{t},\mathtt{a})\big).$$

Remarque 2.8 *1. On déduit facilement de (2.8) et de (2.9), que sous les conditions $\mathbb{E}N(\mathbf{w}) = O(1)$ et $h = o(n)$, la loi du nombre de trains vérifie*

$$d_{vt}\big(\mathcal{L}(\widetilde{N}^\infty(\mathbf{w})), \mathcal{P}((n-h+1)(1-a(\mathbf{w}))\mu(\mathbf{w}))\big) \xrightarrow[n\to\infty]{} 0,$$

où $\mathcal{P}$ désigne la loi de Poisson.

2. Si $\mathbf{w}$ n'est pas recouvrant, $a(\mathbf{w}) = 0$ et la loi de Poisson composée de (2.8) se réduit à une loi de Poisson de paramètre $\mathbb{E}N(\mathbf{w})$.

2.2.3 Lois de la taille et de la longueur d'un train

Lemme 2.9 *(i) Lorsque $a(\mathbf{w}) < 1$, la loi de la taille d'un train de $\mathbf{w}$ dans $\mathbf{X}$ est une loi géométrique de paramètre $a(\mathbf{w})$, i.e. si $\widetilde{Y}_i(\mathbf{w}) := \sum_{k\geq 1}\widetilde{Y}_{i,k}(\mathbf{w})$ désigne la probabilité d'occurrence d'un train à la position i,*

$$\mathbb{P}[\widetilde{Y}_{i,k}(\mathbf{w}) = 1|\widetilde{Y}_i(\mathbf{w}) = 1] = (1-a(\mathbf{w}))(a(\mathbf{w}))^{k-1}.$$

(ii) Pour un mot $\mathbf{w}$ recouvrant (i.e. lorsque $a(\mathbf{w}) > 0$), la loi de la longueur d'un k-train de $\mathbf{w}$ dans $\mathbf{X}$ est donnée par la loi de la variable aléatoire

$$\widetilde{L}_k = \sum_{p\in\mathcal{P}'(\mathbf{w})} pM_p + h, \text{ où } (M_p, p\in\mathcal{P}'(\mathbf{w})) \sim \mathcal{M}\Bigg(k-1, \bigg(\frac{\pi(\mathbf{w}^{(p+1)})}{a(\mathbf{w})}, p\in\mathcal{P}'(\mathbf{w})\bigg)\Bigg),$$

$\mathcal{M}$ désignant la loi multinomiale.

Preuve. La propriété (i) vient simplement du fait que

$$\mathbb{E}\widetilde{Y}_i(\mathbf{w}) = \sum_{k\geq 1} \mathbb{E}\widetilde{Y}_{i,k}(\mathbf{w}) = (1-a)^2 \sum_{k\geq 1} a^{k-1}\mu(\mathbf{w}) = (1-a)\mu(\mathbf{w}).$$

Pour prouver (ii), on définit sur l'événement $\{\widetilde{Y}_{i,k}(\mathbf{w}) = 1\}$ la variable aléatoire $\widetilde{L}_{k,i}$ comme étant la longueur du k-train apparaissant en position i ($\widetilde{L}_{k,i}$ étant définie arbitrairement sur $\{\widetilde{Y}_{i,k}(\mathbf{w}) = 0\}$). Alors, pour toute fonction $f : \mathbb{N} \to \mathbb{R}_+$,

$$\begin{aligned}
\mathbb{E}[f(\widetilde{L}_{k,i})\widetilde{Y}_{i,k}(\mathbf{w})] &= \sum_{\mathbf{bcf}\in\mathcal{C}'_k} f(|\mathbf{c}|)\mathbb{E}Y_{i-h}(\mathbf{bcf}) \\
&= a^{k-1}(1-a)^2\mu(\mathbf{w}) \sum_{p_1,\ldots,p_{k-1}\in\mathcal{P}'(\mathbf{w})} f\left(h + \sum_{\ell=1}^{k-1} p_\ell\right) \prod_{\ell=1}^{k-1} \frac{\pi(\mathbf{w}^{(p_\ell+1)})}{a} \\
&= \mathbb{E}\widetilde{Y}_{i,k}(\mathbf{w})\mathbb{E}f\left(h + \sum_{\ell=1}^{k-1} P_\ell\right) \qquad (2.10)
\end{aligned}$$

où les variables aléatoires $P_1, \ldots, P_{k-1}$ sont définies comme i.i.d avec $\forall p \in \mathcal{P}'(\mathbf{w}), \mathbb{P}(P_1 = p) = \frac{\pi(\mathbf{w}^{(p+1)})}{a}$. Par suite, si on note pour tout $p \in \mathcal{P}'(\mathbf{w})$, $M_p := |\{k' = 1, \ldots, k-1 \mid P_{k'} = p\}|$, le vecteur aléatoire $(M_p, p \in \mathcal{P}'(\mathbf{w}))$ suit une loi multinomiale de paramètres

$$\left(k-1, \left(\pi(\mathbf{w}^{(p+1)})/a, p \in \mathcal{P}'(\mathbf{w})\right)\right).$$

Comme par définition $h + \sum_{\ell=1}^{k-1} P_\ell = h + \sum_{p\in\mathcal{P}'(\mathbf{w})} pM_p$, le résultat découle de l'expression (2.10). ■

2.2.4 Généralisation à l'ordre m

Schbath (1995a) a proposé une généralisation à l'ordre $m \geq 2$ de ces résultats. Elle est directement obtenue à partir des résultats à l'ordre 1 en utilisant une astuce de changement d'alphabet : soit $\mathbf{X} = (X_i)_{i\in\mathbb{Z}}$ une chaîne de Markov homogène d'ordre m et de probabilité de transition π, i.e. un processus tel que chaque état X_i ne dépend dans son passé que des m variables précédentes : $\forall i \in \mathbb{Z}, \forall y_1 \cdots y_m \in \mathcal{A}^m, \forall z \in \mathcal{A}$,

$$\mathbb{P}(X_i = z | (X_j)_{j<i-m}, X_{i-m} \cdots X_{i-1} = y_1 \cdots y_m) = \pi(y_1 \cdots y_m, z).$$

Si on pose $\forall i \in \mathbb{Z}$, $X'_i := X_i \cdots X_{i+m-1}$ et $\forall y_1 \cdots y_m, \forall z_1 \cdots z_m \in \mathcal{A}^m$,

$$\pi'(y_1 \cdots y_m, z_1 \cdots z_m) := \left\{ \begin{array}{cl} \pi(y_1 \cdots y_m, z_m) & \text{si } y_2 \cdots y_m = z_1 \cdots z_{m-1} \\ 0 & \text{sinon} \end{array} \right. ,$$

le processus $\mathbf{X}' = (X'_i)_{i\in\mathbb{Z}}$ est alors une chaîne de Markov homogène d'ordre 1 sur $\mathcal{A}^m$ et de probabilité de transition π' (la matrice associée est notée Π'). D'après la remarque 2.10, la chaîne $\mathbf{X}'$ est de plus irréductible et apériodique, on note sa loi invariante μ'. Comme la chaîne est indexée par $\mathbb{Z}$, les chaînes $\mathbf{X}'$ et $\mathbf{X}$ sont stationnaires. Le modèle résultant pour $\mathbf{X}$ est noté classiquement Mm (Markov d'ordre m).

Par ailleurs, l'occurrence d'un mot $\mathbf{w} = w_1 \cdots w_h$ à la position i dans $\mathbf{X}$ est équivalente à l'occurence d'un mot $\mathbf{w}' = w'_1 \cdots w'_{h-m+1}$ à la position i dans $\mathbf{X}'$, avec $\forall \ell \in \{1, \dots, h-m+1\}$, $w'_\ell := w_\ell \cdots w_{\ell+m-1}$. Ainsi la loi du comptage $\mathbf{w}$ dans $\mathbf{X}$ est égale à la loi du comptage de $\mathbf{w}'$ dans $\mathbf{X}'$. Par suite, en appliquant les résultats précédents à l'ordre 1, on obtient que pour tout h-mot $\mathbf{w}$ vérifiant $\mu^{(m)}(\mathbf{w}) := \mu'(w_1 \dots w_m)\pi(w_1 \dots w_m, w_{m+1}) \times \cdots \times \pi(w_{h-m} \dots w_{h-1}, w_h) > 0$,

$$d_{vt}\Big(\mathcal{L}(N(\mathbf{w})), \mathcal{CP}\big((n-h+m)\big(1-a^{(m)}(\mathbf{w})\big)^2 \big(a^{(m)}(\mathbf{w})\big)^{k-1} \mu^{(m)}(\mathbf{w}), k \geq 1\big)\Big)$$
$$\leq (n-h+m)\mu^{(m)}(\mathbf{w})[C_m(h-m+1)\mu^{(m)}(\mathbf{w}) + C'_m|\alpha'|^{h-m+1}] + 2h\mu^{(m)}(\mathbf{w}),$$

où C_m et C'_m sont deux constantes strictement positives qui ne dépendent que de la matrice Π', α' désigne la seconde plus grande valeur propre en valeur absolue de Π' et où

$$a^{(m)}(\mathbf{w}) := \sum_{p \in \mathcal{P}'(\mathbf{w}),\, p \leq h-m} \prod_{\ell=1}^{p} \pi(w_\ell \cdots w_{\ell+m-1}, w_{\ell+m}).$$

En particulier, si m est fixe, $\mathbb{E}N(\mathbf{w}) = O(1)$ et $h = o(n)$, on a la convergence :

$$d_{vt}\Big(\mathcal{L}(N(\mathbf{w})), \mathcal{CP}\big((n-h+m)\big(1-a^{(m)}\big)^2 \big(a^{(m)}\big)^{k-1} \mu^{(m)}(\mathbf{w}), k \geq 1\big)\Big) \xrightarrow[n\to\infty]{} 0.$$

Remarque 2.10 (Ergodicité de la chaîne $\mathbf{X}'$) *Il est important de vérifier que la chaîne $\mathbf{X}'$ est irréductible et apériodique pour pouvoir utiliser les résultats précédents. On prouve facilement que c'est le cas dès que π satisfait $\forall y_1 \cdots y_m \in \mathcal{A}^m, z \in \mathcal{A}$, $\pi(y_1 \cdots y_m, z) > 0$. En effet, l'irréductibilité vient du fait que pour tout $y_1 \cdots y_m, z_1 \cdots z_m \in \mathcal{A}^m$, $(\pi')^m(y_1 \cdots y_m, z_1 \cdots z_m) \geq \pi(y_1 \cdots y_m, z_1) \times \cdots \times \pi(y_m z_1 \cdots z_{m-1}, z_m) > 0$ et l'apériodicité vient de ce que $\forall y_1 \cdots y_m \in \mathcal{A}^m$, $\forall i \geq 0$, $(\pi')^{m+i}(y_1 \cdots y_m, y_1 \cdots y_m) > 0$.*

Remarque 2.11 (Nombre de trains dans un modèle Mm) *Comme la transformation ci-dessus ne conserve pas le nombre de trains dans la séquence, on se gardera de donner une approximation pour le nombre de trains dans un modèle Mm, $m \geq 2$.*

Remarque 2.12 *A l'origine, le cadre M1 de Schbath (1995a) et Schbath (1995b) était celui où $\forall y, z \in \mathcal{A}, \pi(y, z) > 0$. Cependant, pour faciliter le passage à l'ordre m (la matrice Π' pouvant contenir des 0), on préfère ici seulement supposer qu'à l'ordre 1 la chaîne est irréductible et apériodique.*

Chapitre 3

Cas hétérogène à segmentation fixée

Le but de ce chapitre est d'établir une approximation de Poisson composée pour la loi du comptage d'un mot similaire au chapitre précédent, mais dans une séquence qui suit une chaîne de Markov hétérogène par morceaux. La segmentation ici est **déterministe** et **connue**.

Nous présentons dans la section 3.1 les deux modèles hétérogènes à segmentation fixée dans lesquels nous allons travailler : le modèle PM ("Piece-wise heterogeneous Markov") et le modèle PSM ("Piece-wise Stationary heterogeneous Markov"). Nous définissons dans la section 3.2 les comptages unicolore et bicolore d'un mot. La section 3.3 établit une approximation de Poisson composée générale dans un modèle PM, mais les paramètres de cette loi s'avèrent difficilement calculables en pratique. Nous examinons alors dans la section 3.4 le cas particulier d'un modèle PSM où l'on exploite la stationnarité par morceaux pour trouver deux lois d'approximation explicites pour la loi du comptage : la loi $\mathcal{CP}_{\text{uni}}$ qui donne une approximation valable pour un nombre faible de ruptures dans la segmentation, et la loi $\mathcal{CP}_{\text{bic}}$ qui donne une approximation valable lorsque la longueur minimale des segments est suffisamment grande. La preuve du théorème principal ainsi que l'énoncé et la preuve des lemmes annexes sont effectués dans la section 3.5.

3.1 Présentation des modèles PM et PSM

3.1.1 Segmentation

La segmentation représente l'hétérogénéité du modèle. Elle se définie de la façon suivante. Soit $\mathcal{S} \subset \mathbb{Z}$ un ensemble fini que l'on nomme **espace d'états** ; une **segmentation** est alors définie par une suite de n états $\mathbf{s} = s_1 \cdots s_n$ avec $s_i \in \mathcal{S}$. Elle possède les caractéristiques suivantes :

- Le nombre de ruptures de $\mathbf{s}$ est $\rho = |\{i \in \{2,\dots,n\} \mid s_i \neq s_{i-1}\}|$. Les instants de rupture sont les entiers $\tau_1 < \cdots < \tau_\rho$ tels que $\{\tau_1,\dots,\tau_\rho\} = \{i \in \{2,\dots,n\} \mid s_i \neq s_{i-1}\}$. Nous introduisons également les conventions $\tau_0 = 1$, $\tau_{\rho+1} = n+1$.
- Les $\rho+1$ segments de $\mathbf{s}$ sont les $\mathbf{s}_j = s_{\tau_{j-1}} \dots s_{\tau_j - 1}$ pour $j = 1,\dots,\rho+1$. Pour chaque segment $\mathbf{s}_j$ de $\mathbf{s}$, e_j désigne l'état de $\mathbf{s}_j$ ($e_j = s_{\tau_{j-1}}$).
- La longueur minimum des segments $\mathbf{s}_j$ de $\mathbf{s}$ est notée $L_{\min} = \min_{j\in\{1,\dots,\rho+1\}} |\mathbf{s}_j|$.

Exemple 3.1 *Pour* $\mathcal{S} = \{1,2,3\}$, *et* $\mathbf{s} = 311122311 = 3|111|22|3|11$, *on a* $\rho = 4$, $(\tau_1,\tau_2,\tau_3,\tau_4) = (2,5,7,8)$, $\mathbf{s}_1 = 3, \mathbf{s}_2 = 111, \mathbf{s}_3 = 22, \mathbf{s}_4 = 3, \mathbf{s}_5 = 11$ *et* $L_{\min} = 1$.

Remarque 3.2 *1. Pour donner un sens aux instants de rupture, nous supposerons toujours $\rho \geq 1$.*

2. Comme nous allons considérer plus tard une asymptotique en n, il convient de remarquer qu'ici la segmentation est susceptible de changer entièrement lorsque n varie. Ainsi, pour avoir une notation tout à fait précise, nous devrions écrire la segmentation comme $\mathbf{s}_n = s_{1,n} \cdots s_{n,n}$ et non $\mathbf{s} = s_1 \cdots s_n$. Pour alléger les notations au maximum nous avons choisi la seconde solution.

3. Toujours dans le souci de simplifier les notations, nous avons omis la dépendance en $\mathbf{s}$ dans les quantités ρ, τ_j et $\mathbf{s}_j$.

3.1.2 Modèle PM ("Piece-wise heterogeneous Markov")

Donnons-nous $\{\pi_s\}_{s\in\mathcal{S}}$ une famille de probabilités de transition sur $\mathcal{A}$, c'est-à-dire une famille de fonctions $\pi_s : \mathcal{A} \times \mathcal{A} \to [0, 1]$ avec $\forall s \in \mathcal{S}$, $\forall y \in \mathcal{A}$, $\sum_{z\in\mathcal{A}} \pi_s(y, z) = 1$. Supposons que $\forall s \in \mathcal{S}$, π_s est la probabilité de transition d'une chaîne de Markov (homogène) irréductible apériodique (cf. Section 2.2 pour la définition) et on note μ_s la mesure invariante associée à π_s. En outre, la matrice de transition associée à chaque π_s est notée Π_s. On définit à présent le modèle de Markov hétérogène par morceaux, proche de celui proposé par Robin *et al.* (2003a).

La séquence $\mathbf{X} = (X_i)_{i\in\mathbb{Z}}$ de lettres de $\mathcal{A}$ suit un modèle de **Markov hétérogène par morceaux** ("**P**iece-wise heterogeneous **M**arkov") d'ordre 1 selon la segmentation $\mathbf{s} = s_1 \cdots s_n$ et de paramètres $\{\pi_s\}_{s\in\mathcal{S}}$, si $\mathbf{X}$ est une chaîne de Markov hétérogène dont la probabilité de transition à l'étape i est π_{s_1} si $i < 0$, π_{s_i} si $1 \leq i \leq n$ et π_{s_n} si $i > n$; c'est-à-dire si pour tout $i \in \mathbb{Z}$ et tout $(y_j)_{j\leq i}$ avec $y_j \in \mathcal{A}$, on a

$$\mathbb{P}(X_i = y_i | (X_j)_{j<i} = (y_j)_{j<i}) = \begin{cases} \pi_{s_1}(y_{i-1}, y_i) & \text{si} \quad i < 0 \\ \pi_{s_i}(y_{i-1}, y_i) & \text{si} \quad 1 \leq i \leq n \\ \pi_{s_n}(y_{i-1}, y_i) & \text{si} \quad i > n \end{cases} . \tag{3.1}$$

Ce modèle est noté synthétiquement PM1($\mathbf{s}$,$\{\pi_s\}_s$) et nous omettrons dans la suite ($\mathbf{s}$,$\{\pi_s\}_s$) lorsqu'il n'y aura pas ambiguïté. Le "1" se rapporte à l'ordre de Markov du modèle.

Remarque 3.3 *1. Notons que lorsque pour tout état s les mesures $\pi_s(y, .)$ sont indépendantes de $y \in \mathcal{A}$ (et donc égale à $\mu_s(.)$), les lettres de $\mathbf{X}$ sont indépendantes les unes des autres. On note le modèle correspondant le modèle PM0.*

2. Dans un PM1, la chaîne de Markov $(X_i)_{i\leq 1}$ est une chaîne de Markov ergodique infinie de paramètre π_{s_1}. En particulier, la loi de X_1 est toujours μ_{s_1}.

3. Remarquons que lorsque tous les π_s sont égaux, le modèle PM1 se réduit à un modèle de Markov homogène.

```
3 1 1 1 2 2 3 1 1
a c g t g a t c g
```

$\pi_2(\mathtt{t}, \mathtt{g})$

FIG. 3.1 – Dans un modèle PM1 : $\mathbb{P}(X_5 = \mathtt{g} | X_4 = \mathtt{t}, X_1, X_2, X_3) = \pi_2(\mathtt{t}, \mathtt{g})$.

Nous pouvons définir de manière analogue un modèle PMm($\mathbf{s}$,$\{\pi_s\}_s$) (avec $m \geq 2$) à partir de probabilités de transition $\{\pi_s\}_{s\in\mathcal{S}}$ (supposées strictement positives) de chaînes de Markov homogènes d'ordre m, en remplaçant la condition markovienne (3.1) par celle d'ordre m :

$$\mathbb{P}(X_i = y_i | (X_j)_{j<i} = (y_j)_{j<i}) = \begin{cases} \pi_{s_1}(y_{i-m}\cdots y_{i-1}, y_i) & \text{si} \quad i < 0 \\ \pi_{s_i}(y_{i-m}\cdots y_{i-1}, y_i) & \text{si} \quad 1 \leq i \leq n \\ \pi_{s_n}(y_{i-m}\cdots y_{i-1}, y_i) & \text{si} \quad i > n \end{cases}.$$

Remarque 3.4 (Astuce du changement d'alphabet dans un modèle PM) *De façon similaire à la section 2.2.4, nous pouvons faire le changement d'alphabet suivant : à partir d'une séquence* $\mathbf{X} = (X_i)_{i\in\mathbb{Z}}$ *suivant un modèle PMm($\mathbf{s}$,$\{\pi_s\}_s$), la séquence* $\mathbf{X}' = (X'_i)_{i\in\mathbb{Z}}$ *de lettres dans* $\mathcal{A}^m$ *définie par* $X'_i := X_i \cdots X_{i+m-1}$ *suit un modèle PM1($\mathbf{s}'$,$\{\pi'_s\}_s$), où* $\mathbf{s}' := s_m \cdots s_n$ *et* $\forall y_1 \cdots y_m, z_1 \cdots z_m \in \mathcal{A}^m, \forall s \in \mathcal{S}$,

$$\pi'_s(y_1\cdots y_m, z_1\cdots z_m) := \begin{cases} \pi_s(y_1\cdots y_m, z_m) & \textit{si } y_2\cdots y_m = z_1\cdots z_{m-1} \\ 0 & \textit{sinon} \end{cases}.$$

Cette propriété est illustrée par la FIG. *3.2. Il faut penser à vérifier que comme les probabilités* π_s *sont supposées strictement positives, les* π'_s *associées sont bien des probabilités de transition de chaînes de Markov irréductibles apériodiques sur* $\mathcal{A}^m$ *(cf. la remarque 2.10 de la section 2.2.4). Avec ce changement d'alphabet, les résultats vrais dans un modèle PM1 vont pouvoir se généraliser facilement dans un modèle PMm*

s	`1 1 2 2 1 1 1 2 1`	s'	`1    2    2    1    1    1    2    1`
$X_1 \cdots X_9$	`a t a c a t a t a`	$X'_1 \cdots X'_8$	`(at)(ta)(ac)(ca)(at)(ta)(at)(ta)`
	$\pi_2(\texttt{at},\texttt{a})$		$\pi'_2(\texttt{at},\texttt{ta}) = \pi_2(\texttt{at},\texttt{a})$

FIG. 3.2 – À gauche : un modèle PMm dans $\mathcal{A}$; à droite : le modèle PM1 associé dans $\mathcal{A}^m$; pour $m = 2$, $n = 9$.

À partir de la segmentation $\mathbf{s}$ et d'une séquence aléatoire $\mathbf{X} = (X_i)_{i\in\mathbb{Z}}$, nous pouvons définir les $\rho + 1$ *segments de* $\mathbf{X}$ par $\mathbf{X}_1 = (X_i)_{i<\tau_1}$, $\mathbf{X}_j = (X_i)_{\tau_{j-1}\leq i<\tau_j}$ pour $j \in \{2, \ldots, \rho\}$ et $\mathbf{X}_{\rho+1} = (X_i)_{i\geq\tau_\rho}$. On remarque que lorsque $\mathbf{X}$ suit un modèle PM, les segments $\{\mathbf{X}_j\}$ de $\mathbf{X}$ sont markoviens mais généralement non stationnaires car la loi initiale de $\mathbf{X}_j$ dépend de la dernière lettre du segment $\mathbf{X}_{j-1}$. Ceci peut être évité en choisissant aux ruptures des probabilités de transition qui ne dépendent pas du passé, ce qui va définir le modèle de Markov hétérogène stationnaire par morceaux (PSM).

3.1.3 Modèle PSM ("Piece-wise heterogeneous Stationary Markov")

Une séquence infinie $\mathbf{X} = (X_i)_{i\in\mathbb{Z}}$ de lettres aléatoires de $\mathcal{A}$ suit **un modèle de Markov hétérogène stationnaire par morceaux** ("**P**iece-wise heterogeneous **S**tationary **M**arkov") d'ordre 1 selon la segmentation $\mathbf{s}$ et de paramètres $\{\pi_s\}_{s\in\mathcal{S}}$, si les segments $\{\mathbf{X}_j\}$ de $\mathbf{X}$ sont indépendants et si chaque segment $\mathbf{X}_j$ de $\mathbf{X}$ est une chaîne de Markov (d'ordre 1) homogène stationnaire de probabilité de transition π_{e_j} (e_j étant l'état du segment $\mathbf{s}_j$ de $\mathbf{s}$). Ce modèle est noté synthétiquement PSM1($\mathbf{s}$,$\{\pi_s\}_s$) et nous omettrons dans la suite ($\mathbf{s}$,$\{\pi_s\}_s$) lorsqu'il n'y aura pas ambiguïté.

Remarque 3.5 *1. Lorsque pour tout état s les mesures $\pi_s(y, .)$ sont indépendantes de $y \in \mathcal{A}$ (et donc égale à $\mu_s(.)$), les lettres de $\mathbf{X}$ sont indépendantes les unes des autres. On note le modèle correspondant le modèle PSM0. Remarquons qu'il est égal au modèle PM0.*

2. Étant une concaténation de chaînes de Markov stationnaires indépendantes, le modèle PSM1 est donc "stationnaire par morceaux". A chaque rupture, la chaîne est réinitialisée par la loi invariante correspondant à l'état courant (cf. FIG. *3.3). On remarque en particulier que même si les π_s sont égaux on n'obtient pas, en général, un modèle de Markov homogène.*

```
2 1 1 1 1 1 2 2 2
a c g t g a t c g
```

$\pi_1(\mathtt{c}, \mathtt{g})$ $\quad$ $\mu_2(\mathtt{t})$

FIG. 3.3 – Dans un modèle PSM1 : $\mathbb{P}(X_3 = \mathtt{g} \mid X_2 = \mathtt{c}, X_1) = \pi_1(\mathtt{c}, \mathtt{g})$ et $\mathbb{P}(X_7 = \mathtt{t} \mid X_6 = \mathtt{a}, X_1 \dots X_5) = \mu_2(\mathtt{t})$.

Remarquons que formellement, un modèle PSM1($\mathbf{s}$,$\{\pi_s\}_{s\in\mathcal{S}}$) est un modèle PM1($\mathbf{s}^\star$,$\{\pi^\star_{s^\star}\}_{s^\star\in\mathcal{S}^\star}$) où nous avons posé $\mathcal{S}^\star = \mathcal{S} \cup -\mathcal{S}$, et $\forall i \in \{2, \dots, n\}$, $\forall y, z \in \mathcal{A}$, $\forall s^\star \in \mathcal{S}^\star$,

$$\begin{aligned} s_i^\star &= \begin{cases} s_i & \text{si} \quad s_i = s_{i-1} \\ -s_i & \text{si} \quad s_i \neq s_{i-1} \end{cases} \\ \pi^\star_{s^\star}(y, z) &= \begin{cases} \pi_{s^\star}(y, z) & \text{si} \quad s^\star > 0 \\ \mu_{-s^\star}(z) & \text{si} \quad s^\star < 0 \end{cases} . \end{aligned}$$

La précédente transformation revient juste à remplacer les π_s par des μ_s lorsque s est un état arrivant à une rupture de la segmentation. Par conséquent, quitte à changer la segmentation, tout résultat valable pour un modèle de Markov hétérogène par morceaux est aussi valable pour un modèle de Markov hétérogène stationnaire par morceaux. Cependant, remarquons que si la longueur minimum des segments de $\mathbf{s}$ est plus grande qu'un entier ℓ, cela n'est pas le cas pour $\mathbf{s}^\star$, car la transformation ci-dessus introduit des segments de longueur 1 dans $\mathbf{s}^\star$.

Un modèle PSMm pour $m \geq 2$ se définit de manière similaire à un modèle PSM1, comme une concaténation de chaînes de Markov homogènes stationnaires d'ordre m. Notons cependant que l'astuce du changement d'alphabet ne permet plus de voir un modèle PSMm comme un modèle PSM1 : par exemple, la séquence `acgtt` a une probabilité d'occurrence dans un modèle PSM2 avec la segmentation 11122 égale à $\mu_1(\mathtt{ac})\pi_1(\mathtt{ac}, \mathtt{g})\mu_2(\mathtt{tt})$, alors que la probabilité d'occurrence de la séquence `(ac)(cg)(gt)(tt)` dans un modèle PSM1 avec la segmentation 1122 est $\mu_1(\mathtt{ac})\pi_1(\mathtt{ac}, \mathtt{g})\mu_2(\mathtt{gt})\pi_2(\mathtt{gt}, \mathtt{t})$.

3.2 Comptages colorié, unicolore ou bicolore d'un mot w

Nous fixons ici $\mathbf{X} = (X_i)_{i\in\mathbb{Z}}$ une séquence aléatoire de loi quelconque. Nous lui attachons une segmentation $s_1 \dots s_n$ déterministe donnée. Rappelons qu'avec les notations du chapitre 2, il y a une occurrence de $\mathbf{w}$ dans $\mathbf{X}$ à la position $i \in \{1,\ldots,n-h+1\}$ si et seulement si $X_i \cdots X_{i+h-1} = w_1 \cdots w_h$ c'est-à-dire si et seulement si $Y_i(\mathbf{w}) = \mathbf{1}\{X_i \cdots X_{i+h-1} = w_1 \cdots w_h\} = 1$.

On appelle *coloriage* de $\mathbf{w}$ toute suite de h états $\mathbf{t} = t_1 \cdots t_h$ de $\mathcal{S}$; le coloriage d'une occurrence de $\mathbf{w}$ dans $\mathbf{X}$ à la position i est ainsi la suite (non aléatoire) d'états $s_i \ldots s_{i+h-1}$. On appelle *sous-coloriage* de $\mathbf{w}$ toute suite de r états de $\mathcal{S}$ avec $r \leq h$. Pour un coloriage $\mathbf{t} = t_1 \ldots t_h$, le nombre d'occurrences du mot $\mathbf{w}$ colorié selon $\mathbf{t}$ est

$$N(\mathbf{w},\mathbf{t}) := \sum_{i=1}^{n-h+1} \mathbf{1}\{s_i \cdots s_{i+h-1} = t_1 \cdots t_h\} Y_i(\mathbf{w}).$$

Si le coloriage $\mathbf{t}$ n'est constitué que d'un seul état t, on note $N(\mathbf{w},t)$ au lieu de $N(\mathbf{w},\mathbf{t})$. On remarque que le comptage global de $\mathbf{w}$ est alors égal à la somme des comptages coloriés de $\mathbf{w}$, sur tous les coloriages possibles : $N(\mathbf{w}) = \sum_{\mathbf{t}\in\mathcal{S}^h} N(\mathbf{w},\mathbf{t})$.
Le comptage des occurrences unicolores de $\mathbf{w}$ est défini par

$$N_{\text{uni}}(\mathbf{w}) := \sum_{t\in\mathcal{S}} N(\mathbf{w},t),$$

le comptage des occurrences bicolores à au plus une rupture d'état de $\mathbf{w}$ est défini par

$$N_{\text{bic}}(\mathbf{w}) := \sum_{\mathbf{t}\in\mathcal{S}^h_{\text{bic}}} N(\mathbf{w},\mathbf{t}),$$

avec

$$\mathcal{S}^h_{\text{bic}} := \left\{ s^\ell t^{h-\ell}, (s,t) \in \mathcal{S}^2, s \neq t, \ell \in \{1,\ldots,h\} \right\}.$$

Par ailleurs, si $\{\mathbf{X}_j\}$ désigne les segments de $\mathbf{X}$, on note $N_j(\mathbf{w}) := \sum_{i=\tau_{j-1}}^{\tau_j-h} Y_i(\mathbf{w})$ le nombre d'occurrences de $\mathbf{w}$ dans le segment $\mathbf{X}_j$ (avec la convention $N_j(\mathbf{w}) = 0$ lorsque la longueur de $\mathbf{X}_j$ est strictement inférieure à h).

Remarque 3.6

1. *Le comptage unicolore peut s'écrire* $N_{\text{uni}}(\mathbf{w}) = \sum_{j=1}^{\rho+1} N_j(\mathbf{w}) = \sum_{j=1}^{\rho+1} \mathbf{1}\{|\mathbf{s}_j| \geq h\} N_j(\mathbf{w})$.
2. *De manière générale, on a évidemment* $N_{\text{uni}} \leq N(\mathbf{w})$. *Cependant, l'événement* $\{N(\mathbf{w}) \neq N_{\text{uni}}(\mathbf{w})\}$ *implique au moins une occurrence de* $\mathbf{w}$ *à une position de l'ensemble* $\bigcup_{j=1}^{\rho}\{\tau_j - h+1,\ldots,\tau_j-1\}$. *Ainsi, la distance en variation totale entre la loi de* $N(\mathbf{w})$ *et celle de* $N_{\text{uni}}(\mathbf{w})$ *est bornée par* $\sum_{j=1}^{\rho}\sum_{i=\tau_j-h+1}^{\tau_j-1} \mathbb{E}Y_i(\mathbf{w})$. *Cette quantité tend vers* 0 *dès que* $\rho h \max_i\{\mathbb{E}Y_i(\mathbf{w})\} = o(1)$, *ce qui impose une condition (asymptotique) sur le nombre de ruptures* ρ.
3. *Si la longueur minimum des segments* $L_{\min}$ *est plus grande que* h, *on a* $N(\mathbf{w}) = N_{\text{bic}}(\mathbf{w})$.

3.3 Approximation de Poisson composée dans un modèle PM

On considère dans cette section un mot $\mathbf{w} = w_1 \cdots w_h$ de longueur h sur l'alphabet $\mathcal{A}$ et une séquence $\mathbf{X}$ suivant un modèle PM1. On suppose dans toute la suite que pour tout $\mathbf{s} \in \mathcal{S}$,

$$\forall \ell \in \{2,\ldots,h\}, 0 < \pi_s(w_{\ell-1}, w_\ell) < 1 \tag{3.2}$$

$$\exists \gamma > 0 \text{ (indep. de } n\text{) tel que } \min_{1\leq i\leq n-h+1}\{\mathbb{P}(X_i = w_1)\} \geq \gamma \tag{3.3}$$

$$\exists \zeta < 1 \text{ (indep. de } n\text{) tel que } \max\{a_s(\mathbf{w}), s \in \mathcal{S}\} \leq \zeta, \tag{3.4}$$

où pour $s \in \mathcal{S}$,

$$a_s(\mathbf{w}) := \sum_{p \in \mathcal{P}'(\mathbf{w})} \prod_{\ell=1}^{p} \pi_s(w_\ell, w_{\ell+1}) \tag{3.5}$$

est la probabilité d'auto-recouvrement de $\mathbf{w}$ dans l'état s (lorsqu'il n'y aura pas ambiguïté, on notera simplement a_s au lieu de $a_s(\mathbf{w})$).

Les hypothèses (3.2) et (3.3) garantissent en particulier qu'à n'importe quelle position i, la probabilité d'occurrence de $\mathbf{w}$ est minorée par une quantité strictement positive. Elles seront utiles entre autres pour comprendre le cadre asymptotique de rareté. L'hypothèse (3.3) est vérifiée, par exemple, si pour tout i, $\mathbb{P}(X_i = w_1)$ est égal à une des probabilités $\mu_s(w_1), s \in \mathcal{S}$ et si ces dernières sont supposées strictement positives (c'est notamment le cas lorsque la chaîne est stationnaire par morceaux, cf. Section 3.4). L'hypothèse (3.4) signifie que le mot $\mathbf{w}$ ne peut pas avoir une probabilité d'auto-recouvrement tendant vers 1 dans un état donné. Elle garantit en particulier que le comptage $N^\infty(\mathbf{w})$ (défini page 32) peut être contrôlé en moyenne par le comptage $N(\mathbf{w})$ (cf. Lemme 3.21 page 59).

3.3.1 Probabilité d'occurrence et condition de rareté

Pour tout $i \in \{1, \ldots, n-h+1\}$, la probabilité d'occurrence de $\mathbf{w}$ à la position i dans $\mathbf{X}$ est $\mathbb{E}[Y_i(\mathbf{w})] = \mathbb{P}(X_i \cdots X_{i+h-1} = w_1 \ldots w_h)$, ce qui s'exprime par la propriété de Markov $\mathbb{E}[Y_i(\mathbf{w})] = \mathbb{P}(X_i = w_1)\pi_{s_{i+1} \cdots s_{i+h-1}}(\mathbf{w})$, où l'on a noté pour tout sous-coloriage $\mathbf{t} = t_1 \ldots t_{h-1} \in \mathcal{S}^{h-1}$ de $\mathbf{w}$,

$$\pi_{\mathbf{t}}(\mathbf{w}) := \pi_{t_1}(w_1, w_2) \times \cdots \times \pi_{t_{h-1}}(w_{h-1}, w_h). \tag{3.6}$$

Malheureusement, comme le modèle PM1 n'est en général pas stationnaire par morceaux, la probabilité $\mathbb{P}(X_i = w_1)$ ne dépend en général pas seulement de s_i mais aussi de la position i elle-même. Le comptage attendu s'écrit alors

$$\mathbb{E}[N(\mathbf{w}))] = \sum_{i=1}^{n-h+1} \mathbb{E}Y_i(\mathbf{w}) = \sum_{i=1}^{n-h+1} \mathbb{P}(X_i = w_1)\pi_{s_{i+1} \cdots s_{i+h-1}}(\mathbf{w}).$$

Intéressons-nous à présent à la condition de rareté pour $\mathbf{w}$. On rappelle qu'il s'agit de la condition asymptotique (lorsque $n \to \infty$)

$$\mathbb{E}[N(\mathbf{w})] = O(1).$$

Posons

$$\begin{aligned} \pi_{\min}(\mathbf{w}) &:= \min\{\pi_{\mathbf{t}}(\mathbf{w}), \mathbf{t} \in \mathcal{S}^{h-1}\} \\ \pi_{\max}(\mathbf{w}) &:= \max\{\pi_{\mathbf{t}}(\mathbf{w}), \mathbf{t} \in \mathcal{S}^{h-1}\}, \end{aligned}$$

et

$$\begin{aligned} \delta &:= \min\left\{\{\pi_s(x,y), x, y \in \mathcal{A}, s \in \mathcal{S}\} \cap]0,1]\right\} \\ \Delta &:= \max\left\{\{\pi_s(x,y), x, y \in \mathcal{A}, s \in \mathcal{S}\} \cap [0,1[\right\}. \end{aligned}$$

Notons que par définition $\delta > 0$ et $\Delta < 1$. En vertu des hypothèses (3.2) et (3.3) on obtient pour $\mathbb{E}[N(\mathbf{w})]$ les encadrements :

$$(n-h+1)\delta^{h-1}\gamma \leq (n-h+1)\pi_{\min}(\mathbf{w})\gamma \leq \mathbb{E}(N(\mathbf{w})) \leq (n-h+1)\pi_{\max}(\mathbf{w}) \leq (n-h+1)\Delta^{h-1}.$$

Ainsi, les conditions $\mathbb{E}N(\mathbf{w}) = O(1)$ et $h = o(n)$ imposent que $n\delta^{h-1} = O(1)$ et donc h converge vers l'infini plus vite que $\log(n)$ i.e. $\log(n) = O(h)$. Par suite, on remarque qu'on a $h\pi_{\max}(\mathbf{w}) \leq h\Delta^{h-1} = o(1)$.

3.3.2 Théorème d'approximation

Reprenons les notations de la section 2.1 et rappelons que $N^{\infty}(\mathbf{w})$ est défini par :

$$N^{\infty}(\mathbf{w}) = \sum_{k\geq 1} k\widetilde{N}_k^{\infty}(\mathbf{w}) \quad \text{où} \quad \widetilde{N}_k^{\infty}(\mathbf{w}) = \sum_{i=1}^{n-h+1} \widetilde{Y}_{i,k}(\mathbf{w}),$$

avec $\widetilde{Y}_{i,k}(\mathbf{w})$ qui vaut 1 si il y a une occurrence d'un k-train de $\mathbf{w}$ à la position i dans $\mathbf{X}$ et 0 sinon. Dans le modèle PM1, les comptages $N(\mathbf{w})$ et $N^{\infty}(\mathbf{w})$ vérifient

$$d_{vt}(\mathcal{L}(N(\mathbf{w})), \mathcal{L}(N^{\infty}(\mathbf{w}))) \leq \sum_{i\in\{1,\dots,h-1\}\cup\{n-h+2,n\}} \mathbb{E}Y_i(\mathbf{w}) \leq 2h\pi_{\max}(\mathbf{w}),$$

de sorte que sous la condition de rareté, les comptages $N(\mathbf{w})$ et $N^{\infty}(\mathbf{w})$ sont asymptotiquement équivalents, et nous pouvons considérer $N^{\infty}(\mathbf{w})$ plutôt que $N(\mathbf{w})$.

Nous allons maintenant utiliser le théorème de Chen-Stein (cf. Barbour *et al.* (1992)) pour borner la distance en variation totale entre la loi du processus $(\widetilde{Y}_{i,k}(\mathbf{w}))_{i,k}$ et la loi jointe de variables aléatoires $(Z_{i,k})_{i,k}$ indépendantes de loi de Poisson telles que $\mathbb{E}Z_{i,k} = \mathbb{E}\widetilde{Y}_{i,k}(\mathbf{w})$; on notera ces probabilités $\widetilde{\mu}_{i,k}(\mathbf{w})$. En définissant $Z_k := \sum_{i=1}^{n-h+1} Z_{i,k}$, le théorème de Chen-Stein dit que :

$$d_{vt}\Big(\mathcal{L}\big((\widetilde{N}_k^{\infty}(\mathbf{w}))_k\big), \mathcal{L}\big((Z_k)_k\big)\Big) \leq b_1 + b_2 + b_3, \tag{3.7}$$

où

$$b_1 = \sum_{i=1}^{n-h+1} \sum_{k\geq 1} \sum_{(j,\ell)\in B_{i,k}} \mathbb{E}(\widetilde{Y}_{i,k}(\mathbf{w}))\mathbb{E}(\widetilde{Y}_{j,\ell}(\mathbf{w})) \tag{3.8}$$

$$b_2 = \sum_{i=1}^{n-h+1} \sum_{k\geq 1} \sum_{(j,\ell)\in B_{i,k}\setminus\{(i,k)\}} \mathbb{E}(\widetilde{Y}_{i,k}(\mathbf{w})\widetilde{Y}_{j,\ell}(\mathbf{w})) \tag{3.9}$$

$$b_3 = \sum_{i=1}^{n-h+1} \sum_{k\geq 1} \mathbb{E}\Big|\mathbb{E}\big(\widetilde{Y}_{i,k}(\mathbf{w}) - \widetilde{\mu}_{i,k}(\mathbf{w})|\sigma(\widetilde{Y}_{j,\ell}(\mathbf{w}), (j,\ell)\notin B_{i,k})\big)\Big|, \tag{3.10}$$

et où $B_{i,k} \subset \{1,\dots,n-h+1\}\times\mathbb{N}^{\star}$ est un voisinage de (i,k). Par conséquent, grâce aux propriétés de la distance en variation totale, on a $d_{vt}\big(\mathcal{L}(N^{\infty}(\mathbf{w})), \mathcal{L}\big(\sum_{k\geq 1} kZ_k\big)\big) \leq b_1+b_2+b_3$, et $\sum_{k\geq 1} kZ_k$ suit par définition une loi de Poisson composée de paramètres $\mathbb{E}\widetilde{N}_k^{\infty}(\mathbf{w}) = \sum_{i=1}^{n-h+1} \widetilde{\mu}_{i,k}(\mathbf{w})$, $k \geq 1$. En majorant correctement b_1, b_2 et b_3 (cf. Section 3.5), on obtient le théorème suivant.

Théorème 3.7 *Supposons (3.2), (3.3) et (3.4). Lorsque la séquence* $\mathbf{X}$ *suit un modèle PM1 avec une segmentation fixée* $\mathbf{s}$ *vérifiant* $L_{\min} \geq h$*, on a la majoration suivante :*

$$\begin{aligned} d_{vt}\big(\mathcal{L}(N(\mathbf{w})), \mathcal{CP}(\lambda_k, k \geq 1)\big) &\leq d_{vt}\big(\mathcal{L}((\widetilde{N}_k^\infty(\mathbf{w}))_k), \otimes_k \mathcal{P}(\mathbb{E}\widetilde{N}_k^\infty(\mathbf{w}))\big) \\ &\leq C\mathbb{E}[N(\mathbf{w})]h\pi_{\max}(\mathbf{w}) + C'\mathbb{E}[N(\mathbf{w})]|\alpha_{\max}|^h \\ &\quad + C''[(h\pi_{\max}(\mathbf{w}))^2 + h\pi_{\max}(\mathbf{w})|\alpha_{\max}|^h], \end{aligned} \tag{3.11}$$

avec pour tout $k \geq 1$, $\lambda_k = \mathbb{E}\widetilde{N}_k^\infty(\mathbf{w}) = \sum_{i=1}^{n-h+1} \widetilde{\mu}_{i,k}(\mathbf{w})$ *qui est le nombre attendu de* k*-trains de* $\mathbf{w}$ *dans* $\mathbf{X}$ *commençant à une position de* $\{1, \dots, n-h+1\}$*. En particulier, lorsque le mot* $\mathbf{w}$ *est rare i.e. lorsque* $\mathbb{E}[N(\mathbf{w})] = O(1)$ *et* $h = o(n)$*, on a*

$$d_{vt}\big(\mathcal{L}(N(\mathbf{w})), \mathcal{CP}(\lambda_k, k \geq 1)\big) \xrightarrow[n\to\infty]{} 0. \tag{3.12}$$

Remarque 3.8

1. *On déduit facilement du théorème 3.7 que lorsque* $\mathbb{E}[N(\mathbf{w})] = O(1)$, $h = o(n)$ *et* $L_{\min} \geq h$*, le nombre de trains* $\widetilde{N}^\infty(\mathbf{w})$ *de* $\mathbf{w}$ *dans* $\mathbf{X}$ *vérifie*

$$d_{vt}\big(\mathcal{L}(\widetilde{N}^\infty(\mathbf{w})), \mathcal{P}(\mathbb{E}\widetilde{N}^\infty(\mathbf{w}))\big) \xrightarrow[n\to\infty]{} 0.$$

2. *Dans le théorème 3.7, le nombre de ruptures* ρ *peut être quelconque ; la seule condition est que* $L_{\min} \geq h$.
3. *La condition* $L_{\min} \geq h$ *du théorème 3.7 peut être un peu affaiblie : le théorème reste valide si les segments sont soit plus grands que* $h-1$ *soit de longueur* 1 *et si tous les segments de longueur* 1 *sont espacés d'au moins* $h-1$ *positions. Cela va nous permettre d'utiliser le théorème 3.7 dans le cas particulier où la séquence* $\mathbf{X}$ *suit un PSM, quitte à changer la segmentation* $\mathbf{s}$ *en* $\mathbf{s}^\star$ *et les probabilités de transitions* π_s *en* $\pi^\star_{s^\star}$ *(en utilisant les notations de la section 3.1.3).*
4. *Si* $\mathbf{w}$ *n'est pas recouvrant, la loi de Poisson composée limite se réduit à une loi de Poisson de paramètre* $\mathbb{E}N(\mathbf{w})$.
5. *Dans un modèle PM0,* $b_3 = 0$*, le terme en* $h|\alpha_{\max}|^h$ *disparaît dans la borne (3.11), et le théorème reste valide même sans l'hypothèse* $L_{\min} \geq h$.

La généralisation du théorème 3.7 au cas où $\mathbf{X}$ suit un modèle PMm est immédiate par l'astuce du changement d'alphabet : on applique le théorème 3.7 au mot $\mathbf{w}' = (w'_1 \cdots w'_{h-m+1})$ dans la séquence $\mathbf{X}'$ avec $X'_i := X_i \cdots X_{i+m-1}$ (qui suit un modèle PM1). Comme le comptage de $\mathbf{w}$ dans $X_1 \cdots X_n$ est égal au comptage de $\mathbf{w}'$ dans $X'_1 \cdots X'_{n-m+1}$, on obtient que si $\mathbb{E}[N(\mathbf{w})] = O(1)$ et $h = o(n)$,

$$d_{vt}\big(\mathcal{L}(N(\mathbf{w})), \mathcal{CP}(\lambda'_k, k \geq 1)\big) \xrightarrow[n\to\infty]{} 0, \tag{3.13}$$

où λ'_k est le nombre attendu de k-trains de $\mathbf{w}'$ dans $X'_1 \cdots X'_{n-m+1}$.

Dans les convergences (3.12) (et (3.13)), les paramètres de la loi d'approximation sont les comptages attendus des k-trains. Dans un modèle PM général, on a vu dans la section 3.3.1 qu'il n'y avait pas de simplification pour l'espérance d'un comptage. Ainsi, ces paramètres ne peuvent pas être raisonnablement calculés ici. Dans le cas particulier d'un modèle PSM, l'hypothèse de stationnarité par morceaux nous permet d'avoir une expression plus simple de ces paramètres.

3.4 Approximations de Poisson composée dans un modèle PSM

On considère dans cette section une séquence $\mathbf{X}$ suivant un modèle PSM1 et un mot $\mathbf{w}$ vérifiant les hypothèses suivantes : pour tout $s \in \mathcal{S}$,

$$\forall \ell \in \{2, \ldots, h\}, 0 < \pi_s(w_{\ell-1}, w_\ell) < 1 \tag{3.14}$$

$$\exists \zeta < 1 \text{ (indep. de } n) \text{ tel que } \max\{a_s, s \in \mathcal{S}\} \leq \zeta, \tag{3.15}$$

Ces hypothèses garantissent que les probabilités de transition $\{\pi^\star_{s^\star}\}_{s^\star \in \mathcal{S}^\star}$ associées (cf. notations de la section 3.1.3 page 41) vérifient les hypothèses (3.2), (3.3) et (3.4). Ainsi, dans cette section, nous allons pouvoir appliquer les résultats valables dans le modèle PM.

3.4.1 Probabilité d'occurrence et comptage attendu

Comme $\mathbf{X}$ est stationnaire par morceaux, la probabilité d'occurrence de $\mathbf{w}$ à la position i ne dépend que du coloriage $s_i \cdots s_{i+h-1}$ de $\mathbf{w}$ en i :

$$\mathbb{E}[Y_i(\mathbf{w})] = \mu_{s_i \cdots s_{i+h-1}}(\mathbf{w}),$$

avec $\forall \mathbf{t} = t_1 \cdots t_h$, la probabilité d'occurrence de $\mathbf{w}$ dans le coloriage $\mathbf{t}$ qui est $\mu_{\mathbf{t}}(\mathbf{w}) := \mu_{t_1}(w_1)\pi_{\mathbf{t}}(\mathbf{w})$ où

$$\pi_{\mathbf{t}}(\mathbf{w}) := \prod_{\ell=2}^{h} [\pi_{t_\ell}(w_{\ell-1}, w_\ell)\mathbf{1}\{t_\ell \neq t_{\ell-1}\} + \mu_{t_\ell}(w_\ell)\mathbf{1}\{t_\ell = t_{\ell-1}\}]. \tag{3.16}$$

Lorsque le coloriage $\mathbf{t}$ n'est constitué que d'un seul état t, on note $\mu_t(\mathbf{w})$ au lieu de $\mu_{\mathbf{t}}(\mathbf{w})$. On pose aussi $\mu_{\max}(\mathbf{w}) := \max\{\mu_{\mathbf{t}}(\mathbf{w}), \mathbf{t} \in \mathcal{S}^h\}$.

Exemple 3.9 *La probabilité d'occurrence de* `tga` *en position* 4 *dans la séquence ci-dessous est* $\mu_{122}(\texttt{tga}) = \mu_1(\texttt{t})\pi_{122}(\texttt{tga}) = \mu_1(\texttt{t})\mu_2(\texttt{g})\pi_2(\texttt{g}, \texttt{a})$.

$$\begin{array}{cccccccccccc} 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 1 & 1 \\ a & c & g & \boxed{t & g & a} & t & c & g & a & t & c \end{array}$$

Remarque 3.10 *La notation "$\pi_{\mathbf{t}}(\mathbf{w})$" de la formule (3.16) est la même que dans le cas d'un modèle PM, alors qu'elle désigne a priori une quantité différente (cf. la formule (3.6) page 44). Cet abus n'est en fait pas ambigu à notre sens car dans la formule (3.16),* $\mathbf{t}$ *est de longueur* $|\mathbf{w}|$ *alors que dans la formule (3.6)* $\mathbf{t}$ *est de longueur* $|\mathbf{w}| - 1$. *Remarquons aussi qu'il n'y a pas d'ambiguïté lorsque* $\mathbf{t}$ *n'est constitué que d'un seul état* t*, car dans ce cas il n'y a aucune rupture dans le coloriage :* $\mu_t(\mathbf{w}) = \mu_t(w_1)\pi_t(\mathbf{w})$ *où* $\pi_t(\mathbf{w}) = \pi_t(w_1, w_2) \times \cdots \times \pi_t(w_{h-1}, w_h)$ *pour les deux notations (3.6) et (3.16). Par ailleurs, soulignons le fait que "$\pi_{\max}(\mathbf{w})$" est défini en utilisant des coloriages* $\mathbf{t}$ *dans* $\mathcal{S}^{h-1}$ *donc selon (3.6), alors que "$\mu_{\max}(\mathbf{w})$" est défini avec des coloriages* $\mathbf{t}$ *dans* $\mathcal{S}^h$ *donc selon (3.16).*

En regroupant selon les coloriages possibles pour $\mathbf{w}$, le comptage attendu de $\mathbf{w}$ a l'expression simplifiée suivante :

$$\begin{aligned}
\mathbb{E}(N(\mathbf{w})) &= \sum_{\mathbf{t}\in\mathcal{S}^h} \sum_{i=1}^{n-h+1} \mathbf{1}\{s_i \cdots s_{i+h-1} = t_1 \cdots t_h\} \mu_{\mathbf{t}}(\mathbf{w}) \\
&= \sum_{\mathbf{t}\in\mathcal{S}^h} n_{\mathbf{s}}(\mathbf{t}) \mu_{\mathbf{t}}(\mathbf{w}),
\end{aligned} \tag{3.17}$$

où $n_{\mathbf{s}}(\mathbf{t}) = \sum_{i=1}^{n-|\mathbf{t}|+1} \mathbf{1}\{\mathbf{t} = s_i \cdots s_{i+|\mathbf{t}|-1}\}$ désigne le nombre (déterministe) d'occurrences du sous-coloriage $\mathbf{t}$ dans la segmentation $\mathbf{s}$. Lorsque de plus $L_{\min} \geq h$, la somme se réduit à des coloriages bicolores à au plus une rupture d'état (cf. Section 3.2 page 42), et on a :

$$\begin{aligned}
\mathbb{E}(N(\mathbf{w})) &= \sum_{\mathbf{t}\in\mathcal{S}^h_{\text{bic}}} n_{\mathbf{s}}(\mathbf{t}) \mu_{\mathbf{t}}(\mathbf{w}) \\
&= \sum_{t\in\mathcal{S}} n_{\mathbf{s}}(t^h) \mu_t(\mathbf{w}) + \sum_{s\neq t} n_{\mathbf{s}}(st) \mu_{s,t}(\mathbf{w}),
\end{aligned} \tag{3.18}$$

où $\mu_{s,t}(\mathbf{w}) := \sum_{\ell=1}^{h-1} \mu_{s^\ell t^{h-\ell}}(\mathbf{w})$ pour tout $s \neq t$. La formule (3.18) sera une formule fondamentale pour établir l'approximation de Poisson composée de type "train bicolore" (lors du calcul du comptage attendu des k-trains bicolores à au plus une rupture d'état).

3.4.2 Approximation par $\mathcal{CP}_{\text{uni}}$ pour un nombre faible de ruptures

L'idée exploitée ici est très simple : comme chaque segment $\mathbf{X}_j$ de $\mathbf{X}$ est une chaîne de Markov homogène stationnaire, on peut appliquer directement l'approximation de Poisson composée homogène de Schbath (1995a) sur chacun des segments et regrouper ensuite les paramètres des approximations. Pour cela, on introduit les notations suivantes : pour tout $s \in \mathcal{S}$, α_s désigne la seconde plus grande valeur propre en valeur absolue de la matrice de transition Π_s, $\alpha_{\max} = \max_{s\in\mathcal{S}} \alpha_s$ ($|\alpha_{\max}| < 1$) et on rappelle que la probabilité a_s d'auto-recouvrement de $\mathbf{w}$ dans l'état s est donnée par la formule (3.5) (cf. page 44).

D'après l'inégalité (2.7) page 35, on a pour tout $j \in \{1, \ldots, \rho+1\}$ tel que $|\mathbf{s}_j| \geq h$,

$$\begin{aligned}
&d_{vt}\big(\mathcal{L}(N_j(\mathbf{w})), \mathcal{CP}(\lambda_k^{(j)}, k \geq 1)\big) \\
&\qquad \leq C(|\mathbf{s}_j| - h + 1) h \mu^2_{s_{\tau_{j-1}}}(\mathbf{w}) + C'(|\mathbf{s}_j| - h + 1) \mu_{s_{\tau_{j-1}}}(\mathbf{w}) |\alpha_{s_{\tau_{j-1}}}|^h + 2h\mu_{s_{\tau_{j-1}}}(\mathbf{w}),
\end{aligned}$$

où C et C' sont deux constantes strictement positives qui ne dépendent que des transitions $\{\pi_s\}_s$ et où pour tout $k \geq 1$, $\lambda_k^{(j)} = (|\mathbf{s}_j| - h + 1) a^{k-1}_{s_{\tau_{j-1}}} (1 - a_{s_{\tau_{j-1}}})^2 \mu_{s_{\tau_{j-1}}}(\mathbf{w})$. Rappelons à présent que le comptage unicolore de $\mathbf{w}$ est donné par $N_{\text{uni}}(\mathbf{w}) = \sum_{j=1}^{\rho+1} \mathbf{1}\{|\mathbf{s}_j| \geq h\} N_j(\mathbf{w}) \leq \mathbb{E}[N(\mathbf{w})]$. Par indépendance entre les segments $\mathbf{X}_j$, pour $j \in \{1, \ldots, \rho+1\}$, on obtient

$$\begin{aligned}
&d_{vt}\left(\mathcal{L}(N_{\text{uni}}(\mathbf{w})), \mathcal{CP}\left(\sum_{j=1}^{\rho+1} \mathbf{1}\{|\mathbf{s}_j| \geq h\} \lambda_k^{(j)}, k \geq 1\right)\right) \\
&\qquad \leq C\mathbb{E}[N(\mathbf{w})] h \mu_{\max}(\mathbf{w}) + C'\mathbb{E}[N(\mathbf{w})] |\alpha_{\max}|^h + 2(\rho+1) h \mu_{\max}(\mathbf{w}).
\end{aligned}$$

Afin de simplifier l'expression des paramètres de la loi de Poisson composée, on regroupe les $\lambda_k^{(j)}$ associés aux segments coloriés dans le même état ; on pose pour $s \in \mathcal{S}$ et $k \geq 1$, $\lambda_{k,s} = (n_{\mathbf{s}}(s) - h + 1)a_s^{k-1}(1 - a_s)^2\mu_s(\mathbf{w})$, et on remarque que la distance en variation totale entre les lois $\mathcal{CP}(\sum_{j=1}^{\rho+1} \mathbf{1}\{|\mathbf{s}_j| \geq h\}\lambda_k^{(j)}, k \geq 1)$ et $\mathcal{CP}(\sum_{s\in\mathcal{S}} \lambda_{k,s}, k \geq 1)$ est majorée[1] par

$$\begin{aligned} \sum_{k\geq 1} \left| \sum_{j=1}^{\rho+1} \mathbf{1}\{|\mathbf{s}_j| \geq h\}\lambda_{k,j} - \sum_{s\in\mathcal{S}} \lambda_{k,s} \right| &\leq \sum_{k\geq 1} \max_{s\in\mathcal{S}} \left\{(h-1)(\rho+1)a_s^{k-1}(1-a_s)^2\mu_s(\mathbf{w})\right\} \\ &\leq (\rho+1)h\mu_{\max}(\mathbf{w}). \end{aligned}$$

Finalement, en combinant cela avec la majoration $d_{vt}\big(\mathcal{L}(N(\mathbf{w})), \mathcal{L}(N_{\text{uni}}(\mathbf{w}))\big) \leq h\rho\mu_{\max}(\mathbf{w})$ (cf. la remarque 3.6 page 43), on vient d'établir le résultat suivant.

Proposition 3.11 *Lorsque la séquence suit un modèle PSM1 et pour tout mot* $\mathbf{w}$ *vérifiant l'hypothèse (3.14), on a la majoration suivante :*

$$d_{vt}\big(\mathcal{L}(N(\mathbf{w})), \mathcal{CP}_{\text{uni}}\big) \leq C\mathbb{E}[N(\mathbf{w})]h\mu_{\max}(\mathbf{w}) + C'\mathbb{E}[N(\mathbf{w})]|\alpha_{\max}|^h + 4h(\rho+1)\mu_{\max}(\mathbf{w}), \quad (3.19)$$

$\mathcal{CP}_{\text{uni}}$ *étant la loi de Poisson composée de paramètres définis par :* $\forall k \geq 1$,

$$\sum_{s\in\mathcal{S}} (n_{\mathbf{s}}(s) - h + 1)(a_s(\mathbf{w}))^{k-1}(1 - a_s(\mathbf{w}))^2\mu_s(\mathbf{w}).$$

En particulier, lorsque $\mathbb{E}[N(\mathbf{w})] = O(1)$ *et* $h = o(n)$,

$$d_{vt}\big(\mathcal{L}(N(\mathbf{w})), \mathcal{CP}_{\text{uni}}\big) \xrightarrow[n\to\infty]{} 0, \quad (3.20)$$

dès que le nombre de ruptures ρ *de la segmentation vérifie* $\rho h\mu_{\max}(\mathbf{w}) = o(1)$.

Remarque 3.12

1. *Un nombre de ruptures* ρ *qui vérifie* $\rho h\mu_{\max}(\mathbf{w}) = o(1)$ *est dit* ***faible****. C'est toujours le cas si on a* $\rho = O(1)$. *Si on avait* $n\mu_{\max}(\mathbf{w}) = O(1)$ *(ce qui est une condition plus forte que* $\mathbb{E}N(\mathbf{w}) = O(1)$*), le critère se réduirait à* $\rho h = o(n)$. *Cette dernière condition est plus intuitive et sera utilisée en pratique ; elle indique que la proportion de positions susceptibles de faire apparaître une occurrence non unicolore de* $\mathbf{w}$ *dans la séquence* $X_1 \cdots X_n$ *tend vers 0.*
2. *On montre de même que sous les mêmes conditions que la proposition 3.11, le nombre de trains dans* $\mathbf{X}$ *s'approche par une loi de Poisson de paramètre*

$$\sum_{s\in\mathcal{S}} (n_{\mathbf{s}}(s) - h + 1)(1 - a_s(\mathbf{w}))\mu_s(\mathbf{w}).$$

De manière similaire à la proposition 3.11, on peut montrer que lorsque la séquence suit un modèle PSMm, la loi du comptage d'un mot $\mathbf{w}$ vérifiant $\mathbb{E}[N(\mathbf{w})] = O(1)$ et $h = o(n)$ s'approche par une loi de Poisson composée $\mathcal{CP}_{\text{uni}}^{(m)}$ avec une erreur en variation totale tendant vers 0, lorsque le nombre de ruptures ρ de la segmentation vérifie $\rho h\mu_{\max}^{(m)}(\mathbf{w}) = o(1)$ (la probabilité

[1] La distance en variation totale entre $\mathcal{CP}(x_k, k \geq 1)$ et $\mathcal{CP}(y_k, k \geq 1)$ est majorée par $\sum_{k\geq 1} |x_k - y_k|$ (cf. l'annexe C de Schbath (1995b)).

$\mu_{\max}^{(m)}(\mathbf{w})$ désignant la probabilité maximum d'occurrence de $\mathbf{w}$ à une position de la séquence). Les paramètres de la loi $\mathcal{CP}_{\text{uni}}^{(m)}$ sont alors donnés de la façon suivante : pour tout $k \geq 1$,

$$\sum_{s \in \mathcal{S}} (n_{\mathbf{s}}(s) - h + m)(a_s^{(m)}(\mathbf{w}))^{k-1}(1 - a_s^{(m)}(\mathbf{w}))^2 \mu_s^{(m)}(\mathbf{w}),$$

avec pour tout $s \in \mathcal{S}$

$$a_s^{(m)}(\mathbf{w}) := \sum_{p \in \mathcal{P}'(\mathbf{w}),\, p \leq h-m} \prod_{\ell=1}^{p} \pi_s(w_\ell \cdots w_{\ell+m-1}, w_{\ell+m})$$

$$\mu_s^{(m)}(\mathbf{w}) := \mu_s(w_1 \cdots w_m) \prod_{\ell=1}^{h-m} \pi_s(w_\ell \cdots w_{\ell+m-1}, w_{\ell+m}).$$

Lorsque le terme en $h\rho\mu_{\max}(\mathbf{w})$ (resp. $h\rho\mu_{\max}^{(m)}(\mathbf{w})$ dans le cas PSMm) ne tend plus vers 0, cela signifie qu'il faut alors prendre en compte l'occurrence des mots aux ruptures de la segmentation c'est-à-dire qu'il faut considérer au moins des coloriages "bicolores" ; c'est l'objet du paragraphe suivant.

3.4.3 Approximation par $\mathcal{CP}_{\text{bic}}$ pour un nombre quelconque de ruptures

Comme un modèle PSM1 est un cas particulier d'un modèle PM1 et suivant le point 3 de la remarque 3.8 (page 46), le théorème 3.7 (page 46) établit que sous les hypothèses (3.14) et (3.15), lorsque le mot $\mathbf{w}$ vérifie $\mathbb{E}N(\mathbf{w}) = O(1)$, $h = o(n)$ et si la segmentation $\mathbf{s}$ vérifie $L_{\min} \geq h$, on a :

$$d_{vt}\big(\mathcal{L}(N(\mathbf{w})), \mathcal{CP}(\mathbb{E}\widetilde{N}_k^{\infty}(\mathbf{w}), k \geq 1)\big) \xrightarrow[n \to \infty]{} 0.$$

Si $\mathbf{w}$ n'est pas recouvrant, la loi de Poisson composée se réduit à une loi de Poisson de paramètre $\mathbb{E}N(\mathbf{w})$. On définit donc dans ce cas la loi $\mathcal{CP}_{\text{bic}}$ comme la loi de Poisson de paramètre $\mathbb{E}N_{\text{bic}}(\mathbf{w})$ (l'expression (3.18), page 48, permet de calculer ce paramètre facilement). L'approximation obtenue a alors une erreur en variation totale qui tend vers 0 (sous les conditions ci-dessus) et ce quelque soit le nombre du ruptures ρ. On note que dans ce cas l'approximation est de type "mot bicolore", car basée sur le comptage attendu des mots bicolores à au plus une rupture d'état.

Dans le cas où $\mathbf{w}$ est recouvrant *i.e.* $\mathcal{P}'(\mathbf{w}) \neq \emptyset$, on définit la loi $\mathcal{CP}_{\text{bic}}$ comme la loi de Poisson composée de paramètres : $\forall k \geq 1$,

$$\lambda_{k,\text{bic}} := \lambda_{k,\text{bic}}^{(1)} + \lambda_{k,\text{bic}}^{(2)} \tag{3.21}$$

où $\lambda^{(1)}_{k,\mathrm{bic}}$ et $\lambda^{(2)}_{k,\mathrm{bic}}$ sont obtenus respectivement avec les formules :

$$\lambda^{(1)}_{k,\mathrm{bic}} = \sum_{s\in\mathcal{S}}(1-a_s)^2\mu_s(\mathbf{w})\sum_{m_1+\cdots+m_r=k-1} n_{\mathbf{s}}(s^{3h+\sum_{\ell=1}^r m_\ell p_\ell})(k-1)!\prod_{\ell=1}^r \frac{[\pi_s(\mathbf{w}^{(p_\ell+1)})]^{m_\ell}}{(m_\ell)!}, \tag{3.22}$$

$$\begin{aligned}\lambda^{(2)}_{k,\mathrm{bic}} = & \sum_{s\neq t} n_{\mathbf{s}}(st)\Bigg\{\Bigg[h\mu_t(w_1) - \sum_{\ell=1}^{h}\sum_{p\in\mathcal{P}'(\mathbf{w})}\mu_{(s^\ell t^{h+1-\ell})_{(p+1)}}(\mathbf{w}^{(p+1)})\Bigg]a_t^{k-1}\pi_t(\mathbf{w})(1-a_t)\\ & +(1-a_s)(1-a_t)\mu_s(w_1)\pi_t(\mathbf{w})\Bigg[\sum_{k'=1}^{k-1}a_s^{k'-1}a_t^{k-1-k'}\Bigg]\Bigg[\sum_{p\in\mathcal{P}'(\mathbf{w})}\sum_{\ell=1}^{p}\pi_{s^\ell t^{p+1-\ell}}(\mathbf{w}^{(p+1)})\Bigg]\\ & +(1-a_s)(1-a_t)a_s^{k-1}\mu_s(w_1)\sum_{\ell=1}^{h-1}\pi_{s^\ell t^{h-\ell}}(\mathbf{w})\\ & +(1-a_s)a_s^{k-1}\mu_s(\mathbf{w})\Bigg[h-\sum_{\ell=1}^{h}\sum_{p\in\mathcal{P}'(\mathbf{w})}\pi_{(s^\ell t^{h+1-\ell})^{(p+1)}}(\mathbf{w}_{(p+1)})\Bigg]\Bigg\},\end{aligned} \tag{3.23}$$

où on a noté $\mathcal{P}'(\mathbf{w}) = \{p_1,\ldots,p_r\}$ (avec $r\geq 1$) l'ensemble des périodes principales de $\mathbf{w}$, chaque a_s est donné par la formule (3.5) (page 44), chaque $\pi_{\mathbf{t}}(\mathbf{w})$, $\pi_s(\mathbf{w})$, $\mu_{\mathbf{t}}(\mathbf{w})$ ou $\mu_s(\mathbf{w})$ est donné dans la section 3.4.1 (page 47), s^ℓ désigne la suite formée de ℓ fois l'état s et $(s^\ell t^{h+1-\ell})_{(p+1)}$ (resp. $(s^\ell t^{h+1-\ell})^{(p+1)}$) désigne les $p+1$ derniers (resp. premiers) états du coloriage $s^\ell t^{h+1-\ell}$.

Le terme $\lambda^{(1)}_{k,\mathrm{bic}}$ correspond au comptage attendu des k-trains unicolores et le terme $\lambda^{(2)}_{k,\mathrm{bic}}$ correspond au comptage attendu des k-trains bicolores à une rupture d'état. Ainsi, l'approximation par $\mathcal{CP}_{\mathrm{bic}}$ est de type "train bicolore" lorsque $\mathbf{w}$ est recouvrant. Précisément, la proposition 3.15 (cf. page 52) montre que pour $k\leq \frac{L_{\min}-3h}{\max(\mathcal{P}'(\mathbf{w}))}+1$, le paramètre $\lambda_{k,\mathrm{bic}}$ correspond au nombre attendu de k-trains dans $\mathbf{X}$ commençant à une position de $\{1,\ldots,n-h+1\}$. Ainsi, en utilisant le lemme 3.18 (cf. page 54) et sous les conditions du début de la section, l'approximation $\mathcal{CP}_{\mathrm{bic}}$ a une erreur en variation totale qui tend vers 0 dès que $\frac{L_{\min}-3h}{\max(\mathcal{P}'(\mathbf{w}))}$ tend vers l'infini, pour un nombre de ruptures ρ quelconque.

Nous résumons les propos précédents dans les cas non-recouvrant et recouvrant avec le théorème suivant.

Théorème 3.13 *Dans un modèle PSM1 et sous les hypothèses (3.14) et (3.15), lorsque le mot* $\mathbf{w}$ *est rare i.e. vérifie* $\mathbb{E}N(\mathbf{w}) = O(1)$, $h = o(n)$, *on a :*

$$d_{vt}\big(\mathcal{L}(N(\mathbf{w})),\mathcal{CP}_{\mathrm{bic}}\big)\xrightarrow[n\to\infty]{}0,$$

dans les deux cas suivants :
- $\mathbf{w}$ *est non-recouvrant et* $L_{\min}\geq h$, *avec* $\mathcal{CP}_{\mathrm{bic}} = \mathcal{P}\big(\mathbb{E}N_{\mathrm{bic}}(\mathbf{w})\big)$ *(*$\mathbb{E}N_{\mathrm{bic}}(\mathbf{w})$ *pouvant se calculer avec l'expression (3.18) page 48)*
- $\mathbf{w}$ *est recouvrant et* $\frac{L_{\min}-3h}{\max(\mathcal{P}'(\mathbf{w}))}\to\infty$ *lorsque* $n\to\infty$, *avec* $\mathcal{CP}_{\mathrm{bic}}$ *la loi de Poisson composée de paramètres donnés en (3.21)*

Remarque 3.14 *La formule (3.22) contient une somme avec un nombre de termes de l'ordre de* k^r *(raisonnable en pratique car le nombre de périodes principales* r *est souvent "petit"). Dans* *le cas où* $\mathbf{w}$ *a une unique période principale* ($r = 1$)*, on a* $\mathcal{P}'(\mathbf{w}) = \{p_1\}$ *et la formule (3.22) se simplifie :*

$$\lambda^{(1)}_{k,\mathrm{bic}} = \sum_{s\in\mathcal{S}} (1-a_s)^2 \mu_s(\mathbf{w}) n_{\mathbf{s}}(s^{3h+(k-1)p_1})[\pi_s(\mathbf{w}^{(p_1+1)})]^{k-1}.$$

La formule pour $\lambda^{(2)}_{k,\mathrm{bic}}$ *a une expression compliquée mais son calcul est "direct".*

Par conséquent, dans le cas où $\mathbf{w}$ est recouvrant, pour que l'approximation par $\mathcal{CP}_{\mathrm{bic}}$ soit valide, on n'impose aucune condition sur ρ mais en contrepartie $L_{\min}$ doit être suffisamment grand. Cela permet de traiter des séquences plus segmentées que l'approximation par $\mathcal{CP}_{\mathrm{uni}}$ mais ne permet pas de traiter des séquences "trop segmentées" (par exemple, avec une rupture toutes les $2h$ positions). Cela sera mis en évidence sur des simulations (cf. Section 6.2).

Proposition 3.15 *Supposons que* $\mathbf{w}$ *est recouvrant i.e.* $\mathcal{P}'(\mathbf{w}) \neq \emptyset$*, et notons* $\mathcal{P}'(\mathbf{w}) = \{p_1, \ldots, p_r\}$ *(avec* $r \geq 1$*), l'ensemble des périodes principales de* $\mathbf{w}$*. Dans le modèle PSM1, pour tout* $k \geq 1$ *vérifiant* $3h + (k-1)\max(\mathcal{P}'(\mathbf{w})) \leq L_{\min}$*, on a :*

$$\mathbb{E}\widetilde{N}^{\infty}_k(\mathbf{w}) = \lambda^{(1)}_{k,\mathrm{bic}} + \lambda^{(2)}_{k,\mathrm{bic}}, \tag{3.24}$$

où $\lambda^{(1)}_{k,\mathrm{bic}}$ *et* $\lambda^{(2)}_{k,\mathrm{bic}}$ *sont obtenues par les formules (3.22) et (3.23).*

Remarque 3.16 (Expression directe de $\mathbb{E}\widetilde{N}^{\infty}(\mathbf{w})$) *On peut montrer de manière similaire à la preuve de la proposition 3.15, que si* $L_{\min} \geq 2h$*, le nombre attendu de trains dans* $\mathbf{X}$ *s'écrit*

$$\begin{aligned}\mathbb{E}\widetilde{N}^{\infty}(\mathbf{w}) = & \sum_{s\in\mathcal{S}} n_{\mathbf{s}}(s)(1-a_s)\mu_s(\mathbf{w}) + \sum_{s\neq t} n_{\mathbf{s}}(st)\Bigg\{\Bigg[h\mu_t(w_1) - \\ & \sum_{\ell=1}^{h}\sum_{p\in\mathcal{P}'(\mathbf{w})} \mu_{(s^\ell t^{h+1-\ell})_{(p+1)}}(\mathbf{w}^{(p+1)})\Bigg]\pi_t(\mathbf{w}) + (1-a_s)\mu_s(w_1)\sum_{\ell=1}^{h-1}\pi_{s^\ell t^{h-\ell}}(\mathbf{w})\Bigg\}.\end{aligned}$$

Preuve de la proposition 3.15. Comme un motif $\mathbf{bcf}$ de $\mathcal{C}'_k$ (défini dans la section 2.1.3 page 33) est de longueur au plus $3h + (k-1)\max(\mathcal{P}'(\mathbf{w}))$ et que par hypothèse $L_{\min} \geq 3h + (k-1)\max(\mathcal{P}'(\mathbf{w}))$, toutes les occurrences des motifs $\mathbf{bcf}$ de $\mathcal{C}'_k$ sont bicolores à au plus une rupture d'état. Ainsi, on peut utiliser la décomposition (3.18) (page 48) pour tous les motifs $\mathbf{bcf} \in \mathcal{C}'_k$:

$$\mathbb{E}\widetilde{N}^{\infty}_k(\mathbf{w}) = \underbrace{\sum_{s\in\mathcal{S}}\sum_{\mathbf{bcf}\in\mathcal{C}'_k} n_{\mathbf{s}}(s^{2h+|\mathbf{c}|})\mu_s(\mathbf{bcf})}_{=:T_1} + \underbrace{\sum_{s\neq t} n_{\mathbf{s}}(st)\sum_{\mathbf{bcf}\in\mathcal{C}'_k}\mu_{s,t}(\mathbf{bcf})}_{=:T_2}$$

Il suffit donc de montrer que $T_1 = \lambda^{(1)}_{k,\mathrm{bic}}$ et $T_2 = \lambda^{(2)}_{k,\mathrm{bic}}$. Le terme T_1 se traite simplement en utilisant le fait que la longueur d'un k-train dans une chaîne de Markov stationnaire de probabilité de transition π_s suit la loi de $\sum_{p\in\mathcal{P}'(\mathbf{w})} pM_p + h$, où $(M_p, p \in \mathcal{P}'(\mathbf{w}))$ suit une loi multinomiale de paramètres $k-1$, $\left(\frac{\pi_s(\mathbf{w}^{(p+1)})}{a_s}, p \in \mathcal{P}'(\mathbf{w})\right)$ (cf. le lemme 2.9 page 36 et notamment la relation (2.10)). La démarche pour le terme T_2 est assez technique, et le lecteur pourra éventuellement

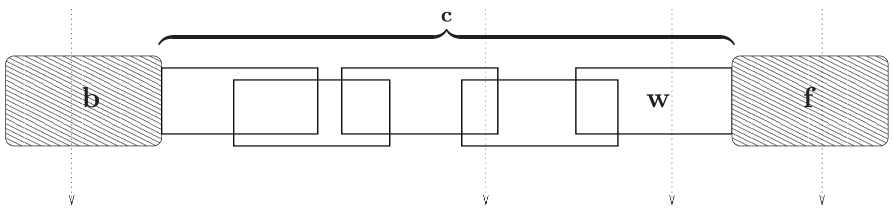


FIG. 3.4 – Les quatre cas de rupture dans un motif de $\mathcal{C}'_k$

s'aider de l'exemple 3.17. Comme l'illustre la figure 3.4, il s'agit de découper chaque motif **bcf** de $\mathcal{C}'_k$ bicolore à exactement une rupture d'état selon l'endroit où la rupture a lieu (dans **b**, dans les chevauchements successifs de **c**, dans la dernière occurrence de **w** de **c**, ou dans **f**).

Par suite,

$$\begin{aligned}
\sum_{\mathbf{bcf}\in\mathcal{C}'_k} \mu_{s,t}(\mathbf{bcf}) &= \sum_{\mathbf{bcf}\in\mathcal{C}'_k} \sum_{\ell=1}^{|\mathbf{bcf}|-1} \mu_{s^\ell t^{|\mathbf{bcf}|-\ell}}(\mathbf{bcf}) \\
&= \underbrace{\sum_{\mathbf{bcf}\in\mathcal{C}'_k} \sum_{\ell=1}^{h} \mu_{s^\ell t^{h+1-\ell}}(\mathbf{b}w_1)\pi_t(\mathbf{cf})}_{=:T_{21}} + \underbrace{\sum_{\mathbf{bcf}\in\mathcal{C}'_k} \sum_{\ell=1}^{|\mathbf{c}|-h} \mu_s(\mathbf{b}w_1)\pi_{s^\ell t^{|\mathbf{c}|-\ell}}(\mathbf{c})\pi_t(w_h\mathbf{f})}_{=:T_{22}} \\
&\quad + \underbrace{\sum_{\mathbf{bcf}\in\mathcal{C}'_k} \sum_{\ell=|\mathbf{c}|-h+1}^{|\mathbf{c}|-1} \mu_s(\mathbf{b}w_1)\pi_{s^\ell t^{|\mathbf{c}|-\ell}}(\mathbf{c})\pi_t(w_h\mathbf{f})}_{=:T_{23}} + \underbrace{\sum_{\mathbf{bcf}\in\mathcal{C}'_k} \sum_{\ell=1}^{h} \mu_s(\mathbf{bc})\pi_{s^\ell t^{h+1-\ell}}(w_h\mathbf{f})}_{=:T_{24}} .
\end{aligned}$$

D'après le cas homogène (cf. Section 2.2.2 page 35), pour un état $u \in \mathcal{S}$ quelconque, on a $\sum_{\mathbf{b}\in\mathcal{B}} \mu_u(\mathbf{b}w_1) = \mu_u(w_1)(1-a_u)$, $\sum_{\mathbf{f}\in\mathcal{F}} \pi_u(w_h\mathbf{f}) = 1-a_u$ et $\sum_{\mathbf{c}\in\mathcal{C}_k} \pi_u(\mathbf{c}) = a_u^{k-1}\pi_u(\mathbf{w})$. Ainsi, on obtient :

$$\begin{aligned}
T_{21} &= \left[\sum_{\ell=1}^{h} \sum_{\mathbf{b}\in\mathcal{B}} \mu_{s^\ell t^{h+1-\ell}}(\mathbf{b}w_1)\right] a_t^{k-1}\pi_t(\mathbf{w})(1-a_t) \\
T_{22} &= (1-a_s)\mu_s(w_1)(1-a_t) \sum_{\mathbf{c}\in\mathcal{C}_k} \sum_{\ell=1}^{|\mathbf{c}|-h} \pi_{s^\ell t^{|\mathbf{c}|-\ell}}(\mathbf{c}) \\
T_{23} &= (1-a_s)\mu_s(w_1)(1-a_t)a_s^{k-1} \sum_{\ell=1}^{h-1} \pi_{s^\ell t^{h-\ell}}(\mathbf{w}) \\
T_{24} &= (1-a_s)a_s^{k-1}\mu_s(\mathbf{w}) \left[\sum_{\ell=1}^{h} \sum_{\mathbf{f}\in\mathcal{F}} \pi_{s^\ell t^{h+1-\ell}}(w_h\mathbf{f})\right].
\end{aligned}$$

Le terme T_{23} a bien la forme voulue. En passant à l'espérance dans les relations (2.4) et (2.5) (page 34), on obtient :

$$\begin{aligned}\sum_{\mathbf{b}\in\mathcal{B}}\mu_{s^\ell t^{h+1-\ell}}(\mathbf{b}w_1) &= \mu_t(w_1)-\sum_{p\in\mathcal{P}'(\mathbf{w})}\mu_{(s^\ell t^{h+1-\ell})_{(p+1)}}(\mathbf{w}^{(p+1)})\\ \sum_{\mathbf{f}\in\mathcal{F}}\pi_{s^\ell t^{h-\ell+1}}(w_h\mathbf{f}) &= 1-\sum_{p\in\mathcal{P}'(\mathbf{w})}\pi_{(s^\ell t^{h-\ell+1})^{(p+1)}}(\mathbf{w}_{(p+1)}),\end{aligned}$$

ce qui permet de conclure pour T_{21} et T_{24}. Pour le terme T_{22}, on utilise la structure d'un motif de $\mathcal{C}_k$:

$$\begin{aligned}\sum_{\mathbf{c}\in\mathcal{C}_k}\sum_{\ell=1}^{|\mathbf{c}|-h}\pi_{s^\ell t^{|\mathbf{c}|-\ell}}(\mathbf{c}) &= \sum_{p_1,\ldots,p_{k-1}\in\mathcal{P}'(\mathbf{w})}\sum_{\ell=1}^{p_1+\cdots+p_{k-1}}\pi_{s^\ell t^{p_1+\cdots+p_{k-1}+h-\ell}}(\mathbf{w}^{(p_1)}\ldots\mathbf{w}^{(p_{k-1})}\mathbf{w})\\ &= \pi_t(\mathbf{w})\sum_{k'=1}^{k-1}a_s^{k'-1}\Bigg[\sum_{p\in\mathcal{P}'(\mathbf{w})}\sum_{\ell=1}^{p}\pi_{s^\ell t^{p-\ell+1}}(\mathbf{w}^{(p+1)})\Bigg]a_t^{k-1-k'}\\ &= \pi_t(\mathbf{w})\Bigg[\sum_{k'=1}^{k-1}a_s^{k'-1}a_t^{k-1-k'}\Bigg]\Bigg[\sum_{p\in\mathcal{P}'(\mathbf{w})}\sum_{\ell=1}^{p}\pi_{s^\ell t^{p-\ell+1}}(\mathbf{w}^{(p+1)})\Bigg].\end{aligned}$$

■

Exemple 3.17 *Pour bien comprendre comment se calcule l'expression* $\sum_{\mathbf{bcf}\in\mathcal{C}'_k}\mu_{s,t}(\mathbf{bcf})$ *dans le terme* T_2 *de la preuve ci-dessus, nous proposons de traiter un exemple. Pour* $\mathbf{w}=\mathtt{agag}$ *et* $k=3$*, on a* $h=4$*,* $\mathcal{P}'(\mathbf{w})=\{2\}$*,* $\mathbf{w}^{(3)}=\mathtt{aga}$*,* $\mathbf{w}_{(3)}=\mathtt{gag}$*,* $\forall s\in\mathcal{S}, a_s=\pi_s(\mathtt{aga})=\pi_s(\mathtt{gag})$*. De plus, les éléments de* $\mathcal{C}'_3$ *sont de la forme* $\mathbf{b}\mathtt{agagagag}\mathbf{f}$ *avec* $\mathbf{b}$ *dans* $\mathcal{B}=\{\mathbf{b}=b_1b_2b_3b_4\,|\,b_3b_4\neq\mathtt{ag}\}$ *et* $\mathbf{f}$ *dans* $\mathcal{F}=\{\mathbf{f}=f_1f_2f_3f_4\,|\,f_1f_2\neq\mathtt{ag}\}$*. Les coloriages bicolores à exactement une rupture possible commençant par l'état s et finissant par l'état t sont les coloriages* $s^\ell t^{16-\ell}$*, pour* $\ell=1,\ldots,15$*. La figure 3.5 donne pour chacun de ces coloriages la probabilité d'occurrence d'un élément de* $\mathcal{C}'_3$*. En sommant sur toutes les lignes de la figure 3.5 (et en regroupant selon les quatre paquets de lignes séparés par un trait), on obtient :*

$$\begin{aligned}\sum_{\mathbf{bcf}\in\mathcal{C}'_3}\mu_{s,t}(\mathbf{bcf}) &= (4\mu_t(\mathtt{a})-2\mu_t(\mathtt{aga})-\mu_{stt}(\mathtt{aga})-\mu_{sst}(\mathtt{aga}))a_t^2\pi_t(\mathbf{w})(1-a_t)\\ &\quad+(1-a_s)\mu_s(\mathtt{a})(a_s+a_t)(\pi_{stt}(\mathtt{aga})+\pi_{sst}(\mathtt{aga}))\pi_t(\mathbf{w})(1-a_t)\\ &\quad+(1-a_s)\mu_s(\mathtt{a})a_s^2(\pi_{sttt}(\mathbf{w})+\pi_{sstt}(\mathbf{w})+\pi_{ssst}(\mathbf{w}))(1-a_t)\\ &\quad+(1-a_s)a_s^2\mu_s(\mathbf{w})(4-\pi_{stt}(\mathtt{gag})-\pi_{sst}(\mathtt{gag})-2a_s),\end{aligned}$$

et on retrouve l'expression dans l'accolade de la formule (3.23).

Lemme 3.18 *Sous les hypothèses et les notations du théorème 3.13, lorsque* $\mathbf{w}$ *est recouvrant*

$$d_{vt}\big(\mathcal{CP}(\mathbb{E}\widetilde{N}_k^{\infty}(\mathbf{w}),k\geq 1),\mathcal{CP}_{\text{bic}}\big)\xrightarrow[n\to\infty]{}0,$$

dès que $K:=\frac{L_{\min}-3h}{\max(\mathcal{P}'(\mathbf{w}))}\to\infty$*.*

La preuve de ce lemme est effectuée à la fin de la section suivante.

b	a g a g a g a g	f	Probabilités d'occurrence correspondantes		
			$\mathbb{P}(\mathbf{ba})$	$\mathbb{P}(\mathtt{agagagag}\mid\mathtt{a})$	$\mathbb{P}(\mathbf{f}\mid\mathtt{g})$
s t t t t	t t t t t t t t	t t t t	$\mu_t(\mathtt{a}) - \mu_t(\mathtt{aga})$	$\pi_t(\mathtt{aga})^2\pi_t(\mathtt{agag})$	$1 - \pi_t(\mathtt{gag})$
s s t t t	t t t t t t t t	t t t t	$\mu_t(\mathtt{a}) - \mu_t(\mathtt{aga})$	$\pi_t(\mathtt{aga})^2\pi_t(\mathtt{agag})$	$1 - \pi_t(\mathtt{gag})$
s s s t t	t t t t t t t t	t t t t	$\mu_t(\mathtt{a}) - \mu_{stt}(\mathtt{aga})$	$\pi_t(\mathtt{aga})^2\pi_t(\mathtt{agag})$	$1 - \pi_t(\mathtt{gag})$
s s s s t	t t t t t t t t	t t t t	$\mu_t(\mathtt{a}) - \mu_{sst}(\mathtt{aga})$	$\pi_t(\mathtt{aga})^2\pi_t(\mathtt{agag})$	$1 - \pi_t(\mathtt{gag})$
s s s s s	t t t t t t t t	t t t t	$(1 - \pi_s(\mathtt{aga}))\mu_s(\mathtt{a})$	$\pi_{stt}(\mathtt{aga})\pi_t(\mathtt{aga})\pi_t(\mathtt{agag})$	$1 - \pi_t(\mathtt{gag})$
s s s s s	s t t t t t t t	t t t t	$(1 - \pi_s(\mathtt{aga}))\mu_s(\mathtt{a})$	$\pi_{sst}(\mathtt{aga})\pi_t(\mathtt{aga})\pi_t(\mathtt{agag})$	$1 - \pi_t(\mathtt{gag})$
s s s s s	s s t t t t t t	t t t t	$(1 - \pi_s(\mathtt{aga}))\mu_s(\mathtt{a})$	$\pi_s(\mathtt{aga})\pi_{stt}(\mathtt{aga})\pi_t(\mathtt{agag})$	$1 - \pi_t(\mathtt{gag})$
s s s s s	s s s t t t t t	t t t t	$(1 - \pi_s(\mathtt{aga}))\mu_s(\mathtt{a})$	$\pi_s(\mathtt{aga})\pi_{sst}(\mathtt{aga})\pi_t(\mathtt{agag})$	$1 - \pi_t(\mathtt{gag})$
s s s s s	s s s s t t t t	t t t t	$(1 - \pi_s(\mathtt{aga}))\mu_s(\mathtt{a})$	$\pi_s(\mathtt{aga})^2\pi_{sttt}(\mathtt{agag})$	$1 - \pi_t(\mathtt{gag})$
s s s s s	s s s s s s t t	t t t t	$(1 - \pi_s(\mathtt{aga}))\mu_s(\mathtt{a})$	$\pi_s(\mathtt{aga})^2\pi_{sstt}(\mathtt{agag})$	$1 - \pi_t(\mathtt{gag})$
s s s s s	s s s s s s s t	t t t t	$(1 - \pi_s(\mathtt{aga}))\mu_s(\mathtt{a})$	$\pi_s(\mathtt{aga})^2\pi_{ssst}(\mathtt{agag})$	$1 - \pi_t(\mathtt{gag})$
s s s s s	s s s s s s s s	t t t t	$(1 - \pi_s(\mathtt{aga}))\mu_s(\mathtt{a})$	$\pi_s(\mathtt{aga})^2\pi_s(\mathtt{agag})$	$1 - \pi_{stt}(\mathtt{gag})$
s s s s s	s s s s s s s s	s t t t	$(1 - \pi_s(\mathtt{aga}))\mu_s(\mathtt{a})$	$\pi_s(\mathtt{aga})^2\pi_s(\mathtt{agag})$	$1 - \pi_{sst}(\mathtt{gag})$
s s s s s	s s s s s s s s	s s t t	$(1 - \pi_s(\mathtt{aga}))\mu_s(\mathtt{a})$	$\pi_s(\mathtt{aga})^2\pi_s(\mathtt{agag})$	$1 - \pi_s(\mathtt{gag})$
s s s s s	s s s s s s s s	s s s t	$(1 - \pi_s(\mathtt{aga}))\mu_s(\mathtt{a})$	$\pi_s(\mathtt{aga})^2\pi_s(\mathtt{agag})$	$1 - \pi_s(\mathtt{gag})$

FIG. 3.5 – Calcul de la probabilité d'occurrence d'un élément de $\mathcal{C}'_3$ selon les 15 coloriages possibles à une rupture commençant par l'état s et finissant par l'état t pour le mot $\mathbf{w} = \mathtt{agag}$

3.5 Preuve du théorème 3.7 et lemmes annexes

Pour borner les termes b_1, b_2 et b_3, on utilise une preuve proche de celle proposée par Schbath (1995a) ; les principales différences sont les suivantes :

1. Pour borner b_1 et b_2, nous n'utilisons pas explicitement la décomposition d'un k-train en $\mathbf{bcf} \in \mathcal{C}'_k$.
2. Le terme b_3 se traite de manière différente puisque la chaîne n'est plus homogène (cf. Lemme 3.19).
3. Nous utilisons la relation
$$\sum_{i=1}^{n-h+1} \sum_{k\geq 1} k\widetilde{\mu}_{i,k} = \mathbb{E}N^{\infty}(\mathbf{w}), \tag{3.25}$$
ce qui nous amène à comparer la différence en espérance entre $N(\mathbf{w})$ et $N^{\infty}(\mathbf{w})$ (cf. Lemme 3.21).

Choix de $B_{i,k}$

On définit pour chaque $(i,k) \in I \times \mathbb{N}^{\star}$, une zone $Z(i,k) \subset \mathbb{Z}$, ensemble qui contient au moins les indices j des lettres X_j impliquées dans la définition de $\widetilde{Y}^{\infty}_{i,k}$. Comme un motif de taille k est de longueur au plus kh et qu'il est nécessaire de connaître les $h-1$ lettres avant i et après le motif pour s'assurer qu'il s'agit d'un train de taille k, on peut prendre $Z(i,k) = \{s \in \mathbb{Z} \text{ tel que } i-h \leq s \leq i+(k+1)h\}$.

Pour pouvoir correctement majorer le terme b_3, on définit (i,k) comme non voisin de (j,ℓ) lorsque les zones associées $Z(i,k)$ et $Z(j,\ell)$ sont distantes d'au moins $2h$ positions, c'est-à-dire soit lorsque $i+(k+3)h < j-h$ soit lorsque $j+(\ell+3)h < i-h$.

Pour finir, le voisinage $B_{i,k}$ de chaque (i,k) est défini comme l'ensemble des (j,ℓ) voisins de (i,k) qui sont dans $\{1,\ldots,n-h+1\}\times\mathbb{N}^\star$:

$$B_{i,k} = \big\{(j,\ell)\in\{1,\ldots,n-h+1\}\times\mathbb{N}^\star \text{ tel que } -(\ell+4)h \le j-i \le (k+4)h\big\}. \tag{3.26}$$

Majoration de b_1

D'après la définition (3.8) (page 45), on a :

$$\begin{aligned} b_1 &= \sum_{i=1}^{n-h+1}\sum_{k\ge 1}\sum_{(j,\ell)\in B_{i,k}} \mathbb{E}(\widetilde{Y}_{i,k}(\mathbf{w}))\mathbb{E}(\widetilde{Y}_{j,\ell}(\mathbf{w})) \\ &\le \sum_{i=1}^{n-h+1}\sum_{k\ge 1}\sum_{\ell\ge 1}\sum_{j=i-(\ell+4)h}^{i+(k+4)h} \widetilde{\mu}_{i,k}(\mathbf{w})\widetilde{\mu}_{i,\ell}(\mathbf{w}). \end{aligned}$$

Si $\widetilde{\mu}_i(\mathbf{w})$ désigne la probabilité d'occurrence d'un train de $\mathbf{w}$ à une position i, on a $\widetilde{\mu}_i(\mathbf{w}) = \sum_{k\ge 1}\widetilde{\mu}_{i,k}(\mathbf{w}) \le \pi_{\max}(\mathbf{w})$. Par symétrie en i et j et en utilisant l'équation (3.25), on déduit :

$$\begin{aligned} b_1 &\le 2\pi_{\max}(\mathbf{w})\sum_{i=1}^{n-h+1}\sum_{k\ge 1}((k+4)h+1)\widetilde{\mu}_{i,k}(\mathbf{w}) \\ &\le 12h\pi_{\max}(\mathbf{w})\sum_{i=1}^{n-h+1}\sum_{k\ge 1} k\widetilde{\mu}_{i,k}(\mathbf{w}) \\ &\le 12\,\mathbb{E}(N^\infty(\mathbf{w}))h\pi_{\max}(\mathbf{w}). \end{aligned} \tag{3.27}$$

Majoration de b_2

D'après la définition (3.9) (page 45), on a :

$$b_2 = \sum_{i=1}^{n-h+1}\sum_{k\ge 1}\sum_{(j,\ell)\in B_{i,k}\setminus\{(i,k)\}} \mathbb{E}(\widetilde{Y}_{i,k}(\mathbf{w})\widetilde{Y}_{j,\ell}(\mathbf{w})).$$

Comme on ne peut pas avoir l'occurrence de deux trains de tailles différentes à la même position, le terme correspondant à $i=j$ disparaît dans la somme, et de nouveau par symétrie on obtient :

$$b_2 \le 2\sum_{i=1}^{n-h+1}\sum_{k\ge 1}\sum_{\ell\ge 1}\sum_{j=i+1}^{i+(k+4)h} \mathbb{E}(\widetilde{Y}_{i,k}(\mathbf{w})\widetilde{Y}_{j,\ell}(\mathbf{w})).$$

Notons $\widetilde{Y}_j(\mathbf{w}) = \sum_{\ell\ge 1}\widetilde{Y}_{j,\ell}(\mathbf{w})$ la variable de Bernouilli qui vaut 1 s'il y a occurrence d'un train de $\mathbf{w}$ à la position j et 0 sinon. Comme $\widetilde{Y}_{i,k}(\mathbf{w}) = \sum_{\mathbf{c}\in\mathcal{C}_k(\mathbf{w})}\widetilde{Y}_{i,k}(\mathbf{w})Y_i(\mathbf{c})$, on a :

$$\begin{aligned} b_2 &\le 2\sum_{i=1}^{n-h+1}\sum_{k\ge 1}\sum_{j=i+1}^{i+(k+4)h} \mathbb{E}(\widetilde{Y}_{i,k}(\mathbf{w})\widetilde{Y}_j(\mathbf{w})) \\ &\le 2\sum_{i=1}^{n-h+1}\sum_{k\ge 1}\sum_{\mathbf{c}\in\mathcal{C}_k(\mathbf{w})}\sum_{j=i+1}^{i+(k+4)h} \mathbb{E}\big(\widetilde{Y}_{i,k}(\mathbf{w})Y_i(\mathbf{c})\widetilde{Y}_j(\mathbf{w})\big). \end{aligned}$$

Comme un train de longueur $|\mathbf{c}|$ qui commence en position i ne peut pas recouvrir un train commençant en position $i+1 \leq j < i+|\mathbf{c}|$, et comme $\widetilde{Y}_j(\mathbf{w}) \leq Y_j(\mathbf{w})$, on déduit que

$$\begin{aligned}
b_2 &\leq 2\sum_{i=1}^{n-h+1}\sum_{k\geq 1}\sum_{\mathbf{c}\in\mathcal{C}_k(\mathbf{w})}\sum_{j=i+|\mathbf{c}|}^{i+(k+4)h}\mathbb{E}\big(\widetilde{Y}_{i,k}(\mathbf{w})Y_i(\mathbf{c})\widetilde{Y}_j(\mathbf{w})\big)\\
&\leq 2\sum_{i=1}^{n-h+1}\sum_{k\geq 1}\sum_{\mathbf{c}\in\mathcal{C}_k(\mathbf{w})}\sum_{j=i+|\mathbf{c}|+h}^{i+(k+4)h}\mathbb{E}\big(\widetilde{Y}_{i,k}(\mathbf{w})Y_i(\mathbf{c})Y_j(\mathbf{w})\big)\\
&\quad +2\sum_{i=1}^{n-h+1}\sum_{k\geq 1}\sum_{\mathbf{c}\in\mathcal{C}_k(\mathbf{w})}\sum_{j=i+|\mathbf{c}|}^{i+|\mathbf{c}|+h-1}\mathbb{E}\big(\widetilde{Y}_{i,k}(\mathbf{w})Y_i(\mathbf{c})Y_j(\mathbf{w})\big).
\end{aligned}$$

Le premier terme (resp. le second terme) du membre de droite est noté b_{21} (resp. b_{22}). Pour majorer b_{21}, on note que la variable aléatoire $\widetilde{Y}_{i,k}(\mathbf{w})Y_i(\mathbf{c})$ dépend seulement des lettres $X_{i-h+1}\dots X_{i+|\mathbf{c}|+h-1}$ alors que $Y_j(\mathbf{w})$ dépend de $X_j\dots X_{j+h-1}$. Donc pour toute position j qui vérifie $j \geq i+|\mathbf{c}|+h$, la propriété de Markov donne :

$$\mathbb{E}\big(\widetilde{Y}_{i,k}(\mathbf{w})Y_i(\mathbf{c})Y_j(\mathbf{w})\big) \leq \pi_{\max}(\mathbf{w})\mathbb{E}\big(\widetilde{Y}_{i,k}(\mathbf{w})Y_i(\mathbf{c})\big).$$

Comme la somme sur j contient moins de $(k+3)h$ termes, on a :

$$\begin{aligned}
b_{21} &\leq 2\pi_{\max}(\mathbf{w})\sum_{i=1}^{n-h+1}\sum_{k\geq 1}(k+3)h\widetilde{\mu}_{i,k}(\mathbf{w})\\
&\leq 8\,\mathbb{E}(N^{\infty}(\mathbf{w}))h\pi_{\max}(\mathbf{w}). \qquad (3.28)
\end{aligned}$$

Pour borner b_{22}, on écrit $\mathbb{E}\big(\widetilde{Y}_{i,k}(\mathbf{w})Y_i(\mathbf{c})Y_j(\mathbf{w})\big) \leq \mathbb{E}\big(\widetilde{Y}_i(\mathbf{w})Y_i(\mathbf{c})Y_j(\mathbf{w})\big)$ et on remarque que la variable aléatoire $\widetilde{Y}_i(\mathbf{w})Y_i(\mathbf{c})$ dépend seulement des lettres $X_{i-h+1}\dots X_{i+|\mathbf{c}|-1}$ alors que $Y_j(\mathbf{w})$ dépend de $X_j\dots X_{j+h-1}$. Par conséquent, pour toute position j telle que $j \geq i+|\mathbf{c}|$, on a :

$$\mathbb{E}\big(\widetilde{Y}_i(\mathbf{w})Y_i(\mathbf{c})Y_j(\mathbf{w})\big) \leq \pi_{\max}(\mathbf{w})\mathbb{E}\big(\widetilde{Y}_i(\mathbf{w})Y_i(\mathbf{c})\big).$$

Ainsi, on obtient la majoration suivante de b_{22} :

$$\begin{aligned}
b_{22} &\leq 2h\pi_{\max}(\mathbf{w})\sum_{i=1}^{n-h+1}\sum_{k\geq 1}\sum_{\mathbf{c}\in\mathcal{C}_k(\mathbf{w})}\mathbb{E}\big(\widetilde{Y}_{i,k}(\mathbf{w})Y_i(\mathbf{c})\big)\\
&\leq 2\,\mathbb{E}(N^{\infty}(\mathbf{w}))h\pi_{\max}(\mathbf{w}). \qquad (3.29)
\end{aligned}$$

En effet,

$$\begin{aligned}
&\sum_{i=1}^{n-h+1}\sum_{k\geq 1}\sum_{\mathbf{c}\in\mathcal{C}_k(\mathbf{w})}\mathbb{E}\big(\widetilde{Y}_i(\mathbf{w})Y_i(\mathbf{c})\big)\\
&= \sum_{i=1}^{n-h+1}\sum_{k\geq 1}\mathbb{P}(\text{un } K\text{-train de } \mathbf{w} \text{ commence en position } i \text{ avec } K \geq k)\\
&= \sum_{i=1}^{n-h+1}\sum_{k\geq 1}\sum_{K\geq k}\widetilde{\mu}_{i,K}(\mathbf{w}) = \sum_{i=1}^{n-h+1}\sum_{K\geq 1}K\widetilde{\mu}_{i,K}(\mathbf{w}) = \mathbb{E}(N^{\infty}(\mathbf{w})).
\end{aligned}$$

Finalement, les inégalités (3.28) et (3.29) donnent

$$b_2 \leq 10\,\mathbb{E}(N^{\infty}(\mathbf{w}))h\pi_{\max}(\mathbf{w}). \tag{3.30}$$

Majoration de b_3

Par définition de (3.10) (page 45), on a :

$$b_3 = \sum_{i=1}^{n-h+1} \sum_{k\geq 1} \mathbb{E}\Big|\mathbb{E}\big(\widetilde{Y}_{i,k}(\mathbf{w}) - \widetilde{\mu}_{i,k}(\mathbf{w})|\sigma(\widetilde{Y}_{j,\ell}(\mathbf{w}),(j,\ell)\notin B_{i,k})\big)\Big|.$$

De plus, pour tout $\mathbf{c} \in \mathcal{C}_k$, on a par définition du voisinage $B_{i,k}$:

$$\sigma(\widetilde{Y}_{j,\ell}(\mathbf{w}),(j,\ell)\notin B_{i,k}) \subset \sigma(\ldots, X_{i-3h-1}, X_{i-3h}, X_{i+|\mathbf{c}|+3h}, X_{i+|\mathbf{c}|+3h+1},\ldots).$$

Ainsi, d'après la décomposition (2.3) et par la propriété de Markov, on a :

$$\begin{aligned} b_3 &\leq \sum_{i=1}^{n-h+1} \sum_{k\geq 1} \sum_{\mathbf{bcf}\in\mathcal{C}'_k} \mathbb{E}\Big|\mathbb{E}\big(Y_{i-h}(\mathbf{bcf}) - \mathbb{E}(Y_{i-h}(\mathbf{bcf}))|\sigma(\ldots, X_{i-3h}, X_{i+|\mathbf{c}|+3h},\ldots)\big)\Big| \\ &\leq \sum_{i=1}^{n-h+1} \sum_{k\geq 1} \sum_{\mathbf{bcf}\in\mathcal{C}'_k} \mathbb{E}\Big|\mathbb{E}\big(Y_{i-h}(\mathbf{bcf}) - \mathbb{E}(Y_{i-h}(\mathbf{bcf}))|X_{(i-h)-2h}, X_{(i-h)+|\mathbf{bcf}|+2h}\big)\Big| \end{aligned}$$

En appliquant le lemme 3.19 on trouve donc

$$b_3 \leq \sum_{i=1}^{n-h+1} \sum_{k\geq 1} \sum_{\mathbf{bcf}\in\mathcal{C}'_k} C'\mathbb{E}(Y_{i-h}(\mathbf{bcf}))|\alpha_{\max}|^h.$$

Finalement, comme on a $\sum_{i=1}^{n-h+1}\sum_{k\geq 1}\sum_{\mathbf{bcf}\in\mathcal{C}'_k}\mathbb{E}(Y_{i-h}(\mathbf{bcf})) = \sum_{i=1}^{n-h+1}\sum_{k\geq 1}\widetilde{\mu}_{i,k}(\mathbf{w}) \leq \mathbb{E}[N^{\infty}(\mathbf{w})]$, on déduit

$$b_3 \quad \leq \quad C'\mathbb{E}(N^{\infty}(\mathbf{w}))|\alpha_{\max}|^h. \tag{3.31}$$

En définitive, les inégalités (3.27), (3.30), (3.31) et le lemme 3.21 prouvent le théorème en prenant $C = 22$ et $C'' = \max\big(22/(1-\zeta), C'/(1-\zeta)\big)$.

Lemme 3.19 *Dans un PM1, pour tout mot $\mathbf{u} = u_1\cdots u_{|\mathbf{u}|}$ et tous entiers j et ℓ et lorsque $L_{\min} \geq \ell$,*

$$\mathbb{E}\big|\mathbb{E}(Y_j(\mathbf{u})|X_{j-2\ell}, X_{j+|\mathbf{u}|+2\ell}) - \mathbb{E}(Y_j(\mathbf{u}))\big| \leq C'\mathbb{E}(Y_j(\mathbf{u}))|\alpha_{\max}|^{\ell},$$

où $C' > 0$ est une constante qui dépend uniquement des $\{\pi_s\}_s$.

Preuve. En notant pour tout $i < i'$, et $a, b \in \mathcal{A}$ $\pi_{i,i'}(a,b) = [\prod_{r=i+1}^{i'}\Pi_{s_r}](a,b)$ la probabilité de passer dans la séquence $\mathbf{X}$ de a en position i à b en position i', on a pour tout $x, y \in \mathcal{A}$,

$$\mathbb{E}(Y_j(\mathbf{u})\mathbf{1}\{X_{j-2\ell} = x, X_{j+|\mathbf{u}|+2\ell} = y\}) = \mathbb{P}(X_{j-2\ell} = x)\pi_{j-2\ell,j}(x,u_1)\pi_j(\mathbf{u})\pi_{j+|\mathbf{u}|-1,j+|\mathbf{u}|+2\ell}(u_{|\mathbf{u}|},y)$$

De plus, comme $L_{\min} \geq \ell$, entre deux positions séparées de 2ℓ positions il apparaît au moins un coloriage composé de ℓ mêmes états successifs, ainsi entre les positions $j-2\ell$ et j (resp. $j+|\mathbf{u}|-1$ et $j+|\mathbf{u}|+2\ell$) on a ℓ mêmes états successifs (une illustration de la situation est donnée sur la figure 3.6). En utilisant les propriétés des matrices de transitions énoncées dans le lemme 3.22 on obtient les relations $\pi_{j-2\ell,j}(x,u_1) = \mathbb{P}(X_j = u_1) + O(|\alpha_{\max}|^\ell)$ et $\pi_{j+|\mathbf{u}|-1,j+|\mathbf{u}|+2\ell}(u_{|\mathbf{u}|},y) = \pi_{j-2\ell,j+|\mathbf{u}|+2\ell}(x,y) + O(|\alpha_{\max}|^\ell)$. Par suite,

$$\begin{aligned}\mathbb{E}(Y_j(\mathbf{u})\mathbf{1}\{X_{j-2\ell}=x, X_{j+|\mathbf{u}|+2\ell}=y\}) &= \mathbb{P}(X_{j-2\ell}=x)\mathbb{P}(X_j=u_1)\pi_j(\mathbf{u})\pi_{j-2\ell,j+|\mathbf{u}|+2\ell}(x,y)\\ &\quad +O(|\alpha_{\max}|^\ell)\\ &= \mathbb{E}(Y_j(\mathbf{u}))[\mathbb{P}(X_{j-2\ell}=x, X_{j+|\mathbf{u}|+2\ell}=y)+O(|\alpha_{\max}|^\ell)],\end{aligned}$$

et on peut conclure en intégrant sur $X_{j-2\ell}$ et $X_{j+|\mathbf{u}|+2\ell}$. ■

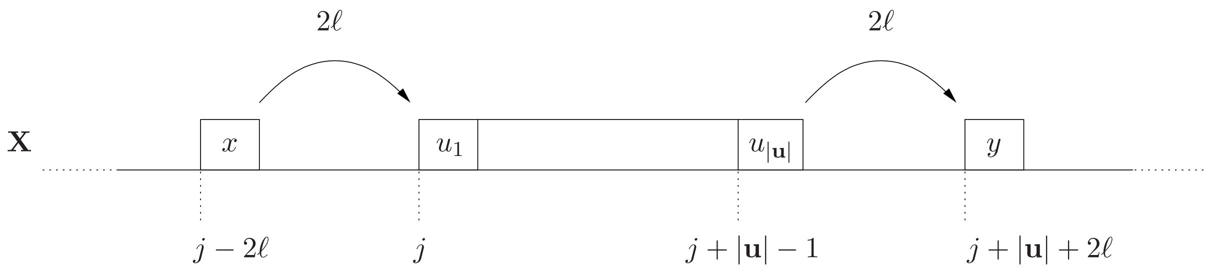


FIG. 3.6 – Illustration de la situation de la preuve du lemme 3.19.

Remarque 3.20 *Quitte à changer la constante C', le lemme 3.19 est encore valable si les segments sont soit plus grands que $\ell-1$ soit de longueur 1 et si tous les segments de longueur 1 sont espacés d'au moins $\ell-1$ positions. En effet, sous ces conditions, il y aura toujours $h-1$ parmi 2ℓ positions qui seront dans le même état.*

Lemme 3.21 *Dans un modèle PM1 et sous l'hypothèse (3.4) (cf. page 43), on a*

$$\mathbb{E}N^\infty(\mathbf{w}) \leq \mathbb{E}N(\mathbf{w}) + \frac{h\pi_{\max}(\mathbf{w})}{1-\zeta}.$$

Preuve. Les occurrences de $\mathbf{w}$ dans $\mathbf{X}$ potentiellement comptées dans $N^\infty(\mathbf{w})$ et pas dans $N(\mathbf{w})$ sont celles contenues dans le (potentiel) motif commençant à une position de $\{n-h+2,\dots,n\}$ et composé de la succession d'occurrences chevauchantes de $\mathbf{w}$ démarrant après la position $n-h+2$. Notons le comptage de ces occurrences $N^\star(\mathbf{w})$ (avec la convention $N^\star(\mathbf{w}) = 0$ s'il n'y a aucune occurrence de $\mathbf{w}$ dans $\{n-h+2,\dots,n\}$). Par suite, par définition des ensembles $\mathcal{C}_k$ et $\mathcal{F}$,

$$\begin{aligned}\mathbb{E}[N^\infty(\mathbf{w}) - N(\mathbf{w})] &\leq \mathbb{E}N^\star(\mathbf{w})\\ &= \sum_{k\geq 1} k\mathbb{P}(N^\star(\mathbf{w}) = k)\\ &\leq \sum_{i=n-h+2}^{n} \sum_{k\geq 1} k \sum_{\mathbf{c}\in\mathcal{C}_k} \sum_{\mathbf{f}\in\mathcal{F}} \mathbb{E}[Y_i(\mathbf{cf})].\end{aligned}$$

Or, comme la séquence $X_{n-h+2}\cdots X_n\cdots$ est homogène stationnaire de probabilité de transition π_{s_n}, nous pouvons utiliser les mêmes calculs que dans la section 2.2.2 pour déduire que $\sum_{\mathbf{c}\in\mathcal{C}_k}\sum_{\mathbf{f}\in\mathcal{F}}\mathbb{E}[Y_i(\mathbf{cf})] = a_{s_n}^{k-1}(1-a_{s_n})\mu_{s_n}(\mathbf{w})$. Finalement, on a :

$$\mathbb{E}[N^\infty(\mathbf{w}) - N(\mathbf{w})] \leq h\pi_{\max}(\mathbf{w})\sum_{k\geq 1} k a_{s_n}^{k-1}(1-a_{s_n}) = h\pi_{\max}(\mathbf{w})/(1-a_{s_n}),$$

et on peut conclure par définition de ζ (cf. page 43).

■

Lemme 3.22 *Pour tout état $s\in S$, si Π_s^∞ désigne la matrice carrée de taille $|\mathcal{A}|$ où toutes les lignes sont égales au vecteur (ligne) $[\mu_s(x)]_{s\in\mathcal{A}}$, on a les propriétés suivantes :*

- $\Pi_s^\infty\Pi_s = \Pi_s^\infty$
- *Pour toute matrice stochastique M, $M\Pi_s^\infty = \Pi_s^\infty$*
- $\left|\left|\Pi_s^h - \Pi_s^\infty\right|\right| = O(|\alpha_s|^h)$, *avec $||.||$ qui désigne la norme infinie $\forall M, ||M|| = \sup_{i,j}|M_{i,j}|$.*

On donne pour finir la preuve du lemme 3.18.

Preuve du lemme 3.18. Comme $\lambda_{k,\text{bic}} = \mathbb{E}\widetilde{N}_k^\infty(\mathbf{w})$ pour $k \leq K = \frac{L_{\min}-3h}{\max(\mathcal{P}'(\mathbf{w}))}$ (cf. Proposition 3.15) et $\lambda_{k,\text{bic}} \leq \mathbb{E}\widetilde{N}_k^\infty(\mathbf{w})$ pour $k > K$, on a

$$\begin{aligned} d_{vt}\big(\mathcal{CP}(\mathbb{E}\widetilde{N}_k^\infty(\mathbf{w}), k\geq 1), \mathcal{CP}_{\text{bic}}\big) &\leq \sum_{k>K}|\mathbb{E}\widetilde{N}_k^\infty(\mathbf{w}) - \lambda_{k,\text{bic}}| \\ &\leq 2\sum_{k>K}\mathbb{E}\widetilde{N}_k^\infty(\mathbf{w}). \end{aligned}$$

Or le lemme 3.21 établit que lorque $\mathbb{E}N(\mathbf{w}) = O(1)$, on a aussi $\mathbb{E}N^\infty(\mathbf{w}) = O(1)$. Comme $\mathbb{E}N^\infty(\mathbf{w}) = \sum_{k\geq 1}\mathbb{E}\widetilde{N}_k^\infty(\mathbf{w})$, la série $\sum\mathbb{E}\widetilde{N}_k^\infty(\mathbf{w})$ est convergente. Le résultat découle donc de ce que $K\to\infty$. ■

Chapitre 4

Cas d'un modèle de Markov caché

Nous examinons dans ce chapitre le cas où la segmentation est aléatoire. Nous supposons pour cela que la séquence suit un modèle de Markov caché, et que les paramètres de ce modèle sont connus a priori. Le but de ce chapitre est de proposer une approximation de Poisson composée pour un mot rare dans une séquence suivant ce modèle. Dans la section 4.1, nous définissons trois lois $\mathcal{CP}'_{\text{uni}}$, $\mathcal{CP}'_{\text{bic}}$ et $\mathcal{CP}'_{\text{mult}}$ pour approcher la loi du comptage. Sous la condition de rareté, l'approximation par $\mathcal{CP}'_{\text{mult}}$ a une erreur qui tend vers 0, mais les paramètres de la loi $\mathcal{CP}'_{\text{mult}}$ sont complexes à calculer (notamment lorsque la segmentation possède beaucoup d'états différents). Les lois $\mathcal{CP}'_{\text{uni}}$ et $\mathcal{CP}'_{\text{bic}}$ sont plus rapides à calculer mais elles approchent moins bien la loi du comptage, car elles introduisent un terme d'erreur supplémentaire qui ne tend pas vers 0.
En section 4.2, nous présentons une approximation de Poisson composée pour les familles de mots rares. Cette approximation est meilleure que celle proposée par Reinert and Schbath (1998), car elle garantit une erreur en variation totale qui converge vers 0, même pour les familles contenant des mots recouvrants. Ce nouveau résultat est l'outil fondamental pour établir les trois approximations par $\mathcal{CP}'_{\text{uni}}$, $\mathcal{CP}'_{\text{bic}}$ et $\mathcal{CP}'_{\text{mult}}$ de la section 4.1.

4.1 Approximations par une loi Poisson composée dans un modèle de Markov caché

4.1.1 Rappels sur le modèle de Markov caché

Définition du modèle

Le modèle de chaîne de Markov caché ("Hidden Markov Model" noté HMM ou M1-Mm) est largement utilisé dans la littérature pour modéliser l'hétérogénéité d'une séquence (cf. par exemple Durbin *et al.* (1998) et Muri (1997)). Rappelons qu'un modèle M1-M1, est défini avec deux processus :

- un processus inobservable (caché) : $\mathbf{S} = (S_i)_{i\in\mathbb{Z}}$ qui suit une chaîne de Markov homogène (stationnaire) sur un espace d'états $\mathcal{S}$,
- un processus observable : $\mathbf{X} = (X_i)_{i\in\mathbb{Z}}$ qui suit une chaîne de Markov hétérogène conditionnellement à la segmentation $\mathbf{S}$.

Plus précisément, on considère une probabilité de transition $\pi_{\mathbf{S}} = \{\pi_{\mathbf{S}}(s,t)\}_{s,t\in\mathcal{S}}$ sur l'ensemble $\mathcal{S}$, supposée strictement positive ; $\forall s,t \in \mathcal{S}$, $\pi_{\mathbf{S}}(s,t) > 0$. On se donne également $\{\pi_s\}_{s\in\mathcal{S}}$ une famille de probabilités de transition sur $\mathcal{A}$. On suppose que $\forall s \in \mathcal{S}$, π_s est la probabilité de transition d'une chaîne de Markov irréductible apériodique et on note μ_s la mesure invariante associée à π_s.

Une séquence infinie $\mathbf{X} = (X_i)_{i\in\mathbb{Z}}$ de lettres aléatoires de $\mathcal{A}$ suit **un modèle M1-M1** selon la segmentation $\mathbf{S} = (S_i)_{i\in\mathbb{Z}}$ et avec les paramètres $\pi_\mathbf{S}$ et $\{\pi_s\}_{s\in\mathcal{S}}$, si $\mathbf{S} = (S_i)_{i\in\mathbb{Z}}$ est une chaîne de Markov homogène d'ordre 1 de probabilité de transition $\pi_\mathbf{S}$ et si pour tout $\mathbf{s} = (s_j)_{j\in\mathbb{Z}}$, pour toute position $i \in \mathbb{Z}$ et tout $(y_j)_{j\leq i}$ avec $y_j \in \mathcal{A}$, on a

$$\mathbb{P}(X_i = y_i \mid (X_j)_{j<i} = (y_j)_{j<i}, \mathbf{S} = \mathbf{s}) = \pi_{s_i}(y_{i-1}, y_i). \tag{4.1}$$

Remarque 4.1 *1.* $\mathbf{X}$ *suit donc un modèle M1-M1 si et seulement si conditionnellement à* $\mathbf{S} = \mathbf{s}$, $\mathbf{X}$ *suit un modèle PM1 avec la segmentation fixée* $\mathbf{s}$ *(au fait près que le support de* $\mathbf{s}$ *est infini).*

2. La segmentation $\mathbf{S}$ *est ici **aléatoire** (d'où la notation en majuscule). Ainsi, son nombre de ruptures* $\rho(\mathbf{S}) = |\{i \in \{2, \ldots, n\} \mid S_i \neq S_{i-1}\}|$ *est une variable aléatoire.*

3. Comme les transitions de $\pi_\mathbf{S}$ *sont positives, la chaîne* $\mathbf{S} = (S_i)_{i\in\mathbb{Z}}$ *est irréductible et apériodique et on notera sa loi invariante* $\mu_\mathbf{S}$*. De plus, on remarque que, comme pour chaque position i la séquence a une infinité d'états avant* S_i, $\mathcal{L}(S_i) = \mu_\mathbf{S}$*, et la séquence* $\mathbf{S}$ *est stationnaire.*

4. On "néglige" ici l'étape d'estimation des paramètres ; on suppose que les paramètres de la chaîne cachée $\{\pi_\mathbf{S}(s,t)\}_{s,t\in\mathcal{S}}$ *(ainsi que les paramètres* $\{\pi_s\}_{s\in\mathcal{S}}$*) sont connus.*

Bien entendu, nous pouvons définir le modèle M1-Mm de la même manière en remplaçant la condition 4.1, par la condition à l'ordre m :

$$\mathbb{P}(X_i = y_i \mid (X_j)_{j<i} = (y_j)_{j<i}, \mathbf{S} = \mathbf{s}) = \pi_{s_i}(y_{i-m} \cdots y_{i-1}, y_i). \tag{4.2}$$

La chaîne $\mathbf{S}$ est quant à elle toujours supposée markovienne d'ordre 1. En remplaçant $\mathcal{A}$ par $\mathcal{A}^m$, le cas d'un M1-Mm peut se ramener au cas d'un M1-M1 en suivant la même astuce de changement d'alphabet que celle de la section 3.1, ce qui va nous permettre de généraliser les résultats valables dans un modèle M1-M1 au cas d'un modèle M1-Mm.

Changement d'alphabet par coloriage

L'astuce suivante va permettre de passer d'un modèle M1-M1 à un modèle M1 (markovien homogène stationnaire d'ordre 1) ; considérons l'alphabet "colorié" $\mathcal{A}^\star := \{(y,s), y \in \mathcal{A}, s \in \mathcal{S}\}$, de cardinal $|\mathcal{S}|.|\mathcal{A}|$ et définissons la nouvelle séquence $\mathbf{X}^\star := (X_i, S_i)_{i\in\mathbb{Z}}$, à valeurs dans l'alphabet $\mathcal{A}^\star$. Lorsque $\mathbf{X}$ suit un modèle M1-M1, il est facile de voir que la séquence $\mathbf{X}^\star$ suit un modèle de Markov homogène d'ordre 1 de probabilités de transition (sur $\mathcal{A}^\star$) : $\forall y, z \in \mathcal{A}$, $\forall s, t \in \mathcal{S}$,

$$\pi^\star((y,s),(z,t)) = \pi_\mathbf{S}(s,t)\pi_t(y,z).$$

De plus, cette nouvelle chaîne de Markov hérite de la récurrence et de l'apériodicité des chaînes de probabilité de transition $\{\pi_s\}_s$: pour tout $y, z \in \mathcal{A}$, $s, t \in \mathcal{S}$, et $r \geq 1$, dès que $(\pi_t)^r(y,z) > 0$, on a

$$(\pi^\star)^r((y,s),(z,t)) \geq \pi_\mathbf{S}(s,t)[\pi_\mathbf{S}(t,t)]^{r-1}(\pi_t)^r(y,z) > 0.$$

Cette chaîne possède donc une mesure invariante $\mu^\star$, et comme $\mathbf{X}^\star$ porte aussi sur tous les indices négatifs, la séquence $\mathbf{X}^\star$ est de plus stationnaire.

Remarque 4.2 *Il est important de noter que $\mu^\star(z,t)$ n'est généralement pas égal au produit $\mu_{\mathbf{S}}(t)\mu_t(z)$ car de manière générale $\mathbb{P}(X_i = z|S_i = t) \neq \mu_t(z)$. Le calcul de $\mu^\star$ doit donc être fait en diagonalisant la matrice $\Pi^\star$ correspondant à la probabilité de transition $\pi^\star$.*

Dans toute la suite, nous fixons $\mathbf{X}$ une séquence suivant un modèle M1-M1 et $\mathbf{X}^\star$ la séquence coloriée correspondante suivant un modèle M1. On se donne également un mot $\mathbf{w} = w_1 \cdots w_h$ vérifiant l'hypothèse classique :

$$\forall \ell \in \{2, \dots, h\}, \forall s \in \mathcal{S}, \quad \pi_s(w_{\ell-1}, w_\ell) > 0. \tag{4.3}$$

Nous cherchons à établir une approximation pour $N(\mathbf{w})$, nombre d'occurrences de $\mathbf{w}$ dans la séquence $\mathbf{X}$. Nous allons utiliser pour cela les comptages coloriés $N_{\text{uni}}(\mathbf{w})$ et $N_{\text{bic}}(\mathbf{w})$. Pour une définition des comptages $N_{\text{uni}}(\mathbf{w})$ et $N_{\text{bic}}(\mathbf{w})$, le lecteur se reportera à la section 3.2 (page 42).

4.1.2 Approximation par $\mathcal{CP}'_{\text{uni}}$

Nous établissons ici une approximation pour la loi du comptage unicolore $N_{\text{uni}}(\mathbf{w})$ et nous en déduisons une approximation de type "mot unicolore" pour la loi de $N(\mathbf{w})$. On pose

$$\mathcal{W}_{\text{uni}} := \{(\mathbf{w}, s^h), s \in \mathcal{S}\},$$

la famille de mots sur l'alphabet $\mathcal{A}^\star$ composée du mot $\mathbf{w}$ colorié de toutes les façons unicolores possibles. Par définition, $N_{\text{uni}}(\mathbf{w})$, nombre d'occurrences unicolores de $\mathcal{W}$, est égal au nombre d'occurrences de la famille de mots $\mathcal{W}_{\text{uni}}$ dans la séquence homogène stationnaire $X_1^\star \cdots X_n^\star$. On remarque de plus que les mots de la famille $\mathcal{W}_{\text{uni}}$ **ne sont pas recouvrants**, car les coloriages considérés dans $\mathcal{W}_{\text{uni}}$ sont seulement unicolores. Pour approcher $N_{\text{uni}}(\mathbf{w})$, il suffit ainsi d'approcher la loi d'une famille de mots rares non-recouvrants dans une chaîne de Markov homogène stationnaire. Nous pouvons ainsi utiliser l'approximation de Poisson composée de Reinert and Schbath (1998), définie comme la somme (indépendante) des lois de Poisson composées relatives aux approximations homogènes de chacun des mots de la famille. De plus, Reinert and Schbath (1998) a prouvé que pour des familles de mots rares non-recouvrantes, l'erreur (en variation totale) de cette approximation tend bien vers 0 ; on en déduit le résultat suivant.

Proposition 4.3 *Soit $\mathbf{X}$ une séquence suivant un modèle M1-M1 et $\mu^\star$ la loi invariante de la séquence coloriée correspondante $\mathbf{X}^\star$. Pour un mot $\mathbf{w}$ vérifiant l'hypothèse (4.3), lorsque $\mathbb{E}N_{\text{uni}}(\mathbf{w}) = O(1)$ et $h = o(n)$, on a*

$$d_{vt}\big(\mathcal{L}(N_{\text{uni}}(\mathbf{w})), \mathcal{CP}'_{\text{uni}}\big) \xrightarrow[n\to\infty]{} 0, \tag{4.4}$$

où $\mathcal{CP}'_{\text{uni}}$ désigne la loi de Poisson composée de paramètres donnés par : $\forall k \geq 1$,

$$(n-h+1)\sum_{s\in\mathcal{S}} (a_s^\star)^{k-1}(1-a_s^\star)^2 \mu^\star(\mathbf{w}, s^h),$$

avec $\mu^\star(\mathbf{w}, s^h) = \mu^\star(w_1, s)(\pi_{\mathbf{S}}(s,s))^{h-1}\prod_{\ell=1}^{h-1}\pi_s(w_\ell, w_{\ell+1})$ qui est la probabilité d'occurrence de $(\mathbf{w}, s^h)$ à une position donnée de $\mathbf{X}^\star$ et avec

$$a_s^\star = \sum_{p\in\mathcal{P}'(\mathbf{w})} (\pi_{\mathbf{S}}(s,s))^p \prod_{\ell=1}^{p} \pi_s(w_\ell, w_{\ell+1})$$

qui est la probabilité d'auto-recouvrement de $(\mathbf{w}, s^h)$ dans $\mathbf{X}^\star$.

A présent, si nous voulons approcher la loi du comptage global $N(\mathbf{w})$ par la loi $\mathcal{CP}'_{\text{uni}}$, on utilise l'inégalité triangulaire :

$$d_{vt}\big(\mathcal{L}(N(\mathbf{w})), \mathcal{CP}'_{\text{uni}}\big) \leq d_{vt}\big(\mathcal{L}(N(\mathbf{w})), \mathcal{L}(N_{\text{uni}}(\mathbf{w}))\big) + d_{vt}\big(\mathcal{L}(N_{\text{uni}}(\mathbf{w})), \mathcal{CP}'_{\text{uni}}\big).$$

Lorsque $\mathbb{E}N(\mathbf{w}) = O(1)$ et $h = o(n)$, l'erreur $d_{vt}\big(\mathcal{L}(N_{\text{uni}}(\mathbf{w})), \mathcal{CP}'_{\text{uni}}\big)$ tend vers 0 d'après la proposition 4.3. Le terme d'erreur $d_{vt}\big(\mathcal{L}(N(\mathbf{w})), \mathcal{L}(N_{\text{uni}}(\mathbf{w}))\big)$ se contrôle de la façon suivante :

$$d_{vt}\big(\mathcal{L}(N(\mathbf{w})), \mathcal{L}(N_{\text{uni}}(\mathbf{w}))\big) \leq \mathbb{P}(N_{\text{uni}}(\mathbf{w}) \neq N(\mathbf{w})) \leq h\pi_{\max}(\mathbf{w})\mathbb{E}[\rho(\mathbf{S})], \tag{4.5}$$

où $\pi_{\max}(\mathbf{w}) := \max\{\pi_{t_2}(w_1, w_2) \times \cdots \times \pi_{t_h}(w_{h-1}, w_h), t_2 \cdots t_h \in \mathcal{S}^{h-1}\}$ et où $\mathbb{E}[\rho(\mathbf{S})]$ est le nombre attendu de ruptures dans $S_1 \cdots S_n$. Par la propriété de Markov, on voit directement que

$$\mathbb{E}[\rho(\mathbf{S})] = (n-1)\sum_{s\in\mathcal{S}}(1 - \pi_{\mathbf{S}}(s, s)).$$

Remarque 4.4 1. *La condition de rareté $\mathbb{E}N(\mathbf{w}) = O(1)$ ($h = o(n)$ et (4.3)) impose que h tende vers l'infini plus vite que $\log(n)$ i.e. $log(n)/h = O(1)$.*

2. *Lorsque $\mathbf{w}$ n'est pas recouvrant i.e. $\mathcal{P}'(\mathbf{w}) = \emptyset$, $\mathcal{CP}'_{\text{uni}}$ se réduit à une loi de Poisson de paramètre $\mathbb{E}N_{\text{uni}}(\mathbf{w}) = \sum_{s\in\mathcal{S}} \mu^\star(\mathbf{w}, s^h)$.*

3. *L'erreur (4.5) tend vers 0 lorsque n tend vers l'infini dès que $\forall s \in \mathcal{S}$, $1 - \pi_{\mathbf{S}}(s, s) = o((hn\pi_{\max}(\mathbf{w}))^{-1})$. Cependant, dans le résultat asymptotique (4.4), les paramètres de la chaîne de Markov $\pi^\star$ (et donc $\pi_{\mathbf{S}}$) sont supposés fixes. Ainsi, on ne peut pas considérer ici une asymptotique garantissant que l'erreur d'approximation par $\mathcal{CP}'_{\text{uni}}$ de la loi du comptage global tende vers 0.*

Le terme d'erreur (4.5) peut devenir grand dès que le nombre de ruptures attendu dans la segmentation augmente, cela est bien entendu dû au fait qu'on ne compte dans $\mathbf{X}^\star$ que les occurrences unicolores de $\mathbf{w}$. Ceci sera résolu dans la section suivante en considérant une approche "multicolore", qui traite la famille constituée du mot $\mathbf{w}$ dans tous les coloriages possibles.

4.1.3 Approximation par $\mathcal{CP}'_{\text{mult}}$

On associe à $\mathbf{w}$ la famille de mots $\mathcal{W}$ sur $\mathcal{A}^\star$ composée du mot $\mathbf{w}$ colorié de toutes les façons possibles :

$$\mathcal{W} = \{(\mathbf{w}, \mathbf{t}), \mathbf{t} \in \mathcal{S}^h\}.$$

On remarque que cette famille contient $|\mathcal{S}|^h$ mots différents. De plus, lorsque $\mathbf{w}$ est recouvrant, les mots de la famille $\mathcal{W}$ sont également recouvrants ; un mot colorié $(\mathbf{w}, \mathbf{t})$ avec $\mathbf{t} = t_1 \cdots t_h \in \mathcal{S}^h$ recouvre chaque mot $\mathbf{w}$ colorié avec un coloriage de la forme $t_{p+1} \cdots t_h t'_1 \cdots t'_p$, avec $p \in \mathcal{P}(\mathbf{w})$, $t'_i \in \mathcal{S}$.

Si $N^\star(\mathcal{W})$ désigne le nombre d'occurrences de la famille $\mathcal{W}$ dans la séquence coloriée $\mathbf{X}^\star$, on a bien évidemment $N(\mathbf{w}) = N^\star(\mathcal{W})$. Ainsi, pour approcher la loi du comptage $N(\mathbf{w})$, il suffit d'approcher la loi d'une famille de mots rare **recouvrante** dans une chaîne de Markov homogène stationnaire. Comme il s'agit d'une famille recouvrante, l'approximation de Poisson composée de Reinert and Schbath (1998) ne garantit plus que l' erreur en variation totale tende vers 0. Nous avons donc amélioré dans un premier temps l'approximation de Reinert and Schbath (1998), pour proposer une approximation dont l'erreur tend vers 0 pour une famille de mots rare quelconque, recouvrante ou non. Cette approximation prend en compte les éventuels recouvrements entre les mots dans une matrice de recouvrement. Ce problème sera présenté dans la section 4.2 et les résultats précis seront énoncés et prouvés dans le chapitre 5.

En utilisant cette nouvelle approximation, (précisément le théorème 5.1), on prouve que pour un mot vérifiant (4.3), si $\mu^\star(\mathcal{W})$ désigne la probabilité d'occurrence de $\mathcal{W}$ dans $\mathbf{X}^\star$ (qui est aussi la probabilité d'occurrence de $\mathbf{w}$ dans $\mathbf{X}$),

$$d_{vt}\big(\mathcal{L}(N(\mathbf{w})), \mathcal{CP}'_{\mathrm{mult}}\big) \leq n\mu^\star(\mathcal{W})[Ch\mu^\star(\mathcal{W}) + C'|\alpha^\star|^h] + 2h\mu^\star(\mathcal{W}), \tag{4.6}$$

où C et C' sont deux constantes qui dépendent uniquement des probabilités de transition $\pi_{\mathbf{S}}$ et $[\pi_s]_{s\in\mathcal{S}}$, et où $\alpha^\star$ désigne la seconde plus grande valeur propre en valeur absolue de la matrice $\Pi^\star$ ($|\alpha^\star| < 1$). Les paramètres de $\mathcal{CP}'_{\mathrm{mult}}$ ont l'expression suivante : $\forall k \geq 1$,

$$(n-h+1)||[A^\star(\mathcal{W})]^{k-1}(I - A^\star(\mathcal{W}))^2\vec{\mu^\star}(\mathcal{W})||_1, \tag{4.7}$$

I étant la matrice identité d'ordre $d = |\mathcal{S}|^h$, $||.||_1$ désignant la norme 1 sur $\mathbb{R}^d$, $\vec{\mu^\star}(\mathcal{W}) = [\mu^\star(\mathbf{w},\mathbf{t})]_{\mathbf{t}\in\mathcal{S}^h}$ étant le vecteur des probabilités d'occurrence des mots de $\mathcal{W}$ et $A^\star(\mathcal{W})$ étant la matrice d'auto-recouvrement de $\mathcal{W}$ (d'ordre d). Pour tout coloriage $\mathbf{t} = t_1\dots t_h, \mathbf{t}' = t'_1\dots t'_h \in \mathcal{S}^h$, l'indice $((\mathbf{w},\mathbf{t}),(\mathbf{w},\mathbf{t}'))$ de cette matrice s'écrit :

$$\frac{\mu^\star(w_1,t_1)}{\mu^\star(w_1,t'_1)} \sum_{p\in\mathcal{P}'(\mathbf{w})\cap\mathcal{P}(\mathbf{t},\mathbf{t}')} \prod_{\ell=1}^{p} \pi_{\mathbf{S}}(t_\ell, t_{\ell+1})\pi_{t_{\ell+1}}(w_\ell, w_{\ell+1}), \tag{4.8}$$

où $\mathcal{P}(\mathbf{t},\mathbf{t}') := \big\{p \in \{1,\dots,h-1\} \mid \forall i \in \{1,\dots,h-p\}, t'_i = t_{i+p}\big\}$ est l'ensemble des distances inférieures à $h-1$ autorisées entre une occurrence du coloriage $\mathbf{t}$ et une occurrence du coloriage $\mathbf{t}'$ (l'analogue de l'ensemble des périodes mais pour les coloriages). La quantité (4.8) représente la probabilité qu'une occurrence du mot $\mathbf{w}$ coloriée avec $\mathbf{t}'$ soit recouvert par une précédente occurrence de $\mathbf{w}$ coloriée avec $\mathbf{t}$, le terme "précédent" étant pris au sens strict car il signifie qu'il est interdit d'avoir une autre occurrence de $\mathbf{w}$ coloriée entre les deux. La matrice de recouvrement $A^\star(\mathcal{W})$ mesure donc la capacité qu'a une occurrence du mot $\mathbf{w}$ coloriée d'une façon à recouvrir une autre occurrence de $\mathbf{w}$ coloriée d'une autre façon. La relation (4.6) donne donc le résultat suivant :

Proposition 4.5 *Soit* $\mathbf{X}$ *une séquence suivant un modèle M1-M1 et* $\mu^\star$ *la loi invariante de la séquence coloriée correspondante* $\mathbf{X}^\star$*. Pour un mot* $\mathbf{w}$ *vérifiant l'hypothèse (4.3), lorsque* $\mathbb{E}N(\mathbf{w}) = O(1)$ *et* $h = o(n)$*, on a*

$$d_{vt}\big(\mathcal{L}(N(\mathbf{w})), \mathcal{CP}'_{\mathrm{mult}}\big) \xrightarrow[n\to\infty]{} 0,$$

où $\mathcal{CP}'_{\mathrm{mult}}$ *est la loi de Poisson composée de paramètres donnés par l'expression (4.7).*

Remarque 4.6
1. *Ici, le cardinal de la famille de mots* $d = |\mathcal{S}|^h$ *dépend de* n *sous la condition de rareté. Par ailleurs, le théorème 5.1 du chapitre 5 est établi pour un nombre de mots* d *fixe avec* n*. Cependant, la preuve de ce théorème s'étend sans difficulté au cas d'une famille de mots dont le cardinal peut varier avec* n*.*
2. *Lorsque* $\mathbf{w}$ *n'est pas recouvrant i.e.* $\mathcal{P}'(\mathbf{w}) = \emptyset$*,* $\mathcal{CP}'_{\mathrm{mult}}$ *se réduit à une loi de Poisson de paramètre* $\mathbb{E}N(\mathbf{w}) = \sum_{\mathbf{t}\in\mathcal{S}^h} \mu^\star(\mathbf{w},\mathbf{t})$*.*

Comme $|\mathcal{W}| = |\mathcal{S}|^h$, la formule (4.7) est assez complexe et donc la loi $\mathcal{CP}'_{\mathrm{mult}}$ peut être longue à calculer en pratique (voire incalculable si $\mathcal{S}$ contient beaucoup d'états différents). Si la segmentation contient beaucoup de ruptures i.e. s'il y a des $\pi_{\mathbf{S}}(s,s)$ "loin" de 1, il n'y a pas de simplification possible et la seule approximation raisonnable est celle utilisant $\mathcal{CP}'_{\mathrm{mult}}$. Dans un cas plus favorable où la segmentation contient beaucoup de segments de longueurs plus grandes que h, alors on peut proposer une approximation de type "mot bicolore".

4.1.4 Approximation par $\mathcal{CP}'_{\mathrm{bic}}$

Nous établissons ici une approximation pour la loi du comptage bicolore $N_{\mathrm{bic}}(\mathbf{w})$. Par suite, nous en déduisons une approximation de type "mot bicolore" pour la loi de $N(\mathbf{w})$, avec un terme d'erreur supplémentaire à contrôler. Définissons la famille $\mathcal{W}_{\mathrm{bic}}$ sur $\mathcal{A}^\star$ par :

$$\mathcal{W}_{\mathrm{bic}} = \{(\mathbf{w}, s^\ell t^{h-\ell}), s, t \in \mathcal{S}, 1 \leq \ell \leq h\}.$$

Remarquons que le nombre de mots dans la famille $\mathcal{W}_{\mathrm{bic}}$ est seulement de $|\mathcal{S}|+(h-1)|\mathcal{S}|(|\mathcal{S}|+1)$ (au lieu de $|\mathcal{S}|^h$ pour $\mathcal{W}$). En utilisant, le théorème 5.1 comme dans la section précédente, on déduit le résultat suivant.

Proposition 4.7 *Soit* $\mathbf{X}$ *une séquence suivant un modèle M1-M1 et* $\mu^\star$ *la loi invariante de la séquence coloriée correspondante* $\mathbf{X}^\star$*. Pour un mot* $\mathbf{w}$ *vérifiant l'hypothèse (4.3), lorsque* $\mathbb{E}N_{\mathrm{bic}}(\mathbf{w}) = O(1)$ *et* $h = o(n)$*, on a*

$$d_{vt}\big(\mathcal{L}(N_{\mathrm{bic}}(\mathbf{w})), \mathcal{CP}'_{\mathrm{bic}}\big) \xrightarrow[n\to\infty]{} 0,$$

où $\mathcal{CP}'_{\mathrm{bic}}$ *est la loi de Poisson composée de paramètres donnés par l'expression (4.7) où on a remplacé* $\mathcal{W}$ *par* $\mathcal{W}_{\mathrm{bic}}$*. Le coefficient d'ordre* $((\mathbf{w},\mathbf{t}),(\mathbf{w},\mathbf{t}'))$ *de la matrice de recouvrement* $A^\star(\mathcal{W}_{\mathrm{bic}})$ *est donné par :*

$$\frac{\mu^\star(w_1,t_1)}{\mu^\star(w_1,t'_1)} \sum_{p\in\mathcal{P}'_{\mathcal{W}_{\mathrm{bic}}}((\mathbf{w},\mathbf{t}),(\mathbf{w},\mathbf{t}'))} \prod_{\ell=1}^{p} \pi_{\mathbf{S}}(t_\ell, t_{\ell+1})\pi_{t_{\ell+1}}(w_\ell, w_{\ell+1}),$$

avec $p \in \mathcal{P}'_{\mathcal{W}_{\mathrm{bic}}}((\mathbf{w},\mathbf{t}),(\mathbf{w},\mathbf{t}'))$ *si et seulement si* $p \in \mathcal{P}(\mathbf{w}) \cap \mathcal{P}(\mathbf{t},\mathbf{t}')$ *et* $\forall 1 \leq \ell \leq h, \forall s \neq t \in \mathcal{S}$, $\forall j \in \mathcal{P}(\mathbf{w}) \cap \mathcal{P}(\mathbf{t}, s^\ell t^{h-\ell})$*, on a* $p - j \notin \mathcal{P}(\mathbf{w}) \cap \mathcal{P}(s^\ell t^{h-\ell}, \mathbf{t}')$*.*

Dans la proposition ci-dessus, l'ensemble $\mathcal{P}'_{\mathcal{W}_{\mathrm{bic}}}((\mathbf{w},\mathbf{t}),(\mathbf{w},\mathbf{t}'))$ peut paraître abstrait. Selon le chapitre 5, il s'appelle l'ensemble des périodes principales de $(\mathbf{w},\mathbf{t}),(\mathbf{w},\mathbf{t}')$ dans la famille $\mathcal{W}_{\mathrm{bic}}$ et se comprend de la façon suivante : fixons une période $p \in \mathcal{P}(\mathbf{w}) \cap \mathcal{P}(\mathbf{t},\mathbf{t}')$, de sorte que nous sommes dans la configuration où le mot colorié $(\mathbf{w},\mathbf{t})$ recouvre le mot colorié $(\mathbf{w},\mathbf{t}')$ sur $h-p$ positions. Le fait que p soit une période principale de $(\mathbf{w},\mathbf{t}),(\mathbf{w},\mathbf{t}')$ dans la famille $\mathcal{W}_{\mathrm{bic}}$ signifie qu'entre les deux occurrences de $(\mathbf{w},\mathbf{t})$ et $(\mathbf{w},\mathbf{t}')$, il ne peut pas y avoir d'autre occurrence de **w bicolore à au plus une rupture d'état**. Par exemple, considérons $\mathbf{w} =$ `acacac` avec $\mathcal{S} = \{1,2\}$, et les coloriages $\mathbf{t} = 111122$ et $\mathbf{t}' = 221111$. On a $\mathcal{P}(\mathbf{w}) = \{2,4\}$, $\mathcal{P}'(\mathbf{w}) = \{2\}$ et $\mathcal{P}(\mathbf{t},\mathbf{t}') = \{4\}$. Le seul candidat pour être dans $\mathcal{P}'_{\mathcal{W}_{\mathrm{bic}}}((\mathbf{w},\mathbf{t}),(\mathbf{w},\mathbf{t}')$ est donc 4. Or, dans le recouvrement faisant intervenir la période 4,

$$\begin{array}{c}\mathtt{ac\underline{acacaca}c}\\ 11\underline{111221}11\end{array},$$

l'occurrence de $\mathbf{w}$ soulignée est coloriée dans un coloriage à deux ruptures ce qui ne fait pas partie de la famille $\mathcal{W}_{\text{bic}}$. Par conséquent, la période 4 est bien une période principale de $(\mathbf{w}, \mathbf{t}), (\mathbf{w}, \mathbf{t}')$ dans la famille $\mathcal{W}_{\text{bic}}$ et $\mathcal{P}'_{\mathcal{W}_{\text{bic}}}((\mathbf{w}, \mathbf{t}), (\mathbf{w}, \mathbf{t}')) = \{4\}$.

A présent, si nous voulons approcher la loi du comptage global $N(\mathbf{w})$ par la loi $\mathcal{CP}'_{\text{bic}}$, on utilise (comme dans le cas unicolore) l'inégalité triangulaire :

$$d_{vt}\big(\mathcal{L}(N(\mathbf{w})), \mathcal{CP}'_{\text{bic}}\big) \leq d_{vt}\big(\mathcal{L}(N(\mathbf{w})), \mathcal{L}(N_{\text{bic}}(\mathbf{w}))\big) + d_{vt}\big(\mathcal{L}(N_{\text{bic}}(\mathbf{w})), \mathcal{CP}'_{\text{bic}}\big).$$

Lorsque $\mathbb{E}N(\mathbf{w}) = O(1)$ et $h = o(n)$, l'erreur $d_{vt}\big(\mathcal{L}(N_{\text{bic}}(\mathbf{w})), \mathcal{CP}'_{\text{bic}}\big)$ tend vers 0 d'après la proposition 4.7. Le terme d'erreur $d_{vt}\big(\mathcal{L}(N(\mathbf{w})), \mathcal{L}(N_{\text{bic}}(\mathbf{w}))\big)$ se contrôle à l'aide du lemme suivant.

Lemme 4.8 *Pour tout mot* $\mathbf{w}$ *dans un modèle* $M1$-$M1$ *:*

$$\mathbb{P}(N(\mathbf{w}) \neq N_{\text{bic}}(\mathbf{w})) \leq h(n-2)\pi_{\max}(\mathbf{w}) \sum_{s \in \mathcal{S}} \mu_{\mathbf{S}}(s)(1 - \pi_{\mathbf{S}}(s,s))(1 - \pi_{\mathbf{S}}(s,s)^{h-1}).$$

Remarque 4.9 *Le majorant de l'erreur* $\mathbb{P}(N(\mathbf{w}) \neq N_{\text{bic}}(\mathbf{w}))$ *ci-dessus est plus petit que celui correspondant à l'approximation par* $\mathcal{CP}'_{\text{uni}}$ $\mathbb{P}(N(\mathbf{w}) \neq N_{\text{uni}}(\mathbf{w}))$ *:*

$$h(n-1)\pi_{\max}(\mathbf{w}) \sum_{s \in \mathcal{S}} (1 - \pi_{\mathbf{S}}(s,s)) \leq h(n-2)\pi_{\max}(\mathbf{w}) \sum_{s \in \mathcal{S}} \mu_{\mathbf{S}}(s)(1 - \pi_{\mathbf{S}}(s,s))(1 - \pi_{\mathbf{S}}(s,s)^{h-1}).$$

Ainsi, lorsque la loi $\mathcal{CP}'_{\text{mult}}$ *est trop longue à calculer, l'approximation* $\mathcal{CP}'_{\text{bic}}$ *est une alternative intéressante ; elle réalise un bon compromis complexité/précision entre les approximations* $\mathcal{CP}'_{\text{uni}}$ *et* $\mathcal{CP}'_{\text{mult}}$.

Preuve du lemme 4.8. On définit $\mathcal{T}$ la famille de mots

$$\mathcal{T} \quad = \quad \{ts^{\ell}t', s, t, t' \in \mathcal{S}, t \neq s, t' \neq s, 1 \leq \ell \leq h-1\}.$$

Alors les occurrences de $\mathbf{w}$ qui ne sont pas bicolores à une rupture d'état apparaissent forcément à une position de la forme $\{i-h+1, \ldots, i\}$ où i est la position d'une occurrence de $\mathcal{T}$ dans la segmentation $\mathbf{S}$:

$$\begin{aligned}
\mathbb{P}(N(\mathbf{w}) \neq N_{\text{bic}}(\mathbf{w})) &\leq \mathbb{E}[N(\mathbf{w}) - N_{\text{bic}}(\mathbf{w})] \\
&\leq \mathbb{E}\Big[\sum_{i=1}^{n-h+1} Y_i^{\mathbf{S}}(\mathcal{T}) \sum_{j=i-h+1}^{i} Y_j(\mathbf{w})\Big] \\
&\leq h\pi_{\max}(\mathbf{w})\mathbb{E}Y_i^{\mathbf{S}}(\mathcal{T}),
\end{aligned}$$

où $Y_i^{\mathbf{S}}(\mathcal{T})$ vaut 1 si il y a une occurrence de $\mathcal{T}$ à la position i dans $\mathbf{S}$ et vaut 0 sinon. Comme $\mathbb{E}Y_i^{\mathbf{S}}(\mathcal{T}) \leq (n-2)\mu_{\mathbf{S}}(\mathcal{T})$, il suffit à présent de calculer $\mu_{\mathbf{S}}(\mathcal{T})$, la probabilité d'occurrence de la famille $\mathcal{T}$ dans la segmentation $\mathbf{S}$. Ceci repose essentiellement sur le fait que le temps de séjour dans un état de **S** suit une loi géométrique :

$$\begin{aligned}
\mu_{\mathbf{S}}(\mathcal{T}) &= \sum_{s\in\mathcal{S}}\sum_{t\neq s}\sum_{t'\neq s}\sum_{\ell=1}^{h-1}\mu_{\mathbf{S}}(ts^{\ell}t') \\
&= \sum_{s\in\mathcal{S}}\left[\sum_{t\neq s}\mu_{\mathbf{S}}(t)\pi_{\mathbf{S}}(t,s)\right]\left[\sum_{t'\neq s}\pi_{\mathbf{S}}(s,t')\right]\left[\sum_{\ell=1}^{h-1}\pi_{\mathbf{S}}(s,s)^{\ell-1}\right] \\
&= \sum_{s\in\mathcal{S}}\left[\sum_{t\in\mathcal{S}}\mu_{\mathbf{S}}(t)\pi_{\mathbf{S}}(t,s)-\mu_{\mathbf{S}}(s)\pi_{\mathbf{S}}(s,s)\right](1-\pi_{\mathbf{S}}(s,s)^{h-1}) \\
&= \sum_{s\in\mathcal{S}}\mu_{\mathbf{S}}(s)(1-\pi_{\mathbf{S}}(s,s))(1-\pi_{\mathbf{S}}(s,s)^{h-1}),
\end{aligned}$$

car $\sum_{t\in\mathcal{S}}\mu_{\mathbf{S}}(t)\pi_{\mathbf{S}}(t,s)=\mu_{\mathbf{S}}(s)$ par définition de la mesure invariante $\mu_{\mathbf{S}}$. ■

4.1.5 Discussion : cas d'une segmentation avec une loi quelconque

Je viens d'explorer le cas où la segmentation suit une chaîne de Markov. Mais ce modèle est parfois trop simple pour modéliser une hétérogénéité ; notamment dans de nombreux exemples biologiques, la longueur des segments ne suit pas une loi géométrique (cf. par exemple Melo De Lima (2005)). Ainsi, je me suis également intéressé à trouver une approximation valable quelque soit la loi sous-jacente de la segmentation (approche "distribution-free"). L'idée est simplement de conditionner par rapport à la segmentation pour se ramener au cas d'une segmentation fixée et d'appliquer les résultats du chapitre 3. L'approximation (de type "mot bicolore") obtenue est alors une loi de Poisson composée de paramètres donnés par les formules (3.24), (3.22) et (3.23) où l'on a remplacé les quantités $n_{\mathbf{s}}(s^{3h+\sum_{\ell=1}^{r}m_\ell p_\ell})$ et $n_{\mathbf{s}}(st)$ par leurs valeurs moyennes respectives $\mathbb{E}N_{\mathbf{S}}(s^{3h+\sum_{\ell=1}^{r}m_\ell p_\ell})$ et $\mathbb{E}N_{\mathbf{S}}(st)$. Cependant, le contrôle que j'ai obtenu de l'erreur de cette approximation n'est pas satisfaisant (notamment parce qu'il suppose l'indépendance des lettres conditionnellement à la segmentation). Cette étude a donc été omise dans le manuscrit final.

4.2 Nouvelle approximation pour le comptage d'une famille de mots rare quelconque

Nous présentons ici une nouvelle approximation pour le comptage d'une famille de mots rare quelconque dans un modèle de Markov **homogène stationnaire**. Les résultats mathématiques seront précisément énoncés et prouvés dans le chapitre 5. La section actuelle a pour but de présenter le résultat principal du chapitre 5, et d'y ajouter une étude de simulation ainsi qu'une application.

Approcher le comptage d'une famille de mots dans une séquence markovienne homogène a été motivé en section 4.1, afin d'approcher la loi du comptage d'un mot dans un modèle HMM. Cependant, ce travail a un intérêt propre ; en effet, les motifs fonctionnels en biologie sont souvent dégénérés au sens ou une (ou plusieurs) lettre(s) du motif peut(peuvent) ne pas être définie(s). Dans ce cas, c'est le nombre d'occurrences de la **famille** de mots qui sera inattendu par rapport à un certain modèle, et c'est donc la loi du comptage de la famille de mots qu'il faut examiner.

4.2.1 Description de la nouvelle approximation par $\mathcal{CP}_{\text{fam}}$

Nous considérons ici une famille $\mathcal{W}$, constituée de mots susceptibles de se recouvrir entre eux, susceptibles de se recouvrir eux-mêmes et de longueurs éventuellement différentes. Nous supposons que la famille $\mathcal{W}$ est réduite, c'est-à-dire qu'aucun mot de $\mathcal{W}$ n'est un sous-mot d'un autre mot de $\mathcal{W}$. Nous considérons le cadre asymptotique où la famille de mots est rare : $\mathbb{E}N(\mathcal{W}) = O(1)$, c'est-à-dire que tous les mots de la famille sont rares. Pour utiliser une approximation de Poisson composée, Reinert and Schbath (1998) utilisent la décomposition du comptage de chaque mot $\mathbf{w} \in \mathcal{W}$ en k-trains, et approchent les comptages des trains de chaque mot $(\widetilde{N}_k(\mathbf{w}), k \geq 1, \mathbf{w} \in \mathcal{W})$ par un processus de Poisson (toujours avec la méthode de Chen-Stein). Ceci revient à considérer les trains du mot $\mathbf{w}_1$ asymptotiquement indépendants des trains du mot $\mathbf{w}_2$ pour deux mots $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}$, et cela ne sera pertinent que si les mots de la famille ne se recouvrent "pas trop". Ainsi, l'approximation de Reinert and Schbath (1998) a un terme d'erreur d'autant plus grand que la famille de mots peut se recouvrir (l'erreur est nulle si $\forall \mathbf{w}, \mathbf{w}' \in \mathcal{W}$, $\mathbf{w}$ ne recouvre pas $\mathbf{w}'$). Une nouvelle idée est alors de considérer les occurrences de mots **ensemble**, en considérant les **trains de la famille de mots** $\mathcal{W}$ définis comme les paquets maximaux d'occurrences de la famille de mots $\mathcal{W}$ (cf. FIG. 4.1). Nous avons la décomposition :

$$N(\mathcal{W}) = \sum_{k \geq 1} k \widetilde{N}_k(\mathcal{W}),$$

où $\widetilde{N}_k(\mathcal{W})$ désigne le nombre d'occurrences des trains de $\mathcal{W}$ de taille k (c'est-à-dire des trains contenant exactement k occurrences de la famille de mots $\mathcal{W}$). On peut montrer, toujours avec la méthode de Chen-Stein, que la loi du comptage $N(\mathcal{W})$ s'approche avec une erreur en variation totale qui tend vers 0 par une loi de Poisson composée de paramètres $(\mathbb{E}\widetilde{N}_k(\mathcal{W}), k \geq 1)$.

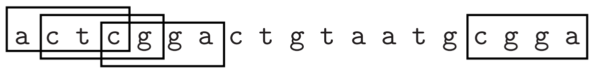


FIG. 4.1 – Cette séquence contient 2 trains de la famille de mots $\mathcal{W} = \{\texttt{actc}, \texttt{ctcg}, \texttt{cgga}\}$: l'un à la position 1 (taille 3), l'autre à la position 16 (taille 1). On doit faire attention à ne pas oublier l'occurrence de `ctcg` à la position 2 dans le train de taille 3.

Le calcul des paramètres s'effectue en définissant les distances typiques qu'il peut y avoir entre deux occurrences successives recouvrantes de deux mots de $\mathcal{W}$ dans un train de $\mathcal{W}$. Ces distances permettent de définir une matrice de recouvrement A carrée, de taille égale au nombre de mots dans la famille et où $A(\mathbf{w}_1, \mathbf{w}_2)$ est défini comme la probabilité que $\mathbf{w}_1$ recouvre $\mathbf{w}_2$ et que ces deux occurrences soient consécutives dans la famille, c'est-à-dire qu'elles arrivent sans qu'aucune autre occurrence d'un mot de la famille n'apparaisse entre $\mathbf{w}_1$ et $\mathbf{w}_2$. Par suite, nous montrons que les paramètres $\mathbb{E}\widetilde{N}_k(\mathcal{W})$ ont une forme similaire au cas d'un mot : pour tout $k \geq 1$,

$$\mathbb{E}\widetilde{N}_k(\mathcal{W}) = ||A^{k-1}(I - A)^2 \vec{\mathbb{E}}N(\mathcal{W})||_1,$$

où I est la matrice identité, $\vec{\mathbb{E}}N(\mathcal{W})$ est le vecteur des comptages attendus $[\mathbb{E}N(\mathbf{w})]_{\mathbf{w} \in \mathcal{W}}$, et où $||.||_1$ est la norme 1. Cette nouvelle approximation est cohérente avec les résultats de Schbath (1995a) lorsque $\mathcal{W} = \{\mathbf{w}\}$ et avec les résultats de Reinert and Schbath (1998) lorsque les mots de la famille ne se recouvrent pas entre eux (A est dans ce cas diagonale). On note cette nouvelle loi de Poisson composée $\mathcal{CP}_{\text{fam}}$.

4.2.2 Qualité de $\mathcal{CP}_{\text{fam}}$ face à l'approximation de Reinert and Schbath (1998)

Nous cherchons maintenant à comparer la qualité de cette approximation avec celle de Reinert and Schbath (1998) sur un exemple concret ; on considère une chaîne de Markov homogène (d'ordre 1) sur $\{\mathtt{a},\mathtt{c},\mathtt{g},\mathtt{t}\}$, dont la matrice de transition a été ajustée à partir des 1000 premières bases du génome complet de *Bacillus subtilis* :

$$\Pi = \begin{pmatrix} 0.351 & 0.174 & 0.181 & 0.294 \\ 0.335 & 0.223 & 0.158 & 0.284 \\ 0.337 & 0.218 & 0.202 & 0.243 \\ 0.203 & 0.190 & 0.223 & 0.384 \end{pmatrix}. \tag{4.9}$$

La loi stationnaire est $\mu = (0.300, 0.197, 0.193, 0.310)$. Considérons la famille de mots $\mathcal{W} = \{\mathtt{acgt},\mathtt{cgta},\mathtt{gtac},\mathtt{tacg}\}$ dans une séquence de longueur $n = 1000$. Le comptage attendu de la famille de mots est alors $\mathbb{E}(N(\mathcal{W})) = (n-3)\mu(\mathcal{W}) = 6.90$, et la famille de mots peut bien être considérée comme rare.

La figure 4.2 (gauche) représente la loi exacte du comptage obtenue à partir de la méthode de Robin and Daudin (1999), la loi de Poisson composée proposée par Reinert and Schbath (1998) et notre nouvelle loi de Poisson composée $\mathcal{CP}_{\text{fam}}$. Clairement, l'approximation par $\mathcal{CP}_{\text{fam}}$ est meilleure que celle de Reinert and Schbath (1998). La distance en variation totale entre la loi exacte et la loi $\mathcal{CP}_{\text{fam}}$ (resp. celle de Reinert and Schbath (1998)) est $8.11\,10^{-3}$ (resp. $9.03\,10^{-2}$). Ces résultats ne sont pas surprenants car l'approximation de Reinert and Schbath (1998) revient à considérer que la matrice de recouvrement A est diagonale, or ici, la famille de mots $\mathcal{W}$ est fortement recouvrante ce qui donne une matrice A loin d'une matrice diagonale.

On peut aussi regarder la robustesse de l'approximation par $\mathcal{CP}_{\text{fam}}$ lorsque le comptage attendu de la famille de mots augmente ; dans une séquence plus longue ($n = 10^4$) dans laquelle le comptage attendu est $\mathbb{E}(N(\mathcal{W})) = 97.15$, la figure 4.2 (droite) montre que la loi $\mathcal{CP}_{\text{fam}}$ a l'air d'approcher encore correctement la loi du comptage.

La figure 4.3 montre que cette amélioration s'étend à toutes les familles du type $\{xyzt, yztx, ztxy, txyz\}$ et de cardinal[1] 4. On trace la dispersion de la distance en variation totale entre la loi du comptage et chacune des deux lois de Poisson composée (dans le cas $n = 1000$).

4.2.3 Application

J'ai participé à l'implémentation de cette nouvelle approximation dans la version 3 du logiciel R'MES[2] (Hoebeke and Schbath (2006)), dédié à la recherche de motifs exceptionnels dans des séquences homogènes. L'utilisateur peut désormais dans cette nouvelle version calculer le score d'exceptionnalité d'une famille de mots rare (recouvrante ou non) selon l'approximation $\mathcal{CP}_{\text{fam}}$. Cela a permis à nos collègues biologistes, dans le cadre du projet MOSAIC[3], d'identifier un motif appelé "parS" qui structure le domaine ORI (782686 pbs) du génome de la bactérie *Bacillus subtilis*. Pour cela ils ont extrait les familles les plus sur-représentées dans le domaine ORI de la forme $\{\mathbf{w}, \overline{\mathbf{w}}\}$ où $\mathbf{w}$ est un mot de longueur 11 et $\overline{\mathbf{w}}$ est son complémentaire inversé. Le motif le plus exceptionnel était alors $\mathcal{W} = (\mathtt{cacgtgtaaca}, \mathtt{tgttacacgtg})$ et une liste de motifs candidats plus longs contenant $\mathcal{W}$ a été testée. Le motif *parS* vient d'être validé expérimentalement (en cours de publication).

[1] Nous ne considérons ni le cas où simultanément $x = z$ et $y = t$, ni celui où $x = y = z = t$

[2] http ://genome.jouy.inra.fr/ssb/rmes/

[3] http ://mig.jouy.inra.fr/recherches/projet/mosaic

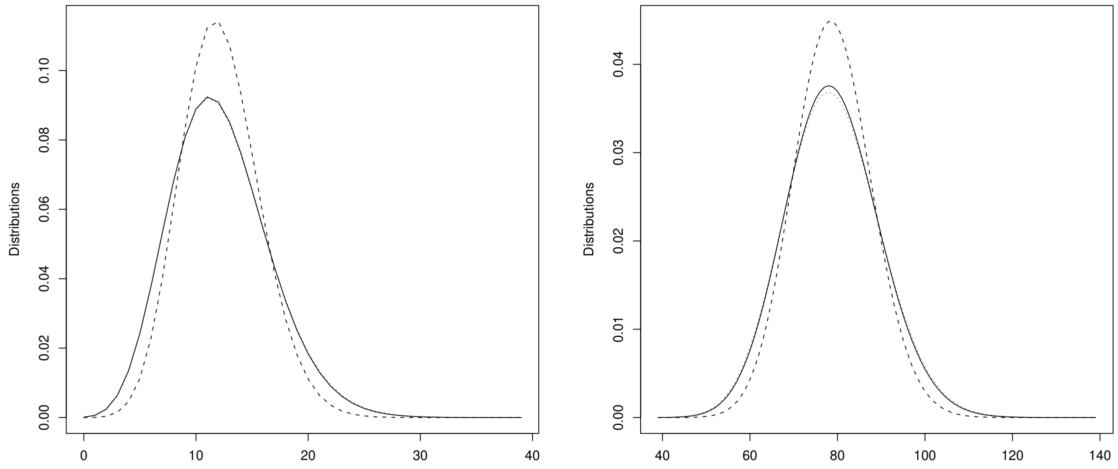


FIG. 4.2 – Comparaison entre la loi du comptage (ligne continue), la loi de Poisson composée de Reinert and Schbath (1998) (ligne discontinue) et la nouvelle loi $\mathcal{CP}_{\mathrm{fam}}$ (ligne pointillés) pour le comptage de la famille de mot $\mathcal{W} = \{\texttt{acgt,cgta,gtac,tacg}\}$ dans une chaîne de Markov de longueur 1000 (gauche) ou 10000 (droite) ; la matrice de transition a été ajustée respectivement à partir des 1000 et 10000 premières bases du génome complet de *Bacillus subtilis*.

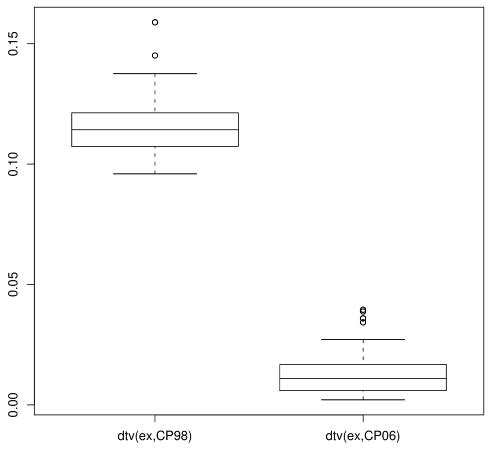


FIG. 4.3 – Boîte de dispersion de la distance en variation totale entre la loi du comptage et chacune des deux lois de Poisson composées, pour toutes les familles du type $\{xyzt, yztx, ztxy, txyz\}$ dans une chaîne de Markov de longueur 1000 (la matrice de transition étant donnée par (4.9)).

Chapter 5

Improved compound Poisson approximation for the number of occurrences of any rare word family in a stationary Markov chain

This chapter is a copy of the paper Roquain and Schbath (2007) (except that the references are joined to the thesis's references).

We derive a new compound Poisson distribution with explicit parameters to approximate the number of overlapping occurrences of any set of words in a Markovian sequence. Using the Chen-Stein method, we provide a bound for the approximation error. This error converges to zero under the rare event condition, even for overlapping families which improves previous results. As a consequence, we also propose Poisson approximations for the declumped count and the number of competing renewals.

5.1 Introduction

Word statistics in random sequences of letters have been popular for a long time because they arise in various application domains. With the huge number of biological sequences now available, genome analysis is an important consumer of probabilistic and statistical results on word occurrences (see chapter 6 of Lothaire (2005) or Reinert *et al.* (2000) for overviews). In particular the number N of occurrences of a given word in a DNA sequence is a quantity of special interest to molecular biologists; Some words, called *motifs*, are recognized by proteins and occur in various biological processes. Over- and under-represented motifs are then looked for in many genomes. Moreover, biological motifs are often degenerated, i.e. some letters are ambiguous, and should be treated as families of fixed words.

The most popular random sequence models are the Markov chain models; They are widely used in genome analysis because they can be used to fit the composition of a DNA sequence in short words of length 1 up to length $(m+1)$ where m is the order of the Markov chain. Various results have been published on the word count distribution in Markov chains. The exact distribution can be obtained through its probability generating function (Régnier (2000)) or by using the distributions of both the waiting time till the first occurrence and the inter-arrival time between two occurrences (Chrysaphinou and Papastavridis (1990), Robin and Daudin (1999)). Several approximations have also been proposed for long sequences. The Gaussian distribution proposed in Prum *et al.* (1995) appears to be a good approximation for words (and word families) having a sufficiently large expected count (Robin and Schbath (2001)). For an expectedly rare word $\mathbf{w}$, i.e. one whose count, $N(\mathbf{w})$, satisfies the rare event condition $\mathbb{E}N(\mathbf{w}) = \mathrm{O}(1)$ as the length n of the sequence tends to infinity, Poisson approximations were first proposed (Godbole (1991)), but compound Poisson approximations appear to be better (Schbath (1995a)). This result is based on the fact that (i) occurrences of a given word occur in clumps, (ii) clumps asymptotically form a Poisson process under the rare event condition, and (iii) the numbers of occurrences per clump are asymptotically independent and identically distributed (with a geometric distribution). The compound Poisson distribution reduces to a Poisson distribution for non overlapping words. For an expectedly rare family of words $\mathcal{W}$, the authors of Reinert and Schbath (1998) proposed that the compound Poisson approximation of Schbath (1995a) be used for each count $N(\mathbf{w}), \mathbf{w} \in \mathcal{W}$, and that $N(\mathcal{W}) = \sum_{\mathbf{w}\in\mathcal{W}} N(\mathbf{w})$ be approximated by the sum of independent compound Poisson variables. Using the Chen-Stein method, a bound for the approximation error was given which explicitly depends on the degree of overlaps between the words of the family $\mathcal{W}$. Unfortunately, this error bound does not converge to zero given that there exists a couple of different words $(\mathbf{w}, \mathbf{w}') \in \mathcal{W}^2$ which overlap.

Also using the Chen-Stein method, we here propose a compound Poisson distribution more suitable to approximate the count $N(\mathcal{W})$ of any expectedly rare word family $\mathcal{W}$. The main difference from Reinert and Schbath (1998) is that we will consider clumps composed of overlapping occurrences of $\mathcal{W}$, instead of separately considering clumps of $\mathbf{w}$ for each word $\mathbf{w} \in \mathcal{W}$. We will then directly adapt the method of Schbath (1995a) for a single word to a word family. The difficulty arises from the structure and the occurrence probabilities of such mixed clumps. The idea of studying mixed clumps has been previously introduced by Chryssaphinou *et al.* (2001) to approximate the count of competing renewals, but the authors there focused only on the event that "a mixed clump starts at a given position". Here, we will also have to take into account the exact size of the mixed clumps.

The paper is organized as follows. In Section 5.2, we state the approximation theorem for the count $N(\mathcal{W})$. The parameters of the limiting compound Poisson distribution will be explicitly derived in Section 5.3, which is the high point of the paper. Section 5.4 contains the proof of the approximation theorem, which uses the Chen-Stein method for Poisson approximations. As a corollary, in Section 5.5 we propose a Poisson approximation for both the number of clumps of a word family $\mathcal{W}$ and the number of competing renewals of $\mathcal{W}$ in a Markov chain. Our contribution, compared to the result of Chryssaphinou *et al.* (2001), is in the explicit formula for the parameter of the limiting Poisson distribution. In Section 5.6 we present generalizations to high-order Markov chains and to hidden Markov models.

5.2 Compound Poisson approximation for $N(\mathcal{W})$

In this paper, we consider a random sequence $\mathbf{X} = (X_i)_{i\in\mathbb{Z}}$, generated by a homogeneous stationary Markov chain of order 1 on a finite alphabet $\mathcal{A}$. The generalization to higher order Markov chains is discussed in the conclusion. The stationary distribution on $\mathcal{A}$ is denoted by μ, and $\Pi = [\pi(x,y)]_{x,y\in\mathcal{A}}$ denotes the transition matrix of the model.

Let $\mathcal{W}$ be a family of d different words $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_d\}$ on the alphabet $\mathcal{A}$ with length at least 2. The length of any word $\mathbf{w}$ will be denoted by $|\mathbf{w}|$ and we define h to be the length of the longest word from the family $\mathcal{W}$, $h := \max\{|\mathbf{w}|, \mathbf{w} \in \mathcal{W}\}$. We make two assumptions on the word family $\mathcal{W}$: (i) it is *reduced*, meaning that, $\forall \mathbf{w} \neq \mathbf{w}' \in \mathcal{W}$, $\mathbf{w}$ is not a substring of $\mathbf{w}'$ (this is a usual assumption when studying occurrences of word families and is immediately satisfied if all the words of $\mathcal{W}$ have the same length), (ii) each word $\mathbf{w} \in \mathcal{W}$ has a non zero probability of occurring in $\mathbf{X}$ (this is a natural assumption). Owing to the Markov property, the occurrence probability of a $|\mathbf{w}|$-letter word $\mathbf{w} = w_1 w_2 \cdots w_{|\mathbf{w}|}$ in $\mathbf{X}$ is given by $\mu(w_1)\prod_{j=1}^{|\mathbf{w}|-1}\pi(w_j, w_{j+1})$ and will be simply denoted by $\mu(\mathbf{w})$ in what follows.

Classically, the number of occurrences $N(\mathcal{W})$ of a word family $\mathcal{W}$ in the finite sequence $X_1 \cdots X_n$ is defined as $N(\mathcal{W}) = \sum_{\mathbf{w}\in\mathcal{W}} \sum_{i=1}^{n-|\mathbf{w}|+1} Y_i(\mathbf{w})$ where $Y_i(\mathbf{w})$ is the Bernoulli variable which is equal to 1 if there is an occurrence of $\mathbf{w}$ starting at position i and is equal to 0 otherwise. Note that we will generalize this Bernoulli variable to $Y_i(\mathcal{W})$ which will be equal to 1 if and only if there exists a word from $\mathcal{W}$ occurring at position i (i.e. if and only if there exists an occurrence of $\mathcal{W}$ occurs at position i). Here we will use another decomposition of the count, based on the occurrences of k-clumps. The notion of clump has no sense out of a sequence: a k-clump of $\mathcal{W}$ in a sequence is a maximal set of k overlapping occurrences of $\mathcal{W}$ in this sequence. Therefore, a k-clump of $\mathcal{W}$ occurs at position i in a sequence if and only if a word composed of exactly k overlapping occurrences of the family $\mathcal{W}$ occurs at position i without overlapping any other occurrence of the family $\mathcal{W}$ in this sequence. For example, for the family $\mathcal{W} = \{$`atta,ttat`$\}$, the sequence `gattagcattattac` has a 1-clump of $\mathcal{W}$ at $i = 2$ and a 3-clump of $\mathcal{W}$ at $i = 8$. We should be careful not to forget the occurrence of `ttat` in the 3-clump `attatta`. Therefore, we have

$$N(\mathcal{W}) = \sum_{k\geq 1} k\widetilde{N}_k(\mathcal{W}),$$

where $\widetilde{N}_k(\mathcal{W})$ is the number of k-clumps of $\mathcal{W}$ in $X_1 \cdots X_n$.

For convenience, we will work in the infinite sequence $\mathbf{X}$. We define $\widetilde{Y}_{i,k}(\mathcal{W})$ to be the Bernoulli variable which is equal to 1 if a k-clump of $\mathcal{W}$ occurs at position i in $\mathbf{X}$, and is equal to 0 otherwise, and we let:

$$N^{\infty}(\mathcal{W}) := \sum_{k\geq 1} k\widetilde{N}_k^{\infty}(\mathcal{W}) \qquad \text{with} \qquad \widetilde{N}_k^{\infty}(\mathcal{W}) := \sum_{i=1}^{n-h+1} \widetilde{Y}_{i,k}(\mathcal{W}). \tag{5.1}$$

Note that the count $N^{\infty}(\mathcal{W})$ can differ slightly from the real observed count ,$N(\mathcal{W})$, of $\mathcal{W}$ in the finite sequence $X_1 \cdots X_n$ because clumps of $\mathcal{W}$ in $\mathbf{X}$ may start before position 1 and/or end after position n, and occurrences of $\mathcal{W}$ in $X_1 \cdots X_n$ may start after position $n - h + 1$ if there exists $\mathbf{w} \in \mathcal{W}$ such that $|\mathbf{w}| \neq h$. However, the occurrence of the event $\{N(\mathcal{W}) \neq N^{\infty}(\mathcal{W})\}$ implies that there exists (at least) one occurrence of $\mathcal{W}$ starting at a position in $\{1, \ldots, h-1\}$ or in $\{n-h+2, \ldots, n\}$. This event occurs with probability less than $2(h-1)\mu(\mathcal{W})$, where $\mu(\mathcal{W}) = \mathbb{E}Y_i(\mathcal{W}) = \sum_{\mathbf{w}\in\mathcal{W}} \mu(\mathbf{w})$ denotes the occurrence probability of $\mathcal{W}$ at a given position. Therefore, the total variation distance[*1] between the distribution of these two counts is bounded

[1*] The total variation distance between two discrete distributions P and P' on $\mathbb{N}$ is defined by $\frac{1}{2}\sum_{x\in\mathbb{N}} |P(x) - P'(x)| \leq \min \mathbb{P}(N \neq N')$, where the minimum ranges over all couplings (N, N') of P and P'.

by $2h\mu(\mathcal{W})$, which tends to zero as n tends to ∞ under both $h = \mathrm{o}(n)$ and the rare event condition. The two counts are then asymptotically equivalent - we will focus on $N^\infty(\mathcal{W})$.

We will now use the Chen-Stein theorem as stated in Arratia *et al.* (1990) to bound the total variation distance, d_{tv}, between the distribution of the vector $(\widetilde{Y}_{i,k}(\mathcal{W}))_{i,k}$ and the joint distribution of independent Poisson variables $(Z_{i,k})_{i,k}$ such that $\mathbb{E}Z_{i,k} = \mathbb{E}\widetilde{Y}_{i,k}(\mathcal{W})$, which expectations will be denoted by $\widetilde{\mu}_k(\mathcal{W})$. With $Z_k := \sum_{i=1}^{n-h+1} Z_{i,k}$, the Chen-Stein theorem states that

$$d_{\text{tv}}\Big(\mathcal{D}\big((\widetilde{N}_k^\infty(\mathcal{W}))_k\big), \mathcal{D}\big((Z_k)_k\big)\Big) \leq b_1 + b_2 + b_3, \tag{5.2}$$

where $\mathcal{D}(.)$ denotes the distribution of its arguments and

$$b_1 = \sum_{i=1}^{n-h+1} \sum_{k\geq 1} \sum_{(j,\ell)\in B_{i,k}} \mathbb{E}(\widetilde{Y}_{i,k}(\mathcal{W}))\mathbb{E}(\widetilde{Y}_{j,\ell}(\mathcal{W})) \tag{5.3}$$

$$b_2 = \sum_{i=1}^{n-h+1} \sum_{k\geq 1} \sum_{(j,\ell)\in B_{i,k}\setminus\{(i,k)\}} \mathbb{E}(\widetilde{Y}_{i,k}(\mathcal{W})\widetilde{Y}_{j,\ell}(\mathcal{W})) \tag{5.4}$$

$$b_3 = \sum_{i=1}^{n-h+1} \sum_{k\geq 1} \mathbb{E}\Big|\mathbb{E}\big(\widetilde{Y}_{i,k}(\mathcal{W}) - \widetilde{\mu}_k(\mathcal{W})|\sigma(\widetilde{Y}_{j,\ell}(\mathcal{W}), (j,\ell)\notin B_{i,k})\big)\Big|, \tag{5.5}$$

and where $B_{i,k} \subset \{1,\dots,n-h+1\}\times\mathbb{N}^\star$ is a neighborhood of (i,k). As we will see, for a particular choice of the neighborhood $B_{i,k}$, the quantities b_1, b_2 and b_3 will tend to zero as n tends to ∞, under both $h = \mathrm{o}(n)$ and the rare event condition $\mathbb{E}N(\mathcal{W}) = \mathrm{O}(1)$ (cf. Section 5.4). It means that the process $(\widetilde{N}_k^\infty(\mathcal{W}))_k$ can be approximated by independent Poisson variables $(Z_k)_k$ with respective expectations $\widetilde{\lambda}_k(\mathcal{W}) := \mathbb{E}\widetilde{N}_k^\infty(\mathcal{W}) = (n-h+1)\widetilde{\mu}_k(\mathcal{W})$. From (5.1) and properties of total variation distance, it also means that, under the same asymptotic conditions, the count $N^\infty(\mathcal{W})$ can be approximated by $\sum_{k\geq 1} kZ_k$, which by definition follows the compound Poisson distribution $\mathcal{CP}(\widetilde{\lambda}_k(\mathcal{W}), k\geq 1)$. We can now state the following approximation theorem.

Theorem 5.1 *For every word family $\mathcal{W}$, the total variation distance between the distribution of $N(\mathcal{W})$ and the compound Poisson distribution with parameters $(\widetilde{\lambda}_k(\mathcal{W}))_{k\geq 1}$ such that $\widetilde{\lambda}_k(\mathcal{W}) = (n-h+1)\widetilde{\mu}_k(\mathcal{W})$, with $\widetilde{\mu}_k(\mathcal{W})$ as given in (5.15), is bounded as follows:*

$$d_{tv}\Big(\mathcal{D}\big(N(\mathcal{W})\big), \mathcal{CP}\big(\widetilde{\lambda}_k(\mathcal{W}), k\geq 1\big)\Big) \leq Cnh\mu^2(\mathcal{W}) + C'n\mu(\mathcal{W})|\alpha|^h + 2h\mu(\mathcal{W}), \tag{5.6}$$

where $C > 0$ and $C' > 0$ are two constants that depend only on the transition matrix Π and α is the eigenvalue of Π second largest in modulus (with $|\alpha| < 1$). Therefore, if $n\mu(\mathcal{W}) = O(1)$ and $h = o(n)$, we have

$$d_{tv}\Big(\mathcal{D}\big(N(\mathcal{W})\big), \mathcal{CP}\big(\widetilde{\lambda}_k(\mathcal{W}), k\geq 1\big)\Big) \xrightarrow[n\to\infty]{} 0. \tag{5.7}$$

The proof is done in Section 5.4.

Remark 5.2 *The condition $\mathbb{E}N(\mathbf{w}) = O(1)$ and $h = o(n)$ imply that $n\mu(\mathcal{W}) = O(1)$, which is equivalent to the condition that $\log(n)/|\mathbf{w}| = O(1)$, $\forall \mathbf{w} \in \mathcal{W}$, which in turn means that the compound Poisson approximation holds for families of long enough words.*

The Chen-Stein method usually does not provide an optimal bound. Our concern here is just to show that the bound given by (5.6) converges to zero when n tends to ∞, $h = \mathrm{o}(n)$ and $n\mu(\mathcal{W}) = \mathrm{O}(1)$.

An important task now is to calculate the parameters of the limiting compound Poisson distribution. We do this in the next section, and then provide an expression for $\widetilde{\mu}_k(\mathcal{W})$ which is the occurrence probability of a k-clump of $\mathcal{W}$ occurring at a given position in the infinite sequence $\mathbf{X}$.

5.3 Occurrence probability of a k-clump of $\mathcal{W}$

We first have to look at the typical distances allowed between successive occurrences of $\mathcal{W}$ in a k-clump i.e. k successive overlapping occurrences of $\mathcal{W}$.

5.3.1 Principal periods

For two words $\mathbf{w} = w_1 \cdots w_{|\mathbf{w}|}$ and $\mathbf{w}' = w'_1 \cdots w'_{|\mathbf{w}'|}$ of $\mathcal{W}$, an integer p, $1 \leq p \leq |\mathbf{w}| - 1$, such that $w'_i = w_{i+p}$ for $i = 1, \ldots, |\mathbf{w}| - p$ is called a **period** of $(\mathbf{w}, \mathbf{w}')$. We denote by $\mathcal{P}(\mathbf{w}, \mathbf{w}')$ the set of periods of $(\mathbf{w}, \mathbf{w}')$. For each couple of words $(\mathbf{w}, \mathbf{w}')$ and each period $p \in \mathcal{P}(\mathbf{w}, \mathbf{w}')$, the prefix $\mathbf{w}^{(p)} := w_1 \ldots w_p$ is called a **root** of $(\mathbf{w}, \mathbf{w}')$. The periods of $(\mathbf{w}, \mathbf{w}')$ are then the distances allowed between an occurrence of $\mathbf{w}$ and a further overlapping occurrence of $\mathbf{w}'$. For instance $\mathcal{P}(\mathtt{taca}, \mathtt{acac}) = \{1, 3\}$.

If we now look at the possible distance between **successive** overlapping occurrences of $(\mathbf{w}, \mathbf{w}')$, it appears that some periods are not possible. For instance, the period $p = 3$ of $(\mathtt{taca}, \mathtt{acac})$ is not possible because an occurrence of $\mathtt{taca}$ at position i and an occurrence of $\mathtt{acac}$ at position $i + 3$ implies an other occurrence of $\mathtt{acac}$ in between (in fact at position $i + 1$). More generally, for two words $\mathbf{w}$ and $\mathbf{w}'$ of $\mathcal{W}$, a period $p \in \mathcal{P}(\mathbf{w}, \mathbf{w}')$ is said to be **principal** with respect to $\mathcal{W}$ if, for all $\mathbf{w}^\star \in \mathcal{W}$ and $j \in \mathcal{P}(\mathbf{w}, \mathbf{w}^\star)$, we have $p - j \notin \mathcal{P}(\mathbf{w}^\star, \mathbf{w}')$. This condition simply means that $\mathcal{W}$ cannot occur between an occurrence of $\mathbf{w}$ at a position i and an occurrence of $\mathbf{w}'$ at position $i + p$. We denote by $\mathcal{P}'_{\mathcal{W}}(\mathbf{w}, \mathbf{w}')$ the set of principal periods of $(\mathbf{w}, \mathbf{w}')$ with respect to $\mathcal{W}$. When there will be no ambiguity, we will omit the subscript $\mathcal{W}$. If $\mathcal{W}$ is composed of a unique word $\mathbf{w}$ then the set $\mathcal{P}'_{\{\mathbf{w}\}}(\mathbf{w}, \mathbf{w})$ coincides with the so-called **principal period set** $\mathcal{P}'(\mathbf{w})$, of $\mathbf{w}$ introduced in Schbath (1995a).

A direct consequence of the definition of a principal period is the following lemma.

Lemma 5.3 *(i) An occurrence of $\mathbf{w}' \in \mathcal{W}$ at position i overlaps an earlier occurrence of $\mathcal{W}$ in the sequence if and only if there exists a word $\mathbf{w} \in \mathcal{W}$ and a principal period $p \in \mathcal{P}'(\mathbf{w}, \mathbf{w}')$ such that there is an occurrence of the principal root $\mathbf{w}^{(p)}$ at position $i - p$ in the sequence.*

(ii) In the previous assertion, the word $\mathbf{w}$ and the period p are unique.

Note that the same result holds for a later occurrence of $\mathcal{W}$ and a suffix $\mathbf{w}_{(p)} := w_{|\mathbf{w}|-p+1} \ldots w_{|\mathbf{w}|}$, with $p \in \mathcal{P}'(\mathbf{w}, \mathbf{w}')$.

5.3.2 Computation of $\widetilde{\mu}_k(\mathcal{W})$

We can now describe more explicitly what is a k-clump of $\mathcal{W}$ in a sequence. Consider a word $\mathbf{c}$ composed of exactly k successive overlapping occurrences $\mathbf{w}_{r_1}, \mathbf{w}_{r_2}, \dots, \mathbf{w}_{r_k}$ of the family $\mathcal{W}$, with $r_1, \dots, r_k \in \{1, \dots, d\}$. Then, for $j \in \{1, \dots, k-1\}$, each occurrence $\mathbf{w}_{r_j}$ overlaps the occurrence $\mathbf{w}_{r_{j+1}}$ with the corresponding period $p_j \in \mathcal{P}(\mathbf{w}_{r_j}, \mathbf{w}_{r_{j+1}})$ (see Figure 5.1). Moreover, the periods p_j are necessarily principal because $\mathbf{c}$ has to contain exactly k overlapping occurrences of $\mathcal{W}$. Therefore, the word $\mathbf{c}$ has the form

$$\mathbf{c} = \mathbf{w}_{r_1}^{(p_1)} \dots \mathbf{w}_{r_{k-1}}^{(p_{k-1})} \mathbf{w}_{r_k}. \tag{5.8}$$

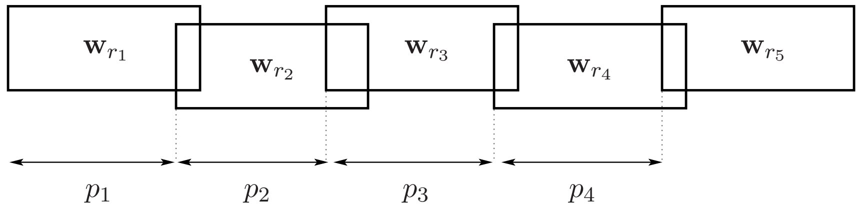


Figure 5.1: Structure of a word composed of exactly five occurrences of $\mathcal{W}$.

To simplify the notations, the first word $\mathbf{w}_{r_1}$ (resp. the second word $\mathbf{w}_{r_2}$, the last word $\mathbf{w}_{r_k}$) of $\mathbf{c}$ is denoted by $\mathbf{u}$ (resp. $\mathbf{v}$, $\mathbf{w}$). We denote by $\mathcal{C}_k(\mathcal{W})$ the set of words of the form (5.8), by $\mathcal{C}_k^{(\mathbf{u};\mathbf{w})}(\mathcal{W})$ the subset of words of $\mathcal{C}_k(\mathcal{W})$ which begin with the word $\mathbf{u}$ and end with $\mathbf{w}$, and by $\mathcal{C}_k^{(\mathbf{u},\mathbf{v})}(\mathcal{W})$ the subset of words of $\mathcal{C}_k(\mathcal{W})$ which have $\mathbf{u}$ and $\mathbf{v}$ as the first two occurrences from $\mathcal{W}$. In the latter notation, when $\mathbf{v}$ is unknown, we replace it by a dot (e.g.we write $\mathcal{C}_k^{(\mathbf{u},.)}(\mathcal{W})$).

A k-clump of $\mathcal{W}$ in $\mathbf{X}$ which begins with $\mathbf{u}$ and ends with $\mathbf{w}$ is then a word $\mathbf{c} \in \mathcal{C}_k^{(\mathbf{u};\mathbf{w})}(\mathcal{W})$ not preceded in $\mathbf{X}$ by any root $\mathbf{u}'^{(p)}$, with $\mathbf{u}' \in \mathcal{W}$ and $p \in \mathcal{P}'(\mathbf{u}', \mathbf{u})$ and not followed by any suffix $\mathbf{w}'_{(q)}$, $\mathbf{w}' \in \mathcal{W}$, $q \in \mathcal{P}'(\mathbf{w}, \mathbf{w}')$. Since the simultaneous occurrence in the sequence of two different elements of $\mathcal{C}_k(\mathcal{W})$ at position i is impossible, using Lemma 5.3, we obtain the following expression for $\widetilde{Y}_{i,k}(\mathcal{W})$:

$$\begin{aligned}
\widetilde{Y}_{i,k}(\mathcal{W}) \;=\; & \sum_{\mathbf{u}\in\mathcal{W}} \sum_{\mathbf{w}\in\mathcal{W}} \sum_{\mathbf{c}\in\mathcal{C}_k^{(\mathbf{u};\mathbf{w})}(\mathcal{W})} \Bigg(Y_i(\mathbf{c}) - \sum_{\mathbf{u}'\in\mathcal{W}} \sum_{p\in\mathcal{P}'(\mathbf{u}',\mathbf{u})} Y_{i-p}(\mathbf{u}'^{(p)}\mathbf{c}) \\
& - \sum_{\mathbf{w}'\in\mathcal{W}} \sum_{q\in\mathcal{P}'(\mathbf{w},\mathbf{w}')} Y_i(\mathbf{c}\mathbf{w}'_{(q)}) \\
& + \sum_{\mathbf{u}'\in\mathcal{W}} \sum_{\mathbf{w}'\in\mathcal{W}} \sum_{p\in\mathcal{P}'(\mathbf{u}',\mathbf{u})} \sum_{q\in\mathcal{P}'(\mathbf{w},\mathbf{w}')} Y_{i-p}(\mathbf{u}'^{(p)}\mathbf{c}\mathbf{w}'_{(q)}) \Bigg).
\end{aligned} \tag{5.9}$$

Thus, by taking the expectation in (5.9), we obtain the equality:

$$\begin{aligned}
\widetilde{\mu}_k(\mathcal{W}) &= \sum_{\mathbf{c}\in\mathcal{C}_k(\mathcal{W})} \mu(\mathbf{c}) - 2 \sum_{\mathbf{c}'\in\mathcal{C}_{k+1}(\mathcal{W})} \mu(\mathbf{c}') + \sum_{\mathbf{c}''\in\mathcal{C}_{k+2}(\mathcal{W})} \mu(\mathbf{c}'') \\
&= p_k(\mathcal{W}) - 2p_{k+1}(\mathcal{W}) + p_{k+2}(\mathcal{W}),
\end{aligned} \tag{5.10}$$

where $p_k(\mathcal{W})$ and $p_k^{(\mathbf{u},.)}(\mathcal{W})$ respectively denote the occurrence probability of a word of $\mathcal{C}_k(\mathcal{W})$ and a word of $\mathcal{C}_k^{(\mathbf{u},.)}(\mathcal{W})$ occurring at a given position. The expression for $\widetilde{\mu}_k(\mathcal{W})$ can thus be deduced from the one of the $p_k(\mathcal{W})$. The computation of $p_k(\mathcal{W})$ is done recursively. For all $k \geq 1$ and $\mathbf{u} = u_1 \dots u_{|\mathbf{u}|} \in \mathcal{W}$,

$$\begin{aligned}
p_1^{(\mathbf{u},.)}(\mathcal{W}) &= \mu(\mathbf{u}) \\
p_{k+1}^{(\mathbf{u},.)}(\mathcal{W}) &= \sum_{\mathbf{v}\in\mathcal{W}} \sum_{\mathbf{c}\in\mathcal{C}_{k+1}^{(\mathbf{u},\mathbf{v})}(\mathcal{W})} \mu(\mathbf{c}) \\
&= \sum_{\mathbf{v}\in\mathcal{W}} \sum_{p\in\mathcal{P}'(\mathbf{u},\mathbf{v})} \sum_{\mathbf{c}'\in\mathcal{C}_k^{(\mathbf{v},.)}(\mathcal{W})} \mu(\mathbf{u}^{(p)}\mathbf{c}') \\
&= \sum_{\mathbf{v}\in\mathcal{W}} \frac{1}{\mu(v_1)} \sum_{p\in\mathcal{P}'(\mathbf{u},\mathbf{v})} \mu(\mathbf{u}^{(p+1)}) \sum_{\mathbf{c}'\in\mathcal{C}_k^{(\mathbf{v},.)}(\mathcal{W})} \mu(\mathbf{c}') \\
&= \sum_{\mathbf{v}\in\mathcal{W}} A_{\mathbf{u},\mathbf{v}} p_k^{(\mathbf{v},.)}(\mathcal{W}), \qquad (5.11)
\end{aligned}$$

where $A_{\mathbf{u},\mathbf{v}}$ is the probability that an occurrence of $\mathbf{v} = v_1 \cdots v_{|\mathbf{v}|}$ overlaps a previous occurrence of $\mathbf{u}$ in the sequence and that there are no other occurrences of $\mathcal{W}$ in between:

$$A_{\mathbf{u},\mathbf{v}} = \frac{\mu(u_1)}{\mu(v_1)} \sum_{p\in\mathcal{P}'(\mathbf{u},\mathbf{v})} \prod_{t=1}^{p} \pi(u_t, u_{t+1}). \qquad (5.12)$$

Therefore, if we introduce the vectorial notations $\vec{p}_k(\mathcal{W})$ for the vector $[p_k^{(\mathbf{u},.)}(\mathcal{W})]_{\mathbf{u}\in\mathcal{W}}$ and A for the matrix $[A_{\mathbf{u},\mathbf{v}}]_{\mathbf{u},\mathbf{v}\in\mathcal{W}}$, (5.11) can be written as follows: $\forall k \geq 1$, $\vec{p}_{k+1}(\mathcal{W}) = A\vec{p}_k(\mathcal{W})$. Similarly, we have $\vec{p}_1(\mathcal{W}) = \vec{\mu}(\mathcal{W}) := [\mu(\mathbf{w})]_{\mathbf{w}\in\mathcal{W}}$, leading to

$$\vec{p}_k(\mathcal{W}) = A^{k-1}\vec{\mu}(\mathcal{W}). \qquad (5.13)$$

Denoting by $||.||_1$ the 1-norm of $\mathbb{R}^d$ defined by $||\vec{z}||_1 = \sum_{r=1}^d |z_r|$ for all$\vec{z} = (z_1, \dots, z_d) \in \mathbb{R}^d$,, we can conclude that

$$\begin{aligned}
p_k(\mathcal{W}) &= ||\vec{p}_k(\mathcal{W})||_1 \\
&= ||A^{k-1}\vec{\mu}(\mathcal{W})||_1. \qquad (5.14)
\end{aligned}$$

Combining relations (5.10) and (5.14) yields our final expression of $\widetilde{\mu}_k(\mathcal{W})$:

$$\widetilde{\mu}_k(\mathcal{W}) = ||A^{k-1}(I - A)^2\vec{\mu}(\mathcal{W})||_1.$$

This establishes the following proposition.

Proposition 5.4 *For all family $\mathcal{W}$, the occurrence probability of a k-clump of $\mathcal{W}$ is given by*

$$\widetilde{\mu}_k(\mathcal{W}) = ||A^{k-1}(I - A)^2\vec{\mu}(\mathcal{W})||_1, \qquad (5.15)$$

where I is the identity matrix, A is the matrix of coefficients $[A_{\mathbf{u},\mathbf{v}}]_{\mathbf{u},\mathbf{v}\in\mathcal{W}}$ defined in (5.12), $\vec{\mu}(\mathcal{W})$ is the vector $[\mu(\mathbf{w})]_{\mathbf{w}\in\mathcal{W}}$, and $||.||_1$ is the 1-norm of $\mathbb{R}^d$.

Remark 5.5 1. *Theorem* 1 *and Proposition 5.4 generalize the Theorem 13 of Schbath (1995a): indeed, for a single word* $\mathcal{W} = \{\mathbf{w}\}$, *(5.15) reduces to* $\widetilde{\mu}_k(\mathbf{w}) = a_{\mathbf{w}}^{k-1}(1-a_{\mathbf{w}})^2\mu(\mathbf{w})$, *where* $a_{\mathbf{w}}$ *is the occurrence probability of two successive overlapping occurrences of* $\mathbf{w}$ *and is given by* $a(\mathbf{w}) = \sum_{p\in\mathcal{P}'(\mathbf{w})}\prod_{t=1}^{p}\pi(w_t, w_{t+1})$ *with* $\mathcal{P}'(\mathbf{w}) := \mathcal{P}'_{\{\mathbf{w}\}}(\mathbf{w},\mathbf{w})$.

2. *For a family* $\mathcal{W}$ *such that, for all* $\mathbf{w} \neq \mathbf{w}' \in \mathcal{W}$, $\mathbf{w}$ *does not overlap* $\mathbf{w}'$ *(i.e.* $\mathcal{P}(\mathbf{w},\mathbf{w}') = \emptyset$*),* A *is a diagonal matrix, and we find that* $\widetilde{\mu}_k(\mathcal{W}) = \sum_{\mathbf{w}\in\mathcal{W}} a_{\mathbf{w}}^{k-1}(1-a_{\mathbf{w}})^2\mu(\mathbf{w})$, *as in Reinert and Schbath (1998).*

3. *From (5.10), we can moreover show that*

$$\sum_{k\geq 1} k\widetilde{\mu}_k(\mathcal{W}) = \mu(\mathcal{W}) \tag{5.16}$$

$$\sum_{k\geq 1} \widetilde{\mu}_k(\mathcal{W}) = ||(I-A)\vec{\mu}(\mathcal{W})||_1. \tag{5.17}$$

5.4 Proof of the approximation theorem

To prove Theorem 5.1, we first have to choose the neighborhoods $B_{i,k}$ for all $(i,k) \in I$, where $I := \{1,\ldots,n-h+1\} \times \mathbb{N}^\star$, and then to bound the three quantities b_1, b_2 and b_3 defined respectively by (5.3),(5.4) and (5.5). To do so, we will adapt the setup presented in Schbath (1995a) for a single word.

5.4.1 Choice of the neighborhood $B_{i,k}$

For each $(i,k) \in I$, we define a set $Z(i,k) \subset \mathbb{Z}$ which contains all the indices j of the letters X_j used in the definition of $\widetilde{Y}_{i,k}(\mathcal{W})$. We can take $Z(i,k) = \{s \in \mathbb{Z}$ such that $i-h \leq s \leq i+(k+1)h\}$, because the length of a k-clump is less than kh and we have to know the $h-1$ letters before and after the clump to ensure that it does not overlap other occurrences. We now define the neighborhood of (i,k) as the set of $(j,\ell) \in I$ such that $Z(i,k)$ and $Z(j,\ell)$ are separated by at most h positions:

$$B_{i,k} = \{(j,\ell) \in I \text{ such that } -(\ell+3)h \leq j-i \leq (k+3)h\}.$$

This implies that if $\widetilde{Y}_{i,k}(\mathcal{W}) = \widetilde{Y}_{j,\ell}(\mathcal{W}) = 1$ with $(j,\ell) \notin B_{i,k}$, then the two clumps will be separated by more than $3h$ letters.

5.4.2 Bounding b_1

From definition (5.3) we have

$$\begin{aligned} b_1 &= \sum_{i=1}^{n-h+1}\sum_{k\geq 1}\sum_{(j,\ell)\in B_{i,k}} \mathbb{E}(\widetilde{Y}_{i,k}(\mathcal{W}))\mathbb{E}(\widetilde{Y}_{j,\ell}(\mathcal{W})) \\ &\leq \sum_{i=1}^{n-h+1}\sum_{k\geq 1}\sum_{\ell\geq 1}\sum_{j=i-(\ell+3)h}^{i+(k+3)h} \widetilde{\mu}_k(\mathcal{W})\widetilde{\mu}_\ell(\mathcal{W}). \end{aligned}$$

Let $\widetilde{\mu}(\mathcal{W})$ be the probability of a clump of $\mathcal{W}$ occurring at a given position; it satisfies $\widetilde{\mu}(\mathcal{W}) = \sum_{k\geq 1}\widetilde{\mu}_k(\mathcal{W}) \leq \mu(\mathcal{W})$. Using the symmetry between i and j and between k and ℓ, and (5.16), we can write

$$\begin{aligned} b_1 &\leq 2\widetilde{\mu}(\mathcal{W}) \sum_{i=1}^{n-h+1} \sum_{k\geq 1}((k+3)h+1)\widetilde{\mu}_k(\mathcal{W}) \\ &\leq 2(n-h+1)\widetilde{\mu}(\mathcal{W})\Big(\big[\mu(\mathcal{W})+3\widetilde{\mu}(\mathcal{W})\big]h+\widetilde{\mu}(\mathcal{W})\Big) \\ &\leq 10nh\mu^2(\mathcal{W}). \end{aligned} \tag{5.18}$$

The last inequality is obtained simply by bounding $\widetilde{\mu}(\mathcal{W})$ by $\mu(\mathcal{W})$.

5.4.3 Bounding b_2

From definition (5.4), we have

$$b_2 = \sum_{i=1}^{n-h+1} \sum_{k\geq 1} \sum_{(j,\ell)\in B_{i,k}\setminus\{(i,k)\}} \mathbb{E}(\widetilde{Y}_{i,k}(\mathcal{W})\widetilde{Y}_{j,\ell}(\mathcal{W}))$$

Since two clumps of different sizes cannot occur at the same position, the term corresponding to $i=j$ disappears in the sum, and, again by symmetry, we obtain

$$b_2 \leq 2\sum_{i=1}^{n-h+1} \sum_{k\geq 1}\sum_{\ell\geq 1} \sum_{j=i+1}^{i+(k+3)h} \mathbb{E}(\widetilde{Y}_{i,k}(\mathcal{W})\widetilde{Y}_{j,\ell}(\mathcal{W})).$$

Let $\widetilde{Y}_j(\mathcal{W}) = \sum_{\ell\geq 1}\widetilde{Y}_{j,\ell}(\mathcal{W})$ denote a Bernoulli variable that is equal to 1 if a clump of $\mathcal{W}$ occurs at position j and is equal to 0 otherwise. Since $\widetilde{Y}_{i,k}(\mathcal{W}) = \sum_{\mathbf{c}\in\mathcal{C}_k(\mathcal{W})}\widetilde{Y}_{i,k}(\mathcal{W})Y_i(\mathbf{c})$, we have

$$\begin{aligned} b_2 &\leq 2\sum_{i=1}^{n-h+1} \sum_{k\geq 1} \sum_{j=i+1}^{i+(k+3)h} \mathbb{E}(\widetilde{Y}_{i,k}(\mathcal{W})\widetilde{Y}_j(\mathcal{W})), \\ &\leq 2\sum_{i=1}^{n-h+1} \sum_{k\geq 1} \sum_{\mathbf{c}\in\mathcal{C}_k(\mathcal{W})} \sum_{j=i+1}^{i+(k+3)h} \mathbb{E}\big(\widetilde{Y}_{i,k}(\mathcal{W})Y_i(\mathbf{c})\widetilde{Y}_j(\mathcal{W})\big). \end{aligned}$$

Since a clump of length $|\mathbf{c}|$ which begins at position i cannot overlap a clump starting at position j, $i+1\leq j< i+|\mathbf{c}|$, and since $\widetilde{Y}_j(\mathcal{W}) \leq Y_j(\mathcal{W})$, it follows that

$$\begin{aligned} b_2 &\leq 2\sum_{i=1}^{n-h+1} \sum_{k\geq 1} \sum_{\mathbf{c}\in\mathcal{C}_k(\mathcal{W})} \sum_{j=i+|\mathbf{c}|}^{i+(k+3)h} \mathbb{E}\big(\widetilde{Y}_{i,k}(\mathcal{W})Y_i(\mathbf{c})\widetilde{Y}_j(\mathcal{W})\big) \\ &\leq 2\sum_{i=1}^{n-h+1} \sum_{k\geq 1} \sum_{\mathbf{c}\in\mathcal{C}_k(\mathcal{W})} \sum_{j=i+|\mathbf{c}|+h}^{i+(k+3)h} \mathbb{E}\big(\widetilde{Y}_{i,k}(\mathcal{W})Y_i(\mathbf{c})Y_j(\mathcal{W})\big) \\ &\quad +2\sum_{i=1}^{n-h+1} \sum_{k\geq 1} \sum_{\mathbf{c}\in\mathcal{C}_k(\mathcal{W})} \sum_{j=i+|\mathbf{c}|}^{i+|\mathbf{c}|+h-1} \mathbb{E}\big(\widetilde{Y}_{i,k}(\mathcal{W})Y_i(\mathbf{c})Y_j(\mathcal{W})\big). \end{aligned}$$

The first term (resp. the second term) in the right-hand side is denoted by b_{21} (resp. b_{22}). Let us bound b_{21}. Note that the random variable $\widetilde{Y}_{i,k}(\mathcal{W})Y_i(\mathbf{c})$ only involves the letters $X_{i-h+1}, \dots, X_{i+|\mathbf{c}|+h-1}$ whereas $Y_j(\mathcal{W})$ involves $X_j, \dots, X_{j+h-1}$. Therefore, for every position j which satisfies $j \geq i + |\mathbf{c}| + h$, the Markov property yields

$$\mathbb{E}\big(\widetilde{Y}_{i,k}(\mathcal{W})Y_i(\mathbf{c})Y_j(\mathcal{W})\big) \leq \frac{\mu(\mathcal{W})}{\mu_{\min}}\mathbb{E}\big(\widetilde{Y}_{i,k}(\mathcal{W})Y_i(\mathbf{c})\big),$$

where $\mu_{\min} = \min_{\mathbf{w}\in\mathcal{W}} \mu(w_1) > 0$. Since the sum over j contains fewer than $(k+2)h$ terms, we get

$$\begin{aligned}
b_{21} &\leq 2(n-h+1)\frac{\mu(\mathcal{W})}{\mu_{\min}}\sum_{k\geq 1}(k+2)h\widetilde{\mu}_k(\mathcal{W}) \\
&\leq 2(n-h+1)\frac{\mu(\mathcal{W})}{\mu_{\min}}\big(\mu(\mathcal{W}) + 2\widetilde{\mu}(\mathcal{W})\big)h \\
&\leq \frac{6nh}{\mu_{\min}}\mu^2(\mathcal{W}). \qquad (5.19)
\end{aligned}$$

To bound b_{22}, we write $\mathbb{E}\big(\widetilde{Y}_{i,k}(\mathcal{W})Y_i(\mathbf{c})Y_j(\mathcal{W})\big) \leq \mathbb{E}\big(\widetilde{Y}_i(\mathcal{W})Y_i(\mathbf{c})Y_j(\mathcal{W})\big)$ and we note that the random variable $\widetilde{Y}_i(\mathcal{W})Y_i(\mathbf{c})$ involves the letters $X_{i-h+1}, \dots, X_{i+|\mathbf{c}|-1}$ whereas $Y_j(\mathcal{W})$ involves $X_j, \dots, X_{j+h-1}$. Therefore, for every position j which satisfies $j \geq i + |\mathbf{c}|$, we have

$$\mathbb{E}\big(\widetilde{Y}_i(\mathcal{W})Y_i(\mathbf{c})Y_j(\mathcal{W})\big) \leq \frac{\mu(\mathcal{W})}{\mu_{\min}}\mathbb{E}\big(\widetilde{Y}_i(\mathcal{W})Y_i(\mathbf{c})\big).$$

Thus, we derive the following bound for b_{22}:

$$\begin{aligned}
b_{22} &\leq \frac{2(n-h+1)h}{\mu_{\min}}\mu(\mathcal{W})\sum_{k\geq 1}\sum_{\mathbf{c}\in\mathcal{C}_k(\mathcal{W})}\mathbb{E}\big(\widetilde{Y}_i(\mathcal{W})Y_i(\mathbf{c})\big) \\
&\leq \frac{2nh}{\mu_{\min}}\mu^2(\mathcal{W}). \qquad (5.20)
\end{aligned}$$

Indeed,

$$\begin{aligned}
&\sum_{k\geq 1}\sum_{\mathbf{c}\in\mathcal{C}_k(\mathcal{W})}\mathbb{E}\big(\widetilde{Y}_i(\mathcal{W})Y_i(\mathbf{c})\big) \qquad (5.21)\\
&= \sum_{k\geq 1}\mathbb{P}(\text{a } K\text{-clump of } \mathcal{W} \text{ with } K \geq k \text{ starts at position } i) \\
&= \sum_{k\geq 1}\sum_{K\geq k}\widetilde{\mu}_K(\mathcal{W}) = \sum_{K\geq 1}K\widetilde{\mu}_K(\mathcal{W}) = \mu(\mathcal{W}).
\end{aligned}$$

Finally, combining equations (5.19) and (5.20) leads to

$$b_2 \leq \frac{8nh}{\mu_{\min}}\mu^2(\mathcal{W}). \qquad (5.22)$$

5.4.4 Bounding b_3

From definition (5.5), we have

$$b_3 = \sum_{i=1}^{n-h+1} \sum_{k\geq 1} \mathbb{E}\Big|\mathbb{E}\big(\widetilde{Y}_{i,k}(\mathcal{W}) - \widetilde{\mu}_k(\mathcal{W})|\sigma(\widetilde{Y}_{j,\ell}(\mathcal{W}), (j,\ell) \notin B_{i,k})\big)\Big|.$$

We denote by $\mathcal{C}'_k$ the set of the words $\mathbf{rcs}$ such that $\mathbf{c} \in \mathcal{C}_k$, $|\mathbf{r}| = |\mathbf{s}| = h$ and $\mathbf{c}$ is a k-clump of $\mathcal{W}$ in the sequence $\mathbf{rcs}$. An occurrence of a word of $\mathcal{C}'_k$ is then equivalent to an occurrence of a k-clump of $\mathcal{W}$: $\widetilde{Y}_{i,k}(\mathcal{W}) = \sum_{\mathbf{rcs}\in\mathcal{C}'_k} Y_{i-h}(\mathbf{rcs})$. Moreover, for all $\mathbf{c} \in \mathcal{C}_k$, we deduce from the definition of the neighborhood $B_{i,k}$ that

$$\sigma(\widetilde{Y}_{j,\ell}(\mathcal{W}), (j,\ell) \notin B_{i,k}) \subset \sigma(\ldots, X_{i-2h-1}, X_{i-2h}, X_{i+|\mathbf{c}|+2h}, X_{i+|\mathbf{c}|+2h+1}, \ldots).$$

Therefore, owing to the Markov property, we have

$$\begin{aligned} b_3 &\leq \sum_{i=1}^{n-h+1} \sum_{k\geq 1} \sum_{\mathbf{rcs}\in\mathcal{C}'_k} \mathbb{E}\Big|\mathbb{E}\big(Y_{i-h}(\mathbf{rcs}) - \mu(\mathbf{rcs})|\sigma(\ldots, X_{i-2h}, X_{i+|\mathbf{c}|+2h}, \ldots)\big)\Big| \\ &\leq \sum_{i=1}^{n-h+1} \sum_{k\geq 1} \sum_{\mathbf{rcs}\in\mathcal{C}'_k} \mathbb{E}\Big|\mathbb{E}\big(Y_{i-h}(\mathbf{rcs}) - \mu(\mathbf{rcs})|X_{(i-h)-h}, X_{(i-h)+|\mathbf{rcs}|+h}\big)\Big|. \end{aligned}$$

Now we use the following result, proved by Schbath (1995b): for all word $\mathbf{w}$ and all integers j and t,

$$\mathbb{E}\big|\mathbb{E}(Y_j(\mathbf{w}) - \mu(\mathbf{w})|X_{j-t}, X_{j+|\mathbf{w}|+t})\big| \leq C'\mu(\mathbf{w})|\alpha|^t, \tag{5.23}$$

where C' is a positive constant that depend only on the matrix Π, and α is the eigenvalue of the matrix Π second largest in modulus (with $|\alpha| < 1$). This leads to

$$b_3 \leq \sum_{i=1}^{n-h+1} \sum_{k\geq 1} \sum_{\mathbf{rcs}\in\mathcal{C}'_k} C'\mu(\mathbf{rcs})|\alpha|^h.$$

Finally, the equality $\sum_{k\geq 1}\sum_{\mathbf{rcs}\in\mathcal{C}'_k} \mu(\mathbf{rcs}) = \widetilde{\mu}(\mathcal{W})$ yields

$$\begin{aligned} b_3 &\leq C'(n-h+1)|\alpha|^h\widetilde{\mu}(\mathcal{W}) \\ &\leq C'n\mu(\mathcal{W})|\alpha|^h. \end{aligned} \tag{5.24}$$

Inequalities (5.18), (5.22) and (5.24) establish Theorem 5.1.

5.5 Clumps and competing renewals

When counting the occurrences of a word or a word family in a finite sequence $X_1 \cdots X_n$, one may be interested in counting only non overlapping occurrences, for instance clumps or renewals. A renewal can be defined as follows: an occurrence is a renewal if and only if either it is the first occurrence or it does not overlap a previous renewal. For a word family, they are called *competing* renewals. Various results have been obtained for the distribution of the number of clumps and the number of competing renewals (see Lothaire (2005), Chapter 6 and references therein). New Poisson approximations directly follows from Theorem 5.1.

First, inequalities (5.2), (5.18), (5.22) and (5.24) lead to

$$d_{\text{tv}}\big(\mathcal{D}(\widetilde{N}^{\infty}(\mathcal{W})), \mathcal{P}(\widetilde{\lambda})\big) \leq Cnh\mu^2(\mathcal{W}) + C'n\mu(\mathcal{W})|\alpha|^h, \tag{5.25}$$

where $\widetilde{N}^{\infty}(\mathcal{W}) := \sum_{k\geq 1} \widetilde{N}_k^{\infty}(\mathcal{W})$, $\mathcal{P}(.)$ denotes the Poisson distribution and, using (5.17),

$$\widetilde{\lambda} = \mathbb{E}(\widetilde{N}^{\infty}(\mathcal{W})) = (n-h+1)||(I-A)\vec{\mu}(\mathcal{W})||_1.$$

Moreover, $\widetilde{N}^{\infty}(\mathcal{W})$ asymptotically has the same distribution as the number, $\widetilde{N}(\mathcal{W})$, of clumps of $\mathcal{W}$ in $X_1 \cdots X_n$: $\mathbb{P}(\widetilde{N}^{\infty}(\mathcal{W}) \neq \widetilde{N}(\mathcal{W})) \leq h\mu(\mathcal{W})$ (by same argument as for N^{∞}). Therefore, under both $h = \mathrm{o}(n)$ and the rare condition $\mathbb{E}N(\mathbf{w}) = \mathrm{O}(1)$, the total variation distance between the distribution of $\widetilde{N}(\mathcal{W})$ and the Poisson distribution $\mathcal{P}(\widetilde{\lambda})$ tends to zero as n tends to ∞.

Second, it can be shown that the distribution of the number, $R(\mathcal{W})$, of competing renewals of $\mathcal{W}$ is asymptotically identical to that of the number of clumps:

$$d_{\text{tv}}\big(\mathcal{D}(R(\mathcal{W})), \mathcal{D}(\widetilde{N}(\mathcal{W}))\big) \leq \mathbb{P}(R(\mathcal{W}) \neq \widetilde{N}(\mathcal{W})) \leq \frac{1}{\mu_{\min}} nh\mu^2(\mathcal{W}), \tag{5.26}$$

where, recall, $\mu_{\min} = \min_{\mathbf{w}\in\mathcal{W}} \mu(w_1) > 0$. Indeed, we note that if all the clumps are such that the occurrence of $\mathcal{W}$ they start with overlaps the occurrence of $\mathcal{W}$ they end with, then $R(\mathcal{W}) = \widetilde{N}(\mathcal{W})$. Thus, if $R(\mathcal{W}) \neq \widetilde{N}(\mathcal{W})$, there exists (at least) one clump whose first and last occurrences from $\mathcal{W}$ do not overlap. Let i be the position of such a clump and let $\mathbf{u}$ be the occurrence from $\mathcal{W}$ it starts with. Then an occurrence of $\mathbf{u}$ starts at position i and an occurrence of $\mathcal{W}$ starts between positions $i + |\mathbf{u}|$ and $i + |\mathbf{u}| + h - 1$; this occurs with probability $h\mu(\mathbf{u})\mu(\mathcal{W})/\mu_{\min}$. Summing over $i \in \{1, \ldots, n-h+1\}$ and $\mathbf{u} \in \mathcal{W}$ leads to inequality (5.26). Owing to the triangular inequality, we then obtain the following Poisson approximation for the number of competing renewals

$$d_{\text{tv}}\big(\mathcal{D}(R(\mathcal{W})), \mathcal{P}(\widetilde{\lambda})\big) = \mathrm{O}(nh\mu^2(\mathcal{W}) + n\mu(\mathcal{W})|\alpha|^h + h\mu(\mathcal{W})). \tag{5.27}$$

If $\mathbb{E}N(\mathbf{w}) = \mathrm{O}(1)$ and $h = \mathrm{o}(n)$, then the total variation distance between the distribution of $R(\mathcal{W})$ and the Poisson distribution $\mathcal{P}(\widetilde{\lambda})$ tends to zero as n tends to ∞. This Poisson distribution is in fact very close to the natural limiting Poisson distribution with parameter $\mathbb{E}R(\mathcal{W})$ proposed by Chryssaphinou *et al.* (2001) because these parameters are asymptotically equivalent under the rare condition and $h = \mathrm{o}(n)$. However, in practice calculating $\mathbb{E}R(\mathcal{W})$ requires solving a system of equations whereas the expression for $\widetilde{\lambda}$ is explicit.

5.6 Generalizations and Conclusion

We have provided a new compound Poisson distribution with explicit parameters to approximate the count of overlapping occurrences of a word family in a stationary Markov chain of length n. The error of approximation converges to zero given that the word family $\mathcal{W}$ is expectedly rare ($\mathbb{E}N(\mathcal{W}) = \mathrm{O}(1)$) and the maximal word length is of order less than n.

Our results can easily be extended to the case of a Markov chain of order m, $2 \leq m \leq \min\{|\mathbf{w}|, \mathbf{w} \in \mathcal{W}\} - 1$. It suffices to consider the sequence $\mathbf{X}^*$ obtained by letting $X_i^* :=$

$X_iX_{i+1}\cdots X_{i+m-1}$, which is a Markov chain of order 1 on the alphabet $\mathcal{A}^* := \mathcal{A}^m$. Moreover, an occurrence of $\mathcal{W}$ in $\mathbf{X}$ corresponds to an occurrence of $\mathcal{W}^*$ in $\mathbf{X}^*$ and vice versa, where $\mathcal{W}^*$ is the word family $\mathcal{W}$ written on the new alphabet $\mathcal{A}^*$. The parameters of the limiting compound Poisson distribution will then be $||A_{(m)}^{k-1}(I - A_{(m)})^2\vec{\mu}(\mathcal{W})||_1$, where $A_{(m)}$ is the matrix whose $(\mathbf{u}, \mathbf{v})$-indexed coefficient is given by

$$\frac{\mu(u_1\cdots u_m)}{\mu(v_1\cdots v_m)} \sum_{\substack{p\in\mathcal{P}'(\mathbf{u},\mathbf{v})\\ p\leq|\mathbf{u}|-m}} \prod_{t=1}^{p} \pi(u_t\dots u_{t+m-1}, u_{t+m}),$$

and $\pi(\cdot,\cdot)$ and $\mu(\cdot)$ respectively denote the transition probabilities and the stationary distribution of the model. This compound Poisson distribution has been included in the *R'MES* software[*2], used to find exceptional motifs in DNA sequences.

Our compound Poisson approximation for the count of any rare word family in a Markov chain, together with a Gaussian approximation or the exact distribution, is extremely useful when one models the sequence as a hidden Markov chain. Indeed, a hidden Markov chain $(\mathbf{X}, \mathbf{S})$ on the alphabet $\mathcal{A}$ with state space $\{1,\dots,s\}$ can be written as a one-order Markov chain $\overline{\mathbf{X}}$ on the alphabet $\mathcal{A}\times\{1,\dots,s\}$ and an occurrence of a given word $\mathbf{w}$ in $\mathbf{X}$ corresponds to an occurrence of a word family $\overline{\mathcal{W}}$ in $\overline{\mathbf{X}}$. For instance, if there are two states 1 and 2, the word family $\overline{\mathcal{W}}$ associated with $\mathbf{w} = \mathtt{aca}$ is $\{\mathtt{a}_1\mathtt{c}_1\mathtt{a}_1, \mathtt{a}_1\mathtt{c}_1\mathtt{a}_2, \mathtt{a}_1\mathtt{c}_2\mathtt{a}_1, \mathtt{a}_2\mathtt{c}_1\mathtt{a}_1, \mathtt{a}_1\mathtt{c}_2\mathtt{a}_2, \mathtt{a}_2\mathtt{c}_1\mathtt{a}_2, \mathtt{a}_2\mathtt{c}_2\mathtt{a}_1, \mathtt{a}_2\mathtt{c}_2\mathtt{a}_2\}$ where $\mathtt{a}_j$ (resp. $\mathtt{c}_j$) stands for the letter $\mathtt{a}$ (resp. $\mathtt{c}$) in state j.

Acknowledgements

The authors thank an anonymous reviewer for his/her helpful comments. This work has been supported by the French Action Concertée Incitative IMPBio.

[2*] http://genome.jouy.inra.fr/ssb/rmes/

Chapitre 6

Mise en oeuvre des lois $\mathcal{CP}_{\text{uni}}$ et $\mathcal{CP}_{\text{bic}}$ pour approcher la loi du comptage

Dans ce chapitre nous mettons en oeuvre les approximations de la loi du comptage d'un mot par les lois de Poisson composées $\mathcal{CP}_{\text{uni}}$ et $\mathcal{CP}_{\text{bic}}$ sur des données. La section 6.2 examine le comportement de ces approximations sur des données simulées, alors que la section 6.3 traite le cas de données réelles. Quelques éléments sur l'implémentation de ces approximations sont donnés dans la section 6.1 et une conclusion est donnée en section 6.4.

6.1 R'MES : logiciel pour la Recherche de Motifs Exceptionnels dans les Séquences

R'MES[1] est un logiciel dédié à la recherche de motifs exceptionnels dans les séquences d'ADN. Ce programme fonctionne de la façon suivante :

– il prend (principalement) en entrée : la séquence d'ADN $\mathbf{X}$, la longueur de motifs h – ou alternativement une liste de motifs –, un ordre de chaîne de Markov m et la méthode statistique utilisée pour le calcul de la loi du comptage (exacte, approximation gaussienne, approximation de Poisson composée).

– il retourne en sortie : la liste de motifs avec pour chaque motif : le comptage observé, le comptage attendu (l'espérance du comptage dans le modèle), le *score* du motif (défini comme la p-value renormalisée par une tranformation quantile d'une loi normale centrée réduite).

Durant ma thèse, j'ai participé à l'élaboration de la version 3 de ce logiciel (cf. Hoebeke and Schbath (2006)), pour l'implémentation de :

- la méthode exacte de Robin *et al.* (2003a) (dans le cas où un motif peut être éventuellement une famille de mots),
- l'approximation de Poisson composée pour les familles de mots rares (cf. chapitre 5).

Cette version 3 traite uniquement le cas des modèles de Markov **homogènes**.

Sous l'encadrement de Mark Hoebeke et de Sophie Schbath, j'ai implémenté les approximations hétérogènes à segmentation fixée $\mathcal{CP}_{\text{uni}}$ et $\mathcal{CP}_{\text{bic}}$ (définies respectivement dans les sections

[1] http ://genome.jouy.inra.fr/ssb/rmes/

3.4.2 et 3.4.3 du chapitre 3) qui figureront dans la prochaine version de R'MES. Le programme prend alors une donnée supplémentaire en entrée, qui est la segmentation **s**, et retourne les statistiques de chaque mot dans un modèle **hétérogène**, calculées selon les approximations $\mathcal{CP}_{\text{uni}}$ ou $\mathcal{CP}_{\text{bic}}$. Pour des raisons de temps de calcul, l'approximation utilisée par défaut est $\mathcal{CP}_{\text{bic}}$ pour les mots non-recouvrants et $\mathcal{CP}_{\text{uni}}$ pour le cas recouvrant (cette dernière approximation est suffisante lorsque la séquence n'est pas "trop" segmentée).

Avec l'aide de cette nouvelle version de R'MES, on cherche dans la suite à mettre en pratique les approximations $\mathcal{CP}_{\text{uni}}$ et $\mathcal{CP}_{\text{bic}}$ à la fois sur des données simulées et réelles.

6.2 Mise en oeuvre sur des données simulées

L'objectif ici est d'évaluer les qualités des approximations de la loi du comptage par $\mathcal{CP}_{\text{uni}}$ et $\mathcal{CP}_{\text{bic}}$, et de les comparer à la qualité de l'approximation homogène du chapitre 2 lorsque les séquences sont simulées selon un modèle hétérogène. La qualité des différentes approximations sera évaluée par comparaison à la loi empirique du comptage obtenue par simulation (soit en distance en variation totale, soit à l'oeil nu avec un graphe des densités).

Rappelons qu'il a été montré en théorie que pour des mots rares **non-recouvrants**, l'approximation par $\mathcal{CP}_{\text{bic}}$ est valide (erreur en variation totale tendant vers 0) dès lors que la longueur minimum des plages $L_{\min}$ est plus grande que h. Pour des mots rares **recouvrants**, la situation est plus complexe : lorsque le nombre de ruptures ρ est faible (c'est-à-dire $\rho h \mu_{\max}(\mathbf{w}) = o(1)$, ou plus grossièrement $h\rho = o(n)$), les approximations par $\mathcal{CP}_{\text{uni}}$ et par $\mathcal{CP}_{\text{bic}}$ sont toutes les deux valides. Par contre, lorsque le nombre de ruptures est quelconque mais la longueur minimum des plages vérifie $\frac{L_{\min}-3h}{\max \mathcal{P}'(\mathbf{w})} \to \infty$, seule l'approximation par $\mathcal{CP}_{\text{bic}}$ est valide. L'approximation homogène n'est quant à elle théoriquement jamais valide s'il y a au moins une rupture et si (au moins) deux probabilités de transitions π_s diffèrent ($\exists s, t \in \mathcal{S}$ tels que $\pi_s \neq \pi_t$).

Ces considérations sont pour l'instant uniquement théoriques, on cherche maintenant à les vérifier sur des simulations. Nous allons mettre en évidence dans un certain cadre de simulation les faits suivants :

1. La loi (empirique) du comptage n'est pas la même sous un modèle PM et sous un modèle de Markov homogène lorsque l'écart entre les $\pi_s, s \in \mathcal{S}$ augmente. L'approximation homogène ne sera donc pas satisfaisante dans ce cas.
2. Pour les mots rares non-recouvrants, l'approximation par $\mathcal{CP}_{\text{bic}}$ (réduite à une loi de Poisson) est satisfaisante.
3. La qualité de l'approximation par $\mathcal{CP}_{\text{uni}}$ est assez bien évaluée avec l'indicateur théorique $h\rho/n$; la loi $\mathcal{CP}_{\text{uni}}$ approchera d'autant mieux la loi du comptage que $h\rho/n$ est petit ($h\rho/n < 1\%$ semble être un bon critère).
4. Dans le cas recouvrant, il existe une bande de valeurs pour ρ dans laquelle l'approximation par $\mathcal{CP}_{\text{bic}}$ est satisfaisante alors que l'approximation par $\mathcal{CP}_{\text{uni}}$ n'est plus valide. Notamment, l'approximation par $\mathcal{CP}_{\text{bic}}$ est valable lorsque $L_{\min}$ est suffisamment grand ($L_{\min} \geq 10h$ par exemple), ce qui laisse penser que la condition théorique $\frac{L_{\min}-3h}{\max \mathcal{P}'(\mathbf{w})} \to \infty$ est un peu stricte et que le domaine de validité de l'approximation par $\mathcal{CP}_{\text{bic}}$ est un peu plus large.

6.2.1 Plan de simulation

Les séquences sont de longueur $n = 100\,000$, dans l'alphabet $\mathcal{A} = \{\mathtt{a,g,c,t}\}$, simulées selon un modèle PM0 (=PSM0) avec une segmentation à deux états $\mathcal{S} = \{1,2\}$ et ρ ruptures régulièrement espacées. Les probabilités d'émissions du modèle sont données par

$$\mu_1(\mathtt{a}) = \mu_1(\mathtt{g}) = 0.25 + \varepsilon, \quad \mu_1(\mathtt{c}) = \mu_1(\mathtt{t}) = 0.25 - \varepsilon,$$
$$\mu_2(\mathtt{a}) = \mu_2(\mathtt{g}) = 0.25 - \varepsilon, \quad \mu_2(\mathtt{c}) = \mu_2(\mathtt{t}) = 0.25 + \varepsilon.$$

Ainsi, ε est un paramètre mesurant l'écart entre μ_1 et μ_2 ; par exemple $\varepsilon = 0$ donne le modèle homogène où les bases sont i.i.d. équidistribuées sur $\{\mathtt{a,g,c,t}\}$. On fera varier les paramètres ρ et ε de la façon suivante : $\rho \in \{10, 100, 1000, 2000, 5000\}$ (ce qui correspond respectivement à $h\rho/n \in \{0.054\%, 0.594\%, 5.99\%, 11.9\%, 29.9\%\}$) et $\varepsilon \in \{0, 0.05, 0.1, 0.15, 0.2\}$. Le nombre de simulations est 100 000. Pour garantir la condition de rareté, les mots considérés sont de longueur 8. Nous avons choisi les 8 mots suivants :
- mots non-recouvrants : `agggact`, `atggacg`, `aaagggg`, `aaagggc`
- mots recouvrants : `aaaaaaa` (période principale 1) ; `agagaga` (période principale 2) ; `aacaaca`, `agaagaa` (période principale 3).

Ces mots ont une composition en bases qui détermine leurs probabilités d'occurrence respectives dans les états 1 et 2 :
- les mots `aaaaaaa`, `aaagggg`, `agagaga` et `agaagaa` ont une occurrence "extrêmement favorisée" dans l'état 1 : la probabilité commune d'occurrence dans l'état 1 est $\mu_1(\mathbf{w}) = (0.25+\varepsilon)^7$ alors que dans l'état 2 elle est de $\mu_2(\mathbf{w}) = (0.25-\varepsilon)^7$.
- le mot `aaagggc` a des probabilités d'occurrence $\mu_1(\mathbf{w}) = (0.25+\varepsilon)^6(0.25-\varepsilon)^1$ et $\mu_2(\mathbf{w}) = (0.25+\varepsilon)^1(0.25-\varepsilon)^6$, son occurrence est donc "très favorisée" dans l'état 1.
- les mots `agggact`, `atggacg` et `aacaaca` ont les probabilités communes d'occurrence $\mu_1(\mathbf{w}) = (0.25+\varepsilon)^5(0.25-\varepsilon)^2$ et $\mu_2(\mathbf{w}) = (0.25+\varepsilon)^2(0.25-\varepsilon)^5$. Chacun de ces mots a une occurrence "favorisée" dans l'état 1.

Remarquons que le mot `agggact` a une probabilité d'occurrence dans l'état 1111122 qui est de $(0.25+\varepsilon)^7$, donc son occurrence est "extrêmement" favorisée aux ruptures. La loi de son comptage va donc être fortement influencée par le nombre de ruptures ρ.

6.2.2 Loi du comptage : homogène contre hétérogène

Ici nous cherchons à établir que la loi empirique du comptage est différente si on la simule avec un modèle homogène ou avec un modèle hétérogène. La figure 6.1 (page 95) représente les comptages moyens empiriques pour les mots `aaaaaaa`, `agggact` et `aacaaca` dans la séquence PSM0, contre la constante $(n-h+1)0.25^7$ (qui correspond aux comptages moyens des mots dans le modèle M0 équidistribué). Nous remarquons que la moyenne du comptage sous le modèle hétérogène peut être bien différente de la moyenne du comptage sous le modèle homogène. Par exemple, le mot `aaaaaaa` devient de plus en plus fréquent lorsque ε augmente (attention l'échelle du graphe de $\mathbb{E}N(\mathtt{aaaaaaa})$ n'est pas la même que pour les autres mots). Cela est dû au fait que la probabilité d'occurrence de `aaaaaaa` dans l'état 1 est "très grande" lorsque ε est "grand". On peut approfondir cette étude en examinant la distance en variation totale entre la loi du comptage empirique sous le modèle M0 (équidistribué) et la loi du comptage empirique sous le modèle PM0 (cf. Fig. 6.2). On remarque que ces deux lois s'éloignent lorsque ε augmente (elles sont dans tous les cas presque étrangères pour $\varepsilon = 0.2$). Ceci signifie que pour étudier la loi du comptage dans une séquence hétérogène, il peut être très faux de se placer dans un modèle homogène.

6.2.3 Qualité des approximations par $\mathcal{CP}_{\text{uni}}$ et $\mathcal{CP}_{\text{bic}}$

Nous évaluons à présent la qualité des approximations par $\mathcal{CP}_{\text{uni}}$ et $\mathcal{CP}_{\text{bic}}$ en calculant la distance en variation totale entre ces lois et la loi empirique du comptage.

Approximation par $\mathcal{CP}_{\text{bic}}$ dans le cas de mots non-recouvrants

D'après la figure 6.3 (page 97), la loi $\mathcal{CP}_{\text{bic}}$ (qui se réduit à une loi de Poisson) semble être une bonne approximation pour les mots non-recouvrants. Ceci s'explique car les segmentations étudiées vérifient bien $L_{\min} \geq h$.

Approximations par $\mathcal{CP}_{\text{uni}}$ et $\mathcal{CP}_{\text{bic}}$ dans le cas de mots recouvrants

Dans le cas de mots recouvrants, la figure 6.4 (page 98) montre que les deux approximations sont bonnes pour un nombre de segments plus petit que 100, mais que pour 1000 segments ou plus, $\mathcal{CP}_{\text{uni}}$ est nettement moins bonne (voire mauvaise), et il faut dans ce cas préférer $\mathcal{CP}_{\text{bic}}$. Cependant, l'approximation par $\mathcal{CP}_{\text{bic}}$ a elle aussi ses limites car elle n'est pas bonne pour 5000 segments ; ce cas correspond à une segmentation avec une rupture toutes les 20 lettres et donc l'hypothèse théorique $\frac{L_{\min}-3h}{\max \mathcal{P}'(\mathbf{w})} \to \infty$ n'est bien entendu plus vérifiée. Pour mieux visualiser la qualité des approximations par $\mathcal{CP}_{\text{uni}}$ et $\mathcal{CP}_{\text{bic}}$, on trace sur la figure 6.5 la densité des lois $\mathcal{CP}_{\text{uni}}$, $\mathcal{CP}_{\text{bic}}$ et de la loi empirique du comptage pour le mot `aaaaaaa` dans plusieurs situations. Le cas où $\varepsilon = 0.05$ et le nombre de segments est égal à 1000 ($\rho h/n \simeq 6\%$) correspond à un cas assez "réaliste" ; $\mathcal{CP}_{\text{bic}}$ semble dans ce cas plus proche de la loi empirique que $\mathcal{CP}_{\text{uni}}$, mais ces lois sont suffisamment proches pour que $\mathcal{CP}_{\text{uni}}$ reste assez satisfaisante. Nous avons aussi ajusté une loi de Poisson sur la loi empirique, pour se convaincre qu'elle ne pouvait pas être utilisée dans ce cas recouvrant.

Rappelons que les conclusions générales de ces simulations sont données en début de section.

6.3 Mise en oeuvre sur des données réelles

Dans cette section, nous recherchons les mots $\mathbf{w}$ de fréquence exceptionnelle dans de vraies séquences d'ADN et nous considérons plusieurs types d'hétérogénéités biologiques (à deux états). Nous examinons seulement les mots rares, c'est-à-dire avec une longueur choisie pour que le comptage attendu de chaque mot soit suffisamment petit.

6.3.1 Scores homogènes et hétérogènes

Étant données une séquence et une segmentation, pour chaque mot $\mathbf{w}$ d'une longueur donnée h, nous définissons un **score homogène** et un **score hétérogène** de la façon suivante : nous approchons la loi de $N(\mathbf{w})$ par la loi de Poisson composée homogène proposée par Schbath (1995a) (cf. chapitre 2) et par la loi de Poisson composée hétérogène $\mathcal{CP}_{\text{uni}}$ pour les mots recouvrants et $\mathcal{CP}_{\text{bic}}$ pour les mots non-recouvrants (pour les définitions de $\mathcal{CP}_{\text{uni}}$ et $\mathcal{CP}_{\text{bic}}$ voir chapitre 3). Par suite, pour chacune des approximations, nous calculons la p-value $p_{\mathbf{w}} \in [0, 1]$

de $\mathbf{w}$, que l'on normalise en un score dans $\mathbb{R}$ selon la transformation quantile $1 - \Phi^{-1}(p_{\mathbf{w}})$ où Φ désigne la fonction de répartition de la loi normale centrée réduite.

Remarque 6.1 (Choix de m) *Remarquons qu'ainsi définis, les scores homogène et hétérogène dépendent de l'ordre m de la chaîne de Markov sous-jacente. Rappelons également que le choix de m définit l'a priori que nous nous donnons sur la séquence : dans le cas homogène, il s'agit de la composition de la séquence en $(m+1)$-mots ; dans le cas hétérogène, il s'agit de la composition de la séquence dans chaque état en $(m+1)$-mots (composition "coloriée"). Nous insistons aussi sur le fait que dans la suite on ne doit pas choisir des valeurs de m trop grandes sous peine d'avoir des problèmes d'estimation. En pratique, la valeur la plus grande utilisée pour m est telle que $4^{m+1}/n$ soit plus petit que $1/100$.*

Afin d'évaluer la pertinence des scores homogènes lorsque la séquence présente une hétérogénéité, nous allons à présent comparer sur plusieurs exemples l'ensemble des scores homogènes et l'ensemble des scores hétérogènes (pour tous les mots d'une longueur donnée).

6.3.2 Cas hétérogènes "dégénérés"

En calculant les scores homogènes et hétérogènes sur des séquences réelles, j'ai essentiellement détecté deux cas où les scores homogènes étaient très (trop) proches des scores hétérogènes :

(i) Lorsque **les compositions en $(m+1)$-mots dans les différents états sont très proches**.

(ii) Lorsque **la segmentation a un état "dominant",** c'est-à-dire que la segmentation possède un état bien plus fréquent que les autres états.

Dans ces cas, les scores homogènes et hétérogènes sont proches simplement parce que les modèles de Markov homogène et hétérogène sous-jacents sont très proches. Le cas (i) arrive par exemple lorsque que l'on regarde le génome complet d'*Escherichia coli* ; il est connu qu'il y a un biais de composition sur le brin direct entre la zone "Ori-Ter" et la zone "Ter-Ori" (cf. Fig. 6.6 page 100). Dans la zone "Ori-Ter" la fréquence en (`a,c,g,t`) est $(0.248, 0.262, 0.245, 0.246)$ alors que dans la zone "Ter-Ori" la fréquence en (`a,c,g,t`) est $(0.242, 0.249, 0.264, 0.245)$. Ce biais en `g-c` est important d'un point de vue biologique, mais il est trop faible au niveau du score d'exceptionnalité pour que les scores homogènes et hétérogènes soient différents (cela correspondrait au cas $\varepsilon = 0.008$ dans les simulations de la section 6.2). Nous examinons dans la suite des cas réels que nous estimons "non-dégénérés".

6.3.3 Analyse du phage *Lambda*

Nous examinons ici le cas du phage *Lambda* qui est un phage de la bactérie *Escherichia coli*. Son génome est de taille $n = 48502$; il est composé de nombreuses parties codantes (gènes) sur le brin direct ou sur le brin complémentaire. Nous considérons ici la segmentation à deux états :

- 1 pour "codant dans le sens direct"
- 2 pour "non codant dans le sens direct".

Cette segmentation est représentée sur la figure 6.7 (page 100). Le nombre de ruptures dans la segmentation est alors $\rho = 36$.

Nous recherchons les mots de longueur 5 qui sont de fréquence exceptionnelle dans ce génome. Comme la séquence est ici assez courte, les mots de longueur 5 peuvent être considérés comme rares et nous pouvons utiliser une approximation poissonienne du comptage. Remarquons également que l'approximation par $\mathcal{CP}_{uni}$ est valide pour la segmentation considérée car $36 * 5/48502 \simeq 0.0037$ (cf. Section 6.2). On représente[2] les scores hétérogènes en fonction des scores homogènes (cf. FIG. 6.8 page 101) pour différents choix de l'ordre de Markov m. Nous remarquons que certains 5-mots sont éloignés de la première bissectrice, par exemple :

- le mot `atttt` est davantage sur-représenté dans le modèle homogène que dans le modèle hétérogène pour $m = 0$. On en déduit que ce mot s'explique mieux à partir de la composition en nucléotides "coloriés" dans les parties codantes et non-codantes, qu'à partir de la composition globale en nucléotide.
- le mot `gcaat` est davantage sur-représenté dans le modèle hétérogène que dans le modèle homogène pour $m = 3$. On en déduit que ce mot est plus exceptionnel par rapport à la composition en 4-mots "coloriés" dans les parties codantes et non-codantes, que par rapport à la composition globale en 4-mots.

Pour mesurer de manière concise et globale la différence entre les scores homogènes et hétérogènes, nous avons calculé dans le tableau 6.1 la corrélation des valeurs et leurs corrélations, à la fois sur les valeurs et sur les rangs. La corrélation sur les rangs est mesurée à l'aide du coefficient de Kendall qui mesure le nombre de paires concordantes entre les deux scores (rapporté dans $[-1, 1]$). On voit ainsi que les scores homogènes et hétérogènes sont globalement fortement corrélés, mais il semble que ces corrélations diminuent avec m. Ceci est dû au fait que le contraste entre les compositions coloriées et non-coloriées de la séquence en $(m + 1)$-mots augmente avec m. Nous insistons sur le fait que les valeurs du tableau 6.1 indiquent une tendance **globale** sur l'ensemble des 5-mots mais elles ne mesurent pas le comportement pour chaque mot **individuellement** ; notamment les scores homogènes et hétérogènes des mots dit "exceptionnels" peuvent être différents même si la corrélation globale est proche de 1 (c'est le cas pour `gcaat` par exemple).

	$m = 0$	$m = 1$	$m = 3$
corrélation	0.9924	0.9916	0.9813
coeff. Kendall	0.9291	0.9201	0.8780

TAB. 6.1 – Coefficients de corrélation entre les scores des 1024 5-mots calculés dans un modèle homogène (Mm) et dans un modèle hétérogène (PSMm). Cas du phage *Lambda*.

6.3.4 Cas d'un mélange *Escherichia coli* — *Haemophilus influenzae*

Une conséquence indirecte — mais non moins intéressante — de l'approche hétérogène, est de pouvoir calculer l'exceptionnalité de mots dans un ensemble de séquences donné. Dans ce cas, on concatène simplement les séquences, et les états de la segmentation correspondent simplement aux différentes séquences.

Remarque 6.2 *Bien sûr, dans ce cadre, seules les occurrences unicolores des mots ont un intérêt, et on doit utiliser $\mathcal{CP}_{\mathrm{uni}}$. Cependant, on remarque que comme le nombre de ruptures est faible, $\mathcal{CP}_{\mathrm{bic}}$ est très proche de $\mathcal{CP}_{\mathrm{uni}}$, ce qui permet d'utiliser également $\mathcal{CP}_{\mathrm{bic}}$. En définitive,*

[2]Les scores ont été tracés en utilisant le logiciel R'MESPlot

la manière dont nous avons calculé le score hétérogène est également appropriée dans le cas de séquences concaténées.

Nous avons concaténé ici les 100 000 premières bases du génome de *Escherichia coli* avec les 100 000 premières bases du génome de *Haemophilus influenzae*. Comme le montre la figure 6.9 (page 102), la composition en $(m+1)$-mots de la première séquence est bien différente de celle de la seconde séquence (pour $m = 0$, les fréquences des nucléotides valent $(0.304, 0.201, 0.188, 0.307)$ pour la première et $(0.235, 0.271, 0.251, 0.243)$ pour la seconde, dans l'ordre `a,g,c,t`).

Le modèle hétérogène correspondant et donc a priori assez différent du modèle homogène. Nous considérons les mots de longueur 8 (mots rares pour ces données). Les scores homogènes et hétérogènes sont représentés sur la figure 6.10 (page 103) et les corrélations sont indiquées dans la table 6.2. Les remarques sont les mêmes qu'au paragraphe précédent : les scores, qui se comportent globalement de la même façon, peuvent différer pour certains mots. Le score homogène, qui ne tient pas compte du fait que l'on examine deux génomes différents, peut donc faire des erreurs pour certains motifs et il faut préférer le score hétérogène.

Nous avons examiné en particulier le motif Chi `gctggtgg`, car il est impliqué dans la réparation du chromosome à la fois de *E.coli* et de *H. influenzae*. Ses scores homogènes et hétérogènes sont calculés dans la table 6.3 ; on voit là aussi (par exemple à l'ordre $m = 3$) que ses scores peuvent changer lorsque l'on tient compte de l'hétérogénéité.

	$m=0$	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
corrélation	0.9903	0.9897	0.9859	0.9841	0.9823	0.9750
coeff. Kendall	0.9135	0.9006	0.8812	0.8735	0.8657	0.8436

TAB. 6.2 – Coefficients de corrélation entre les scores calculés dans un modèle homogène (Mm) et dans un modèle hétérogène (PSMm). Cas du mélange *E. coli – H. influenzae.*

m	score homogène	score hétérogène
0	7.9043	8.0832
1	8.0217	8.0899
2	6.2246	5.6670
3	4.3787	3.8298
4	4.8376	4.4353
5	3.2902	< 3
6	< 3	< 3

TAB. 6.3 – Scores homogène (Mm) et hétérogène (PSMm) du mot `gctggtgg`. Cas du mélange *E. coli – H. influenzae.*

6.4 Conclusion

Au vu des résultats obtenus sur les données simulées et réelles, nous pouvons affirmer que le score calculé sous un modèle homogène peut différer du score calculé sous un modèle hétérogène lorsque les deux conditions suivantes sont vérifiées :

1. Les compositions en $(m+1)$-mots dans les différents états sont "suffisamment éloignées".
2. La segmentation ne possède pas un état "dominant", c'est-à-dire que le nombre d'occurrences de chaque état dans la segmentation est "du même ordre".

Lorsque la séquence présente une hétérogénéité connue qui est susceptible de vérifier les deux points ci-dessus, il est fortement recommandé d'utiliser la méthode hétérogène.

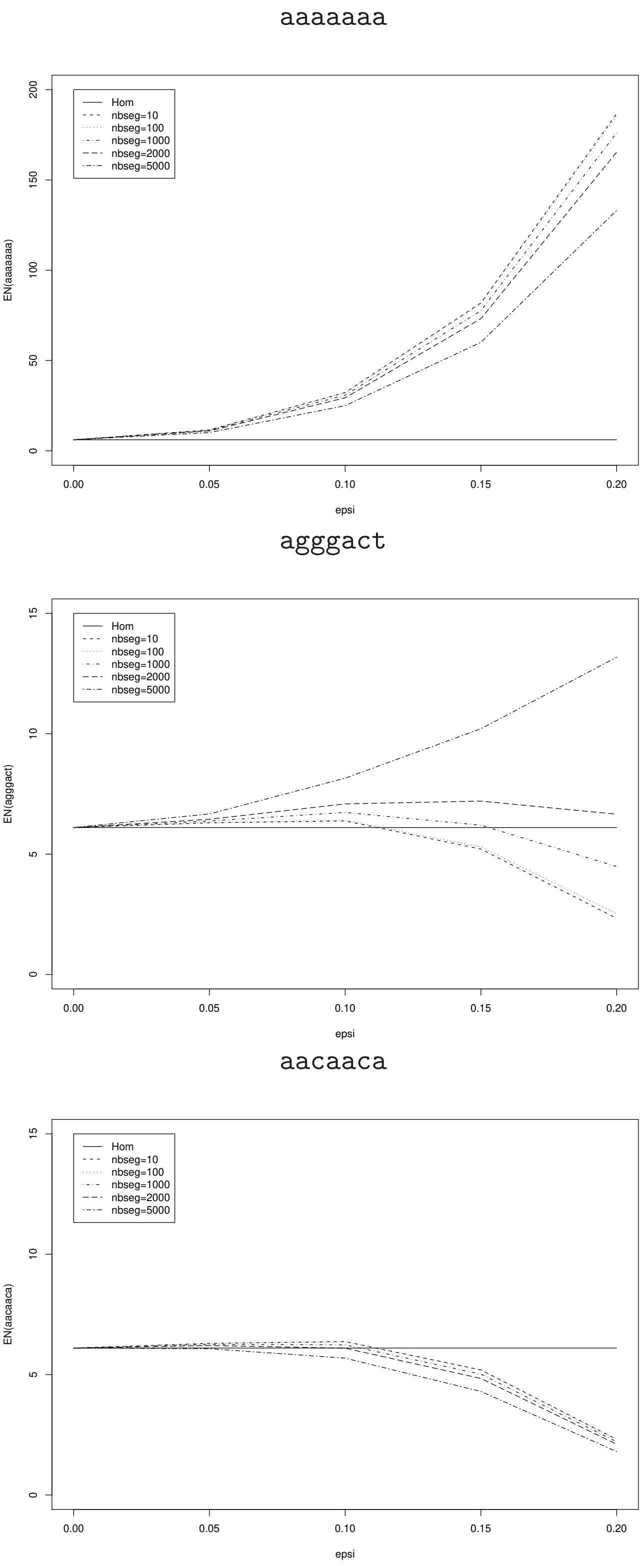


FIG. 6.1 – Espérance du comptage pour les mots `aaaaaaa`, `agggact` et `aacaaca`, selon ε et pour plusieurs nombres de segments.

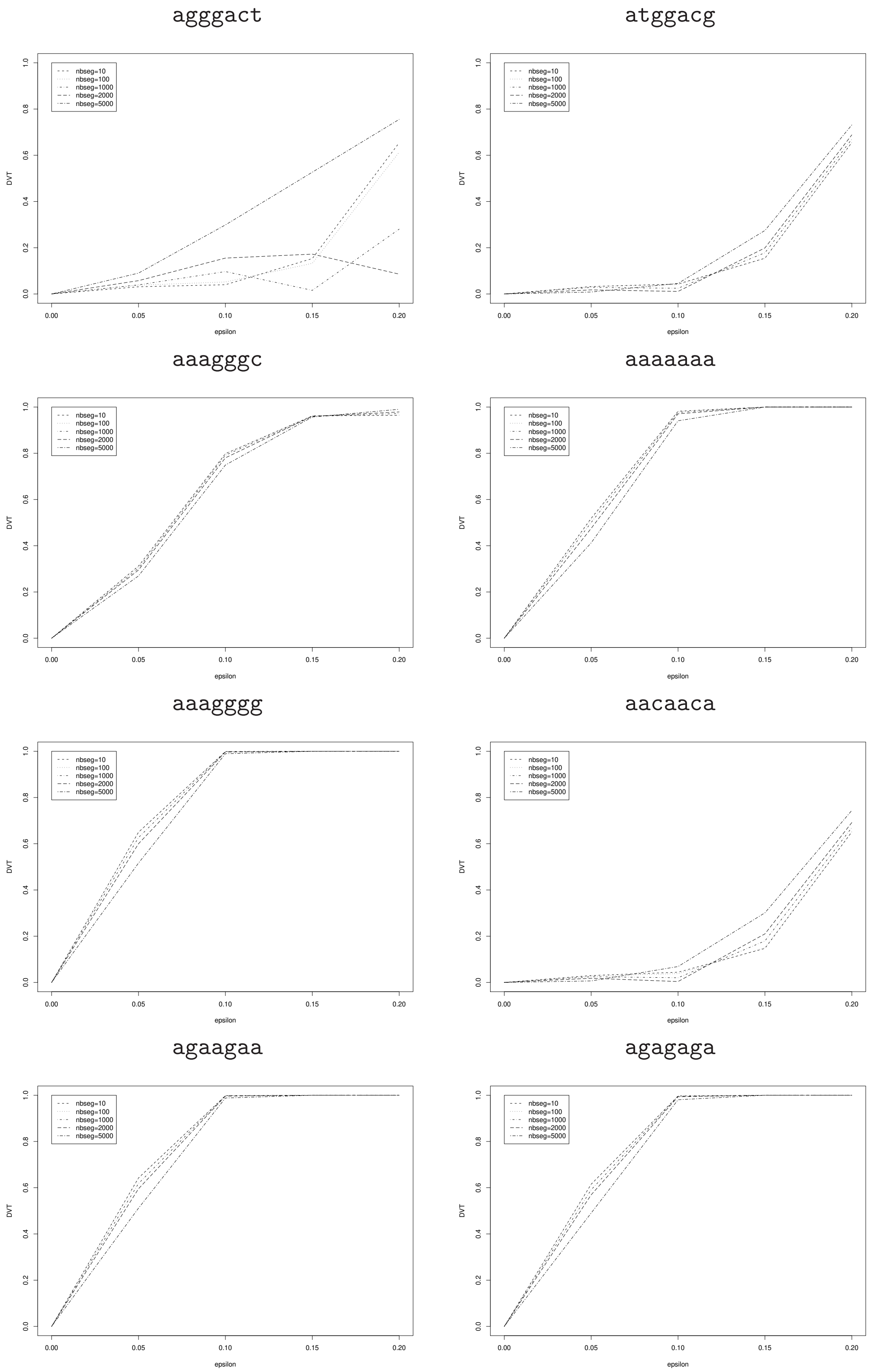


FIG. 6.2 – Distance en variation totale entre la loi du comptage (empirique) dans M0 et la loi du comptage (empirique) dans PM0 pour plusieurs mots, en fonction de ε, et pour plusieurs nombres de segments.

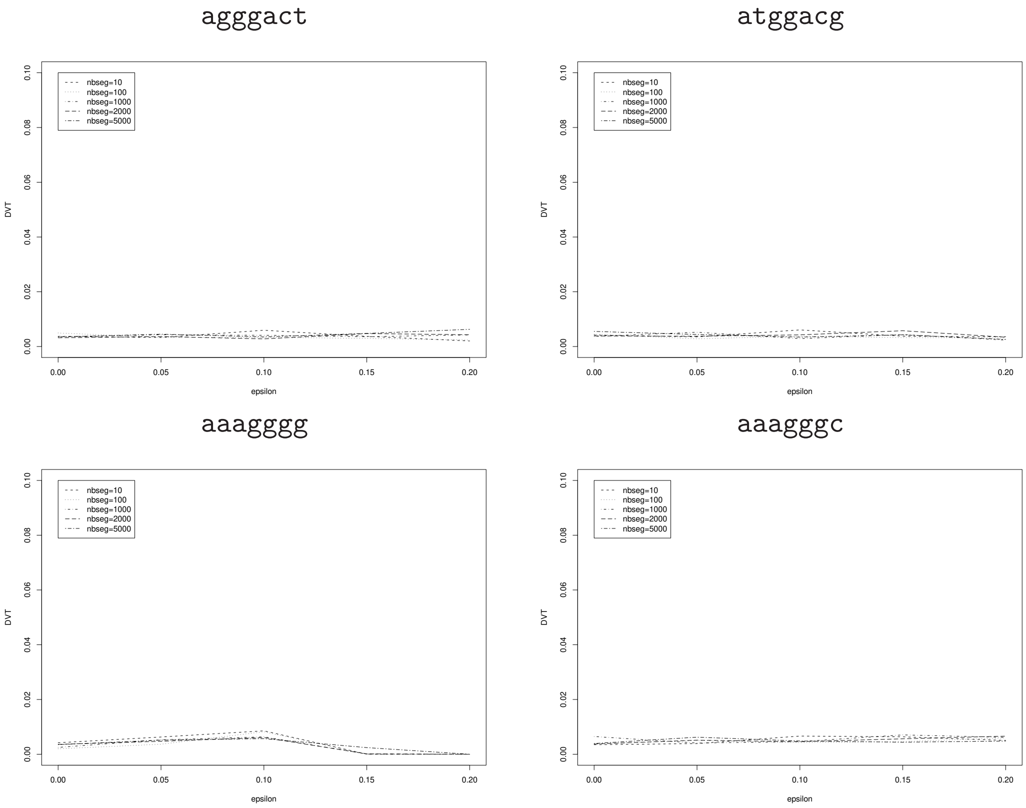


FIG. 6.3 – Pour les mots non-recouvrants : distance en variation totale entre la loi empirique du comptage et la loi $\mathcal{CP}_{\mathrm{bic}}$ (réduite à une loi de Poisson) dans un modèle PM0, selon ε et pour plusieurs nombres de segments.

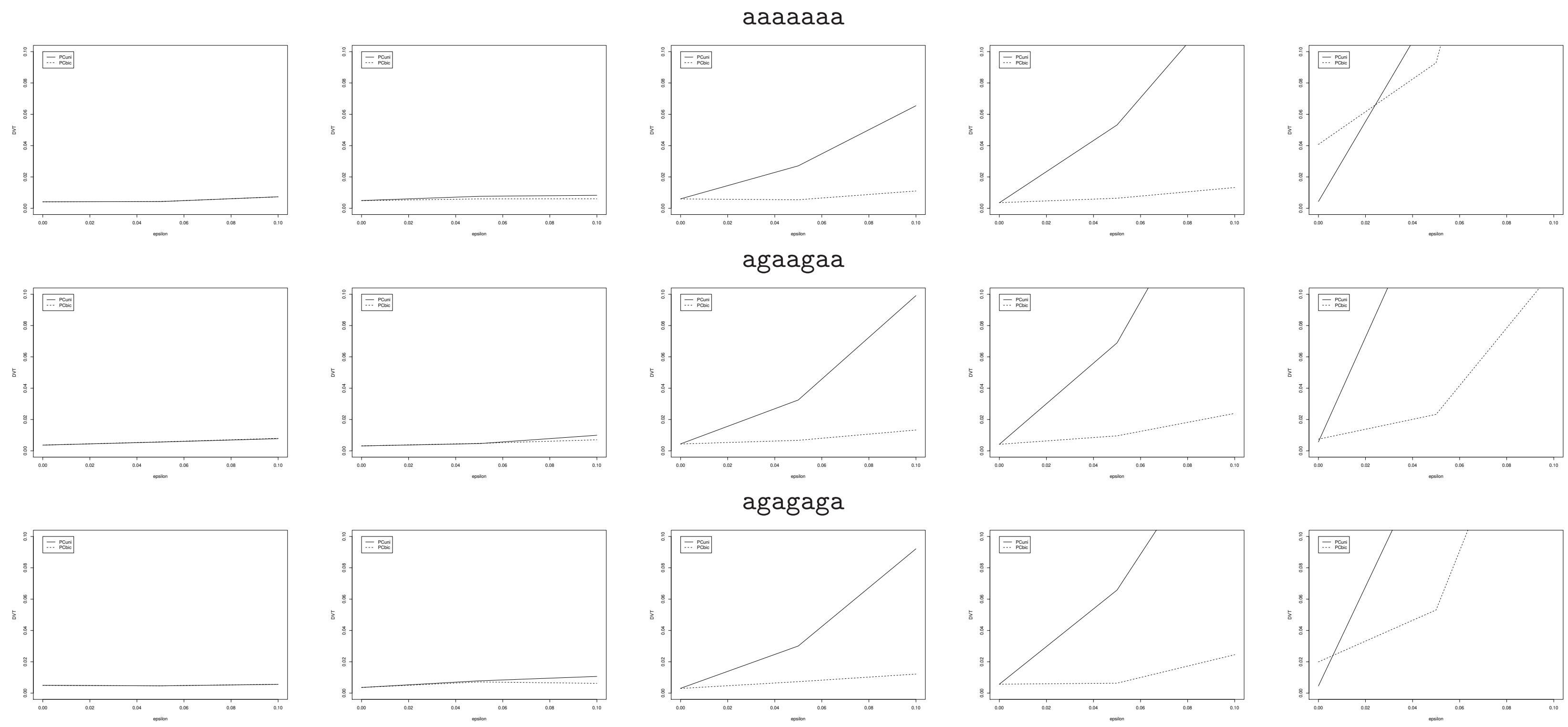


FIG. 6.4 – Pour chaque mot recouvrant (de haut en bas) : chaque graphe représente la distance en variation totale entre la loi empirique du comptage et les lois de Poisson composée $\mathcal{CP}_{\text{uni}}$ et $\mathcal{CP}_{\text{bic}}$ dans PM0 en fonction de $\varepsilon = 0, 0.05, 0.1$. La courbe pleine correspond à $\mathcal{CP}_{\text{uni}}$ et la courbe en pointillés correspond à $\mathcal{CP}_{\text{bic}}$. Le nombre de segments prend les valeurs (de gauche à droite) 10, 100, 1000, 2000 et 5000.

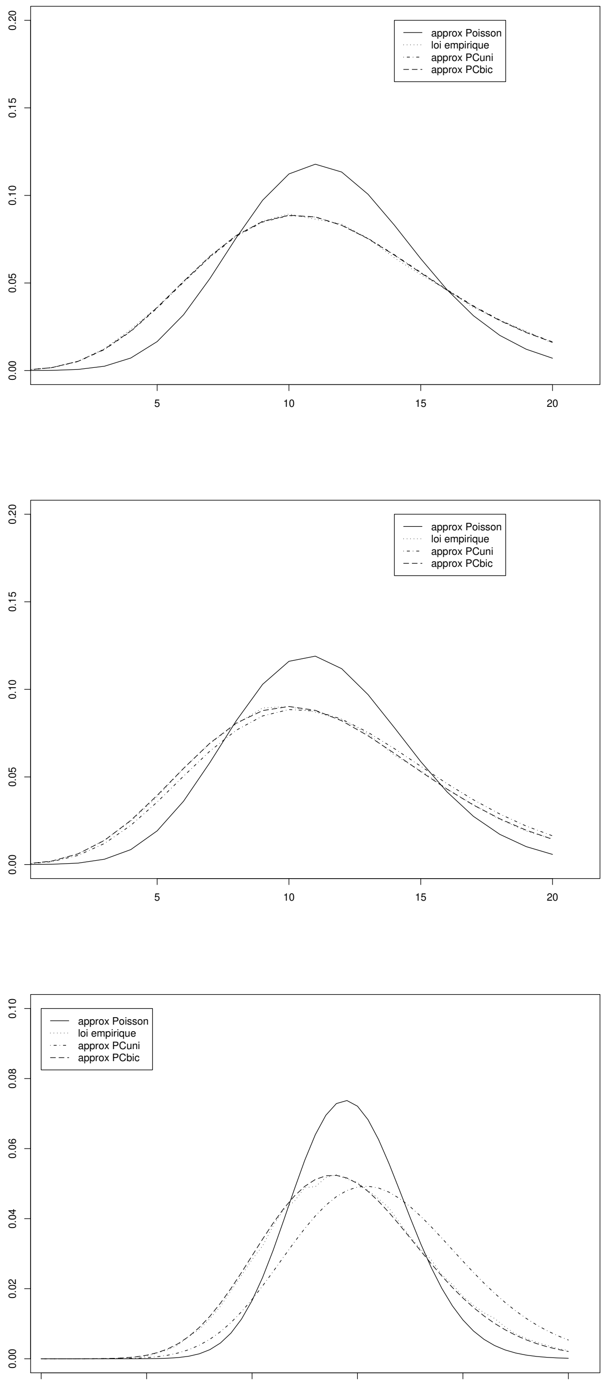


FIG. 6.5 – Densités des deux lois $\mathcal{CP}_{\text{uni}}$ ("approx. PCuni") et $\mathcal{CP}_{\text{bic}}$ ("approx. PCbic"), de la loi de Poisson ajustée ("approx. Poisson") et de la loi empirique du comptage ("loi empirique") pour le mot `aaaaaaa`. Pour le graphe du haut : $\varepsilon = 0.05$ et le nombre de segments est 100 ; le graphe du milieu : $\varepsilon = 0.05$ et le nombre de segments est 1000 ; le graphe du bas : $\varepsilon = 0.1$ et le nombre de segments est 2000.

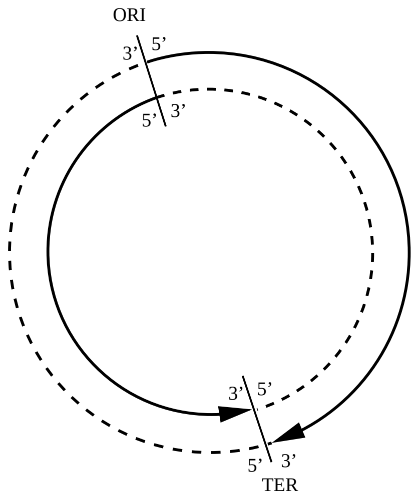


Fig. 6.6 – Génome circulaire de *Escherichia coli*. "Ori" désigne l'origine de réplication, "Ter" désigne la terminaison de réplication, le sens de réplication ($5' \rightarrow 3'$) est indiqué par les flèches. Sur le brin d'ADN direct (à l'extérieur), la zone "Ori-Ter" est représentée en trait plein et la zone "Ter-Ori" est représentée en pointillés.

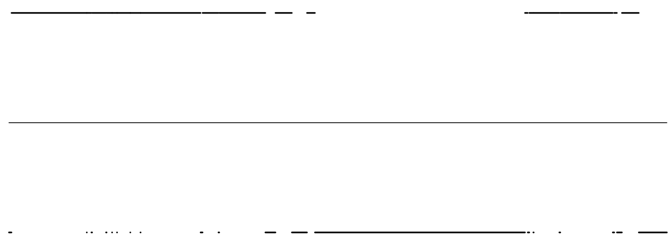


Fig. 6.7 – Segmentation du génome du phage *Lambda* (ligne du milieu) selon les états "codant dans le sens direct" (segments du haut) et "non codant dans le sens direct" (segments du bas).

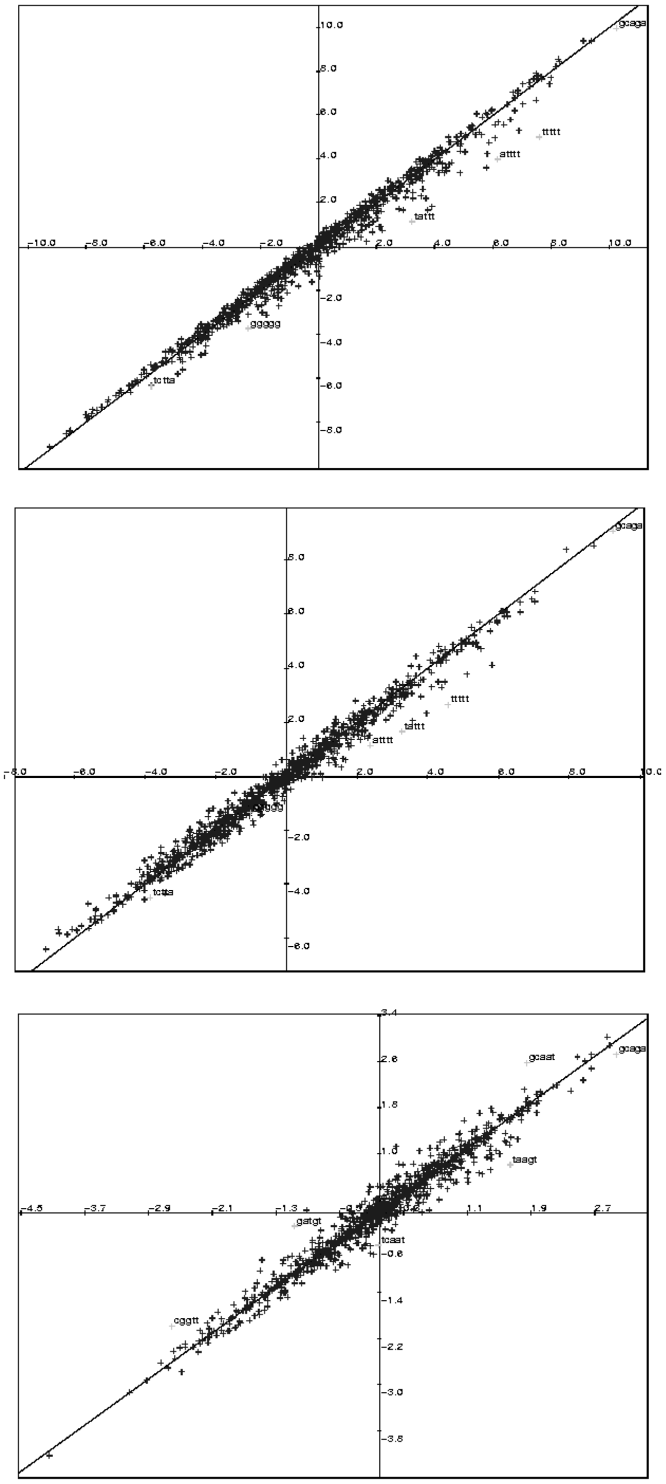


FIG. 6.8 – Score de tous les 5-mots : en abscisse le score est calculé dans un modèle homogène Mm, en ordonnée le score est calculé dans un modèle hétérogène PSMm. Le graphe du haut (resp. milieu, bas) correspond à $m = 0$ (resp. $m = 1$, $m = 3$). Cas du phage *Lambda*.

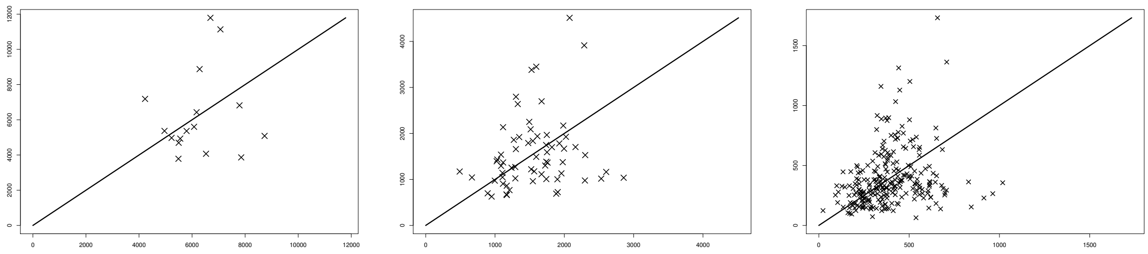


FIG. 6.9 – Comptages des $(m+1)$-mots dans le génome de *Haemophilus influenzae* (en ordonnée) et dans le génome de *Escherichia coli* (en abscisse). À gauche $m = 1$, au centre $m = 2$ et à droite $m = 3$. Les comptages sont effectués en restriction aux 100 000 premières bases de chaque génome.

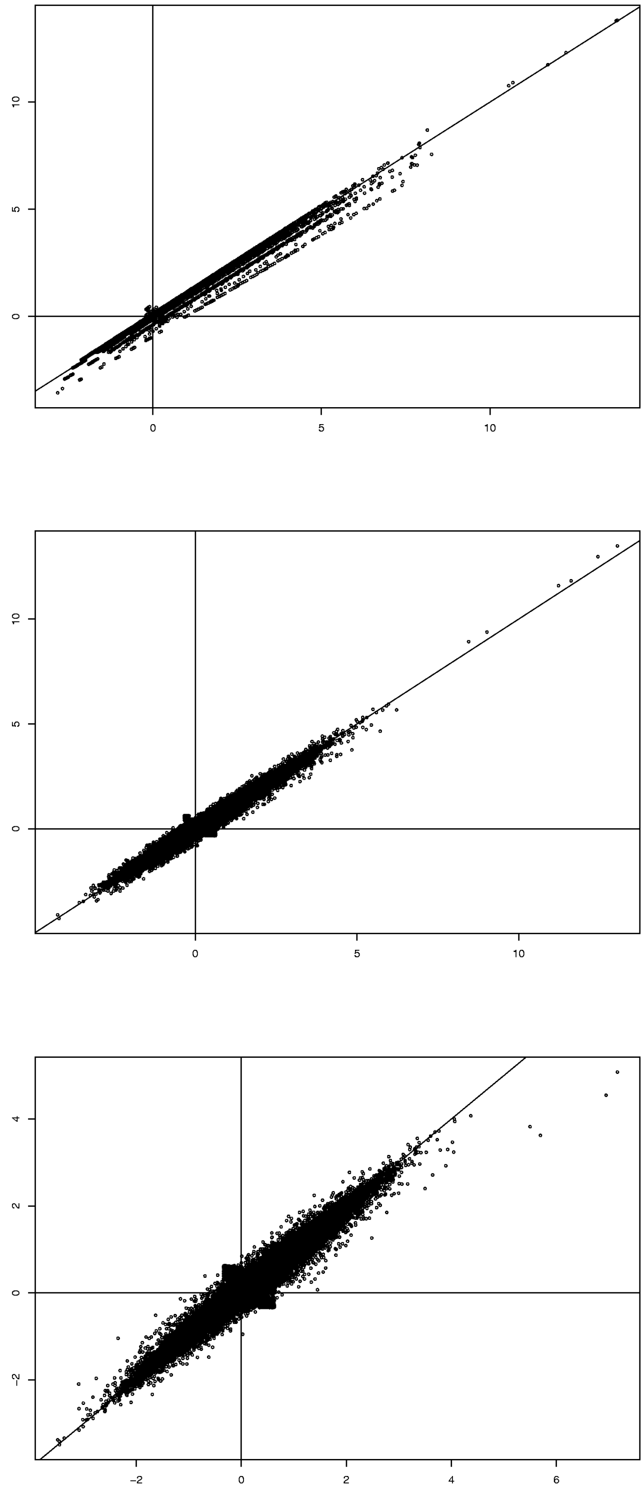


FIG. 6.10 – Score de tous les 8-mots : en abscisse le score est calculé dans un modèle homogène Mm, en ordonnée le score est calculé dans un modèle hétérogène PSMm. Le graphe du haut (resp. milieu, bas) correspond à $m = 0$ (resp. $m = 2$, $m = 5$). Cas du mélange *E. coli* – *H. influenzae*.

Chapitre 7

Compléments

Dans ce chapitre, nous travaillons avec une segmentation déterministe et connue a priori. Dans la section 7.1, nous proposons deux approximations gaussiennes pour la loi du comptage de mots fréquents : l'une est de type "mot unicolore" et a une erreur qui tend vers 0 dès que le nombre de ruptures ρ est fixe avec n ; l'autre est une approximation gaussienne de type "mot multicolore" qui est basée sur le calcul de l'espérance conditionnelle et de la variance conditionnelle du comptage dans le cas indépendant (sans garantie sur l'erreur d'approximation). La section 7.2 présente une méthode pour calculer la loi exacte du comptage, qui pourra être utilisée pour des séquences courtes. Nous examinerons dans la section 7.3 le problème de l'estimation des paramètres dans le modèle PM : nous calculerons dans un premier temps les estimateurs du maximum de vraisemblance ; par la suite nous montrerons que la loi $\mathcal{CP}_{\text{uni}}$ du chapitre 3, utilisée avec les paramètres estimés, a une erreur d'approximation qui tend encore vers 0 (dans le cadre de rareté et pour ρ fixe).

Au cours de ce chapitre, nous allons considérer deux asymptotiques (avec n) possibles pour la segmentation :

$$\rho \text{ fixe }, \quad \forall j, |\mathbf{s}_j| \to \infty, \quad \forall j, \frac{|\mathbf{s}_j|}{n} \to q_j \in [0,1] \tag{7.1}$$

$$\forall s \in \mathcal{S}, \frac{n_{\mathbf{s}}(s)}{n} \to q_s \in]0,1] \quad ; \quad \forall k \geq 2, \ \forall \mathbf{t} = t_1 \cdots t_k, \ \frac{n_{\mathbf{s}}(\mathbf{t})}{n} \to q_{\mathbf{t}} \in [0,1] \tag{7.2}$$

L'asymptotique (7.1) est propre à une approximation de type "mot unicolore" ; elle est légèrement plus restrictive que celle proposée dans le chapitre 3 pour l'approximation $\mathcal{CP}_{\text{uni}}$, car le nombre de ruptures est supposé fixe avec n. L'hypothèse $\forall j, |\mathbf{s}_j| \to \infty$ est uniquement ajoutée pour faciliter la démarche ; les résultats proposés seront toujours valides si certains segments ont une longueur bornée avec n. Dans cette asymptotique, q_j représente la proportion du segment $\mathbf{s}_j$ dans la segmentation lorsque n tend vers l'infini (ainsi on a $\sum_j q_j = 1$). Nous noterons également q_s la limite de la proportion $n_{\mathbf{s}}(s)/n$ du nombre d'états s dans la segmentation $\mathbf{s}$ (de sorte que $\sum_{s\in\mathcal{S}} q_s = 1$).

L'asymptotique (7.2) sera utilisée pour l'approximation de type mot "multicolore" ; elle est très générale et signifie juste que chaque coloriage $\mathbf{t}$ arrive asymptotiquement dans une certaine proportion $q_{\mathbf{t}}$ (cette proportion devant être strictement positive lorsque $|\mathbf{t}| = 1$). En particulier, les coloriages avec de nombreuses ruptures peuvent être présents dans $\mathbf{s}$ (aucune hypothèse n'est faite sur $L_{\min}$).

7.1 Approximations gaussiennes dans un modèle PSM

Nous considérons ici le cas où $\mathbf{w}$ est fixe avec n et devient donc de plus en plus **fréquent** lorsque la longueur n de la séquence augmente. Nous cherchons à généraliser l'approche gaussienne conditionnelle proposée par Prum *et al.* (1995). Nous proposons dans la section 7.1.1 une approximation de type "mot unicolore" dans l'asymptotique (7.1). Par la suite, nous proposerons une approximation de type multicolore dans l'asymptotique (7.2) et dans le cas où les lettres de la séquence sont indépendantes.

7.1.1 Approximation gaussienne de type "mot unicolore"

Considérons une séquence $\mathbf{X}$ qui suit un modèle PSMm ($m \geq 1$), avec des probabilités de transition strictement positives. Nous montrons ici que dans l'asymptotique (7.1), on peut approcher la loi du comptage par une loi gaussienne obtenue simplement à partir de la somme des approximations gaussiennes sur chaque segment de $\mathbf{X}$.

Plaçons-nous dans l'asymptotique (7.1) et appliquons sur chaque segment $\mathbf{X}_j$, $1 \leq j \leq \rho+1$, l'approximation gaussienne conditionnelle proposée par Prum *et al.* (1995) (et généralisée à l'ordre m par Schbath (1995b)) ; le comptage $N_j(\mathbf{w})$ de $\mathbf{w}$ sur le segment $\mathbf{X}_j$ vérifie :

$$|\mathbf{s}_j|^{-1/2}\big(N_j(\mathbf{w}) - \hat{\mathbb{E}}_{m,j}(\mathbf{w})\big) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2_{m,e_j}(\mathbf{w})),$$

où e_j désigne l'état du segment $\mathbf{X}_j$, où

$$\hat{\mathbb{E}}_{m,j}(\mathbf{w}) := \frac{N_j(w_1 \cdots w_{m+1}) \times \cdots \times N_j(w_{h-m} \cdots w_h)}{N_j(w_2 \cdots w_{m+1}) \times \cdots \times N_j(w_{h-m} \cdots w_{h-1})}, \tag{7.3}$$

et où pour tout $s \in \mathcal{S}$,

$$\begin{aligned}\sigma^2_{m,s}(\mathbf{w}) := \mu_s(\mathbf{w}) + 2 \sum_{p \in \mathcal{P}(\mathbf{w}), p \leq h-m-1} \mu_s(\mathbf{w}^{(p)}\mathbf{w}) + [\mu_s(\mathbf{w})]^2 \Bigg(\sum_{y_1 \cdots y_m} \frac{[n_{\mathbf{w}}(y_1 \cdots y_m +)]^2}{\mu_s(y_1 \cdots y_m)} \\ - \sum_{y_1 \cdots y_{m+1}} \frac{[n_{\mathbf{w}}(y_1 \cdots y_{m+1})]^2}{\mu_s(y_1 \cdots y_{m+1})} + \frac{1 - 2n_{\mathbf{w}}(w_1 \cdots w_m +)}{\mu_s(w_1 \cdots w_m)} \Bigg).\end{aligned} \tag{7.4}$$

Rappelons que dans l'expression (7.4), $n_{\mathbf{w}}(y_1 \cdots y_m +)$ désigne le comptage de $y_1 \cdots y_m$ suivi d'une lettre quelconque, dans le mot $\mathbf{w}$.

Remarque 7.1 *Dans le cas où $m = 0$, $\hat{\mathbb{E}}_{0,j}(\mathbf{w})$ vaut $\frac{N_j(w_1)}{n} \times \cdots \times \frac{N_j(w_h)}{n}$ et la dernière grande parenthèse de l'égalité (7.4) est égale à $(h-1)^2 - \sum_{x \in \mathcal{A}} n_{\mathbf{w}}(x)/\mu_s(x)$.*

Par suite, comme les comptages N_j sont indépendants, on obtient,

$$\begin{aligned}\frac{\sum_{j=1}^{\rho+1} N_j(\mathbf{w}) - \sum_j \hat{\mathbb{E}}_{m,j}(\mathbf{w})}{\sqrt{n}} &= \sum_{j=1}^{\rho+1} \left(\frac{|\mathbf{s}_j|}{n}\right)^{1/2} \frac{N_j(\mathbf{w}) - \hat{\mathbb{E}}_{m,j}(\mathbf{w})}{\sqrt{|\mathbf{s}_j|}} \\ &\xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \sum_{j=1}^{\rho+1} q_j \sigma^2_{m,e_j}(\mathbf{w})\right).\end{aligned}$$

De plus, comme le nombre d'occurrences non-unicolores de $\mathbf{w}$ est plus petit que $h\rho = O(1)$, on a $N(\mathbf{w}) = \sum_{j=1}^{\rho+1} N_j(\mathbf{w}) + O(1)$ et nous venons de montrer la proposition suivante.

Proposition 7.2 *Supposons que* $\mathbf{X}$ *suit un modèle PSMm,* $0 \leq m \leq h-2$*. Lorsque la segmentation* $\mathbf{s}$ *suit l'asymptotique (7.1), on a*

$$\frac{N(\mathbf{w}) - \sum_{j=1}^{\rho+1} \hat{\mathbb{E}}_{m,j}}{\sqrt{n}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\Big(0, \sum_{s\in\mathcal{S}} q_s \sigma^2_{m,s}(\mathbf{w})\Big),$$

où $\hat{\mathbb{E}}_{m,j}$ *et* $\sigma^2_{m,s}(\mathbf{w})$ *sont donnés respectivement par (7.3) et (7.4).*

Remarque 7.3
1. *L'estimation dans le terme d'espérance se fait sur chaque segment* $\mathbf{X}_j$ *; on ne peut donc pas tirer profit pour ce terme d'un regroupement éventuel entre les segments de même état.*
2. *Les seuls segments qui comptent dans la loi limite sont ceux pour lesquels* $q_j > 0$*.*
3. *On a un résultat analogue pour l'approche gaussienne dite "martingale" de Prum* et al. *(1995).*

Nous essaierons dans la suite d'adopter une démarche où le nombre de ruptures est quelconque. Comme la combinatoire est plus complexe, nous nous placerons dans le cas de bases indépendantes.

7.1.2 Approximation gaussienne générale (i.e. de type "mot multicolore") dans le cas indépendant

On suppose que la séquence suit un modèle PM0 de paramètre $\theta = \{\mu_s\}_{s\in\mathcal{S}}$ vérifiant $\forall s \in \mathcal{S}$, $\forall x \in \mathcal{A}$, $\mu_s(x) > 0$. Nous notons $\mathcal{N} = (N(x_s), x \in \mathcal{A}, s \in \mathcal{S})$ la statistique exhaustive du modèle exponentiel PM0. Ici, pour simplifier les notations on note x_s la lettre x coloriée dans l'état s (plutôt que (x,s)). Lorsque nous ne faisons pas d'hypothèse particulière quant au nombre de ruptures de la segmentation, l'occurrence du mot peut avoir lieu à des positions dans des coloriages très différents et on ne dispose pas a priori de suffisamment de répétitions pour que les théorèmes ergodiques classiques puissent s'appliquer. Par conséquent, dire que la loi du comptage a un comportement asymptotiquement gaussien reste simplement une heuristique. Nous supposerons donc que la démarche de Prum *et al.* (1995) "s'étend" dans le cadre hétérogène, à savoir que :

Heuristique 7.4 $\mathcal{L}\Big(n^{-1/2}\big(N(\mathbf{w}) - \mathbb{E}[N(\mathbf{w})|\mathcal{N}]\big)\Big) \simeq \mathcal{N}(0, \sigma^2_{\mathbf{w}})$*, avec* $\sigma^2_{\mathbf{w}} = \lim_{n\to\infty}(\mathbb{V}[N(\mathbf{w})|\mathcal{N}]/n)$*.*

Cette démarche est intéressante, car elle prend implicitement en compte l'estimation des paramètres : l'espérance conditionnelle est l'estimateur de $\mathbb{E}N(\mathbf{w})$ de variance minimale parmi la classe des estimateurs sans biais (d'après le lemme de Lehmann Scheffe ; voir par exemple le théorème 2 p.92 et le théorèmes 4-5 p.61-62 de Monfort (1997)), et la variance conditionnelle mesure précisément l'écart entre $N(\mathbf{w})$ et $\mathbb{E}[N(\mathbf{w})|\mathcal{N}]$. Nous nous proposons ainsi d'évaluer $\mathbb{E}[N(\mathbf{w})|\mathcal{N}]$ et $\mathbb{V}(N(\mathbf{w})|\mathcal{N})$ en suivant la démarche de Schbath (1995b). Notons $\hat{\mathbb{E}}(N(\mathbf{w}))$ et

$\hat{\mathbb{V}}(N(\mathbf{w}))$ les estimateurs "plug-in" de l'espérance et de la variance de $\mathbf{w}$:

$$\begin{aligned}
\hat{\mathbb{E}}(N(\mathbf{w})) &:= \sum_{\mathbf{t}=t_1\dots t_h} n_{\mathbf{s}}(\mathbf{t})\hat{\mu}_{\mathbf{t}}(\mathbf{w}),\\
\hat{\mathbb{V}}(N(\mathbf{w})) &:= \sum_{\mathbf{t}=t_1\dots t_h} n_{\mathbf{s}}(\mathbf{t})\hat{\mu}_{\mathbf{t}}(\mathbf{w})(1-\hat{\mu}_{\mathbf{t}}(\mathbf{w}))\\
&+ 2\sum_{p=1}^{h-1}\sum_{\mathbf{t}=t_1\dots t_{h+p}} n_{\mathbf{s}}(\mathbf{t})\Big[\mathbf{1}\{p\in\mathcal{P}(\mathbf{w})\}\hat{\mu}_{\mathbf{t}}(\mathbf{w}^{(p)}\mathbf{w}) - \hat{\mu}_{\mathbf{t}^{(h)}}(\mathbf{w})\hat{\mu}_{\mathbf{t}_{(h)}}(\mathbf{w})\Big],
\end{aligned}$$

où $\hat{\mu}_{\mathbf{t}}(\mathbf{w}) = \hat{\mu}_{t_1}(w_1)\dots\hat{\mu}_{t_h}(w_h)$ et $\hat{\mu}_s(x) = N(x_s)/n_{\mathbf{s}}(s)$ pour $x\in\mathcal{A}$ et $s\in\mathcal{S}$. La proposition suivante met en évidence des équivalents de $\mathbb{E}[N(\mathbf{w})|\mathcal{N}]$ et $\mathbb{V}(N(\mathbf{w})|\mathcal{N})$ (lorsque n tend vers l'infini), fonctions respectivement de $\hat{\mathbb{E}}(N(\mathbf{w}))$ et $\hat{\mathbb{V}}(N(\mathbf{w}))$.

Théorème 7.5 *Considérons un mot $\mathbf{w}$ fixe avec n. Pour toute segmentation $\mathbf{s}$ suivant l'asymptotique (7.2), on a :*

$$\begin{aligned}
\mathbb{E}[N(\mathbf{w})|\mathcal{N}] &= \hat{\mathbb{E}}(N(\mathbf{w})) + O(1)\\
\mathbb{V}[N(\mathbf{w})|\mathcal{N}] &= \hat{\mathbb{V}}(N(\mathbf{w})) + n\sum_{s\in\mathcal{S}}\frac{n}{n_{\mathbf{s}}(s)}\Bigg[\Bigg(\frac{1}{n}\sum_{\mathbf{t}=t_1\dots t_h} n_{\mathbf{s}}(\mathbf{t})\hat{\mu}_{\mathbf{t}}(\mathbf{w})n_{\mathbf{t}}(s)\Bigg)^2\\
&\quad -\sum_{x\in\mathcal{A}}\frac{n_{\mathbf{s}}(s)}{N(x_s)}\Bigg(\frac{1}{n}\sum_{\mathbf{t}=t_1\dots t_h} n_{\mathbf{s}}(\mathbf{t})\hat{\mu}_{\mathbf{t}}(\mathbf{w})n_{\mathbf{w},\mathbf{t}}(x_s)\Bigg)^2\Bigg] + O(1). \qquad (7.5)
\end{aligned}$$

Remarque 7.6 1. *Ce résultat est général dans le sens où le nombre de ruptures dans la segmentation peut être quelconque.*

2. *Les $O(\cdot)$ sont valables point par point en $\mathbf{X}$ (dire que les O sont valables presque sûrement n'a pas vraiment d'utilité puisque tous les événements non vides sont de probabilités strictement positives).*

Corollaire 7.7 *Dans les conditions du théorème 7.5, la variance conditionnelle vérifie la convergence ponctuelle :*

$$\begin{aligned}
&\lim_{n\to\infty}\left(\frac{1}{n}\mathbb{V}[N(\mathbf{w})|\mathcal{N}]\right)\\
&= \sum_{\mathbf{t}=t_1\dots t_h} q_{\mathbf{t}}\mu_{\mathbf{t}}(\mathbf{w})(1-\mu_{\mathbf{t}}(\mathbf{w}))\\
&\quad + 2\sum_{p=1}^{h-1}\sum_{\mathbf{t}=t_1\dots t_{h+p}} q_{\mathbf{t}}\Big[\mathbf{1}\{p\in\mathcal{P}(\mathbf{w})\}\mu_{\mathbf{t}}(\mathbf{w}^{(p)}\mathbf{w}) - \mu_{\mathbf{t}^{(h)}}(\mathbf{w})\mu_{\mathbf{t}_{(h)}}(\mathbf{w})\Big],\\
&\quad + \sum_{s\in\mathcal{S}}\frac{1}{q_s}\Bigg[\Bigg(\sum_{\mathbf{t}=t_1\dots t_h} q_{\mathbf{t}}\mu_{\mathbf{t}}(\mathbf{w})n_{\mathbf{t}}(s)\Bigg)^2 - \sum_{x\in\mathcal{A}}\frac{1}{\mu_s(x)}\Bigg(\sum_{\mathbf{t}=t_1\dots t_h} q_{\mathbf{t}}\mu_{\mathbf{t}}(\mathbf{w})n_{\mathbf{w},\mathbf{t}}(x_s)\Bigg)^2\Bigg]. \qquad (7.6)
\end{aligned}$$

Remarque 7.8 *1. Dans le cas homogène $|\mathcal{S}| = 1$, l'expression (7.6) se réduit à celle mentionnée par Schbath (1995b) (p154) :*

$$\begin{aligned}\lim_{n\to\infty}\left(\frac{1}{n}\mathbb{V}[N(\mathbf{w})|\mathcal{N}]\right) &= \mu(\mathbf{w}) + 2\sum_{p\in\mathcal{P}(\mathbf{w})}\mu(\mathbf{w}^{(p)}\mathbf{w}) + \mu(\mathbf{w})^2\left[(h-1)^2 - \sum_{x\in\mathcal{A}}\frac{n_{\mathbf{w}}(x)^2}{\mu(x)}\right] \\ &= \lim_{n\to\infty}\big(\mathbb{V}[N(\mathbf{w})]/n\big) + \mu(\mathbf{w})^2\left[h^2 - \sum_{x\in\mathcal{A}}\frac{n_{\mathbf{w}}(x)^2}{\mu(x)}\right].\end{aligned}$$

2. Dans le cas où le nombre de ruptures ρ est fixe avec n, pour tout coloriage $\mathbf{t}$ qui n'est pas unicolore, on a $q_{\mathbf{t}} = 0$, et alors $\mathbb{V}[N(\mathbf{w})|\mathcal{N}]/n \to \sum_{s\in\mathcal{S}} q_s\sigma^2_{0,s}(\mathbf{w})$. On retrouve ainsi la variance limite de l'approximation gaussienne de type "mot unicolore" de la proposition 7.2 dans le cas indépendant.

Dans l'expression de la variance conditionnelle (7.6), nous pouvons montrer que le dernier terme est strictement négatif. En effet, comme $\forall x \in \mathcal{A}$, $\forall s \in \mathcal{S}$, $\mu_s(x) \in]0,1[$ et $\sum_{x\in\mathcal{A}} n_{\mathbf{w},\mathbf{t}}(x_s) = n_{\mathbf{t}}(s)$, on a par stricte convexité de $z \in \mathbb{R} \mapsto z^2$,

$$\begin{aligned}\sum_{x\in\mathcal{A}}\frac{1}{\mu_s(x)}\left(\sum_{\mathbf{t}=t_1\dots t_h} q_{\mathbf{t}}\mu_{\mathbf{t}}(\mathbf{w})n_{\mathbf{w},\mathbf{t}}(x_s)\right)^2 &= \sum_{x\in\mathcal{A}}\mu_s(x)\left(\frac{\sum_{\mathbf{t}=t_1\dots t_h} q_{\mathbf{t}}\mu_{\mathbf{t}}(\mathbf{w})n_{\mathbf{w},\mathbf{t}}(x_s)}{\mu_s(x)}\right)^2 \\ &> \left(\sum_{\mathbf{t}=t_1\dots t_h} q_{\mathbf{t}}\mu_{\mathbf{t}}(\mathbf{w})n_{\mathbf{t}}(s)\right)^2.\end{aligned}$$

Par conséquent, comme la somme des deux premiers termes de (7.6) est égale à $\lim_{n\to\infty}(\mathbb{V}[N(\mathbf{w})]/n)$ nous venons de prouver que, de manière similaire au cas homogène, le fait de prendre en compte l'estimation des paramètres du modèle (par la statistique exhaustive du modèle) affecte la variance asymptotiquement :

$$\lim_{n\to\infty}\left(\frac{\mathbb{V}[N(\mathbf{w})|\mathcal{N}]}{n}\right) < \lim_{n\to\infty}\left(\frac{\mathbb{V}[N(\mathbf{w})]}{n}\right).$$

Dans le reste de la section, nous prouvons le théorème 7.5 ainsi que quelques lemmes annexes.

Preuve du théorème 7.5. Traitons d'abord le cas de l'espérance conditionnelle. On insiste sur le fait que comme on travaille conditionnellement à $\mathcal{N}$, la composition en lettres coloriées $\mathcal{N}$ est fixée.

$$\begin{aligned}\mathbb{E}(N(\mathbf{w})|\mathcal{N}) &= \sum_{i=1}^{n-h+1}\mathbb{E}(Y_i(\mathbf{w})|\mathcal{N}) \\ &= \sum_{\mathbf{t}=t_1\cdots t_h}\sum_{i=1}^{n-h+1}\mathbf{1}\{\mathbf{t}\text{ apparaît dans }\mathbf{s}\text{ en }i\}\mathbb{E}(Y_i(\mathbf{w})|\mathcal{N}).\end{aligned}$$

On note $\mathcal{X}$ l'ensemble des séquences qui ont $\mathcal{N}$ pour composition en lettres colorées, c'est-à-dire

$$\mathcal{X} = \big\{\mathbf{x} = x_1\dots x_n \text{ telle que } \forall s\in\mathcal{S}, x\in\mathcal{A}, n_{\mathbf{x},\mathbf{s}}(x_s) = N(x_s)\big\}.$$

Remarquons que, vue la vraisemblance du modèle PM0, la séquence $X_1 \cdots X_n$ conditionnellement à $\mathcal{N}$ suit une loi uniforme sur l'ensemble (fini) $\mathcal{X}$. Ainsi pour toute position i,

$$\mathbb{E}(Y_i(\mathbf{w})|\mathcal{N}) \quad = \quad \frac{\big|\{\text{séquences } \mathbf{x} \text{ de } \mathcal{X} \text{ qui ont } \mathbf{w} \text{ à la position } i\}\big|}{|\mathcal{X}|}.$$

Notons $\mathcal{X}\backslash^i\mathbf{w}$ l'ensemble des séquences $\mathbf{x} = x_1 \dots x_n$ de $\mathcal{X}$ qui ont $\mathbf{w}$ en i et auxquelles on a enlevé les bases $x_i \dots x_{i+h-1}$, c'est-à-dire,

$$\mathcal{X}\backslash^i\mathbf{w} = \big\{x_1 \dots x_{i-1}x_{i+h} \dots x_n, \mathbf{x} = x_1 \dots x_n \in \mathcal{X} \text{ qui a } \mathbf{w} \text{ en } i\big\}.$$

Définissons aussi $\mathbf{s}^i$ la segmentation $\mathbf{s}$ à laquelle on a enlevé les couleurs $s_i \dots s_{i+h-1}$, c'est-à-dire,

$$\mathbf{s}^i = s_1 \dots s_{i-1}s_{i+h} \dots s_n.$$

Alors, comme les lettres $x_i \dots x_{i+h-1}$ sont déterminées pour toutes les séquences qui ont $\mathbf{w}$ en i, on a :

$$\big|\{\text{séquences } \mathbf{x} \text{ de } \mathcal{X} \text{ qui ont } \mathbf{w} \text{ à la position } i\}\big| = \big|\mathcal{X}\backslash^i\mathbf{w}\big|,$$

et on obtient l'expression suivante.

$$\mathbb{E}(N(\mathbf{w})|\mathcal{N}) = \sum_{\mathbf{t}=t_1\cdots t_h} \sum_{i=1}^{n-h+1} \mathbf{1}\{\mathbf{t} \text{ apparaît dans } \mathbf{s} \text{ en } i\} \frac{|\mathcal{X}\backslash^i\mathbf{w}|}{|\mathcal{X}|}.$$

On applique ensuite le lemme 7.9 (page 111) pour conclure :

$$\begin{aligned}\mathbb{E}(N(\mathbf{w})|\mathcal{N}) \quad &= \quad \sum_{\mathbf{t}=t_1\cdots t_h} N_{\mathbf{s}}(\mathbf{t})\hat{\mu}_{\mathbf{t}}(\mathbf{w})\bigg(1 + \frac{\Delta_{\mathbf{t}}}{n} + O\big(1/n^2\big)\bigg) \\ &= \quad \sum_{\mathbf{t}=t_1\cdots t_h} N_{\mathbf{s}}(\mathbf{t})\hat{\mu}_{\mathbf{t}}(\mathbf{w}) + O(1),\end{aligned}$$

car $\Delta_{\mathbf{t}} = O(1)$ pour tout coloriage $\mathbf{t}$.
Le cas de la variance conditionnelle se traite de manière similaire ; on part de la relation

$$\mathbb{V}(N(\mathbf{w})|\mathcal{N}) = \sum_{i,j}\big(\mathbb{E}(Y_i(\mathbf{w})Y_j(\mathbf{w})|\mathcal{N}) - \mathbb{E}(Y_i(\mathbf{w})|\mathcal{N})\mathbb{E}(Y_j(\mathbf{w})|\mathcal{N})\big).$$

Puis, en séparant les cas "$i = j$", "$0 < |i-j| < h$" et "$|i-j| \geq h$", on obtient :

$$\begin{aligned}\mathbb{V}(N(\mathbf{w})|\mathcal{N}) \quad = \quad & \sum_{i=1}^{n-h+1} \mathbb{E}(Y_i(\mathbf{w})|\mathcal{N})\big(1 - \mathbb{E}(Y_i(\mathbf{w})|\mathcal{N})\big) \\ & +2\sum_{p=1}^{h-1}\sum_{i=1}^{n-h-p+1} \big(\mathbb{E}(Y_i(\mathbf{w}^{(p)}\mathbf{w})|\mathcal{N}) - \mathbb{E}(Y_i(\mathbf{w})|\mathcal{N})\mathbb{E}(Y_{i+p}(\mathbf{w})|\mathcal{N})\big) \\ & + \sum_{|i-j|\geq h} \big[\mathbb{E}(Y_i(\mathbf{w})Y_j(\mathbf{w})|\mathcal{N}) - \mathbb{E}(Y_i(\mathbf{w})|\mathcal{N})\mathbb{E}(Y_j(\mathbf{w})|\mathcal{N})\big].\end{aligned}$$

Le premier et le second terme se traitent exactement comme pour l'espérance conditionnelle. Pour traiter le dernier terme, nous calculons $\mathbb{E}(Y_i(\mathbf{w})Y_j(\mathbf{w})|\mathcal{N})$ pour $|i-j| \geq h$ de la façon suivante : notons $\mathcal{X}\backslash^{i,j}\mathbf{w}$ l'ensemble des séquences $\mathbf{x} = x_1 \dots x_n$ de $\mathcal{X}$ qui ont $\mathbf{w}$ en i et j et auxquelles on a enlevé les bases $x_i \dots x_{i+h-1}$ et $x_j \dots x_{j+h-1}$. Notons $\mathbf{s}^{i,j}$ la segmentation $\mathbf{s}$ à laquelle on a enlevé $s_i \dots s_{i+h-1}$ et $s_j \dots s_{j+h-1}$. Comme les séquences de $\mathcal{X}$ sont équiprobables et que le nombre de séquences de $\mathcal{X}$ qui ont $\mathbf{w}$ en i et j est $|\mathcal{X}\backslash^{i,j}\mathbf{w}|$, on déduit :

$$\mathbb{E}(Y_i(\mathbf{w})Y_j(\mathbf{w})|\mathcal{N}) = \frac{|\mathcal{X}\backslash^{i,j}\mathbf{w}|}{|\mathcal{X}|}.$$

Ainsi, en utilisant les lemmes 7.9 (page 111) et 7.10 (page 112), et en notant $\mathbf{t} = s_i \dots s_{i+h-1}$ et $\mathbf{t}' = s_j \dots s_{j+h-1}$, on obtient :

$$\begin{aligned}
&\mathbb{E}(Y_i(\mathbf{w})Y_j(\mathbf{w})|\mathcal{N}) - \mathbb{E}(Y_i(\mathbf{w})|\mathcal{N})\mathbb{E}(Y_j(\mathbf{w})|\mathcal{N}) \\
&= \hat{\mu}_{\mathbf{t}}(\mathbf{w})\hat{\mu}_{\mathbf{t}'}(\mathbf{w})\left[\frac{\Delta_{\mathbf{t},\mathbf{t}'} - \Delta_{\mathbf{t}} - \Delta_{\mathbf{t}'}}{n} + O\big(1/n^2\big)\right].
\end{aligned}$$

Par suite,

$$\begin{aligned}
&\sum_{|i-j|\geq h} \big[\mathbb{E}(Y_i(\mathbf{w})Y_j(\mathbf{w})|\mathcal{N}) - \mathbb{E}(Y_i(\mathbf{w})|\mathcal{N})\mathbb{E}(Y_j(\mathbf{w})|\mathcal{N})\big] \\
&\quad = \frac{1}{n}\sum_{\mathbf{t}}\sum_{\mathbf{t}'}\left[\sum_{|i-j|\geq h} \mathbf{1}\{\mathbf{t} \text{ et } \mathbf{t}' \text{ apparaissent dans } \mathbf{s} \text{ respectivement aux positions } i \text{ et } j\}\right] \\
&\quad\quad \times \hat{\mu}_{\mathbf{t}}(\mathbf{w})\hat{\mu}_{\mathbf{t}'}(\mathbf{w})\Big[\Delta_{\mathbf{t},\mathbf{t}'} - \Delta_{\mathbf{t}} - \Delta_{\mathbf{t}'}\Big] + O(1) \\
&\quad = \frac{1}{n}\sum_{\mathbf{t}}\sum_{\mathbf{t}'} n_{\mathbf{s}}(\mathbf{t})n_{\mathbf{s}}(\mathbf{t}')\hat{\mu}_{\mathbf{t}}(\mathbf{w})\hat{\mu}_{\mathbf{t}'}(\mathbf{w})\Big[\Delta_{\mathbf{t},\mathbf{t}'} - \Delta_{\mathbf{t}} - \Delta_{\mathbf{t}'}\Big] + O(1).
\end{aligned}$$

Nous utilisons alors la relation $\forall x, y \in \mathbb{R}, \frac{(x+y)(x+y-1)}{2} - \frac{x(x-1)}{2} - \frac{y(y-1)}{2} = xy$ pour déduire :

$$\Delta_{\mathbf{t},\mathbf{t}'} - \Delta_{\mathbf{t}} - \Delta_{\mathbf{t}'} = \sum_{\mathbf{s}\in\mathcal{S}} n_{\mathbf{t}}(s)n_{\mathbf{t}'}(s)\frac{n}{n_{\mathbf{s}}(s)} - \sum_{s\in\mathcal{S}}\sum_{x\in\mathcal{A}} n_{(\mathbf{w},\mathbf{t})}(x_s)n_{(\mathbf{w},\mathbf{t}')}(x_s)\frac{n}{N(x_s)},$$

et nous regroupons ensuite les sommes sur $\mathbf{t}$ et $\mathbf{t}'$ pour conclure. ∎

Lemme 7.9 *En notant* $\mathbf{t} = s_i \dots s_{i+h-1}$,

$$\frac{|\mathcal{X}\backslash^{i}\mathbf{w}|}{|\mathcal{X}|} = \hat{\mu}_{\mathbf{t}}(\mathbf{w})\left(1 + \frac{\Delta_{\mathbf{t}}}{n} + O\big(1/n^2\big)\right),$$

où

$$\Delta_{\mathbf{t}} = \sum_{s\in\mathcal{S}} \frac{n_{\mathbf{t}}(s)(n_{\mathbf{t}}(s)-1)}{2}\frac{n}{n_{\mathbf{s}}(s)} + \sum_{s\in\mathcal{S}}\sum_{x\in\mathcal{A}} \frac{n_{(\mathbf{w},\mathbf{t})}(x_s)(n_{(\mathbf{w},\mathbf{t})}(x_s)-1)}{2}\frac{n}{N(x_s)}$$

Preuve. On applique le lemme 7.11 (page 112) simultanément à l'ensemble de séquences $\mathcal{X}$ avec la segmentation $\mathbf{s}$ et à l'ensemble de séquences $\mathcal{X}\backslash^{i}\mathbf{w}$ avec pour segmentation $\mathbf{s}^i$, en remarquant que le nombre d'occurrences d'une lettre coloriée x_s commun aux séquences de $\mathcal{X}\backslash^{i}\mathbf{w}$ est $N(x_s) - n_{(\mathbf{w},\mathbf{t})}(x_s)$ et le nombre d'occurrences de l'état s dans $\mathbf{s}^i$ est $n_{\mathbf{s}}(s) - n_{\mathbf{t}}(s)$ et on obtient

$$\frac{|\mathcal{X}\backslash^{i}\mathbf{w}|}{|\mathcal{X}|} = \prod_{s\in\mathcal{S}}\left[\frac{(n_{\mathbf{s}}(s) - n_{\mathbf{t}}(s))!}{n_{\mathbf{s}}(s)!}\prod_{x\in\mathcal{A}}\frac{N(x_s)!}{(N(x_s) - n_{(\mathbf{w},\mathbf{t})}(x_s))!}\right].$$

Par suite, comme on a pour tout d entier fixé, $\frac{m!}{(m-d)!} = m^d\left(1 - \frac{d(d-1)}{2m} + O\big(\frac{1}{m^2}\big)\right)$,

$$\begin{aligned}
\frac{|\mathcal{X}\backslash^i \mathbf{w}|}{|\mathcal{X}|} &= \prod_{s\in\mathcal{S}}\Bigg[\Bigg\{n_{\mathbf{s}}(s)^{n_{\mathbf{t}}(s)}\Bigg(1 - \frac{n_{\mathbf{t}}(s)(n_{\mathbf{t}}(s)-1)}{2n_{\mathbf{s}}(s)} + O\big(1/n_{\mathbf{s}}^2(s)\big)\Bigg)\Bigg\}^{-1} \\
&\quad \times \prod_{x\in\mathcal{A}} N(x_s)^{n_{(\mathbf{w},\mathbf{t})}(x_s)}\Bigg(1 - \frac{n_{(\mathbf{w},\mathbf{t})}(x_s)(n_{(\mathbf{w},\mathbf{t})}(x_s)-1)}{2N(x_s)} + O\big(1/N^2(x_s)\big)\Bigg)\Bigg].
\end{aligned}$$

Par hypothèse, $n_{\mathbf{s}}(s) \sim q_s n$, et donc comme $N(x_s) \sim \mu_s(x) n_{\mathbf{s}}(s)$ ($\mu_s(x) > 0$), on a $N(x_s) \sim q_s \mu_s(x) n$. On obtient donc :

$$\begin{aligned}
\frac{|\mathcal{X}\backslash^i \mathbf{w}|}{|\mathcal{X}|} &= \Bigg[\prod_{s\in\mathcal{S}} \frac{\prod_{x\in\mathcal{A}} N(x_s)^{n_{(\mathbf{w},\mathbf{t})}(x_s)}}{n_{\mathbf{s}}(s)^{n_{\mathbf{t}}(s)}}\Bigg] \prod_{s\in\mathcal{S}}\Bigg[\Bigg(1 + \frac{n_{\mathbf{t}}(s)(n_{\mathbf{t}}(s)-1)}{2n_{\mathbf{s}}(s)} + O\big(1/n^2\big)\Bigg) \\
&\quad \times \Bigg(1 - \sum_{x\in\mathcal{A}} \frac{n_{(\mathbf{w},\mathbf{t})}(x_s)(n_{(\mathbf{w},\mathbf{t})}(x_s)-1)}{2N(x_s)} + O\big(1/n^2\big)\Bigg)\Bigg] \\
&= \hat{\mu}_{\mathbf{t}}(\mathbf{w}) \prod_{s\in\mathcal{S}}\Bigg(1 + \frac{n_{\mathbf{t}}(s)(n_{\mathbf{t}}(s)-1)}{2n_{\mathbf{s}}(s)} - \sum_{x\in\mathcal{A}} \frac{n_{(\mathbf{w},\mathbf{t})}(x_s)(n_{(\mathbf{w},\mathbf{t})}(x_s)-1)}{2N(x_s)} + O\big(1/n^2\big)\Bigg) \\
&= \hat{\mu}_{\mathbf{t}}(\mathbf{w})\Bigg(1 + \frac{\Delta_{\mathbf{t}}}{n} + O\big(1/n^2\big)\Bigg).
\end{aligned}$$

■

Lemme 7.10 *En notant* $\mathbf{t} = s_i \dots s_{i+h-1}$ *et* $\mathbf{t}' = s_j \dots s_{j+h-1}$ *(avec* $|i-j| \geq h$*),*

$$\frac{|\mathcal{X}\backslash^{i,j} \mathbf{w}|}{|\mathcal{X}|} = \hat{\mu}_{\mathbf{t}}(\mathbf{w})\hat{\mu}_{\mathbf{t}'}(\mathbf{w})\Bigg(1 + \frac{\Delta_{\mathbf{t},\mathbf{t}'}}{n} + O\big(1/n^2\big)\Bigg),$$

où

$$\begin{aligned}
\Delta_{\mathbf{t},\mathbf{t}'} &= \sum_{s\in\mathcal{S}} \frac{(n_{\mathbf{t}}(s) + n'_{\mathbf{t}}(s))(n_{\mathbf{t}}(s) + n'_{\mathbf{t}}(s) - 1)}{2} \frac{n}{n_{\mathbf{s}}(s)} \\
&\quad + \sum_{s\in\mathcal{S}}\sum_{x\in\mathcal{A}} \frac{(n_{(\mathbf{w},\mathbf{t})}(x_s) + n_{(\mathbf{w},\mathbf{t}')}(x_s))(n_{(\mathbf{w},\mathbf{t})}(x_s) + n_{(\mathbf{w},\mathbf{t}')}(x_s) - 1)}{2} \frac{n}{N(x_s)}.
\end{aligned}$$

Preuve. La preuve est totalement analogue à celle du lemme 7.9. ■

Lemme 7.11 *Pour un ensemble* $\mathcal{X}$ *de séquences ayant une composition commune de lettres coloriées égale à* $(N(x_s), x \in \mathcal{A}, s \in \mathcal{S})$ *et pour segmentation associée* $\mathbf{s}$*,*

$$|\mathcal{X}| = \prod_{s\in\mathcal{S}} \frac{n_{\mathbf{s}}(s)!}{\prod_{x\in\mathcal{A}} N(x_s)!}$$

Preuve. Travaillons avec l'état s ; il y a $n_{\mathbf{s}}(s)$ lettres au total qui sont dans l'état s, parmi celles-ci on doit choisir pour chaque $x \in \mathcal{A}$ $N(x_s)$ positions. Ceci fait $\dfrac{n_{\mathbf{s}}(s)!}{\prod_{x\in\mathcal{A}} N(x_s)!}$ possibilités. Il reste juste à faire le produit sur les états possibles. ■

7.2 Calcul exact pour la loi du comptage

Le calcul exact de la loi du comptage d'un mot est basé sur des astuces algorithmiques. Nous allons suivre la méthode utilisée par Robin *et al.* (2003a) (voir aussi Blom and Thorburn (1982) et Chrysaphinou and Papastavridis (1990)), qui établit dans le cadre homogène une récurrence pour calculer la loi du comptage. J'en propose ici une adaptation lorsque la séquence $\mathbf{X} = (X_i)_{i\in\mathbb{Z}}$ suit un modèle PSM1. Fixons un mot quelconque $\mathbf{w} = w_1 \dots w_h$ et notons $T_\ell(\mathbf{w})$ la position aléatoire de la ℓ-ième occurrence de $\mathbf{w}$ dans la séquence $(X_i)_{i\geq 1}$ (avec la convention $T_\ell(\mathbf{w}) = \infty$ si la ℓ-ième occurrence n'apparaît jamais). Pour $i \geq 1$ et $\ell \geq 1$, notons $p(i,\ell)$ la probabilité que la ℓ-ième occurrence de $\mathbf{w}$ soit en position i, c'est-à-dire $p(i,\ell) = \mathbb{P}(T_\ell(\mathbf{w}) = i)$. Le lien entre la loi de $T_\ell(\mathbf{w})$ et la loi de $N(\mathbf{w})$ est donnée par la relation suivante : pour tout $\ell \geq 1$,

$$\mathbb{P}(N(\mathbf{w}) \geq \ell) = \mathbb{P}(T_\ell(\mathbf{w}) \leq n-h+1) = \sum_{i=1}^{n-h+1} p(i,\ell). \tag{7.7}$$

Une relation de récurrence sur les $p(i,\ell)$ peut s'obtenir à partir de la remarque suivante : pour $\ell \geq 1$ et $i \geq 1$ fixés, si on a une occurrence de $\mathbf{w}$ à la position i, soit cette occurrence est la k-ième avec $1 \leq k \leq \ell$, soit la ℓ-ième occurrence de $\mathbf{w}$ a eu lieu à une position j, $\ell \leq j < i$:

$$Y_i(\mathbf{w}) = \sum_{k=1}^{\ell} \mathbf{1}\{T_k(\mathbf{w}) = i\} + \sum_{j=\ell}^{i-1} \mathbf{1}\{T_\ell(\mathbf{w}) = j\} Y_i(\mathbf{w}),$$

et donc en intégrant,

$$\mathbb{E}Y_i(\mathbf{w}) = \sum_{k=1}^{\ell-1} p(i,k) + p(i,\ell) + \sum_{j=\ell}^{i-1} p(j,\ell)\mathbb{P}(Y_i(\mathbf{w}) = 1 | T_\ell(\mathbf{w}) = j),$$

Or, par la propriété de Markov (valable aussi dans un modèle PSM1), $\mathbb{P}(Y_i(\mathbf{w}) = 1 | T_\ell(\mathbf{w}) = j) = \mathbb{P}(Y_i(\mathbf{w}) = 1 | Y_j(\mathbf{w}) = 1)$. On a donc la relation :

$$\mathbb{P}(Y_i(\mathbf{w}) = 1 | Y_j(\mathbf{w}) = 1) = \left\{ \begin{array}{lll} \pi_{s_{j+h-1}\cdots s_{i+h-1}}(\mathbf{w}_{(i-j+1)}) \mathbf{1}\{i-j \in \mathcal{P}(\mathbf{w})\} & \text{si} & i-j < h \\ \mathbb{P}(X_i = w_1 | X_{j+h-1} = w_h) \pi_{s_i \cdots s_{i+h-1}}(\mathbf{w}) & \text{si} & i-j \geq h \end{array} \right. .$$

Ainsi, en effectuant les changements de variables $''q = i-j''$ et $''d = i-j-h''$, nous obtenons la formule de récurrence :

$$\begin{aligned} p(i,\ell) \quad = \quad & \mu_{s_i\cdots s_{i+h-1}}(\mathbf{w}) - \sum_{k=1}^{\ell-1} p(i,k) - \sum_{q\in\mathcal{P}(\mathbf{w})} p(i-q,\ell) \pi_{s_{i-q+h-1}\cdots s_{i+h-1}}(\mathbf{w}_{(q+1)}) \\ & - \sum_{d=0}^{i-\ell-h} p(i-d-h,\ell) \mathbb{P}(X_i = w_1 | X_{i-d-1} = w_h) \pi_{s_i\cdots s_{i+h-1}}(\mathbf{w}), \end{aligned} \tag{7.8}$$

où la quantité $\mathbb{P}(X_i = w_1 | X_{i-d-1} = w_h)$ se calcule en faisant le produit des matrices de transition correspondantes et où l'on a adopté la convention $p(j,\ell) = 0$ pour tout $j \leq 0$. Notons que l'on a $p(i,\ell) = 0$ dès que $\ell > i$. Avec l'initialisation $p(1,1) = \mu_{s_1\cdots s_h}(\mathbf{w})$, la relation (7.8) donne un algorithme explicite pour calculer les $p(i,\ell)$, et donc pour calculer la loi du comptage selon (7.7).

Cependant, $p(i,\ell)$ s'exprime en fonction des $p(i,k)$, $k<\ell$ et des $p(j,\ell)$, $j<i$, ce qui rend l'algorithme trop complexe en pratique : on explore des indices d bien trop grands. On propose ainsi d'utiliser l'approximation

$$\mathbb{P}(X_i = w_1 | X_{i-d-1} = w_h) \simeq \mu_{s_i}(w_1),$$

pour une distance $d \geq d^\star$, où $d^\star$ est un seuil à fixer. Elle se justifie de la manière suivante : si $i-d-1$ et i sont dans des segments différents, alors on a indépendance entre X_i et X_{i-d-1} et $\mathbb{P}(X_i = w_1|X_{i-d-1} = w_h) = \mu_{s_i}(w_1)$. Sinon, $i-d-1$ et i sont dans le même segment, et le théorème ergodique nous garantit que $\mathbb{P}(X_i = w_1|X_{i-d-1} = w_h) \to \mu_{s_i}(w_1)$ lorsque d tend vers l'infini. Dans ce cas, on suppose donc que la convergence s'effectue à distance finie $d^\star$. La récurrence (7.8) où l'on a remplacé $\mathbb{P}(X_i = w_1|X_{i-d-1} = w_h)$ par $\mu_{s_i}(w_1)$ pour une distance $d > d^\star$ ne donnant pas exactement les $p(i,\ell)$, on note donc $\widetilde{p}(i,\ell)$ les probabilités vérifiant cette nouvelle récurrence :

$$\begin{aligned}
\widetilde{p}(i,\ell) \quad = \quad & \mu_{s_i\cdots s_{i+h-1}}(\mathbf{w}) - \sum_{k=1}^{\ell-1} \widetilde{p}(i,k) - \sum_{q\in\mathcal{P}(\mathbf{w})} \widetilde{p}(i-q,\ell)\pi_{s_{i-q+h-1}\cdots s_{i+h-1}}(\mathbf{w}_{(q+1)}) \\
& - \sum_{d=0}^{d^\star-1} \widetilde{p}(i-d-h,\ell)\mathbb{P}(X_i = w_1|X_{i-d-1} = w_h)\pi_{s_i\cdots s_{i+h-1}}(\mathbf{w}) \\
& - \mu_{s_i\cdots s_{i+h-1}}(\mathbf{w}) \sum_{d=d^\star}^{i-\ell-h} \widetilde{p}(i-d-h,\ell).
\end{aligned}$$

Pour diminuer la complexité de la récurrence utilisant la formule ci-dessus, nous allons chercher à réduire le support de la somme $\sum_{d=d^\star}^{i-\ell-h} \widetilde{p}(i-d-h,\ell)$. Pour cela, l'idée est d'exprimer $\widetilde{p}(i,\ell)$ en fonction de $\widetilde{p}(i^\star,\ell)$, où $\mu_{s_i\cdots s_{i+h-1}}(\mathbf{w}) = \mu_{s_{i^\star}\cdots s_{i^\star+h-1}}(\mathbf{w})$; supposons $i \geq 2$ et notons

$$i^\star = \max\{j < i \mid s_j \cdots s_{j+h-1} = s_i \cdots s_{i+h-1} \text{ ou } j = 1\}$$

la première position j antérieure à i pour laquelle les h-coloriages présents en i et j coïncident (ou bien $i^\star = 1$ s'il n'en existe pas). On remarque que $i^\star$ est bien entendu fonction de i mais, par souci de clarté, on ne le fait pas apparaître explicitement dans les notations. Nous obtenons ainsi $\widetilde{p}(i,\ell)$ en fonction de $\widetilde{p}(i^\star,\ell)$ de la façon suivante :

$$\begin{aligned}
\widetilde{p}(i,\ell) \quad = \quad & \widetilde{p}(i^\star,\ell) + \sum_{k=1}^{\ell-1} \left[\widetilde{p}(i^\star,k) - \widetilde{p}(i,k)\right] - \mu_{s_i\cdots s_{i+h-1}}(\mathbf{w}) \sum_{j=i^\star-d^\star-h+1}^{i-d^\star-h} \widetilde{p}(j,\ell) \\
& + \sum_{q\in\mathcal{P}(\mathbf{w})} \left[\widetilde{p}(i^\star-q,\ell)\pi_{s_{i^\star-q+h-1}\cdots s_{i^\star+h-1}}(\mathbf{w}_{(q+1)}) - \widetilde{p}(i-q,\ell)\pi_{s_{i-q+h-1}\cdots s_{i+h-1}}(\mathbf{w}_{(q+1)})\right] \\
& + \sum_{d=0}^{d^\star-1} \left[\widetilde{p}(i^\star-d-h,\ell)\mathbb{P}(X_{i^\star} = w_1|X_{i^\star-d-1} = w_h)\pi_{s_{i^\star}\cdots s_{i^\star+h-1}}(\mathbf{w}) \right. \\
& \left. - \widetilde{p}(i-d-h,\ell)\mathbb{P}(X_i = w_1|X_{i-d-1} = w_h)\pi_{s_i\cdots s_{i+h-1}}(\mathbf{w})\right]. \qquad (7.9)
\end{aligned}$$

Dans cette expression, $\widetilde{p}(i,\ell)$ ne s'exprime plus qu'en fonction de $\widetilde{p}(i^\star,k)$ et $\widetilde{p}(i,k)$ pour $1 \leq k \leq \ell-1$ et de $\widetilde{p}(j,\ell)$ pour $i^\star - d^\star - h + 1 \leq j \leq i-1$. Ainsi, à condition que $i^\star$ soit "proche" de i, la complexité de la récurrence utilisant l'expression (7.9) est bien plus petite que celle de l'expression (7.8). Par exemple :

- Lorsque la position i n'est pas située en début de segment et appartient à un segment de longueur plus grande que h, $i^\star = i - 1$ et la somme indexée par j dans (7.9) se réduit au terme $\widetilde{p}(i - d^\star - h, \ell)$.
- Si la segmentation est périodique de période p, alors le même coloriage apparaît régulièrement toutes les p positions et $i - i^\star = p$.

De manière générale, la récurrence sera "efficace" si la plupart des coloriages apparaissent plusieurs fois et de manière "bien répartie" dans la segmentation $\mathbf{s}$.

Remarque 7.12 *1. Comme on n'a pas d'idée théorique de l'écart entre $\widetilde{p}(i, \ell)$ et $p(i, \ell)$, il faudrait faire une étude pratique pour valider l'algorithme. Cependant l'algorithme proposé est très proche de celui utilisé dans le cas homogène, et ce dernier donne de très bons résultats en pratique à condition que $d^\star$ soit suffisamment grand.*

2. Même en utilisant la formule simplifiée (7.9), le temps de calcul de cet algorithme ne sera raisonnable que pour des séquences courtes (par exemple $n \leq 10000$).

3. Lorsqu'on utilise cette méthode, l'estimation des paramètres se fait de manière "plug-in" et l'influence de l'estimation dans la procédure est négligée.

4. Cette méthode est généralisable au cas des familles de mots.

7.3 Estimation dans un modèle PM

Nous avons supposé dans le chapitre 3 que les paramètres du modèle PM étaient connus. Nous proposons ici de traiter le problème de l'estimation des paramètres dans un modèle PM. Dans un premier temps, nous calculerons l'estimateur du maximum de vraisemblance du modèle PM1 et PSM1. Par suite, nous estimerons les paramètres de la loi $\mathcal{CP}_{\text{uni}}$ (cf. Section 3.4.2) et nous montrerons que la loi $\widehat{\mathcal{CP}}_{\text{uni}}$ obtenue avec les paramètres estimés a encore une erreur d'approximation qui tend vers 0 lorsque le mot est rare.

7.3.1 Maximum de vraisemblance dans un modèle PM1

Rappelons que $N(xy)$ désigne le nombre d'occurrences de xy dans la séquence et que $N(x+) = \sum_y N(xy)$. Il est connu que dans un modèle de Markov homogène, l'estimateur du maximum de vraisemblance de la probabilité de transition est donné par $\hat{\pi}(x, y) = N(xy)/N(x+)$ lorsque $N(x+) > 0$ (et $\hat{\pi}(x, .)$ quelconque si $N(x+) = 0$) (cf. Robin *et al.* (2003a) ou encore Dacunha-Castelle and Duflo (1983) pour une étude générale). La mesure stationnaire (qui existe en supposant la chaîne irréductible) est généralement estimée par ailleurs avec l'estimateur $\hat{\mu}(x) = N(x)/n$. Nous allons ici généraliser ces résultats au cas d'un modèle PM1.

Le modèle PM1 (cf. Section 3.1.2 du chapitre 3) est défini à partir d'une segmentation $\mathbf{s} = s_1 \cdots s_n$ fixe et d'une famille de probabilités de transitions $\{\pi_s\}_{s \in \mathcal{S}}$, comme un modèle de Markov hétérogène où la i-ème probabilité de transition est donnée par $\pi_{s_{i+1}}$. La loi initiale est définie comme la (supposée existante) mesure stationnaire associée à la probabilité de transition π_{s_1}. Ici, pour simplifier le problème, on considérera que la loi initiale est une loi γ qui n'est pas une fonction des probabilités de transition π_s, $s \in \mathcal{S}$. Ainsi, on ne fera pas d'hypothèses de récurrence et d'apériodicité concernant les chaînes de Markov relatives aux transitions π_s, $s \in \mathcal{S}$. L'espace des paramètres Θ du modèle est donc défini par :

$$\Theta \; := \; \left\{ (\gamma, \{\pi_s\}_{s\in\mathcal{S}}) \in [0,1]^{\mathcal{A}} \times [0,1]^{\mathcal{A}^2\times\mathcal{S}} \middle| \sum_{x\in\mathcal{A}} \gamma(x) = 1, \forall x \in \mathcal{A}, \forall s \in \mathcal{S}, \sum_{y\in\mathcal{A}} \pi_s(x,y) = 1 \right\}. \tag{7.10}$$

La vraisemblance du modèle a l'expression suivante : $\forall \theta = (\gamma, \{\pi_s\}_{s\in\mathcal{S}}) \in \Theta$, $\forall \mathbf{x} = x_1 \cdots x_n \in \mathcal{A}^n$,

$$\begin{aligned} \mathbb{P}_\theta(\mathbf{x}) &= \gamma(x_1) \prod_{i=2}^{n} \pi_{s_i}(x_{i-1}, x_i) \\ &= \gamma(x_1) \prod_{(x,y)\in\mathcal{A}^2, s\in\mathcal{S}} (\pi_s(x,y))^{N(xy_s)}, \end{aligned} \tag{7.11}$$

où $N(xy_s)$ désigne le nombre d'occurrences du 2-mot xy dans la séquence $\mathbf{x}$ (segmentée selon $\mathbf{s}$), avec y dans l'état s : $N(xy_s) = \sum_{i=2}^{n} \mathbf{1}\{x_{i-1}x_i = xy, s_i = s\}$. La log-vraisemblance s'écrit donc

$$\log(\mathbb{P}_\theta(\mathbf{x})) = \log(\gamma(x_1)) + \sum_{(x,y)\in\mathcal{A}^2, s\in\mathcal{S}} N(xy_s) \log(\pi_s(x,y)). \tag{7.12}$$

Il s'agit donc d'un modèle exponentiel qui a pour statistique exhaustive $(x_1, (N(xy_s))_{(x,y)\in\mathcal{A}^2, s\in\mathcal{S}})$.

On note que Θ est un fermé borné d'un espace de dimension finie, donc Θ est compact ; comme la log-vraisemblance est une fonction continue de θ (à $\mathbf{x}$ fixé), elle possède au moins un extremum global $\hat{\theta}$. L'expression du maximum de vraisemblance $\hat{\theta} = (\hat{\gamma}, \{\hat{\pi}_s\}_{s\in\mathcal{S}})$ est donnée par le lemme suivant.

Lemme 7.13 *Le maximum de vraisemblance $\hat{\theta} = (\hat{\gamma}, \{\hat{\pi}_s\}_{s\in\mathcal{S}})$ du modèle $(\mathcal{A}^n, \mathcal{P}(\mathcal{A}^n), \{\mathbb{P}_\theta\}_{\theta\in\Theta})$, avec $\mathbb{P}_\theta$ défini en (7.11) et Θ défini en (7.10), est donné par $\forall x \in \mathcal{A}$, $\hat{\gamma}(x) = \delta_{x_1}(x)$ (la mesure de Dirac en x_1), et pour tout $x \in \mathcal{A}$ et $s \in \mathcal{S}$ tel que $N(x+_s) > 0$,*

$$\forall y \in \mathcal{A}, \;\; \hat{\pi}_s(x,y) = \frac{N(xy_s)}{N(x+_s)},$$

où $N(xy_s) = \sum_{i=2}^{n} \mathbf{1}\{x_{i-1}x_i = xy, s_i = s\}$ et $N(x+_s) = \sum_{i=2}^{n} \mathbf{1}\{x_{i-1} = x, s_i = s\}$.

Pour prouver ce lemme, il suffit de regarder les points où les dérivées partielles s'annulent. Cependant, pour que cette dérivation soit licite, il faut s'assurer que le point où l'on dérive est bien à l'intérieur de Θ.

Preuve. Le fait que $\log(\gamma(x_1))$ soit maximum en $\gamma = \delta_{x_1}(.)$ est évident. Si on fixe $x \in \mathcal{A}$ et $s \in \mathcal{A}$, il reste à maximiser la fonction

$$f(\pi) := \sum_{y\in\mathcal{A}} N(xy_s) \log(\pi_s(x,y))$$

en $\pi = [\pi_s(x,y)]_{y\in\mathcal{A}} \in [0,1]^{\mathcal{A}}$ sous la contrainte $\sum_{y\in\mathcal{A}} \pi_s(x,y) = 1$. Comme on cherche à maximiser une fonction continue sur un compact, l'existence d'un point π^0 réalisant le maximum global est garantie. On note :

$$\Lambda = \{y \in \mathcal{A} \mid N(xy_s) > 0\}.$$

Si $\Lambda = \emptyset$ (i.e. $N(x+_s) = 0$), alors la fonction à maximiser vaut constamment 0 et n'importe quel π^0 maximise la fonction. Supposons donc $\Lambda \neq \emptyset$. Alors, comme f est croissante coordonnée par coordonnée en $\pi_s(x,y)$, on a $\forall y \notin \Lambda$, $\pi_s^0(x,y) = 0$. Ainsi, π_0 (restreint à ces coordonnées y dans Λ) maximise la fonction

$$f(\pi) = \sum_{y\in\Lambda} N(xy_s)\log(\pi_s(x,y))$$

en $\pi = [\pi_s(x,y)]_{y\in\Lambda} \in [0,1]^{\Lambda}$ sous la contrainte $\sum_{y\in\Lambda} \pi_s(x,y) = 1$. On distingue à présent les deux cas suivants :

- 1er cas : $|\Lambda| = 1$. Alors en notant y^0 l'unique élément de Λ tel que $N(xy_s^0) > 0$, le maximum vaut $\forall y \in \Lambda$, $\hat{\pi}_s^0(x,y) = \mathbf{1}\{y = y^0\} = N(xy_s)/N(x+_s)$.
- 2nd cas : $|\Lambda| \geq 2$. Alors on montre que $\forall y \in \Lambda$, $\pi_s^0(x,y) \in]0,1[$: dans le cas contraire, il existe un $y \in \Lambda$ tel que $\pi_s^0(x,y) = 0$ ou 1. Le cas $\pi_s^0(x,y) = 0$ est exclu car π_0 est un maximum pour f. Le cas $\pi_s^0(x,y) = 1$ donne l'existence d'un autre élément $y' \in \Lambda$ avec $\pi_s^0(x,y') = 0$ et donc cela contredit de nouveau le fait que π_0 soit un maximum pour f. Ainsi, le maximum de f est atteint en $\pi_0 \in]0,1[^{\Lambda}$, où $]0,1[^{\Lambda}$ est un ensemble **ouvert**. Donc π_0 est un point critique et il vérifie la propriété suivante : le vecteur $\left[\frac{N(xy_s)}{\pi_s^0(x,y)}\right]_{y\in\Lambda}$ est colinéaire au vecteur $[1]_{y\in\Lambda}$. Par conséquent, pour tout $y, y' \in \Lambda$,

  $$\frac{N(xy_s)}{\pi_s^0(x,y)} = \frac{N(xy'_s)}{\pi_s^0(x,y')}.$$

  De la relation $\frac{x}{y} = \frac{x'}{y'} \Rightarrow \frac{x}{y} = \frac{x'}{y'} = \frac{x+x'}{y+y'}$, on déduit $\forall y \in \Lambda$,

  $$\frac{N(xy_s)}{\pi_s^0(x,y)} = \frac{\sum_{y'\in\Lambda} N(xy'_s)}{\sum_{y'\in\Lambda} \pi_s^0(x,y')} = N(x+_s).$$

Finalement, le maximum de vraisemblance vérifie pour tout $x \in \mathcal{A}$ et $s \in \mathcal{S}$: si $N(x+_s) > 0$ alors $\pi_s^0(x,y) = N(xy_s)/N(x+_s)$ (cas 1 et 2). ■

Remarque 7.14

1. *S'il existe un $x \in \mathcal{A}$ et $s \in \mathcal{S}$ avec $N(x+_s) = 0$, les coordonnées $[\hat{\pi}_s(x,y)]_{y\in\mathcal{A}}$ de l'estimateur du maximum de vraisemblance peuvent être définies de manière quelconque (il suffit juste que $\sum_y \hat{\pi}_s(x,y) = 1$). On peut prendre par exemple $\forall y$, $\hat{\pi}_s(x,y) = 1/|\mathcal{A}|$. Aussi, il est un peu abusif de parler de **l'**estimateur du maximum de vraisemblance mais on devrait plutôt parler d'**un** estimateur du maximum de vraisemblance, car il n'est pas forcément unique. Cependant, comme dans le cas (fréquent) où pour tout $x \in \mathcal{A}$ et $s \in \mathcal{S}$ on a $N(x+_s) > 0$ il y a unicité, on a préféré garder la première terminologie.*
2. *On ne traite pas ici le cas où la loi de X_1 est la loi invariante associée à la probabilité de transition π_{s_1}. Pour estimer cette loi invariante on peut utiliser l'estimateur classique $\hat{\mu}_{s_1}(x) = N(x_{s_1})/n_{\mathbf{s}}(s_1)$ où $n_{\mathbf{s}}(s_1)$ (> 0) est le nombre d'occurrences de l'état s_1 dans la segmentation $\mathbf{s}$.*

Cas d'un modèle stationnaire par morceaux

Si on se place dans un modèle de Markov stationnaire par morceaux (cf. Section 3.1.2 du chapitre 3), la séquence est définie comme une concaténation de segments markoviens de probabilité de transition π_s, fonction de l'état s dans lequel se trouve chaque segment. Dans le modèle que nous avons proposé, nous considérons que chaque segment a pour loi initiale la loi stationnaire. Ici, de manière similaire au paragraphe précédent, si on considère la loi initiale de chaque segment comme libre, on peut trouver facilement l'expression du maximum de vraisemblance pour les transitions ; en appliquant la même méthode que dans la preuve du lemme 7.13, on prouve aisément que le maximum de vraisemblance pour les transitions s'écrit pour tout $x \in \mathcal{A}$ et $s \in \mathcal{S}$ tel que $N(x_s+_s) > 0$,

$$\forall y \in \mathcal{A}, \ \hat{\pi}_s(x,y) = \frac{N(x_sy_s)}{N(x_s+_s)},$$

où $N(x_sy_s) = \sum_{i=2}^n \mathbf{1}\{x_{i-1}x_i = xy, s_{i-1} = s_i = s\}$ et $N(x_s+_s) = \sum_{i=2}^n \mathbf{1}\{x_{i-1} = x, s_{i-1} = s_i = s\}$. Par ailleurs, la loi initiale d'un segment dans l'état s peut s'estimer par $\hat{\mu}_s(x) = N(x_s)/n_{\mathbf{s}}(s)$ pour $x \in \mathcal{A}$.

7.3.2 Estimation des paramètres de la loi $\mathcal{CP}_{\text{uni}}$

Considérons un modèle PSM1 avec une segmentation (déterministe) vérifiant l'asymptotique (7.1) (définie page 105). Dans la section 3.4.2 du chapitre 3, on a montré que lorsque un h-mot $\mathbf{w}$ vérifie l'hypothèse (3.14) (page 47) et sous les conditions de rareté $\mathbb{E}N(\mathbf{w}) = O(1)$ et $h = o(n)$, la distance en variation totale $d_{vt}\big(\mathcal{L}(N(\mathbf{w})), \mathcal{CP}_{\text{uni}}\big)$ tend vers 0 (lorsque n tend vers l'infini). Rappelons que les paramètres de la loi de Poisson composée $\mathcal{CP}_{\text{uni}}$ sont donnés par : $\forall k \geq 1$

$$\lambda_{k,\text{uni}} := \sum_{s \in \mathcal{S}} (n_{\mathbf{s}}(s) - h + 1) a_s^{k-1} (1 - a_s)^2 \mu_s(\mathbf{w}).$$

En pratique, les paramètres $\lambda_{k,\text{uni}}$ sont bien entendu inconnus, et on propose ici de les estimer en utilisant une méthode "plug in" à partir des estimateurs $\hat{\pi}_s(x,y)$ et $\hat{\mu}_s(x)$ du paragraphe précédent (cas PSM). Pour cela, nous définissons les estimateurs de $\mu_s(\mathbf{w})$ et a_s respectivement par :

$$\begin{aligned} \hat{\mu}_s(\mathbf{w}) &:= \hat{\mu}_s(w_1)\hat{\pi}_s(w_1,w_2) \times \cdots \times \hat{\pi}_s(w_{h-1},w_h), \\ \hat{a}_s &:= \sum_{p \in \mathcal{P}'(\mathbf{w})} \hat{\pi}_s(w_1,w_2) \times \cdots \times \hat{\pi}_s(w_p,w_{p+1}), \end{aligned}$$

avec $\forall x,y \in \mathcal{A}$, $\forall s \in \mathcal{S}$, $\hat{\mu}_s(x) = N(x_s)/n_{\mathbf{s}}(s)$ et $\hat{\pi}_s(x,y) = N(x_sy_s)/N(x_s+_s)$. Finalement, nous définissons $\widehat{\mathcal{CP}}_{\text{uni}}$ la loi de Poisson composée de paramètres $\hat{\lambda}_{k,\text{uni}}$ avec : $\forall k \geq 1$

$$\hat{\lambda}_{k,\text{uni}} := \sum_{s \in \mathcal{S}} (n_{\mathbf{s}}(s) - h + 1) \hat{a}_s^{k-1} (1 - \hat{s}_t)^2 \hat{\mu}_s(\mathbf{w}).$$

Le théorème suivant établit que pour approcher la loi de $N(\mathbf{w})$, la loi $\widehat{\mathcal{CP}}_{\text{uni}}$ peut être utilisée au même titre que $\mathcal{CP}_{\text{uni}}$ (sous certaines conditions).

Théorème 7.15 *Considérons une segmentation vérifiant l'asymptotique (7.1) (page 105). Pour un h-mot $\mathbf{w}$ vérifiant les hypothèses (3.14) et (3.15) (page 47), et tel que $\mathbb{E}N(\mathbf{w}) = O(1)$ et $\forall s \in \mathcal{S}$, $h^4 = o\big(\sqrt{n_{\mathbf{s}}(s)/\log\log n_{\mathbf{s}}(s)}\big)$, la distance en variation totale entre les lois de Poisson composée $\mathcal{CP}_{\text{uni}}$ et $\widehat{\mathcal{CP}}_{\text{uni}}$ tend vers 0 lorsque n tend vers l'infini. En particulier :*

$$d_{vt}\big(\mathcal{L}(N(\mathbf{w})), \widehat{\mathcal{CP}}_{\text{uni}}\big) \xrightarrow[n\to\infty]{} 0.$$

Le reste de cette section est consacrée à la preuve de ce théorème. Nous allons pour cela nous inspirer de la démarche de l'annexe C de Schbath (1995b). Pour tout segment $\mathbf{s}_j$ ($j = 1, \ldots, \rho{+}1$) de la segmentation $\mathbf{s}$ dans l'état s : pour tout $x, y \in \mathcal{A}$ d'après Meyn and Tweedie (1993) et Senoussi (1990),

$$\frac{N_j(x)}{|\mathbf{s}_j|} = \mu_s(x) + O\left(\frac{\sqrt{\log\log|\mathbf{s}_j|}}{\sqrt{|\mathbf{s}_j|}}\right) \text{ p.s.} \tag{7.13}$$

$$\frac{N_j(xy)}{N_j(x+)} = \pi_s(x,y) + O\left(\frac{\sqrt{\log\log|\mathbf{s}_j|}}{\sqrt{|\mathbf{s}_j|}}\right) \text{ p.s.} \tag{7.14}$$

Remarquons que (7.13) et $\mu_s(x) > 0$ (chaîne irréductible) impliquent que $N_j(x+)$ est non nul à partir d'un certain rang, ce qui donne en particulier un sens à (7.14). En regroupant les segments de même état, nous obtenons à partir de (7.13) et (7.14) les résultats suivants.

Lemme 7.16 *Pour tout $x, y \in \mathcal{A}$, pour tout $s \in \mathcal{S}$,*

(i) $\frac{N(x_s)}{n_{\mathbf{s}}(s)} = \mu_s(x) + O\left(\frac{\sqrt{\log\log n_{\mathbf{s}}(s)}}{\sqrt{n_{\mathbf{s}}(s)}}\right)$ *p.s.*

(ii) $\frac{N(x_s y_s)}{N(x_s+_s)} = \pi_s(x,y) + O\left(\frac{\sqrt{\log\log n_{\mathbf{s}}(s)}}{\sqrt{n_{\mathbf{s}}(s)}}\right)$ *p.s.*

(iii) Sous l'hypothèse $\mathbb{E}N(\mathbf{w}) = O(1)$ et si pour tout $s \in \mathcal{S}$, $h = o\big(\sqrt{n_{\mathbf{s}}(s)/\log\log n_{\mathbf{s}}(s)}\big)$, on a

$$(n_{\mathbf{s}}(s) - h + 1)\hat{\mu}_s(\mathbf{w}) = (n_{\mathbf{s}}(s) - h + 1)\mu_s(\mathbf{w}) + O\left(h\frac{\sqrt{\log\log n_{\mathbf{s}}(s)}}{\sqrt{n_{\mathbf{s}}(s)}}\right) \textit{ p.s.}$$

(iv) Si pour tout $s \in \mathcal{S}$, $h^2 = o\big(\sqrt{n_{\mathbf{s}}(s)/\log\log n_{\mathbf{s}}(s)}\big)$, on a

$$\hat{a}_s = a_s + O\left(h^2\frac{\sqrt{\log\log n_{\mathbf{s}}(s)}}{\sqrt{n_{\mathbf{s}}(s)}}\right) \textit{ p.s.}$$

Preuve du lemme 7.16. En considérant uniquement les indices j correspondant aux segments $\mathbf{s}_j$ dans l'état s :

$$\begin{aligned}\frac{N(x_s)}{n_{\mathbf{s}}(s)} &= \sum_j \frac{|\mathbf{s}_j|}{n_{\mathbf{s}}(s)}\frac{N_j(x)}{|\mathbf{s}_j|} \\ &= \sum_j \frac{|\mathbf{s}_j|}{n_{\mathbf{s}}(s)}\left[\mu_s(x) + O\left(\frac{\sqrt{\log\log|\mathbf{s}_j|}}{\sqrt{|\mathbf{s}_j|}}\right)\right] \\ &= \mu_s(x) + O\left(\sum_j \frac{|\mathbf{s}_j|}{n_{\mathbf{s}}(s)}\frac{\sqrt{\log\log|\mathbf{s}_j|}}{\sqrt{|\mathbf{s}_j|}}\right).\end{aligned}$$

Puis on obtient (i) par concavité de la fonction racine carrée $\sqrt{.}$:

$$\sum_j \frac{|\mathbf{s}_j|}{n_{\mathbf{s}}(s)} \frac{\sqrt{\log\log|\mathbf{s}_j|}}{\sqrt{|\mathbf{s}_j|}} \leq \sqrt{\sum_j \frac{\log\log|\mathbf{s}_j|}{n_{\mathbf{s}}(s)}} \leq \sqrt{\frac{(\rho+1)\log\log n_{\mathbf{s}}(s)}{n_{\mathbf{s}}(s)}}.$$

Nous traitons le point (ii) de manière similaire (les indices j correspondant toujours aux segments $\mathbf{s}_j$ dans l'état s) :

$$\begin{aligned} \frac{N(x_s y_s)}{N(x_s +_s)} &= \sum_j \frac{N_j(x+)}{N(x_s +_s)} \frac{N_j(xy)}{N_j(x+)} \\ &= \pi_s(x,y) + O\left(\sum_j \frac{N_j(x+)}{N(x_s +_s)} \frac{\sqrt{\log\log|\mathbf{s}_j|}}{\sqrt{|\mathbf{s}_j|}}\right). \end{aligned}$$

Puis, en notant que $N_j(x+) \leq |\mathbf{s}_j|$ et que par (i) $N(x_s +_s) \sim n_{\mathbf{s}}(s)\mu_s(x)$, on a :

$$\sum_j \frac{N_j(x+)}{N(x_s +_s)} \frac{\sqrt{\log\log|\mathbf{s}_j|}}{\sqrt{|\mathbf{s}_j|}} \leq \sqrt{\sum_j \frac{N_j(x+)}{N(x_s +_s)} \frac{\log\log|\mathbf{s}_j|}{|\mathbf{s}_j|}} = O\left(\sqrt{\frac{(\rho+1)\log\log n_{\mathbf{s}}(s)}{n_{\mathbf{s}}(s)}}\right).$$

Le point (iii) s'obtient en utilisant (i) et (ii) :

$$\begin{aligned} \hat{\mu}_s(\mathbf{w}) =& \left[\mu_s(w_1) + O\left(\frac{\sqrt{\log\log n_{\mathbf{s}}(s)}}{\sqrt{n_{\mathbf{s}}(s)}}\right)\right]\left[\pi_s(w_1,w_2) + O\left(\frac{\sqrt{\log\log n_{\mathbf{s}}(s)}}{\sqrt{n_{\mathbf{s}}(s)}}\right)\right] \times \cdots \times \\ & \left[\pi_s(w_{h-1},w_h) + O\left(\frac{\sqrt{\log\log n_{\mathbf{s}}(s)}}{\sqrt{n_{\mathbf{s}}(s)}}\right)\right]. \end{aligned}$$

Pour développer ce produit de h termes (attention h tend ici vers l'infini), nous utilisons le lemme 7.17 avec $\varepsilon_n = \sqrt{\log\log n_{\mathbf{s}}(s)}/\sqrt{n_{\mathbf{s}}(s)}$ et $u_n = h$ pour déduire :

$$\hat{\mu}_s(\mathbf{w}) = \mu_s(\mathbf{w}) + O\left(\mu_s(\mathbf{w}) h \frac{\sqrt{\log\log n_{\mathbf{s}}(s)}}{\sqrt{n_{\mathbf{s}}(s)}}\right),$$

ce qui prouve (iii). Le point (iv) est analogue à (iii). ■

Nous pouvons à présent prouver le théorème 7.15 :

Preuve du théorème 7.15. La distance en variation totale entre les deux lois de Poisson composées est majorée par

$$\begin{aligned} &\sum_{k\geq 1} |\lambda_{k,\text{uni}} - \hat{\lambda}_{k,\text{uni}}| \\ &\leq \sum_{t\in\mathcal{S}} \sum_{k\geq 1} \left| a_t^{k-1}(1-a_t)^2 (n_{\mathbf{s}}(t)-h+1)\mu_t(\mathbf{w}) - \hat{a}_t^{k-1}(1-\hat{a}_t)^2 (n_{\mathbf{s}}(t)-h+1)\hat{\mu}_t(\mathbf{w}) \right|. \end{aligned}$$

En utilisant plusieurs fois l'inégalité triangulaire et d'après le (iii) et le (iv) du lemme 7.16, il suffit de prouver que pour tout $s \in \mathcal{S}$,

$$\sum_{k \geq 1} |\hat{a}_s^{k-1} - a_s^{k-1}| \to 0.$$

Ceci s'obtient en découpant la somme ci-dessus selon les indices $k \leq h$ et $k > h$; pour la somme finie on utilise le lemme 7.17 et le (iv) du lemme 7.16 pour établir $\forall k \leq h$, $\hat{a}_s^{k-1} = a_s^{k-1} + O\left(h^3 \frac{\sqrt{\log\log n_{\mathbf{s}}(s)}}{\sqrt{n_{\mathbf{s}}(s)}}\right)$ p.s. ; pour la somme infinie on utilise que $a_s \leq \zeta$ avec $\zeta < 1$ (d'après l'hypothèse (3.15)). ■

Lemme 7.17 *Considérons une suite entière (u_n) et une suite réelle (ε_n). Pour tout $(x_{i,n})_{i \leq n}$, $n \geq 1$ avec $x_{i,n} \in [c, 1[$ (et $c > 0$ constante), dès que $u_n \varepsilon_n \to 0$, on a*

$$\prod_{i=1}^{u_n} (x_{i,n} + \varepsilon_n) = \left(\prod_{i=1}^{u_n} x_{i,n} \right) (1 + O(u_n \varepsilon_n)).$$

Preuve. En notant $x_n = \prod_{i=1}^{u_n} x_{i,n}$, on développe simplement le produit :

$$\prod_{i=1}^{u_n} (x_{i,n} + \varepsilon_n) \leq x_n + \sum_{i=1}^{u_n} \binom{u_n}{i} (\varepsilon_n)^i \frac{x_n}{c^i} \leq x_n + x_n \sum_{i=1}^{u_n} \left(\frac{u_n \varepsilon_n}{c} \right)^i = x_n (1 + O(u_n \varepsilon_n)).$$

■

Chapter 8

Testing simultaneously the exceptionality of several motifs

While computing the score of significance of several motifs in an observed sequence, a major issue is to select the real significant motifs. This chapter presents a solution in the multiple testing framework. We propose to use the Bonferroni's and k-min procedures while controlling the probability of making at least k errors. These approaches are performed and compared on a practical case. This chapter also connects the two parts of my thesis.

8.1 Framework

8.1.1 Number of occurrences of words in a random sequence

Let $\mathbf{X} = X_1 \cdots X_n$ be a random sequence of length n, composed of letters X_i in a finite alphabet $\mathcal{A}$. Denote the general distribution of $\mathbf{X}$ by P. In the case where $\mathbf{X}$ is generated by an homogeneous stationary Markov chain of order $m \geq 1$, with known transition matrix Π and stationary measure μ, the distribution of $\mathbf{X}$ is denoted by P_M and called the null distribution of $\mathbf{X}$. Fix a word length $h \geq m+2$, and denote by

$$\mathcal{W} = \big\{ \mathbf{w} = w_1 \cdots w_h, w_i \in \mathcal{A}, 1 \leq i \leq h \big\}$$

the set of words of length h on the alphabet $\mathcal{A}$. For each word $\mathbf{w} = w_1 \cdots w_h \in \mathcal{W}$, define the number of occurrences of $\mathbf{w}$ in $\mathbf{X}$ by

$$N(\mathbf{w}) = \sum_{i=1}^{n-h+1} \mathbf{1}\{X_i \cdots X_{i+h-1} = w_1 \cdots w_h\}.$$

The goal here is to test simultaneously for each word $\mathbf{w}$ of length h, if its count distribution $L_{\mathbf{w}}$ is in accordance with the one imposed by the underlying Markovian model P_M. We then test

$$H_{\mathbf{w}}\text{:“}L_{\mathbf{w}} = L_{M,\mathbf{w}}\text{” against } A_{\mathbf{w}}\text{:“}L_{\mathbf{w}} \neq L_{M,\mathbf{w}}\text{”, for all } \mathbf{w} \in \mathcal{W},$$

where $L_{M,\mathbf{w}}$ denotes the distribution of $N(\mathbf{w})$ when the sequence $\mathbf{X}$ follows the null distribution P_M.

8.1.2 Single testing

Consider first the single testing problem of testing the null hypothesis $H_{\mathbf{w}}$:“$L_{\mathbf{w}} = L_{M,\mathbf{w}}$” against the alternative $A_{\mathbf{w}}$:“$L_{\mathbf{w}} \neq L_{M,\mathbf{w}}$” for a fixed word $\mathbf{w}$ in $\mathcal{W}$. We can define a p-value for the above single test by

$$p_{\mathbf{w}} := \mathbb{P}_{N \sim L_{M,\mathbf{w}}}(N \geq N(\mathbf{w})),$$

which is defined for each realization of $\mathbf{X}$ as the probability that the number of occurrences of $\mathbf{w}$ in a sequence following the null model is larger than the one in $\mathbf{X}$. For each realization of $\mathbf{X}$, the p-value $p_{\mathbf{w}}$ measures the over-representation[1] of the word $\mathbf{w}$ in the sequence $\mathbf{X}$. Therefore, as soon as $H_{\mathbf{w}}$ is true (i.e $N(\mathbf{w}) \sim L_{M,\mathbf{w}}$), we have (by definition of the p-value):

$$\forall \alpha \in (0,1), \quad \mathbb{P}(p_{\mathbf{w}} \leq \alpha) \leq \alpha,$$

which means that the distribution of $p_{\mathbf{w}}$ is stochastically lower bounded by a uniform distribution. This implies that the test which rejects $H_{\mathbf{w}}$ when $p_{\mathbf{w}} \leq \alpha$ has a probability of rejecting a true null hypothesis controlled by α.

As we have seen in the previous chapters, the computation of the p-value $p_{\mathbf{w}}$ is not trivial because it depends on the distribution $L_{M,\mathbf{w}}$. To compute the latter distribution, we can use an exact approach (e.g. Chrysaphinou and Papastavridis (1990), Robin and Daudin (1999)) but it is time-consuming as n becomes large, and difficult to compute when $m \geq 2$. Alternatively, as n tends to infinity, several approximations are valid: when $\mathbf{w}$ has a sufficiently large expected count, we can use a Gaussian approximation (see Prum *et al.* (1995)) and when $\mathbf{w}$ has a bounded expected count, we rather use a Compound Poisson approximation (see Schbath (1995a)). We will suppose in the following that for each word $\mathbf{w}$ the corresponding p-value $p_{\mathbf{w}}$ is known and computable exactly, so-neglecting the possible errors of approximations (and of estimations).

8.1.3 Multiple testing

Define $\mathcal{W}_0$ as the set of words $\mathbf{w}$ such that $H_{\mathbf{w}}$ is true, that is, the set of words $\mathbf{w}$ for which the count follows the null distribution $L_{M,\mathbf{w}}$. Since we have to perform a huge number of tests simultaneously ($|\mathcal{W}| = |\mathcal{A}|^h$), we can not perform the single tests individually at level α, without generating too many false positives (words $\mathbf{w} \in \mathcal{W}_0$ from whose $H_{\mathbf{w}}$ is rejected). Indeed, the expected number of such false positives would be then exactly $|\mathcal{W}_0|\alpha$. Therefore, a multiple testing approach is needed.

A multiple testing procedure is defined by a measurable function R of the set of p-values $\mathbf{p} = (p_{\mathbf{w}}, \mathbf{w} \in \mathcal{W})$ that returns a subset of words $R(\mathbf{p}) \subset \mathcal{W}$, corresponding to the rejected null hypotheses. The quality of such multiple testing procedure can be measured with the k-FWER (“k-family wise error rate”), defined as the probability that the procedure makes at least k wrong rejections (see *e.g.* Lehmann and Romano (2005a)):

$$k\text{-FWER}(R) = \mathbb{P}(|\mathcal{W}_0 \cap R| \geq k).$$

In the case $k = 1$, this quantity reduces to the well known “family wise error rate” (FWER).

While controlling a type I error rate like the k-FWER in a multiple testing problem, the assumptions made on the dependencies between the p-values are a major point (see the second part of this thesis). Here, we are in the most general case where the p-values have unspecified

[1]The under-representation of $\mathbf{w}$ would be measured similarly by $\mathbb{P}_{N \sim L_{M,\mathbf{w}}}(N \leq N(\mathbf{w}))$.

dependencies (potentially non-positively correlated, see the computation of the covariances in Robin *et al.* (2003b))

8.2 Multiple testing procedures that control the k-FWER

8.2.1 The k-Bonferroni procedure

Lehmann and Romano (2005a) proposed to control the k-FWER with an "extended Bonferroni procedure", called here the "k-Bonferroni procedure", that rejects the null hypotheses corresponding to the words $\mathbf{w}$ such that $p_{\mathbf{w}} \leq \alpha k/d$ i.e.

$$R_{B,k} = \{\mathbf{w} \in \mathcal{W} \mid p_{\mathbf{w}} \leq \alpha k/d\},$$

where $d = |\mathcal{A}|^h$ is the cardinal[2] of $\mathcal{W}$. Note that in the special case where $k = 1$, it reduces to the Bonferroni procedure R_B. Following Lehmann and Romano (2005a), the procedure $R_{B,k}$ achieves the right control simply by applying Markov's inequality:

$$\mathbb{P}(|\mathcal{W}_0 \cap R_{B,k}| \geq k) \leq \frac{\mathbb{E}|\mathcal{W}_0 \cap R_{B,k}|}{k} = \sum_{\mathbf{w}\in\mathcal{W}_0} \frac{\mathbb{P}(p_{\mathbf{w}} \leq \alpha k/d)}{k} \leq \frac{|\mathcal{W}_0|}{d}\alpha \leq \alpha.$$

This control holds for any dependency structure between the p-values. Therefore, the k-Bonferroni procedure can be used in our setting. However, since Markov's inequality is a quite conservative device, the control $\mathbb{P}(|\mathcal{W}_0 \cap R_{B,k}| \geq k) \leq \alpha$ can sometimes be conservative, which results in a loss of power of the k-Bonferroni procedure. For general p-values, this loss depends on k and on the dependency structure between the p-values:

- When the p-values are independent, the k-Bonferroni procedure performs well for $k = 1$ but is too conservative for $k \geq 2$ (especially for large values of k, see Table 8.1).
- When the p-values are all equal (and when all the null hypotheses are true), the k-FWER is bounded above by $\alpha k/d$ (with equality if the p-values are uniformly distributed), so that the k-Bonferroni procedure is conservative for small values of k.

Among the two above extreme dependency cases, we will see in Section 8.3 that the word testing case seems to be closer to the independent situation.

In the next paragraph, we present a procedure less conservative than the k-Bonferroni procedure (for all k) and adjusted to the dependency structure between the p-values.

8.2.2 The k-min procedure

The k-min procedure (see *e.g.* Dudoit *et al.* (2004) and Romano and Wolf (2007)) is defined by adjusting the threshold to the $1 - \alpha$ quantile of the distribution of the k-th minimum among all the p-values. The latter distribution is usually estimated with resampling techniques, which provide an asymptotic control of the k-FWER (see Romano and Wolf (2007)). Here, this estimation step is particularly simple, because we can simulate data from the null distribution

[2] Warning : the number of null hypotheses will be denoted by m in the part II of this thesis.

Ratio	$k=1$	$k=2$	$k=3$	$k=10$
$\alpha = 0.01$	0.995	0.0197	4.38×10^{-4}	$< 10^{-16}$
$\alpha = 0.05$	0.975	0.0935	1.00×10^{-2}	3.28×10^{-9}
$\alpha = 0.1$	0.952	0.175	3.59×10^{-2}	1.07×10^{-6}

Table 8.1: Values of the ratio $\mathbb{P}(Y \geq k)/\alpha$ for a random variable Y following a binomial distribution with parameters $(|\mathcal{W}_0|, \alpha k/d)$ and $|\mathcal{W}_0| = d = 1000$. The ratio measure the accuracy of the k-FWER control when using the k-Bonferroni procedure in the case where the p-values are independent.

P_M. We describe now precisely what is the k-min procedure. Let us order the p-values with a given permutation:

$$p_{(1)} \leq p_{(2)} \leq \dots p_{(d)}.$$

Denote by k-min$\{p_{\mathbf{w}}, \mathbf{w} \in \mathcal{W}\}$ the k-th smaller value of the $p_{\mathbf{w}}, \mathbf{w} \in \mathcal{W}$, so that k-min$\{p_{\mathbf{w}}, \mathbf{w} \in \mathcal{W}\} = p_{(k)}$. For any subset S of $\mathcal{W}$ and $\alpha \in (0,1)$, put $q(\alpha, k, S)$ the α-quantile of the distribution of k-min$\{p_{\mathbf{w}}, \mathbf{w} \in S\}$ when the underlying sequence $\mathbf{X}$ follows the null distribution P_M. That is:

$$q(\alpha, k, S) = \inf\left\{x \mid \mathbb{P}_{\mathbf{X}\sim P_M}(k\text{-min}\{p_{\mathbf{w}}, \mathbf{w} \in S\} \leq x) \geq \alpha\right\}. \tag{8.1}$$

Remark that $q(\alpha, k, S)$ is non-increasing in S: for a fixed α, and given subsets S and S' of $\mathcal{W}$,

$$S \subset S' \Rightarrow q(\alpha, k, S') \leq q(\alpha, k, S). \tag{8.2}$$

The following result holds (see *e.g.* Romano and Wolf (2007)).

Theorem 8.1 *Consider the* k-min *procedure*

$$R_{k\text{-min}} = \{\mathbf{w} \in \mathcal{W} \mid p_{\mathbf{w}} \leq q(1-\alpha, k, \mathcal{W})\},$$

where $q(1-\alpha, k, \mathcal{W})$ is given by (8.1). Then we have k-FWER$(R_{k\text{-min}}) \leq \alpha$.

Remark 8.2 *Here, the point is that the quantiles $q(\alpha, k, \mathcal{W})$ can be easily estimated, because it is easy to simulate a Markov chain with given parameters and to get the empirical distribution of $q(\alpha, k, \mathcal{W})$.*

Proof of Theorem 8.1. It is a direct consequence of (8.2):

$$\begin{aligned}\mathbb{P}(|\mathcal{W}_0 \cap R_{k\text{-min}}| \geq k) &= \mathbb{P}\Big(k\text{-min}\{p_{\mathbf{w}}, \mathbf{w} \in \mathcal{W}_0\} \leq q(1-\alpha, k, \mathcal{W})\Big) \\ &\leq \mathbb{P}\Big(k\text{-min}\{p_{\mathbf{w}}, \mathbf{w} \in \mathcal{W}_0\} \leq q(1-\alpha, k, \mathcal{W}_0)\Big) \leq \alpha.\end{aligned}$$

■

8.3 Application to find exceptional words in DNA sequences

In this section, we compare the k-Bonferroni and the k-min procedures to find exceptional words in DNA sequences. Under the null distribution, the sequence $\mathbf{X}$ is supposed to follow a Markov model of order 1 on the DNA alphabet $\mathcal{A} = \{\mathtt{a},\mathtt{c},\mathtt{g},\mathtt{t}\}$ with a transition matrix Π derived from the complete genome of *Haemophilus influenzae*:

$$\Pi = \begin{pmatrix} 0.382 & 0.155 & 0.164 & 0.299 \\ 0.343 & 0.187 & 0.276 & 0.254 \\ 0.270 & 0.264 & 0.197 & 0.269 \\ 0.230 & 0.160 & 0.220 & 0.390 \end{pmatrix}.$$

The corresponding stationary distribution is $\mu = (0.305, 0.184, 0.198, 0.313)$ and the length of the sequence is $n = 1830140$. We have simulated Markov chains with the above parameters and we have computed the p-values of each word of length h using the R'MES[3] software. More precisely, for $h = 3, 4$ we have performed the Gaussian approximation of Prum *et al.* (1995), which is valid when h is "small" (short words), and for $h = 6, 7, 8$ we have performed the compound Poisson approximation of Schbath (1995a), which is valid when h is "large" (long words). We have then computed the quantile $q(1-\alpha, k, \mathcal{W})$ for $\alpha = 0.05$, $k = 1, 10, 100$ and $\mathcal{W}$ being all the words of length h. In order to get more interpretable values, we finally have transformed these probabilities into *thresholds*, using the $\mathcal{N}(0,1)$-quantile transformation: $p \in [0,1] \mapsto$ the $1-p$ quantile of a standard Gaussian distribution. The resulting thresholds are given in Table 8.2. Recall that the best threshold among two thresholds which provide the same k-FWER control is simply the smaller, because it will reject more null hypotheses with the same type I error rate control. Therefore, we see that the k-min procedure is much better than the k-Bonferroni procedure when $k = 10, 100$. However, for $k = 1$ (i.e. for the FWER control), the 1-min procedure gives just a slight improvement with respect to the Bonferroni approach.

8.4 Some conclusions and future works

When we want to find exceptional words among all the words of a given size in a DNA sequences, preliminary experiments show that the k-min procedure is much better than the k-Bonferroni procedure while controlling the k-FWER (at the price of a longer calculation). However, for controlling the FWER, since the Bonferroni procedure is faster than the 1-min procedure and seems to perform almost as well, the Bonferroni procedure can be an interesting alternative in practice.

This chapter gives exciting direction for future works:

- For a given observed DNA sequence, it is reasonable to think that many null hypotheses are false. Hence, to find more exceptional words while controlling the FWER, the step-down procedures of Romano and Wolf (2005) and Romano and Wolf (2007) can be used.
- We could try to test simultaneously a set of degenerated words (that is, words with unspecified letters). In this case, since the structure of dependencies between the p-values should be different, the two approaches proposed here would maybe have different behaviors.

[3]http://genome.jouy.inra.fr/ssb/rmes

k-Bonf	$h = 3$	$h = 4$	$h = 6$	$h = 7$	$h = 8$
$k = 1$	3.163	3.546	4.220	4.523	4.808
$k = 10$	2.418	2.886	3.668	4.009	4.325
$k = 100$	–	2.064	3.031	3.427	3.787

k-min	$h = 3$	$h = 4$	$h = 6$	$h = 7$	$h = 8$
$k = 1$	3.159	3.545	4.17	4.44	4.71
$k = 10$	1.340	2.034	2.97	3.35	3.68
$k = 100$	–	0.365	2.00	2.51	2.93

Table 8.2: Top: thresholds for the k-Bonferroni procedure ($\mathcal{N}(0,1)$-quantile transformation of $k\alpha/4^h$). Bottom: thresholds for the k-min procedure ($\mathcal{N}(0,1)$-quantile transformation of $q(1-\alpha, k, \mathcal{W})$). $\mathcal{W}$ is the set of the words of length h. For $h = 3, 4$ we used the Gaussian approximation to compute the p-values and we performed 10 000 simulations. For $h = 6, 7, 8$ we used the compound Poisson approximation and we performed 1 000 simulations. The global confidence level is $\alpha = 0.05$. We did not compute the case where $h = 3$ and $k = 100$ because it is not relevant ($4^3 < 100$).

Part II

Contributions to theory and methodology of multiple testing

Notations of the part II

$\mathbf{1}\{E\}$, $\lvert E \rvert$	indicator function, cardinal of the set E
$\mathbb{R}^+$	set of the non-negative real numbers
$\mathcal{D}(X)$, $\mathbb{E}(X)$	distribution, expectation of X
Φ	standard Gaussian cumulative distribution function
$\overline{\Phi}$	standard Gaussian upper tail function
h, $\mathcal{H}$	null hypothesis, set of null hypotheses
m (or K in Chapter 12)	number of null hypotheses
$\mathcal{H}_0$, m_0	set, number of true null hypotheses
π_0	proportion of true null hypotheses
$\mathcal{H}_1$, m_1	set, number of false null hypotheses
p_h, $\mathbf{p} = (p_h, \in \mathcal{H})$	p-value, collection of the p-values
$R(\mathbf{p})$	multiple testing procedure (= set of the rejected null hypotheses)
Δ	threshold collection
α, π, β	confidence level, weight function, shape function
ν	prior distribution
$F(\mathbf{p})$, $G(\mathbf{p})$	estimators of π_0^{-1}

Chapter 9

Presentation of part II

This part is a joint work with Gilles Blanchard[1].

9.1 Biological motivations

To put some intuition behind the multiple testing problem, we detail how it is formulated in several biological frameworks — microarray data, neuroimaging, DNA sequences — in which the objects of interest are genes, spatial points or words respectively.

- **Analysis of microarray data (objects = genes)**: a microarray is a collection of several microscopic spots, each one measuring the expression level of a single gene in a certain experimental condition. We look for the genes which have a significantly different expression level in comparison to a control experimental condition. Since the gene expression levels fluctuate naturally (not to speak of other sources of fluctuation introduced by the experimental protocol), it is appropriate to perform a statistical test on each single gene. But the point is that the number of genes m can be large (for instance several thousands), so that non-differentially expressed genes can have a high score of significance by chance, and a non-corrected procedure is likely to select a lot of non-differentially expressed genes (usually called "false positives" or "type I errors"). A multiple testing procedure is a procedure that tests *simultaneously* the expression level of all the genes and that controls in a specific way the type I errors and also the "type II errors" (defined as the non-selected differentially expressed genes). The goal of such a procedure is to select a set of genes as "close" as possible to the set of truly differentially expressed genes. We remark that, since the expression levels of several genes can be related, correlations may exist between the single tests. Moreover, these dependencies are often complex or/and unknown.

  For a specific study of multiple testing problems in microarray experiments we refer the reader for instance to Dudoit *et al.* (2003) and Ge *et al.* (2003).

- **Analysis of neuroimaging (objects = spatial points)**: different neuroimaging techniques are available to measure the brain activity during an experiment (MEG : Magnetoencephalography; fMRI : Functional magnetic resonance imaging). The goal is then to detect the activated areas (spatial points) in a brain map. Again, this generates a large

[1]Fraunhofer FIRST.IDA, Berlin, Germany.

multiplicity problem because we want to make a decision for a large number m of spatial points simultaneously. In this setting, we note that the single tests are moreover spatially correlated, with possibly unknown correlations.

A study of multiple testing procedures in this neuroimaging setting was made for instance by Perone Pacifico *et al.* (2004). For more applied studies, we refer the reader for instance to Pantazis *et al.* (2005), Darvas *et al.* (2005) and Jerbi *et al.* (2007).

- **Finding over-represented words in DNA sequences (objects = words)**: the data are given by a DNA sequence, in which we want to detect words (i.e. short succession of letters in {`a`,`c`,`g`,`t`}) which have a particular biological function. The significance of each word can be computed by counting the number of occurrences of the word in the observed sequence and by comparing it to the corresponding count in a random sequence (this step is not trivial, and is examined in the first part of this thesis). Since we test a huge number of words simultaneously (for instance $m = 16384$ for all the words of length 7), a multiple testing procedure is needed to infer a decision. Note that the scores of significance are in this case also correlated.

  One solution to this specific multiple testing problem is given in Chapter 8.

9.2 Framework: from single testing to multiple testing

9.2.1 Single testing framework

We recall here the setting of single testing theory (see *e.g.* Lehmann and Romano (2005b)). Suppose that the observed data are generated from a probability space $(\mathcal{X}, \mathfrak{X}, P)$, where P is an unknown underlying probability distribution belonging to a subset $\mathcal{M}$ (model) of probability distributions on $(\mathcal{X}, \mathfrak{X})$. We are interested in determining whether the distribution P satisfies or not certain "properties" called *null hypotheses*.

Formally, a *null hypothesis* h is a subset of $\mathcal{M}$. We say "P satisfies the null hypothesis h" or "h is true" whenever $P \in h$. An *alternative* of h is given by any subset of $\mathcal{M} \backslash h$, where $\mathcal{M} \backslash h$ denotes the complementary of h in $\mathcal{M}$. For simplicity, we will consider in what follows that the alternative of h is the whole set $\mathcal{M} \backslash h$ (otherwise we can reduce $\mathcal{M}$).

Example 9.1 (Gaussian single null hypothesis (one-sided)) *Consider $\mathcal{X} = \mathbb{R}$ and $\mathcal{M}$ the set of all the Gaussian measures on $\mathbb{R}$ with a fixed variance σ_0^2. For a given mean $\mu_0 \in \mathbb{R}$, the set of Gaussian measures with mean smaller than μ_0 and variance σ_0^2 defines a null hypothesis h. This is denoted classically*

$$h : \text{“ } P \text{ is Gaussian with mean } \mu \text{ and variance } \sigma_0^2 \text{, with } \mu \leq \mu_0 \text{”.}$$

The complementary of h in $\mathcal{M}$ is then " P is Gaussian with mean μ and variance σ_0^2, with $\mu > \mu_0$".
For instance, in a problem where we observe the expression level of a single gene, the null hypothesis h can mean "the gene's expression level is not significantly larger than the control level" (assuming that the data are Gaussian with known variance).

Let X be a random variable taking values in $(\mathcal{X}, \mathfrak{X})$ such that $X \sim P$. For a given hypothesis h, the goal of hypothesis testing is to take a decision about whether P satisfies h, based on a realization x of X. We subsume under this framework the case where we have, for instance, n repeated i.i.d. observations of the same random variable, in which case X represents the whole sample and the null hypotheses are implicitly of the form "P is a product distribution of the form $Q^{\otimes n}$, and Q satisfies a certain property".

Given a null hypothesis h, the decision is made with a *(single) test*, defined as a measurable function $T : (\mathcal{X}, \mathfrak{X}) \to \{0,1\}$, where "$T = 1$" codes for "$h$ is rejected" and "$T = 0$" codes for "h is not rejected". Such a decision can make two kinds of errors: a *type I error* arises when T rejects h although h is true and a *type II error* arises whenever T does not reject h although h is false. Following Neyman-Pearson approach, these two error probabilities are not equivalent. A test T should first control its probability of type I error:

$$\forall P \in h, \quad \mathbb{P}(T(X) = 1),$$

by a given confidence level α, and then, provided that the previous control holds, T should minimize the probability of type II error (which is $\forall P \in \mathcal{M}\backslash h$, $\mathbb{P}(T(X) = 0)$).

Given a confidence level α, a test T for h is often of the form $\mathbf{1}\{p \leq \alpha\}$, where p is a *p-value* function for h, that is, a measurable function $p : (\mathcal{X}, \mathfrak{X}) \to [0,1]$, such that the distribution of $p(X)$ is stochastically lower bounded by a uniform random variable whenever h is true:

$$\forall P \in h, \quad \forall t \in [0,1], \quad \mathbb{P}(p(X) \leq t) \leq t.$$

Therefore, the test $T = \mathbf{1}\{p \leq \alpha\}$ that rejects h whenever the p-value p is less than or equal to α, has a probability of type I error smaller than α.

Example 9.2 (Gaussian single null hypothesis (one-sided) — continued) *We consider the null hypothesis h of Example 9.1 and we denote by $\overline{\Phi}$ the standard Gaussian upper distribution tail function. Then the function*

$$p : x \in \mathbb{R} \mapsto \overline{\Phi}((x - \mu_0)/\sigma_0),$$

defines a p-value for h, and the test which rejects h whenever $\overline{\Phi}((x-\mu_0)/\sigma_0) \leq \alpha$ has a probability of type I error less than or equal to α (note that this probability can be strictly smaller than α when the mean μ of P is strictly smaller than μ_0).

We will focus in the sequel on test T of the form $\mathbf{1}\{p \leq \alpha\}$, so that we will always assume the existence of a p-value. The following lemma shows that this is not a major restriction.

Lemma 9.3 *Every family $T = (T_\alpha)_{\alpha \in [0,1]}$ which satisfies*

(i) $\forall \alpha \in [0,1]$, T_α is a test for h with a probability of type I error less than or equal to α,

(ii) $\alpha \mapsto T_\alpha$ is right-continuous and non-decreasing (pointwise),

is of the form $T_\alpha = \mathbf{1}\{p \leq \alpha\}$, where $p = \inf\{\alpha \in [0,1] \mid T_\alpha = 1\}$ is a p-value function for h.

Proof. Point (ii) implies pointwise $T_\alpha = \mathbf{1}\{p \leq \alpha\}$. Therefore, it is sufficient to prove that p is a p-value function for h; for all $t \in [0,1]$, we have $\{x \in \mathcal{X} \mid p(x) \leq t\} = \{x \in \mathcal{X} \mid T_t(x) = 1\} \in \mathfrak{X}$, which implies that p is measurable. Moreover, from (i) we get: $\forall t \in [0,1]$, $\mathbb{P}[p(X) \leq t] = \mathbb{P}[T_t(X) = 1] \leq t$. ■

9.2.2 Multiple testing framework

While the single testing framework deals with one null hypothesis for the distribution P at a time, multiple testing is concerned with a whole set of such hypotheses. We consider a set of null hypotheses, denoted by $\mathcal{H}$, which is supposed to be finite of cardinal m.

Example 9.4 (Gaussian multiple null hypotheses (one-sided)) *Let $\mathcal{X} = \mathbb{R}^m$ and denote by X_i the projection of X on the i-th coordinate in $\mathbb{R}^m$. Consider the model $\mathcal{M}$ of Gaussian measures on $\mathbb{R}^m$ where each X_i has variance $\sigma_{0,i}^2$. Given a set of means $\mu_{0,i}, i \in \{1,\dots,m\}$, a classical set of null hypotheses is given by*

$$\mathcal{H} = \left\{ \text{“}P \textit{ is Gaussian, } X_i \textit{ has mean } \mu_i \textit{ and variance } \sigma_{0,i}^2, \textit{ with } \mu_i \leq \mu_{0,i}\text{”}, \; i \in \{1,\dots,m\} \right\}.$$

For instance, in a problem where we observe simultaneously the expression levels of m genes, this set of null hypotheses allows to test simultaneously for each i “the i-th gene's expression level is not significantly larger than the control level” against “the i-th gene's expression level is significantly larger than the control level” (assuming that the data are Gaussian with known variances).

The underlying distribution P being fixed in the model $\mathcal{M}$, we denote by

$$\mathcal{H}_0 := \{h \in \mathcal{H} \mid P \text{ satisfies } h\}$$

the set of true null hypotheses and we put $m_0 := |\mathcal{H}_0|$ the number of true null hypotheses. We also denote by $\mathcal{H}_1 := \mathcal{H} \backslash \mathcal{H}_0$ the set of false null hypotheses and we put $m_1 := m - m_0$ the number of false null hypotheses. A quantity of interest which will appear later is $\pi_0 := m_0/m$, which is the proportion of true null hypotheses. Since P is unknown, the quantities $\mathcal{H}_0$, m_0 ($\mathcal{H}_1$, m_1) and π_0 are of course unknown.

A decision in this multiple testing context is a procedure that returns a subset of rejected[2] null hypotheses. A *multiple testing procedure* is defined as a function

$$R : x \in \mathcal{X} \mapsto R(x) \subset \mathcal{H},$$

such that for any $h \in \mathcal{H}$, the function $x \in \mathcal{X} \mapsto \mathbf{1}\{h \in R(x)\}$ is measurable, and where $R(x)$ corresponds to the set of the rejected null hypotheses for the procedure R given the realization x of X. Such a multiple decision is generally built from the individual decision of each single null hypothesis, and more precisely, from the individual p-value of each single null hypothesis. Therefore, we will consider in this work that for each null hypothesis $h \in \mathcal{H}$, there exists a p-value p_h for h, i.e. a measurable function $p_h : (\mathcal{X}, \mathfrak{X}) \to [0,1]$ such that: if $h \in \mathcal{H}_0$,

$$\forall t \in [0,1], \; \mathbb{P}(p_h(X) \leq t) \leq t.$$

For clarity reasons, we will now drop the explicit dependence in X in our notations.

All the multiple testing procedures R that we will consider are supposed to be measurable functions of *the set of p-values* $\mathbf{p} = (p_h, h \in \mathcal{H})$ *i.e.* are of the form $R = \widetilde{R}(\mathbf{p})$, where $\widetilde{R}$ is a measurable function from $[0,1]^{\mathcal{H}}$ to the set of the subsets of $\mathcal{H}$. We will always identify R and

[2]Remember that the procedure selects the objects which correspond to the “rejected” null hypotheses.

$\widetilde{R}$ in our notations. Therefore, in this work, any multiple testing procedure R can be written as $R(\mathbf{p})$, where $\mathbf{p} = (p_h, h \in \mathcal{H})$ is the set of p-values.

Below, we give simple examples of multiple testing procedures (more sophisticated examples will be given in Section 9.4). To make a choice among all the possible multiple testing procedures, we must define precisely a criterion of quality, which is discussed in the following section.

Example 9.5 (Simple examples of multiple testing procedures)

1. *If we just perform for each null hypothesis h the corresponding single test at level α, we can choose to reject all the p-values less than or equal to α; this gives the non-corrected multiple testing procedure, defined in our setting by $R = \{h \mid p_h \leq \alpha\}$.*

2. *If we "correct" the individual levels by α/m, this defines the Bonferroni (or Bonferroni-corrected) multiple testing procedure: $R_B = \{h \mid p_h \leq \alpha/m\}$.*

3. *More generally, the multiple testing procedure with threshold t rejects all the p-values smaller than t:*
$$R = \{h \mid p_h \leq t\}.$$
Note that the above definition still makes sense if t depends on the set of p-values.

Remark 9.6 (Our choice for $\mathcal{M}$) *The choice of the model $\mathcal{M}$ depends on the assumptions made on the underlying distribution P. In this work, we will make the following choices:*

1. *In Chapters 10 and 11, the distribution assumptions will only concern the dependency structure between the p-values (and also of course the existence of the p-values). In particular, we will make no assumption on the (marginal) distribution of p_h when $h \in \mathcal{H}_1$.*

2. *In Chapter 12, we will make more specific assumptions on the distribution model: $\mathcal{M}$ will be taken equal to a set of Gaussian measures, or to a set of bounded symmetric distributions.*

Remark 9.7 (Other existing multiple testing frameworks)

1. *In our framework, P is fixed and $\mathcal{H}_0$ is not random. Another classical framework is to use a random effects model, in which each null hypothesis can be true or false with a certain probability. In this model, the resulting p-values are generated from a mixture model (see for instance Efron* et al. *(2001), Storey (2003) and Genovese and Wasserman (2004)).*

2. *Another multiple testing framework close in spirit to model selection is considered by Baraud* et al. *(2003, 2005). They consider only one null hypothesis which is tested against several alternative hypotheses.*

9.3 Quality of a multiple testing procedure R

A multiple testing procedure R can make two kinds of errors for a given null hypothesis h:

- A *type I error* arises for h whenever R rejects h although h is true, that is, $h \in \mathcal{H}_0 \cap R$.
- A *type II error* arises for h whenever R does not reject h although h is false, that is, $h \in \mathcal{H}_1 \backslash R$.

Following Neyman-Pearson approach, the first concern is to build a multiple testing procedure which makes "not too many" type I errors. To quantify this precisely, several type I error rates can be proposed, each of them measuring the type I errors in a specific way.

9.3.1 Type I error rates

Here are the most standard type I error rates:

- The *Per-comparison error rate* (PCER), defined as the average number of type I errors divided by m:
$$\mathrm{PCER}(R) := \mathbb{E}|\mathcal{H}_0 \cap R|/m.$$

- The *Per-family error rate* (PFER), defined as the average number of type I errors :
$$\mathrm{PFER}(R) := \mathbb{E}|\mathcal{H}_0 \cap R|.$$

- The *family-wise error rate* (FWER), defined as the probability that at least one type I error occurs:
$$\mathrm{FWER}(R) := \mathbb{P}(|\mathcal{H}_0 \cap R| > 0).$$

- The *false discovery rate* (FDR) (see Benjamini and Hochberg (1995)), defined as the average proportion of type I errors among the rejected null hypotheses:
$$\mathrm{FDR}(R) := \mathbb{E}\bigg(\frac{|\mathcal{H}_0 \cap R|}{|R|}\mathbf{1}\{|R| > 0\}\bigg).$$
Note that, in the above expectation, the indicator means that the ratio is equal to 0 when $|R| = 0$.

More recently, the following generalizations of the FWER and FDR have been proposed:

- The *k-family-wise error rate* (k-FWER) , defined as the probability that at least k type I error occur:
$$k\text{-FWER}(R) := \mathbb{P}(|\mathcal{H}_0 \cap R| \geq k).$$

- The *k-false discovery rate* (k-FDR) (see Sarkar and Guo (2006)), defined as the average proportion of k or more type I errors among the rejected null hypotheses:
$$k\text{-FDR}(R) := \mathbb{E}\bigg(\frac{|\mathcal{H}_0 \cap R|}{|R|}\mathbf{1}\{|\mathcal{H}_0 \cap R| \geq k\}\bigg).$$

Finally, the *false discovery proportion* (FDP) is defined by $\mathrm{FDP}(R) := \frac{|\mathcal{H}_0 \cap R|}{|R|}\mathbf{1}\{|R| > 0\}$ and a standard associated type I error rate is $\mathbb{P}(\mathrm{FDP} > \gamma)$, for a given parameter $\gamma \in [0, 1)$ (see *e.g.* Lehmann and Romano (2005a)).

Remark 9.8 *1. The following relations hold: $PCER(R) \leq FDR(R) \leq FWER(R) \leq PFER(R)$. Similarly k-$FDR(R) \leq k$-$FWER(R)$ for $k \geq 1$.*

2. When all of the null hypotheses are true, i.e. $\mathcal{H} = \mathcal{H}_0$, we have $FDR(R) = FWER(R)$.

3. *The control of* $\mathbb{P}(FDP(R) > \gamma)$ *at level* $1/2$ *implies that the median of the* $FDP(R)$ *is smaller than* γ*.*

Remark 9.9 *Throughout this work we will use the following convention: whenever there is an indicator function inside an expectation, this has logical priority over any other factor appearing in the expectation. What we mean is that if other factors include expressions that may not be defined (such as the ratio* $\frac{0}{0}$*) outside of the set defined by the indicator, this is safely ignored. In other terms, any indicator function implicitly entails that we perform integration over the corresponding set only. This results in more compact notations, such as in the above definitions.*

As we can see, several choices are possible for measuring the type I errors of a multiple testing procedure. Of course, this choice depends on what the user wants to control in practice. In actual applications, the most popular error rates are those which are related to the FDP, because they are more permissive and therefore often allow to reject larger number of null hypotheses. When the user prefers a stricter criterion, the FWER (or alternatively k-FWER) can be used. In this thesis, we will mainly focus on the FDR and the FWER (the k-FWER is also considered in Chapter 8).

9.3.2 Controlling a type I error rate

Choose E_1 equal to one of the previous type I error rates. Given a confidence level $\alpha \in (0,1)$, we want to build a multiple testing procedure R which controls the error rate E_1 at level α, i.e. such that

$$E_1(R) \leq \alpha. \tag{9.1}$$

We emphasize that the latter control has to hold here for all the possible set $\mathcal{H}_0$ (and not only for $\mathcal{H}_0 = \mathcal{H}$). This is commonly called a *strong control*. Moreover, we will focus in this thesis on *non-asymptotic* controls, that is, controls that hold for any fixed value m (when $m \to \infty$, asymptotic controls have been proposed for instance by Genovese and Wasserman (2002, 2004), Storey *et al.* (2004) and Farcomeni (2007)).

In our setting, the level α is fixed and we look for a procedure R satisfying (9.1). A reverse approach consists in fixing the procedure R (for instance a procedure based on a fixed threshold t) and in estimating the corresponding type I error rate $E_1(R)$. This reverse approach has been first proposed by Storey (2002) with the FDR, and has been widely used since (see e.g. Robin *et al.* (2007) and van de Wiel and In Kim (2007)). A link between the approach "type I error rate control" and the reverse approach "type I error rate estimation" is pointed out by Storey *et al.* (2004).

Going back to the control of $E_1(R)$, one trivial fact is that the procedure $R = \emptyset$ (which rejects no null hypothesis) satisfies trivially $E_1(R) = 0 \leq \alpha$. The point is that such a procedure is not interesting because it will never reject any false null hypotheses. Therefore, we have to add some constraints to the simple control (9.1), using a type II error rate.

9.3.3 Type II error rates while controlling a type I error rate

Similarly to type I error rates, there are also many type II error rates that can be defined. Typically, some can be obtained simply by replacing $\mathcal{H}_0$ by $\mathcal{H}_1$ and R by $\mathcal{H}\backslash R$ in all the above type I error rates (for instance, the FDR becomes the FNR as introduced by Genovese and Wasserman (2002)). A common choice of type II error rate is $E_2 = \mathbb{E}|\mathcal{H}_1 \backslash R| = m_1 - \mathbb{E}|\mathcal{H}_1 \cap R|$, where the quantity $\mathbb{E}|\mathcal{H}_1 \cap R|$ is usually called the *power* of R.

Such a type II error rate E_2 being fixed, we want to find R such that $E_2(R)$ is minimum provided that the control (9.1) holds. Of course, finding such an "optimal" procedure is a difficult task (the interested reader can find some elements in Storey (2005), Lehmann *et al.* (2005) and Wasserman and Roeder (2006)). On the other hand, it is relatively easy to compare two procedures that control the type I error rate at the same level: for two procedures R and R' such that $E_1(R) \leq \alpha$ and $E_1(R') \leq \alpha$, we prefer R to R' when $E_2(R') \geq E_2(R)$. Using this criterion with the type II error rate $E_2 = \mathbb{E}|\mathcal{H}_1 \backslash R|$, we obtain the following criterion.

Definition 9.10 *Given two multiple testing procedures R and R' such that $E_1(R) \leq \alpha$ and $E_1(R') \leq \alpha$, R is said more powerful than R' whenever R as a larger expected number of rejected false null hypotheses, that is $\mathbb{E}|\mathcal{H}_1 \cap R| \geq \mathbb{E}|\mathcal{H}_1 \cap R'|$.*

We emphasize that the comparison in terms of power can only be made if both procedures control the type I error rate at level α (under the same assumptions). Since the power of a given procedure is often hard to compute exactly (it is usually estimated with simulations), one interesting remark is that R is more powerful than R' as soon as $R' \subset R$ pointwise, that is, when the rejected null hypotheses of R' are always contained in those of R. The latter comparison criterion is quite restrictive but practical and common.

Definition 9.11 *Given two multiple testing procedures R and R' such that $E_1(R) \leq \alpha$ and $E_1(R') \leq \alpha$, R is said less conservative than R' , if for all set $\mathbf{p}$ of p-values we have $R'(\mathbf{p}) \subset R(\mathbf{p})$.*

Therefore, if R is less conservative than R', R is also more powerful than R'. However, given two procedures R and R', it can be the case that neither is less conservative than the other.

Remark 9.12 *In order to get a powerful procedure R, the rate $E_1(R)$ should be close to α. However, this condition is not sufficient: consider the case where R is the procedure that rejects all the null hypotheses with probability α and that rejects no null hypotheses otherwise; such a procedure satisfies $FDR(R) = FWER(R) = \alpha$, but rejects (on average) only α percent of the false null hypotheses.*

We present now a popular type of multiple testing procedures which are known to control some of the proposed type I error rates: the step-down and step-up multiple testing procedures.

9.4 Step-down and step-up multiple testing procedures

9.4.1 Definition

These procedures are defined by comparing the ordered p-values: $p_{(1)} \leq \cdots \leq p_{(m)}$ to a (non-negative) *threshold collection* $\Delta(i), i \in \{1, \dots, m\}$.

A *step-down* procedure starts by comparing the most significant p-values. The procedure is defined with the following iterative algorithm:

- Step 1: if $p_{(1)} > \Delta(1)$ stop and reject no null hypothesis, otherwise go to step 2.

- Step i ($i \geq 2$): if $p_{(i)} > \Delta(i)$ stop and reject the $i-1$ null hypotheses corresponding to the first $i-1$ ordered p-values, otherwise go to step $i+1$ (if $i = n$ stop and reject all the null hypotheses).

Using our notations, the step-down procedure with the threshold collection Δ is thus defined as

$$R = \{h \in \mathcal{H} \mid p_h \leq p_{(k)}\}, \text{ where } k = \max\big\{i \in \{0, \cdots, m\} \mid \forall j \leq i, p_{(j)} \leq \Delta(j)\big\},$$

where we have put $p_{(0)} := 0$ (so that $R = \emptyset$ whenever $k = 0$).

A *step-up* procedure starts by comparing the least significant p-values. The procedure is defined with the following iterative algorithm:

- Step 1: if $p_{(m)} \leq \Delta(m)$ stop and reject all the null hypotheses, otherwise go to step 2.
- Step i ($i \geq 2$): if $p_{(m-i+1)} \leq \Delta(m-i+1)$ stop and reject the $m-i+1$ null hypotheses corresponding to the first $m-i+1$ ordered p-values, otherwise go to step $i+1$ (if $i = n$ stop and reject no null hypothesis).

Using our notations, the step-up procedure with the threshold collection Δ is thus defined as

$$R' = \{h \in \mathcal{H} \mid p_h \leq p_{(k')}\}, \text{ where } k' = \max\big\{i \in \{0, \cdots, m\} \mid p_{(i)} \leq \Delta(i)\big\}.$$

Remark 9.13 1. *The rejection set of a step-down (resp. step-up) procedure is a non-decreasing function of the threshold collection: if two fixed threshold collections Δ and Δ' satisfy $\forall i, \Delta(i) \geq \Delta'(i)$, the step-down (resp. step-up) procedure based on Δ is always less conservative than the one based on Δ'.*

2. *For a fixed threshold collection Δ, the corresponding step-up procedure R always rejects more null hypotheses than the corresponding step-down R' (simply because $k \leq k'$). This implies that for the same control of a type I error rate, R' is always more conservative than R. However, controlling the type I error rate for the step-up procedures often requires stricter assumptions on the distribution of p-values. Therefore, both step-up and step-down procedures have their own interest.*

3. *In this work, we will always consider non-decreasing threshold collections Δ. Note that some authors have considered non-monotonous threshold collections (e.g. Finner and Roters (1998)).*

For a step-up or step-down method, we want to find a threshold collection which allows to control a given type I error rate while being as "large" as possible.

9.4.2 Example: constant threshold collection

As a first example, we propose to detail the case where the p-values are just compared to a constant threshold (this is a limiting case of both step-up and step-down procedure, where the threshold collection is constant). The probably most well-known multiple testing procedure of this kind is the *Bonferroni procedure* R_B that rejects all the p-values smaller than $\Delta = \alpha/m$. This procedure controls at level α the PFER and thus the FWER too:

$$\mathrm{FWER}(R_B) \leq \mathrm{PFER}(R_B) = \sum_{h \in \mathcal{H}_0} \mathbb{P}(p_h \leq \alpha/m) \leq \alpha m_0/m \leq \alpha.$$

This control holds under no particular assumptions on the dependencies between the p-values. Therefore, this control is called "distribution-free" (d.f. in short) or more explicitly "with unspecified dependencies".

If we suppose that the p-values are independent, the *Sidak procedure* R_S rejecting all the p-values smaller than the (constant) threshold collection $\Delta = 1 - (1-\alpha)^{1/m}$ $(\geq \alpha/m)$ controls the FWER at level α:

$$\begin{aligned}\mathrm{FWER}(R_S) &= \mathbb{P}(\exists h \in \mathcal{H}_0 \mid p_h \leq 1-(1-\alpha)^{1/m}) = 1 - \mathbb{P}(\forall h \in \mathcal{H}_0, p_h > 1-(1-\alpha)^{1/m}) \\ &= 1 - \prod_{h\in\mathcal{H}_0} \mathbb{P}(p_h > 1-(1-\alpha)^{1/m}) \leq 1-(1-\alpha)^{m_0/m} \leq \alpha.\end{aligned}$$

Moreover, this control still holds if the p-values are not longer supposed independent, but satisfy instead the "positive quadrant dependence condition": for all $c > 0$, $\mathbb{P}(\forall h \in \mathcal{H}_0, p_h > c) \geq \prod_{h\in\mathcal{H}_0} \mathbb{P}(p_h > c)$. The latter condition is satisfied when the p-values satisfy a certain type of positive dependencies (e.g. Karlin and Rinott (1980) proved that the classical MTP_2 condition implies the positive quadrant dependence condition).

The above example illustrates a common situation in multiple testing:

1. A conservative procedure satisfies a type I error control under unspecified dependencies between the p-values.

2. Under independence, the latter procedure can be improved while the type I error control still holds.

3. The latter improvement is still valid under a kind of positive dependencies between the p-values.

9.4.3 Some classical choices for Δ with type I error rate control

We give in Table 9.1 and Table 9.2 some of the most classical step-up and step-down procedures with the associated type I error rate control. We do not give here an exhaustive review of all the existing step-up and step-down procedures (more procedures will be considered in the different following chapters). For instance, the first line of Table 9.1 means that Holm (1979) proved that the step-down procedure with threshold collection $\Delta(i) = \alpha/(m-i+1)$ has a FWER controlled by α under unspecified dependencies between the p-values. The assumptions "indep" means that the p-values are independent. The assumptions "pos-dep" means that the p-values are positively dependent. The exact notion of positive dependency can be different in each case since there exists several such notions. Precisely, in Table 9.1 it refers to MTP_2 condition (see *e.g.* Karlin and Rinott (1980) or Sarkar (2002)), whereas in Table 9.2 it refers to PRDS condition (see *e.g.* Benjamini and Yekutieli (2001)).

As mentioned by Finner and Roters (1998), in order to control the FWER under independence , both step-up or step-down methods can be used. Since $1-(1-\alpha)^{1/(m-i+1)} \geq \alpha/(m-i+1)$ and since $1-(1-\alpha)^{1/(m-i+1)}$ corresponds to the step-down procedure whereas $\alpha/(m-i+1)$ corresponds to the step-up one, no procedure always outperforms the other. However, it may be argued that step-up is better when both threshold collections are sufficiently "uniformly close".

Type I error rate (control at level α)	Choices for the threshold collection $\Delta(i)$ in a step-down procedure	Assumptions on p-values
FWER	$\frac{\alpha}{m-i+1}$	d.f. (Holm79)
	$1-(1-\alpha)^{1/(m-i+1)}$	indep (Holm79), pos-dep
FDR	$1-\left[1-\min\left(1,\frac{\alpha m}{m-i+1}\right)\right]^{1/(m-i+1)}$	indep (BL99), pos-dep (Sar02)

Table 9.1: Classical step-down procedures with type I error rate control: (Holm79) corresponds to Holm (1979), (BL99) corresponds to Benjamini and Liu (1999b), (Sar02) corresponds to Sarkar (2002).

Type I error rate (control at level α)	Choices for the threshold collection $\Delta(i)$ in a step-up procedure	Assumptions on p-values
FWER	$\frac{\alpha}{m-i+1}$	indep (Hoch88)
FDR	$\frac{\alpha i}{m(1+1/2+\cdots+1/m)}$	d.f. (BY01)
	$\frac{\alpha i}{m}$	indep (BH95), pos-dep (BY01)

Table 9.2: Classical step-up procedures with type I error rate control: (Hoch88) corresponds to Hochberg (1988), (BY01) corresponds to Benjamini and Yekutieli (2001), (BH95) corresponds to Benjamini and Hochberg (1995).

In order to control the FDR under independence, the step-up method of Benjamini and Hochberg (1995) and the step-down method of Benjamini and Liu (1999b) can be proposed. However, as discussed by Benjamini and Liu (1999b), the procedure of Benjamini and Hochberg (1995) always seems to outperform the one of Benjamini and Liu (1999b) except in very particular cases (when there is a large proportion of false null hypotheses and a small number of null hypotheses).

9.4.4 Resampling-based multiple testing procedures

Until now, when the p-values have unspecified dependencies, we have only considered procedures that control a type I error rate without taking into account the potential dependencies between the p-values. Therefore, the threshold collection is intrinsically adjusted to the "worst case of dependencies", and not to the specific case of dependencies contained in the data. For instance, consider the Bonferroni procedure with the FWER: it is adjusted to the "worst case" of dependencies (corresponding to negative dependencies between the p-values[3]). If we perform this method on data where all the p-values are equal (the opposite extreme case), this procedure thresholds at level α/m while a threshold α will be sufficient, and this results in a huge loss of power.

Therefore, in order to improve the power of a multiple testing procedure when the p-values may have some dependencies, we have to take into account these specific dependencies in the procedure. Moreover, as shown in Section 9.1, the dependencies are often unknown, so that the procedure has to be "adaptive" to this unknown dependency structure. This can be done using resampling-based methods.

Such methods have been proposed for instance by Westfall and Young (1993), Yekutieli and Benjamini (1999), Pollard and van der Laan (2003), Ge *et al.* (2003) and Romano and Wolf (2005, 2007). Given a whole i.i.d. n-sample of the data, the principle of resampling (see *e.g.* Efron (1979) and Arlot (2007)) is to build new (re)samples from the original sample, simply obtained by drawing randomly some data points of the original sample. The rationale is that the resampled data should mimic the variations that we would observed if we had new independent samples. While using these resampled data, we are often able to estimate more precisely the ideal threshold needed to achieve a given type I control. For now, there are to our knowledge two ways to prove that resampling-based procedures really provide a correct control:

1. Asymptotic approaches (when the sample size n tends to infinity), based on the fact that the bootstrap process is asymptotically close to the original empirical process (see for instance van der Vaart and Wellner (1996)). However, these approaches typically assume that the number of null hypotheses m is fixed while n goes to infinity. Whether this type of result still holds when the dimension m grows with n with $m(n) \gg n$ is up to our knowledge an area where only very little is known.

2. Exact approaches (including permutation methods) with n fixed, based on an invariance of the null distribution of the sample under a given transformation (see *e.g.* Romano and Wolf (2005)).

9.5 Presentation of our results

The goals of our work are:

1. To propose a new, synthetic point of view on existing type I error rate controls, providing concise proofs in the three cases of dependencies between the p-values: unspecified dependencies, positive dependencies and independence.

[3]When $m = m_0$, the bound on FWER is achieved for the Bonferroni procedure with the following choice for the joint distribution of the p-values: take $K \sim (1-\alpha)\delta_0 + \alpha\delta_1$. Then, if $K = 0$ take (independently) all the p-values uniformly in $(\alpha/m, 1]$; if $K = 1$ choose (all independently) h uniformly in $\mathcal{H}$ and p_h uniformy in $[0, \alpha/m]$ whereas for $h' \neq h$, $p_{h'}$ is taken uniformly in $(\alpha/m, 1]$.

2. To extend some existing procedures to more general threshold collection or to weaker assumptions.

3. To propose new multiple testing procedures that improve or are competitive with existing ones.

Our results are presented in three chapters 10,11 and 12. These chapters are largely independent.

9.5.1 Chapter 10: "A set-output point of view on FDR control in multiple testing"

We propose in this work a "set-output point of view" on multiple testing and FDR control. We introduce a type of "self-consistency condition" on the set rejected by the procedure. Different versions of this condition imply the control of the FDR respectively under independence, positive dependencies (PRDS) or unspecified dependencies between the p-values. We prove then that the step-up procedures satisfy a "self-consistency condition", implying that we recover in particular the results of Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001) through synthetic and simple proofs.

This work is in part based on the work of Blanchard and Fleuret (2007), which proved in a learning theory context, that a whole family of step-up procedures can control the FDR under unspecified dependencies between the p-values. Namely, the FDR is smaller than $\alpha m_0/m \leq \alpha$ for all the step-up procedures with threshold collection $\Delta(i) = \alpha\beta(i)/m$, where β is of the form

$$\beta(i) = \sum_{1\leq k\leq i} k\nu(\{k\}) \tag{9.2}$$

and ν is some distribution on $\{1,\ldots,m\}$. We called here β the *(threshold) shape function.* The distribution ν represents a prior distribution on the final number of rejections of the procedures, and taking $\nu(\{k\}) = k^{-1}(1 + 1/2 + \cdots + 1/m)^{-1}$ recovers the distribution-free procedure of Benjamini and Yekutieli (2001) (see Table 9.2). A form of "self-consistency condition" has been implicitly introduced by Blanchard and Fleuret (2007). Our contribution with respect to Blanchard and Fleuret (2007) is to extend the "self-consistency condition" to independent or PRDS p-values. To be as exhaustive as possible, we will also detail the above distribution-free procedure.

This set-output point of view allows also to integrate naturally the two following generalizations:

- The multiple testing procedures considered can be "weighted" i.e they can use a *weight function* $\pi(h)$ in the threshold collection.

- This approach allows to build procedures that control the "modified FDR" $\mathbb{E}\big[\frac{|\mathcal{H}_0\cap R|}{|R|}\mathbf{1}\{|R|>0\}\big]$ in which $|\cdot|$ denotes a general finite volume measure on $\mathcal{H}$.

This set-output point of view will be useful to prove FDR controls in Chapter 11.

9.5.2 Chapter 11: "New adaptive step-up procedures that control the FDR under independence and dependence"

A consequence of Chapter 10 is that the step-up procedure of threshold collection $\alpha\beta(i)/m$ has a FDR smaller than $\alpha m_0/m$ in either of the following situations:

- with the shape function $\beta(i) = i$ when the p-values are independent or have positive dependencies (PRDS) (recovering results of Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001)),
- with a shape function β of the form (9.2) when the dependencies between the p-values are unspecified (recovering the result of Blanchard and Fleuret (2007)).

Denoting the proportion of true null hypotheses by $\pi_0 = m_0/m$, the above controls hold at level $\pi_0\alpha$, which is smaller than α. Consequently, the above procedures are — inevitably — conservative when π_0 is small. To solve this problem, some authors (e.g. Benjamini and Hochberg (2000), Storey (2002) and Black (2004)) proposed to adjust the previous threshold collections with an estimator of π_0^{-1}, resulting in so-called (π_0-)*adaptive step-up procedures*. In a recent work, Benjamini *et al.* (2006) proposed such an adaptive procedure that rigorously controls the FDR at level α under independence. It is called "two-stage" because it is based on a first step of estimation of π_0^{-1}. We propose in this work the two following new step-up procedures:

- The "one-stage" procedure R_0' with threshold collection $\Delta_0'(i) = \frac{\alpha}{1+\alpha}\min\left(\frac{i}{m-i+1}, 1\right)$.
- The "two-stage" procedure R_1': first perform the above procedure R_0', and then perform the step-up procedure with threshold collection $\Delta(i) = \frac{\alpha}{1+\alpha}\frac{i}{m-|R_0'|+1}$.

The first procedure is said "adaptive one-stage" because it contains implicitly an estimation step of π_0^{-1}. We prove that both procedures control rigorously the FDR at level α when the p-values are independent. Moreover, up to some marginal cases, R_0' is always less conservative than the one of Benjamini and Hochberg (1995) and the new two-stage procedure is always less conservative than the one of Benjamini *et al.* (2006). We also perform a simulation study under independence and positive dependencies, where we also compare our new adaptive procedures with the one of Storey (2002).

We also propose adaptive step-up procedures that control the FDR when the p-values may have some dependencies. Denote by R_0 (resp. $R_{0,\beta}$) the step-up procedure of threshold collection $\frac{\alpha}{4}\frac{i}{m}$ (resp. $\frac{\alpha}{4}\frac{\beta(i)}{m}$) and put $F(x) = (1 - \sqrt{(2x-1)_+})^{-1}$, where $(\cdot)_+$ is the positive part function. We propose the following new two-stage step-up procedures:

- The two-stage procedure with threshold collection $\Delta(i) = \frac{\alpha}{2}\frac{i}{m}F(|R_0|/m)$.
- The family of two-stage procedures with threshold collection $\Delta(i) = \frac{\alpha}{2}\frac{\beta(i)}{m}F(|R_{0,\beta}|/m)$, with a shape function β of the form (9.2).

We prove that the first procedure and the second procedure family control the FDR at level α, respectively under positive dependencies (PRDS) and unspecified dependencies between the p-values. Compared to the independent case, we lose drastically in the "adaptive part" of these procedures: the level α is divided by 4 in a first step and by 2 in the second step and moreover $F(x)$ is uniformly dominated by $1/(1-x)$. This loss is due to the fact that we use Markov's inequality as a tool in the proofs, which is quite a conservative device. However, if the number of rejections is sufficiently large, our adaptive procedures can outperform the corresponding non-adaptive ones. The interest of the latter procedures is mainly theoretical, but it shows in principle that adaptivity can improve performance of step-up procedures in a theoretically rigorous way even under dependence.

9.5.3 Chapter 12: "Resampling-based confidence regions and multiple tests for a correlated random vector"

Following the description of resampling-based multiple testing procedures of Section 9.4.4, this work gives elements for a third approach that may be called "non-asymptotic approximated approach". Recall that the "exact approach" (see *e.g.* Romano and Wolf (2005)) is based on the fact that the null distributions of the sample are invariant under a transformation. In a particular setting, we show here that a non-asymptotic approach is still possible even if the latter invariance is not exactly satisfied, up to the price of a remaining term.

Motivation

In this work, we observe $\mathbf{Y} := (\mathbf{Y}^1, \dots, \mathbf{Y}^n)$ a $n \geq 2$ i.i.d. sample of integrable random vectors $\mathbf{Y}^i \in \mathbb{R}^K$, supposed to be symmetric around their common mean μ, i.e. $\mathbf{Y}^i - \mu \sim \mu - \mathbf{Y}^i$. We consider the two following multiple testing problems:

- *One-sided problem*: test H_k : "$\mu_k \leq 0$" against A_k : "$\mu_k > 0$", $1 \leq k \leq K$
- *Two-sided problem*: test H_k : "$\mu_k = 0$" against A_k : "$\mu_k \neq 0$", $1 \leq k \leq K$,

in which we want to build multiple testing procedures[4] $R \subset \{1, \dots, K\}$ that control the FWER. Denote by $\overline{\mathbf{Y}} = \frac{1}{n}\sum_{i=1}^n \mathbf{Y}^i$ the empirical mean of the sample $\mathbf{Y}$ and by $[x]$ either x in the one-sided context or $|x|$ in the two-sided context. We easily see that the procedure R rejecting all the H_k such that $\overline{\mathbf{Y}}_k$ is larger than a threshold t_α has a FWER smaller than

$$\mathbb{P}\left(\sup_{k \in \mathcal{H}_0} [\overline{\mathbf{Y}}_k - \mu_k] > t_\alpha\right) , \tag{9.3}$$

where $\mathcal{H}_0 = \{k \mid H_k \text{ is true }\}$. Since μ_k can be unknown if H_k is true, we propose to find confidence regions for μ to bound the above probability.

Two new resampling-based approaches for confidence regions

The main goal of Chapter 12 is to find general $(1-\alpha)$-confidence regions for μ of the form

$$\left\{x \in \mathbb{R}^K \mid \phi\left(\overline{\mathbf{Y}} - x\right) \leq t_\alpha(\mathbf{Y})\right\} \ , \tag{9.4}$$

where $\phi : \mathbb{R}^K \to \mathbb{R}$ is a measurable function and $t_\alpha : \left(\mathbb{R}^K\right)^n \to \mathbb{R}$ is a measurable threshold.

We propose to approach t_α by some resampling scheme, following the heuristics that the distribution of $\overline{\mathbf{Y}} - \mu$ is "close" to the one of

$$\overline{\mathbf{Y}}_{[W-\overline{W}]} := \frac{1}{n}\sum_{i=1}^n (W_i - \overline{W})\mathbf{Y}^i \ ,$$

conditionally to $\mathbf{Y}$, where $(W_i)_{1 \leq i \leq n}$ are real random variables independent of $\mathbf{Y}$ called the *resampling weights* and $\overline{W} = n^{-1}\sum_{i=1}^n W_i$. We propose two different approaches to obtain non-asymptotic confidence regions (9.4):

[4] Note that the number of tests is here denoted by K and that the set of null hypotheses is identified with $\{1, \dots, K\}$.

1. A concentration approach, where we consider $t_\alpha(\mathbf{Y})$ of the form
$$t_\alpha(\mathbf{Y}) = C\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)|\mathbf{Y}\right] + \varepsilon(\alpha, n),$$
with an explicit constant $C > 0$.

2. A quantile approach, where we consider $t_\alpha(\mathbf{Y})$ of the form
$$t_\alpha(\mathbf{Y}) = q_{\alpha'}(\mathbf{Y} - \overline{\mathbf{Y}}) + \varepsilon'(\alpha, n),$$
where $q_{\alpha'}(\mathbf{Y} - \overline{\mathbf{Y}})$ denotes the $(1-\alpha)$-quantile of the distribution of $\overline{\mathbf{Y}}_{[W-\overline{W}]}$ conditionally to $\mathbf{Y}$, and where α' is a level "slightly smaller" than α.

In both approaches, $\varepsilon(\alpha, n)$ and $\varepsilon'(\alpha, n)$ are remaining terms. We prove that both methods provide $(1-\alpha)$-confidence regions for μ as soon as the $\mathbf{Y}^i$ are bounded symmetric vectors or Gaussian vectors. The first method allows us to deal with a very large class of resampling weights W. Moreover, we show that it can be "mixed" with the Bonferroni method (in the Gaussian case). The second method are restricted to the Rademacher weights, because it uses a symmetrization trick (sign-flipping). However, since the second method is adjusted on a quantile, it is generally more accurate than the first method.

Application to multiple testing

Applying these confidence regions with $\phi = \sup_{\mathcal{H}_0}(.)$ or $\phi = 0 \vee \sup_{\mathcal{H}_0}(.)$ (one sided case) and $\phi = \sup_{\mathcal{H}_0}|.|$ (two-sided case), we can use a method developed by Romano and Wolf (2005) to derive new step-down resampling-based multiple testing procedures that control the FWER. Since these procedures use translation-invariant thresholds, the number of iterations in the step-down algorithm is expected to be small. Because of the remaining terms, these procedures are quite conservative, but we show on simulations that they can outperform Holm's procedure when the coordinates of the observed vector has strong enough correlations.

In the two-sided context, since the probability in (9.3) does not depend on the unknown parameter μ anymore, an exact step-down procedure is valid. Moreover, the latter procedure is more accurate than the above methods because it has no remainder term. However, this exact method needs generally more iterations in the step-down algorithm than the above translation-invariant methods. Therefore, we propose to combine our quantile approach with the latter exact method to get a faster procedure. This new "mixed" method can be useful in situation where the non-zero means have a very wide dynamic range (this will be illustrated with a simulation study).

This work has been motivated by neuroimaging data. In this context, computation time is an interesting issue because an iteration of the step-down resampling algorithm sometimes takes more than one day and several iterations of the algorithm can be needed. This is the case in neuroimaging experiment where a "multi-scale" signal has to be detected, for instance when large areas of the brain are strongly activated while many other interesting areas have a small signal. Therefore, one of our future works will be to perform our "mixed" approach on real neuro-imaging data to see if the computation time improvement is really significant in this case.

Chapter 10

A set-output point of view on FDR control in multiple testing

We adopt a set-output point of view of multiple testing and FDR control. We introduce a "self-consistency condition" on the set of hypotheses rejected by the procedure, which implies the control of the corresponding false discovery rate (FDR) under various conditions on the distribution of the p-values. Maximizing the size of the rejected null hypotheses set under the self-consistency condition constraint, we recover various step-up procedures. This way, we recover previous results through synthetic and simple proofs.

10.1 Introduction

In this chapter, we present a survey of some of the most prominent procedures with a controlled FDR. We put a particular emphasis on presenting these results in a "set-output" point of view. While the essence of the arguments used is generally unchanged, we believe this point of view often allows for a more direct and general approach.

This work is in part based on the work of Blanchard and Fleuret (2007), where the authors have shown a link between randomized estimation in learning theory and multiple testing procedures. Under unspecified dependencies between the p-values, they extend significatively the step-up procedure of Benjamini and Yekutieli (2001), by showing that there exists an entire family of related procedures depending on a "prior distribution". To prove this result, they implicitly used a form of "self-consistency condition" in the distribution-free context. In this chapter we propose to extend this "self-consistency condition" to independent or PRDS p-values, where less conservative procedures can be investigated. To be as exhaustive as possible, we will also detail the above result of Blanchard and Fleuret (2007).

In this work, we show that the control of the FDR is deeply connected with a probabilistic inequality over two real variables taking the following form:

$$\mathbb{E}\left(\frac{\mathbf{1}\{U \leq cV\}}{V}\right) \leq c\,, \tag{10.1}$$

where U has a distribution stochastically lower bounded by a uniform distribution, V is a nonnegative random variables and c is a given constant. Moreover, the way in which U and V are related is directly linked to the case of dependencies between the p-values:

(i) in the independent case: V is a non-increasing function of U,

(ii) in the PRDS case: the conditional distribution of V given $U \leq u$ is stochastically decreasing in u,

(iii) in the distribution-free case: the dependence between U and V is unspecified and V is replaced by $\beta(V)$ in the indicator of (10.1), where the shape function β has a specific form.

We prove that (10.1) holds in each of the above contexts in a separated section (Section 10.6). This results in synthetic and simple proofs of FDR controls throughout this chapter. As an illustration, we also show in Section 10.7 a different application of (10.1), which allows to prove that a step-down procedure proposed by Benjamini and Liu (1999a) (see also Romano and Shaikh (2006a)) has FDR control under the PRDS condition (this is as far as we know a novel result).

This chapter is organized as follows: after some preliminaries in Section 10.2, we introduce a type of "self-consistency condition" in Section 10.3, and we prove that different versions of this condition implies the FDR control when the p-values are independent, PRDS or have unspecified dependencies, respectively. In Section 10.4, we show that the step-up procedures satisfy a form of "self-consistency condition", which implies that step-up procedures can control the FDR in all the latter cases of dependencies. We give a conclusion in Section 10.5. Section 10.6 presents technical lemmas and in particular the probabilistic lemmas proving (10.1) in the cases (i), (ii), (iii) described above.

10.2 Preliminaries

We consider the multiple testing framework of Section 9.2.2 (Chapter 9) where it is given a set of *p-values* $\mathbf{p} = (p_h, h \in \mathcal{H})$ for a set of null hypotheses $\mathcal{H}$. Remember that, for a multiple testing procedure R, the *false discovery rate* is defined as the average proportion of true null hypotheses in the set of all the rejected hypotheses:

$$\mathrm{FDR}(R) = \mathbb{E}\left(\frac{|R \cap \mathcal{H}_0|}{|R|}\mathbf{1}\{|R|>0\}\right).$$

10.2.1 Heuristics for FDR control

It is commonly the case that multiple testing procedures are defined as level sets of the p-values:

$$R = \{h \in \mathcal{H} \mid p_h \leq t\}, \tag{10.2}$$

where t is a given (possibly data-dependent) threshold. The FDR of such a threshold-based multiple testing procedure is given by:

$$\mathrm{FDR}(R) = \mathbb{E}\left(\frac{|R \cap \mathcal{H}_0|}{|R|}\mathbf{1}\{|R|>0\}\right) = \sum_{h\in\mathcal{H}_0} \mathbb{E}\left(\frac{\mathbf{1}\{p_h \leq t\}}{|R|}\mathbf{1}\{|R|>0\}\right).$$

At an intuitive level, if the goal is to upper bound the above quantity by a constant, we can be more lax in the choice of the threshold t when the number of rejections $|R|$ is larger. Therefore, a natural idea is to choose a threshold $t = \Delta(h, |R|)$ as a non-decreasing function of

$|R|$ (possibly depending on h). However, one problem with this heuristics is that it apparently leads to a problematic self-referring definition of the procedure (10.2).

In order to formalize this approach, we first introduce in the next section a notion of thresholding-based multiple testing procedures which generalizes the form (10.2) to the case where the threshold t can depend on h and on a global "rejection level" parameter r. Then, in Section 10.3, we introduce a notion of "self-consistency condition" which avoids the self-referring problem mentioned above.

10.2.2 Thresholding-based multiple testing procedures

Definition 10.1 (Threshold collection) *A threshold collection Δ is a function*

$$\Delta : (h, r) \in \mathcal{H} \times \mathbb{R}^+ \mapsto \Delta(h, r) \in \mathbb{R}^+,$$

which is non-decreasing in its second variable. A factorized threshold collection is a threshold collection Δ with the particular form: $\forall (h, r) \in \mathcal{H} \times \mathbb{R}^+$,

$$\Delta(h, r) = \alpha\pi(h)\beta(r)\,,$$

where $\pi : \mathcal{H} \to (0, 1]$ is called the weight function and $\beta : \mathbb{R}^+ \to \mathbb{R}^+$ is a non-decreasing function called the shape function.

When a threshold collection Δ does not depend on h we write just $\Delta(r)$ instead of $\Delta(h, r)$.

Definition 10.2 (Thresholding-based multiple testing procedure) *Given a threshold collection Δ, the Δ-thresholding-based multiple testing procedure at rejection level r is defined as*

$$L_\Delta(r) := \{h \in \mathcal{H} \mid p_h \leq \Delta(h, r)\}. \tag{10.3}$$

10.3 The self-consistency condition in FDR control

Definition 10.3 (Self-consistency condition) *Given a threshold collection Δ, a multiple testing procedure R satisfies the self-consistency condition for the threshold collection Δ, denoted by* $\mathbf{SC}(\Delta)$, *if*

$$R \subset L_\Delta(|R|).$$

The self-consistency condition has a following post-hoc interpretation (close to the reasoning proposed in Section 3.3 of Benjamini and Hochberg (1995)): take R of the form $\{h \in \mathcal{H} \mid p_h \leq t\}$ and suppose that the number of rejected null hypotheses is known by advance and is equal to a deterministic integer $|R| = C$. Consider the problem of choosing the threshold t such that the FDR of R is less than or equal to α: since the expected number of false rejections of R is bounded above by tm, we want to choose $t \leq \alpha C/m$. Therefore, we get $R \subset \{h \in \mathcal{H} \mid p_h \leq \alpha C/m\}$ i.e. $R \subset L_\Delta(|R|)$ for $\Delta(h, r) = \alpha r/m$.

Obviously, since R is a random variable, the above reasoning cannot be applied and proving rigorously the FDR control from the self-consistency condition will be one of the main task of this chapter. Namely, our main results will be to prove that self-consistent procedures have a FDR controlled by α for certain choices of factorized threshold collections $\Delta(h, r) = \alpha\pi(h)\beta(r)$. The choice for the shape function β will depend on the assumptions on the dependency structure between the p-values. Under independence or positive dependencies (PRDS), β can be taken equal to identity: $\beta(r) = r$. Under unspecified dependencies, β is chosen under a particular form (see (10.7) below).

In Section 10.4, we will see that the "step-up" multiple testing procedures are a particular case of self-consistent procedures.

10.3.1 Independent case

We first consider the case where the family of p-values $\mathbf{p} = (p_h, h \in \mathcal{H})$ is independent.

Proposition 10.4 *Assume that the collection of p-values $\mathbf{p} = (p_h, h \in \mathcal{H})$ forms an independent family of random variables. Let R be a multiple testing procedure such that $|R(\mathbf{p})|$ is non-increasing in each p-value p_h with $h \in \mathcal{H}_0$, and satisfies the self-consistency condition $\mathbf{SC}(\Delta)$ with $\Delta(h, r) = \alpha\pi(h)r$. Then R has a FDR less than or equal to $\alpha\pi(\mathcal{H}_0)$, with $\pi(\mathcal{H}_0) := \sum_{h\in\mathcal{H}_0} \pi(h)$.*

The proof of the above result is particularly simple.

Proof. For each $h \in \mathcal{H}$, we denote by $\mathbf{p}_{-h}$ the collection of p-values $(p_{h'}, h' \neq h)$. From the definition of the FDR and using $\mathbf{SC}(\Delta)$, we get:

$$\begin{aligned}
\mathrm{FDR}(R) &\leq \sum_{h\in\mathcal{H}_0} \mathbb{E}\left(\frac{\mathbf{1}\{p_h \leq \alpha\pi(h)|R|\}}{|R|}\right) && (10.4)\\
&= \sum_{h\in\mathcal{H}_0} \mathbb{E}\left(\mathbb{E}\left(\frac{\mathbf{1}\{p_h \leq \alpha\pi(h)|R(\mathbf{p})|\}}{|R(\mathbf{p})|}\middle|\mathbf{p}_{-h}\right)\right)\\
&\leq \sum_{h\in\mathcal{H}_0} \alpha\pi(h).
\end{aligned}$$

For the last step, we use that the distribution of p_h conditionally to $\mathbf{p}_{-h}$ is stochastically lower bounded by a uniform distribution (because of the independence assumption), so that we can apply Lemma 10.17 with $U = p_h$, $g(U) = |R(\mathbf{p}_{-h}, U)|$ (the value of $\mathbf{p}_{-h}$ being fixed) and $c = \alpha\pi(h)$. ■

Remark 10.5 *In the above proof, note that the expectations are well defined because the event $\{p_h = 0\}$ is of measure zero.*

Remark 10.6 *Note that Proposition 10.4 is still valid under the slightly weaker assumption where for all $h \in \mathcal{H}_0$, p_h is independent from $(p_{h'}, h' \neq h)$ (in particular, the p-values of $(p_h, h \in \mathcal{H}_1)$ may not be mutually independent).*

10.3.2 Case of positive dependencies

We now consider an extension of the previous result where instead of requiring $|R|$ to be a non-increasing function of $\mathbf{p}$ and the p-values to be independent, we reach the same conclusion as previously under the weaker hypothesis that $|R|$ is stochastically decreasing with respect to each p-value associated to a true null hypothesis.

Proposition 10.7 *Let R be a multiple testing procedure satisfying the self-consistency condition* $\mathbf{SC}(\Delta)$ *with $\Delta(h,r) = \alpha\pi(h)r$, and such that for any $h \in \mathcal{H}_0$, the conditional distribution of $|R|$ given $p_h \leq u$ is stochastically decreasing in u:*

$$\text{for any } r \geq 0\,,\ \text{the function } u \mapsto \mathbb{P}(|R| < r \mid p_h \leq u) \text{ is non-decreasing}. \tag{10.5}$$

Then, we have $FDR(R) \leq \alpha\pi(\mathcal{H}_0)$.

Proof. We use (10.4) and we can conclude with Lemma 10.18 applied with $U = p_h$, $V = |R|$ and $c = \alpha\pi(h)$. ∎

We give now conditions providing that R satisfies (10.5). For this, we recall the definition of positive regression dependency on each one from a subset (PRDS) (see *e.g.* Benjamini and Yekutieli (2001)). Remember that a subset $D \subset [0,1]^{\mathcal{H}}$ is called *non-decreasing* if for all $\mathbf{z}, \mathbf{z}' \in [0,1]^{\mathcal{H}}$ such that $\mathbf{z} \leq \mathbf{z}'$ (i.e. $\forall h \in \mathcal{H}, z_h \leq z'_h$), we have $\mathbf{z} \in D \Rightarrow \mathbf{z}' \in D$.

Definition 10.8 *For $\mathcal{H}'$ a subset of $\mathcal{H}$, the p-values of $\mathbf{p} = (p_h, h \in \mathcal{H})$ are said to be positively regressively dependent on each one from $\mathcal{H}'$ (denoted in short by PRDS on $\mathcal{H}'$), if for any non-decreasing set $D \subset [0,1]^{\mathcal{H}}$, and for any $h \in \mathcal{H}'$, the function*

$$u \in [0,1] \mapsto \mathbb{P}(\mathbf{p} \in D \mid p_h = u) \tag{10.6}$$

is non-decreasing.

Note that in expression (10.6), the conditional probability is well defined because it can be seen as a conditional expectation (it is then defined almost surely with respect to the distribution of p_h). We can now state that the self-consistency condition implies the FDR control under positive dependencies:

Corollary 10.9 *Suppose that the p-values of $\mathbf{p} = (p_h, h \in \mathcal{H})$ are PRDS on $\mathcal{H}_0$, and consider a multiple testing procedure R such that $|R(\mathbf{p})|$ is non-increasing in each p-value. If R satisfies the self-consistency condition* $\mathbf{SC}(\Delta)$ *with $\Delta(h,r) = \alpha\pi(h)r$, then $FDR(R) \leq \alpha\pi(\mathcal{H}_0)$.*

Proof. We merely check that condition (10.5) of Proposition 10.7 is satisfied. For any fixed $r \geq 0$, put $D = \left\{\mathbf{z} \in [0,1]^{\mathcal{H}} \mid |R(\mathbf{z})| < r\right\}$. It is clear from the assumptions on R that D is a non-decreasing set. Under the PRDS condition, for all $u \leq u'$, putting $\gamma = \mathbb{P}\left[p_h \leq u \mid p_h \leq u'\right]$,

$$\begin{aligned}
\mathbb{P}\left[\mathbf{p} \in D \mid p_h \leq u'\right] &= \mathbb{E}\left[\mathbb{P}\left[\mathbf{p} \in D \mid p_h\right] \mid p_h \leq u'\right] \\
&= \gamma \mathbb{E}\left[\mathbb{P}\left[\mathbf{p} \in D \mid p_h\right] \mid p_h \leq u\right] + (1-\gamma)\mathbb{E}\left[\mathbb{P}\left[\mathbf{p} \in D \mid p_h\right] \mid u < p_h \leq u'\right] \\
&\geq \mathbb{E}\left[\mathbb{P}\left[\mathbf{p} \in D \mid p_h\right] \mid p_h \leq u\right] = \mathbb{P}\left[\mathbf{p} \in D \mid p_h \leq u\right].
\end{aligned}$$

∎

10.3.3 Case of unspecified dependencies

We now consider a setting with no assumptions on the dependency structure between the p-values. In order to ensure FDR control in this situation, we require a more restrictive self-consistency condition, using a shape function $\beta(r) \leq r$ of a particular form.

Proposition 10.10 *Consider a multiple testing procedure R satisfying the self-consistency condition* $\mathbf{SC}(\Delta)$ *with* $\Delta(h,r) = \alpha\pi(h)\beta(r)$, *and a shape function* β *of the form:* $\forall r \geq 0$,

$$\beta(r) = \int_0^r u d\nu(u)\,, \tag{10.7}$$

where ν *is a probability distribution on* $(0,\infty)$ *. Then we have* $FDR(R) \leq \alpha\pi(\mathcal{H}_0)$.

Proof. Using (10.4), we apply Lemma 10.19 with $U = p_h$, $V = |R|$ and $c = \alpha\pi(h)$. ■

10.4 Step-up multiple testing procedures in FDR control

10.4.1 A general definition of the step-up procedures

We give here a general definition of a step-up procedure, connected to the point of view of Theorem 2 of Benjamini and Hochberg (1995): the step-up procedure is defined as the maximal set satisfying the self-consistency condition $\mathbf{SC}(\Delta)$, which can be caracterized as follows:

Definition 10.11 (Step-up procedure) *Let* Δ *be a threshold collection. The step-up multiple testing procedure* R *associated to* Δ *, is given by either of the following equivalent definitions:*

$$(i)\ \ R = L_\Delta(\hat{r})\,,\ \ \textit{where } \hat{r} := \max\{r \geq 0 \mid |L_\Delta(r)| \geq r\}$$
$$(ii)\ \ R = \bigcup \big\{A \subset \mathcal{H} \mid A \textit{ satisfies } \boldsymbol{SC}(\Delta)\big\}$$

Proof of the equivalence between (i) **and** (ii)**.** Note that since Δ is assumed to be non-decreasing in its second variable, $L_\Delta(r)$ is a non-decreasing set as a function of $r \geq 0$. Therefore, $|L_\Delta(r)|$ is a non-decreasing function of r and the supremum appearing in (i) is indeed a maximum i.e. $|L_\Delta(\hat{r})| \geq \hat{r}$. Hence $L_\Delta(\hat{r}) \subset L_\Delta(|L_\Delta(\hat{r})|)$, so $L_\Delta(\hat{r})$ is included in the set union appearing in (ii). Conversely, for any set A satisfying $A \subset L_\Delta(|A|)$, we have $|L_\Delta(|A|)| \geq |A|$, so that $|A| \leq \hat{r}$ and $A \subset L_\Delta(\hat{r})$. ■

The decision point $\hat{r}$ is obtained easily from the "last right crossing" point between the (non-decreasing) number of rejected hypotheses $|L_\Delta(.)|$ and the identity function (see an illustration on Figure 10.1).

When the threshold collection $\Delta(h,r) = \alpha\pi(h)\beta(r)$ is factorized, Definition 10.11 is equivalent to the classical "re-ordering-based" definition of a step-up procedure: for any $h \in \mathcal{H}$, denote by $p'_h := p_h/(m\pi(h))$ the *weighted p-value* of h, and consider a permutation $i \in \{1,\ldots,m\} \mapsto (i) \in \mathcal{H}$ ordering the weighted p-values i.e. such that

$$p'_{(1)} \leq p'_{(2)} \leq \cdots \leq p'_{(m)}.$$

Since $L_\Delta(r) = \{h \in \mathcal{H} \mid p'_h \leq \alpha\beta(r)/m\}$, the condition $|L_\Delta(r)| \geq r$ is equivalent to $p'_{(r)} \leq \alpha\beta(r)/m$. Hence, the step-up procedure associated to Δ defined in Definition 10.11 rejects all the $\hat{r}$ smallest weighted p-values, where $\hat{r}$ corresponds to the "last right crossing" point between the ordered weighted p-values $p'_{(\cdot)}$ and the threshold $\alpha\beta(\cdot)/m$ (see Figure 10.2 for an illustration):

$$\hat{r} \;=\; \max\big\{r \in \{0,\ldots,m\} \mid p'_{(r)} \leq \alpha\beta(r)/m\big\},$$

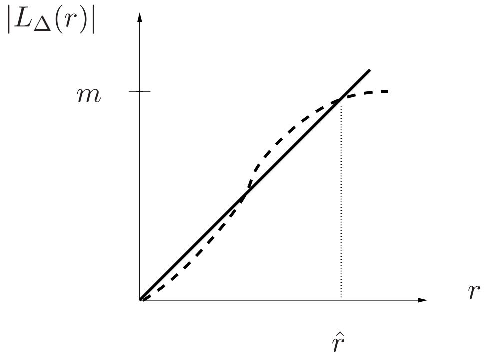


Figure 10.1: Graphs of $|L_\Delta(\cdot)|$ (dashed line) and of the identity function (solid line). The number of rejected hypotheses of the step-up procedure $\hat{r}$ is obtained on the X-axis. On the Y-axis, m ensures that the crossing point corresponding to $\hat{r}$ is the last.

with $p'_{(0)} := 0$. If for each hypothesis $h \in \mathcal{H}$, $\pi(h) = 1/m$, we note that the weighted p-values $(p'_h)_h$ are just the p-values $(p_h)_h$. In particular:

- The step-up procedure associated to $\Delta(h, r) = \alpha r/m$ is the so-called linear step-up procedure of Benjamini and Hochberg (1995) (corresponding to the shape function $\beta(r) = r$).

- The step-up procedure associated to $\Delta(h, r) = \alpha r/(m(1 + 1/2 + \cdots + 1/m))$ is the distribution-free step-up procedure of Benjamini and Yekutieli (2001) (corresponding to the shape function $\beta(r) = r/(1 + 1/2 + \cdots + 1/m)$).

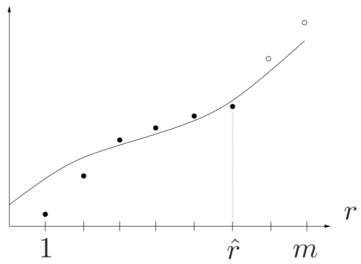


Figure 10.2: Comparison between the ordered weighted p-values $p'_{(\cdot)}$ (points) and the threshold $\alpha\beta(.)/m$ (solid line). Here the step-up procedure rejects the 6 hypotheses corresponding to the 6 smallest reweighted p-values (solid points) ($\hat{r} = 6$).

Remark 10.12 *Using the weight function π, the step-up procedures that we consider here can be "weighted". The interest of the "weighted" procedures is that choosing a particular π can increase their performance (see for instance Genovese* et al. *(2006) and Wasserman and Roeder (2006)).*

10.4.2 Classical FDR control with some extensions

A direct consequence of Definition 10.11 is that the step-up procedure associated to Δ satisfies the self-consistency condition $\mathbf{SC}(\Delta)$. Therefore, we can apply the results of Section 10.3 to derive FDR control theorems. We first consider the *weighted linear step-up procedure*, that is, the step-up procedure associated to the threshold collection $\Delta(h, r) = \alpha\pi(h)r$.

Theorem 10.13 *The weighted linear step-up procedure R satisfies the FDR control: FDR$(R) \leq \pi(\mathcal{H}_0)\alpha$, where $\pi(\mathcal{H}_0) := \sum_{h\in\mathcal{H}_0} \pi(h)$, in either of the following cases:*

- *the p-values of $\mathbf{p} = (p_h, h \in \mathcal{H})$ are independent.*
- *the p-values of $\mathbf{p} = (p_h, h \in \mathcal{H})$ are PRDS on $\mathcal{H}_0$.*

Moreover, in the independent case, if the p-values p_h with $h \in \mathcal{H}_0$ are exactly distributed like a uniform distribution, the linear step-up procedure (uniformly weighted i.e. with $\forall h \in \mathcal{H}, \pi(h) = 1/m$) has a FDR exactly equal to $m_0\alpha/m$.

The two first points of Theorem 10.13 were initialy proved by Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001) with a uniform π. The additional assertion was proved by Finner and Roters (2001). A proof for FDR$(R) \leq \pi(\mathcal{H}_0)\alpha$ with a general π and in the independent case was investigated by Genovese *et al.* (2006). Here the set-ouput framework gives a general version of these results with a concise proof.

Proof. The two first FDR controls are both a direct consequence of Proposition 10.4 and Corollary 10.9. It remains to prove the additional assertion: for each null hypothesis h, denote R'_{-h} the step-up procedure associated to the threshold collection $\Delta'(r) = \alpha(r+1)/m$ and restricted to the hypotheses of $\mathcal{H}\backslash\{h\}$. Lemma 10.20 states that

$$\begin{aligned} h \in R &\Leftrightarrow R = R'_{-h} \cup \{h\} \\ &\Leftrightarrow p_h \leq \alpha(|R'_{-h}| + 1)/m. \end{aligned}$$

Therefore,

$$\begin{aligned} \mathrm{FDR}(R) &= \sum_{h\in\mathcal{H}_0} \mathbb{E}\left(\frac{\mathbf{1}\{h \in R\}}{|R|}\right) \\ &= \sum_{h\in\mathcal{H}_0} \mathbb{E}\left(\frac{\mathbf{1}\{p_h \leq \alpha(|R'_{-h}| + 1)/m\}}{|R'_{-h}| + 1}\right) \\ &= \sum_{h\in\mathcal{H}_0} \mathbb{E}\left(\frac{\mathbb{P}\left[p_h \leq \alpha(|R'_{-h}| + 1)/m \;\middle|\; |R'_{-h}|\right]}{|R'_{-h}| + 1}\right). \end{aligned}$$

For any $h \in \mathcal{H}_0$, we use simultaneously:

- $|R'_{-h}|$ depends only on the p-values of $(p_{h'}, h' \neq h)$,
- p_h has a uniform distribution conditionally to $\mathbf{p}_{-h}$ (independence assumption)
- $\alpha(|R'_{-h}| + 1)/m \leq 1$,

to deduce that $\mathbb{P}\left[p_h \leq \alpha(|R'_{-h}|+1)/m) \;\middle|\; |R'_{-h}|\right] = \alpha(|R'_{-h}|+1)/m$. The result follows. ■

We now consider the case where the p-values have unspecified dependencies (the first part of the following theorem was first proved by Blanchard and Fleuret (2007) and the second part was established by Lehmann and Romano (2005a) in relation to Hommel's inequality):

Theorem 10.14 *Consider R the step-up procedure associated to the factorized threshold collection $\Delta(h,r) = \alpha\pi(h)\beta(r)$, where the shape function β has the following specific form: for each $r \geq 0$,*

$$\beta(r) = \int_0^r u d\nu(u)\,, \tag{10.8}$$

where ν is some probability distribution on $(0,\infty)$. Then R has its FDR controlled at level $m_0\alpha/m$.
Moreover, when $m_0 = m$, if the distribution ν has its support included in $\{1,\ldots,m\}$, and $\pi(h) = 1/m$ for each $h \in \mathcal{H}$, the above control is sharp, meaning that there exist a joint distribution for the p-values such that $FDR(R) = \alpha$.

Remark 10.15 *Theorem 10.14 can be seen as an extension to the FDR and in a continuous case of a celebrated inequality due to Hommel (1983), which has been widely used in the multiple testing literature (see e.g. Lehmann and Romano (2005a); Romano and Shaikh (2006a,b)). Namely, when ν has discrete support $\{1,\ldots,m\}$ and $m = m_0$, the above result recovers Hommel's inequality. Note that this specific case corresponds to a "weak control" where we assume that all null hypotheses are true; in this situation the FDR is equal to the FWER.*

Proof. The first part of Theorem 10.14 is a direct application of Proposition 10.10. Let us prove the last part of the theorem. We build the joint distribution of the p-values in the following way: take a random variable K such that for all $k \in \{1,\ldots,m\}$, $\mathbb{P}(K = k) = \alpha\nu(k)$ and $\mathbb{P}(K = 0) = 1-\alpha$. Conditionnally to K, we choose I a subset of $\mathcal{H}$ uniformly distributed among the subsets of K elements of $\mathcal{H}$. Conditionally to I (and K), we choose (all independently)

$$\begin{aligned}
&\forall h \in I, p_h \text{ uniform in } [\alpha\beta(K-1)/m, \alpha\beta(K)/m)\\
&\forall h \notin I, p_h \text{ uniform in } [\alpha\beta(m)/m, 1],
\end{aligned}$$

We check that unconditionally, each p_h is uniform on $[0,1]$: for all $k \in \{1,\ldots,m\}$,

$$\mathbb{P}(p_h \in [\alpha\beta(k-1)/m, \alpha\beta(k)/m)) = \mathbb{P}(K = k, h \in I) = \mathbb{P}(h \in I \mid K = k)\alpha\nu(k) = \alpha k\nu(k)/m,$$

and by definition of β, $\beta(k) - \beta(k-1) = k\nu(k)$. Finally, we just have to remark that the step-up procedure R rejects exactly the K hypotheses in $[\alpha\beta(K-1)/m, \alpha\beta(K)/m)$ to conclude $\mathrm{FDR}(R) = \mathrm{FWER}(R) = \mathbb{P}(K > 0) = \alpha$. ■

Theorem 10.14 establishes that, under unspecified dependencies between the p-values, there exists a family of step-up procedures that control the false discovery rate. This family depends on the shape function β which itself depends on the distribution ν. The distribution ν represents a prior belief on the final number of rejections of the procedure. For instance,

- If we do not have any prior belief, we can choose ν uniform on $\{1,\ldots,m\}$ and this gives a quadratic shape function:
$$\beta(r) = r(r+1)/2m\,.$$
Note that the corresponding step-up procedure has already been proposed by Sarkar (2006).

- If we are in a problem where a small number of rejections is expected (for instance m_0 "large"), we can choose $\nu(k) = C^{-1}k^{-1}$ for $k \in \{1,\ldots,m\}$ with the normalization constant $C = \sum_{k=1}^{m} 1/k$. This gives
$$\beta(r) = r/C\,,$$
and we find the distribution-free procedure of Benjamini and Yekutieli (2001). In particular, Theorem 10.14 is a generalization of Theorem 1.3 of Benjamini and Yekutieli (2001) (we also note that here the procedure may be weighted with π).

- If we expect a large number of rejections (for instance m_0 "small"), we can choose $\nu(k) = 2k/(m(m+1))$ for $k \in \{1,\ldots,m\}$, which leads to
$$\beta(r) = r(r+1)(2r+1)/(3m(m+1))\,.$$

Of course, many other choices for ν are possible. In Example 10.16, we give several choices of continuous ν, and we plot the graphs of the corresponding shape functions in Figure 10.3 (page 165). We could also use the discretized versions of the proposed continuous distributions ν; it leads to slightly smaller functions β, but generally with more complex expressions. From Figure 10.3, it is clear that the choice of the prior has a large impact on the final number of rejections of the procedure. Moreover, since no shape function dominates the others, there is no optimal choice among these prior distributions (the performance of a given procedure will depend on the data). It is then tempting to choose ν in a data-dependent way: showing that the corresponding FDR is still well controlled is an interesting open problem for future research.

Example 10.16 (Some choices for ν and corresponding shape functions β)

1. *Dirac distributions: $\nu = \delta_\lambda$, with $\lambda > 0$. $\beta(r) = \lambda \mathbf{1}\{r \geq \lambda\}$.*

2. *(Truncated-) Gaussian distributions: ν equals to the distribution of $\max(X,1)$, where X follows a Gaussian distribution with mean μ and variance σ^2.*
$$\beta(r) = \big[\Phi((r-\mu)/\sigma) - \Phi((1-\mu)/\sigma)\big]\mu + \sigma\big[\exp(-(1-\mu)^2/(2\sigma^2)) - \exp(-(r-\mu)^2/(2\sigma^2))\big]/\sqrt{2\pi},$$
where Φ is the standard Gaussian cumulative distribution function: $\forall y \in \mathbb{R}$, $\Phi(y) = \mathbb{P}(Y \leq y)$, where $Y \sim \mathcal{N}(0,1)$.

3. *Distributions with a power function density: $\forall r \geq 0, d\nu(r) = r^\gamma \mathbf{1}\{r \in [1,m]\}dr / \int_1^m u^\gamma du$, $\gamma \in \mathbb{R}$.*
$$\beta(r) = \begin{cases} \frac{\gamma+1}{\gamma+2}\frac{r^{\gamma+2}-1}{m^{\gamma+1}-1} & \text{if} \quad \gamma \neq -1,-2 \\ \frac{r-1}{\log(m)} & \text{if} \quad \gamma = -1 \\ \frac{\log(r)}{1-1/m} & \text{if} \quad \gamma = -2 \end{cases}.$$
As a particular case, when $\gamma = 0$, ν is uniformly distributed on $[1,m]$ and $\beta(r) = (r^2 - 1)/(2(m-1))$.

4. Exponential distributions: $d\nu(r) = (1/\lambda)\exp(-r/\lambda)dr$, *with* $\lambda > 0$.

$$\beta(r) = \lambda(1 - \exp(-r/\lambda)) - r\exp(-r/\lambda).$$

10.5 Conclusion

Extending the work of Blanchard and Fleuret (2007), we demonstrated in this chapter that the set-output point view — through the self-consistency condition — is a pratical approach to prove that a procedure controls the FDR, when the p-values are independent, PRDS or when they have unspecified dependencies. This point of view is very flexible, because the FDR control is provided as soon as a suitable self-consistency condition is satisfied: for instance, it is easy to check that the step down procedures satisfy a self-consistency condition, which implies directly that all the FDR control results in this chapter hold with "step-up" replaced by "step-down".

Moreover, the self-consistency condition, as well as the general definition of a step-up procedure that we give here can be extended to the case where $|\cdot|$ is a general volume measure on $\mathcal{H}$. While we choose to present the case where $|\cdot|$ is the counting measure just for clarity, all the FDR control results presented here can be extended to a general measure. The FDR appears then as the expected ratio between the volume of the rejected true null hypotheses and the volume of the rejected null hypotheses. This can be useful in practice, if we want to give more weights to some null hypotheses, and if we want to control the corresponding ratio.

Furthermore, in a spirit close to Perone Pacifico *et al.* (2004), we believe that some of the results presented here may be extended to the case where $\mathcal{H}$ is a possibly continuous measurable space, endowed with a proper σ-algebra $\mathfrak{H}$. This would allow us to test continuous sets of null hypotheses, so that we would be able to deal with the problem detection of non-zero mean of a continuous process, while controlling the "rate" of false discovery. While it is clear that the independent assumption can not be extended to continuous $\mathcal{H}$, it is legitimate to think that this will be the case for PRDS and unspecified dependencies. This is an exciting direction for future work.

10.6 Technical lemmas

Lemma 10.17 *Let* $g : [0,1] \to (0,\infty)$ *be a non-increasing function. Let* U *be a random variable which has a distribution stochastically lower bounded by a uniform distribution, that is,* $\forall u \in [0,1]$, $\mathbb{P}(U \le u) \le u$ *. Then, for any constant* $c > 0$, *we have*

$$\mathbb{E}\left(\frac{\mathbf{1}\{U \le cg(U)\}}{g(U)}\right) \le c\,.$$

Proof. We let $\mathcal{U} = \{u \mid cg(u) \ge u\}$, $u^* = \sup \mathcal{U}$ and $C^\star = \inf\{g(u) \mid u \in \mathcal{U}\}$. It is not difficult to check that $u^* \le cC^*$ (for instance take any non-decreasing sequence $u_n \in \mathcal{U} \nearrow u^*$, so that $g(u_n) \searrow C^*$). If $C^* = 0$, then $u^* = 0$ and the result is trivial. Otherwise, we have

$$\mathbb{E}\left(\frac{\mathbf{1}\{U \le cg(U)\}}{g(U)}\right) \le \frac{\mathbb{P}(U \in \mathcal{U})}{C^*} \le \frac{\mathbb{P}(U \le u^*)}{C^*} \le \frac{u^*}{C^*} \le c.$$

■

Lemma 10.18 *Let U, V be two non-negative real variables. Assume the following:*

1. *The distribution of U is stochastically lower bounded by a uniform distribution, that is, $\forall u \in [0,1],\ \mathbb{P}(U \leq u) \leq u$.*

2. *The conditional distribution of V given $U \leq u$ is stochastically decreasing in u, that is, for any $v \geq 0$, the function $u \mapsto \mathbb{P}(V < v \mid U \leq u)$ is non-decreasing.*

Then, for any constant $c > 0$, we have

$$\mathbb{E}\left(\frac{\mathbf{1}\{U \leq cV\}}{V}\right) \leq c\,.$$

Proof. Fix some $\varepsilon > 0$ and some $\rho \in (0,1)$ and choose K big enough so that $\rho^K < \varepsilon$. Put $v_0 = 0$ and $v_i = \rho^{K+1-i}$ for $1 \leq i \leq 2K+1$. Therefore,

$$\begin{aligned}
\mathbb{E}\left(\frac{\mathbf{1}\{U \leq cV\}}{V \vee \varepsilon}\right) &\leq \sum_{i=1}^{2K+1} \frac{\mathbb{P}(U \leq cv_i; V \in [v_{i-1}, v_i))}{v_{i-1} \vee \varepsilon} + \varepsilon \\
&\leq c \sum_{i=1}^{2K+1} \frac{\mathbb{P}(U \leq cv_i; V \in [v_{i-1}, v_i))}{\mathbb{P}(U \leq cv_i)} \frac{v_i}{v_{i-1} \vee \varepsilon} + \varepsilon \\
&\leq c\rho^{-1} \sum_{i=1}^{2K+1} \mathbb{P}(V \in [v_{i-1}, v_i) \mid U \leq cv_i) + \varepsilon \\
&= c\rho^{-1} \sum_{i=1}^{2K+1} \big(\mathbb{P}(V < v_i \mid U \leq cv_i) - \mathbb{P}(V < v_{i-1} \mid U \leq cv_i)\big) + \varepsilon \\
&\leq c\rho^{-1} \sum_{i=1}^{2K+1} \big(\mathbb{P}(V < v_i \mid U \leq cv_i) - \mathbb{P}(V < v_{i-1} \mid U \leq cv_{i-1})\big) + \varepsilon \\
&\leq c\rho^{-1} + \varepsilon\,.
\end{aligned}$$

We obtain the conclusion by letting $\rho \to 1$, $\varepsilon \to 0$ and applying the monotone convergence theorem. ■

Lemma 10.19 *Let U, V be two non-negative real variables and β be a function of the form (10.7). Assume that the distribution of U is stochastically lower bounded by a uniform distribution, that is, $\forall u \in [0,1],\ \mathbb{P}(U \leq u) \leq u$. Then, for any constant $c > 0$, we have*

$$\mathbb{E}\left(\frac{\mathbf{1}\{U \leq c\beta(V)\}}{V}\right) \leq c\,.$$

Proof. First note that since $\beta(0) = 0$, the expectation is always well defined. Since for any

$z > 0$, $\int_0^{+\infty} v^{-2}\mathbf{1}\{v \geq z\}dv = 1/z$ and so using Fubini's theorem:

$$\begin{aligned}\mathbb{E}\left(\frac{\mathbf{1}\{U \leq c\beta(V)\}}{V}\right) &= \mathbb{E}\Big(\int_0^{+\infty} v^{-2}\mathbf{1}\{v \geq V\}\mathbf{1}\{U \leq c\beta(V)\}dv\Big)\\ &= \int_0^{+\infty} v^{-2}\mathbb{E}[\mathbf{1}\{v \geq V\}\mathbf{1}\{U \leq c\beta(V)\}]dv\\ &\leq \int_0^{+\infty} v^{-2}\mathbb{P}\big(U \leq c\beta(v)\big)dv\\ &\leq c\int_0^{+\infty} v^{-2}\beta(v)dv,\end{aligned}$$

and we conclude because any function β of the form (10.7) satisfies $\int_0^{+\infty} v^{-2}\beta(v)dv = 1$. ■

Lemma 10.20 *Let R be a step-up procedure associated to a threshold collection Δ . Note for any $h \in \mathcal{H}$, R'_{-h} the step-up procedure associated to the threshold collection $\Delta'(h,r) = \Delta(h,r{+}1)$ and restricted to the null hypotheses of $\mathcal{H}\backslash\{h\}$. Then the three following conditions are equivalent:*

(i) $h \in R$

(ii) $R = R'_{-h} \cup \{h\}$

(iii) $p_h \leq \Delta(h, |R'_{-h}| + 1)$

Proof. Let us denote by $\mathbf{SC}(\Delta)$ the self-consistency condition $A \subset \{h' \in \mathcal{H} \mid p_{h'} \leq \Delta(h', |A|)\}$ (satisfied by R) and by **SC'**(Δ') the self-consistency condition $A \subset \{h' \in \mathcal{H}\backslash\{h\} \mid p_{h'} \leq \Delta'(h', |A|)\}$ (satisfied by R'_{-h}). We first prove the equivalence between (i) and (ii): $(ii) \Rightarrow (i)$ is trivial. Let us prove $(i) \Rightarrow (ii)$. Suppose that $h \in R$. We first prove $R \subset R'_{-h} \cup \{h\}$ by showing that $R\backslash\{h\} \subset R'_{-h}$: for this we just see that $R\backslash\{h\}$ satisfies the self-consistency condition **SC'**(Δ'):

$$\begin{aligned}R\backslash\{h\} &\subset \{h' \in \mathcal{H}\backslash\{h\} \mid p_{h'} \leq \Delta(h', |R|)\}\\ &= \{h' \in \mathcal{H}\backslash\{h\} \mid p_{h'} \leq \Delta(h', |R\backslash\{h\}| + 1)\}\\ &= \{h' \in \mathcal{H}\backslash\{h\} \mid p_{h'} \leq \Delta'(h', |R\backslash\{h\}|)\}.\end{aligned}$$

To prove $R'_{-h} \cup \{h\} \subset R$, we remark that the set R'_{-h} satisfies

$$\begin{aligned}R'_{-h} &\subset \{h' \in \mathcal{H}\backslash\{h\} \mid p_{h'} \leq \Delta'(h', |R'_{-h}|)\}\\ &= \{h' \in \mathcal{H}\backslash\{h\} \mid p_{h'} \leq \Delta(h', |R'_{-h}| + 1)\}\\ &= \{h' \in \mathcal{H}\backslash\{h\} \mid p_{h'} \leq \Delta(h', |R'_{-h} \cup \{h\}|)\}.\end{aligned}$$

Moreover, since h is such that $p_h \leq \Delta(h, |R|) \leq \Delta(h, |R'_{-h} \cup \{h\}|)$, the set $R'_{-h} \cup \{h\}$ satisfies $\mathbf{SC}(\Delta)$ and $R'_{-h} \cup \{h\} \subset R$.

It is clear that $((i)$ and $(ii)) \Rightarrow (iii)$. Finally, $(iii) \Rightarrow (i)$ holds because when we proved $R'_{-h} \cup \{h\} \subset R$, we only used $p_h \leq \Delta(h, |R'_{-h} \cup \{h\}|)$. ■

10.7 Appendix: another consequence of the probabilistic lemmas

In this section, we propose another application of Lemma 10.18 by showing that the step-down procedure proposed by Benjamini and Liu (1999a) and Romano and Shaikh (2006a) controls the FDR under the PRDS assumption.

Consider $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ the ordered p-values of $(p_h, h \in \mathcal{H})$, and put $p_{(0)} = 0$. Given a non-decreasing threshold collection $i \mapsto \Delta(i)$ (independent of h), remember that the *step-down procedure* of threshold collection Δ is defined as $R = \{h \in \mathcal{H} \mid p_h \leq p_{(k)}\}$, where

$$k = \max\{i \mid \forall j \leq i, p_{(j)} \leq \Delta(j)\}. \tag{10.9}$$

Benjamini and Liu (1999a) and Romano and Shaikh (2006a) have introduced the step-down procedure with the threshold collection $\Delta(i) = \frac{\alpha m}{(m-i+1)^2}$. They proved that this procedure controls the FDR at level α if for each $h \in \mathcal{H}_0$, p_h is independent of the collection of p-values $(p_{h'}, h' \in \mathcal{H}_1)$ (in fact Romano and Shaikh (2006a) used a slightly weaker assumption: see 3 of Remark 10.22 below). Here, we give a proof valid under the more general PRDS assumption. First, we extend slightly the notion of "PRDS on $\mathcal{H}_0$" given in Definition 10.8: the p-values of $(p_h, h \in \mathcal{H})$ are said to be *PRDS from $\mathcal{H}_1$ to $\mathcal{H}_0$*, if for all non-decreasing set $D \subset [0,1]^{\mathcal{H}_1}$ and for all $h \in \mathcal{H}_0$, the function

$$u \mapsto \mathbb{P}\big((p_{h'})_{h' \in \mathcal{H}_1} \in D \mid p_h = u\big)$$

is non-decreasing. Note that the latter condition is obviously satisfied when p_h is independent of $(p_{h'}, h' \in \mathcal{H}_1)$. We give now the main result of this section.

Theorem 10.21 *Suppose that the p-values of $(p_h, h \in \mathcal{H})$ are PRDS from $\mathcal{H}_1$ to $\mathcal{H}_0$. Then the step-down procedure of threshold collection $\Delta(i) = \frac{\alpha m}{(m-i+1)^2}$ has a FDR less than or equal to α.*

Proof. Assume $m_0 > 0$ (otherwise the result is trivial). Denote by j_0 the (data-dependent) smallest integer $j \geq 1$ for which $p_{(j)}$ corresponds to a true null hypothesis. Denote by R_1 the step-down procedure of threshold collection Δ and restricted to the set of the false null hypotheses $\mathcal{H}_1$. First note that the two following points hold:

(*i*) $|R \cap \mathcal{H}_0| > 0 \Rightarrow p_{(j_0)} \leq \frac{\alpha m}{(m-j_0+1)^2}$

(*ii*) $|R \cap \mathcal{H}_0| > 0 \Rightarrow j_0 - 1 \leq |R_1|$

To prove this, suppose that $|R \cap \mathcal{H}_0| > 0$, so that the null hypothesis corresponding to $p_{(j_0)}$ is rejected by R. Hence, from the definition of a step-down procedure we have $p_{(j_0)} \leq \Delta(j_0)$ and (*i*) holds. Moreover, since $\forall j \leq j_0 - 1$, we have $p_{(j)} \leq \Delta(j)$ and $p_{(j)}$ corresponds to a false null hypothesis, R_1 necessarily rejects all the null hypotheses corresponding to $p_{(j)}, j \leq j_0 - 1$, and we get (ii). Therefore,

$$\begin{aligned}
\mathrm{FDR}(R) &= \mathbb{E}\left(\frac{|R\cap\mathcal{H}_0|}{|R|}\mathbf{1}\{|R\cap\mathcal{H}_0|>0\}\right)\\
&= \mathbb{E}\left(\frac{|R\cap\mathcal{H}_0|}{|R\cap\mathcal{H}_0|+|R\cap\mathcal{H}_1|}\mathbf{1}\{|R\cap\mathcal{H}_0|>0\}\right)\\
&\le \mathbb{E}\left(\frac{m_0}{m_0+|R\cap\mathcal{H}_1|}\mathbf{1}\{|R\cap\mathcal{H}_0|>0\}\right)\\
&\le \sum_{h\in\mathcal{H}_0}\mathbb{E}\left(\frac{m_0}{m_0+|R_1|}\mathbf{1}\{p_h\le(\alpha m/m_0)(m-|R_1|)^{-1}\}\right),
\end{aligned}$$

where for the first inequality, we used that fact that for each fixed $a \ge 0$, $x \mapsto \frac{x}{x+a}$ is a non-decreasing function on $\mathbb{R}^+\backslash\{0\}$. For the second inequality, we used simultaneously that $|R_1| \le |\mathcal{H}_1 \cap R|$ and the points (i) and (ii) above. Since the function $x \mapsto \frac{m_0}{m_0+x}\frac{m}{m-x}$ is log-convex on $[0, m_1]$ and takes values 1 in $x = 0$ and $x = m_1$, we have pointwise

$$\frac{m_0}{m_0+|R_1|}\frac{m}{m-|R_1|}\le 1\,.$$

Therefore, we get

$$\begin{aligned}
\mathrm{FDR}(R) &\le \frac{1}{m}\sum_{h\in\mathcal{H}_0}\mathbb{E}\left(\frac{\mathbf{1}\{p_h\le(\alpha m/m_0)(m-|R_1|)^{-1}\}}{(m-|R_1|)^{-1}}\right)\\
&\le \frac{1}{m}\sum_{h\in\mathcal{H}_0}\alpha m/m_0=\alpha\,.
\end{aligned}$$

In the last inequality we used Lemma 10.18 with $c = \alpha m/m_0$, $U = p_h$ and $V = (m - |R_1|)^{-1}$, because for any $v > 0$, $D = \{\mathbf{z} \in [0, 1]^{\mathcal{H}_1} \mid (m - |R_1(\mathbf{z})|)^{-1} < v\}$ is a non-decreasing set. ■

Remark 10.22 *1. Benjamini and Liu (1999b) proposed a slightly less conservative step-down procedure: the step-down procedure with the threshold collection*

$$\Delta(i)=1-\left[1-\min\left(1,\frac{\alpha m}{m-i+1}\right)\right]^{1/(m-i+1)}.$$

Benjamini and Liu (1999b) proved that this procedure controls the FDR at level α as soon as the p-values are independent. More recently, a proof of this result is given by Sarkar (2002) when the p-values are MTP_2 (see the definition there) and if the p-values corresponding to true null hypotheses are exchangeable. However, the latter conditions are more restrictive than the PRDS assumption of Theorem 10.21.

2. Remember that if the p-values are PRDS on $\mathcal{H}_0$ (which implies PRDS from $\mathcal{H}_1$ to $\mathcal{H}_0$), the linear step-up (LSU) procedure of Benjamini and Hochberg (1995) controls the FDR at level α (see Theorem 10.13). The procedure of Theorem 10.21 is often more conservative than the LSU procedure. First because the LSU procedure is a step-up procedure and secondly *because the threshold collection of the LSU procedure is most of the time larger. However, in some specific cases (m small and large number of rejections), the threshold collection of Theorem 10.21 can be larger than the one of the LSU procedure: this is the case for instance if* $m = 50$ *and if the LSU procedure rejects more than* 44 *hypotheses.*

3. *Romano and Shaikh (2006a) proved that the result of Theorem 10.21 holds if for each* $h \in \mathcal{H}_0$*, and for all* $u \in [0, 1]$,

$$\mathbb{P}(p_h \le u \mid (p_{h'})_{h' \in \mathcal{H}_1}) \le u\,. \tag{10.10}$$

This condition is slightly weaker than "for each $h \in \mathcal{H}_0$*,* p_h *is independent of* $(p_{h'}, h' \in \mathcal{H}_1)$*". However, when for all* $h \in \mathcal{H}_0$*,* p_h *is exactly distributed like a uniform distribution, the two above conditions are equivalent: to see this, integrate the two sides of inequality* (10.10) *with respect to* $(p_{h'}, h' \in \mathcal{H}_1)$ *and note that both integrated quantities are equal. Therefore,* (10.10) *is an equality a.s. in* u*, and the distribution of* p_h *conditionally to* $(p_{h'}, h' \in \mathcal{H}_1)$ *is uniform.*

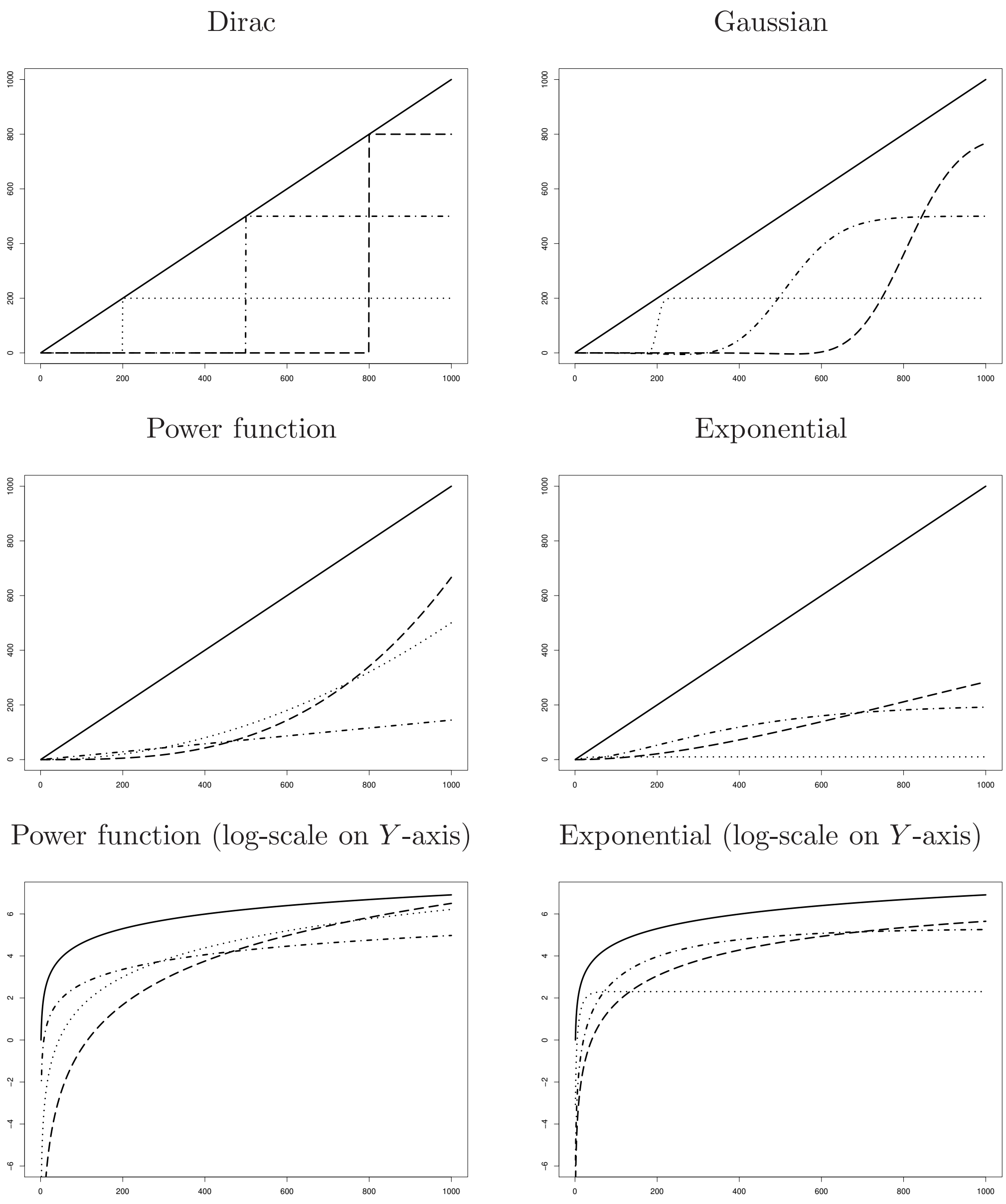


Figure 10.3: For $m = 1000$ hypotheses, this figure shows several shape functions $\beta(\cdot)$ associated to different prior distributions on $\mathbb{R}^+$ (according to expression (10.8), see Example 10.16 for the formulas). The top-left graph corresponds to Dirac distributions ($\lambda = 200$ (dotted), $\lambda = 500$ (dashed-dotted), $\lambda = 800$ (dashed)). The top-right graph correspond to Gaussian distributions ($\mu = 200, \sigma = 10$ (dotted); $\mu = 500, \sigma = 100$ (dashed-dotted); $\mu = 800, \sigma = 100$ (dashed)). The middle-left graph (and bottom-left graph with log-scale) corresponds to distributions with a power function density ($\gamma = 0$ (dotted); $\gamma = -1$ (dashed-dotted); $\gamma = 1$ (dashed)). The middle-right graph (and bottom-right graph with log-scale) corresponds to exponential distributions ($\lambda = 10$ (dotted); $\lambda = 200$ (dashed-dotted); $\lambda = 800$ (dashed)). Finally, we plot in solid line the identity function (which is the shape function of the linear step-up procedure).

Chapter 11

New adaptive step-up procedures that control the FDR under independence and dependence

The proportion π_0 of true null hypotheses is a quantity that often appears explicitly in the FDR control bounds. In order to obtain more powerful procedures, recent research has focussed on finding ways to estimate this quantity and incorporate it in a meaningful way in multiple testing procedures, leading to so-called "adaptive" procedures. We present here new adaptive multiple testing procedures with control of the false discovery rate (FDR) respectively under independence, positive dependencies (PRDS) or unspecified dependencies between the p-values. First, we present a new "one-stage" adaptive procedure and a new "two-stage" adaptive procedure that control the FDR in the independent context. Up to some marginal cases, the latter "two-stage" procedure is less conservative than a recent adaptive procedure proposed by Benjamini *et al.* (2006). Second, we propose adaptive versions of the linear step-up procedures of Benjamini and Hochberg (1995) and of the step-up procedures of Blanchard and Fleuret (2007), that control the FDR under positive dependencies and unspecified dependencies respectively. The latter adaptive procedures are not uniformly better than the non-adaptive ones, but we show that they can significantly outperform the latter when the number of rejected hypotheses is large.

11.1 Introduction

In this work, we focus on building procedures that control the false discovery rate (FDR), which is defined as the expected proportion of rejected true null hypotheses among all the rejected null hypotheses. Benjamini and Hochberg (1995) proposed a powerful procedure, called *the linear step-up* (LSU) procedure, that controls the FDR under independence between the p-values. Later, Benjamini and Yekutieli (2001) proved that the LSU procedure still controls the FDR when the p-values have positive dependencies (more precisely a specific form of positive dependency called PRDS). Under unspecified dependencies, the same authors have shown that the FDR control still holds if the threshold collection of the LSU procedure is divided by a factor $1 + 1/2 + \cdots + 1/m$, where m is the total number of null hypotheses to test. More recently, the latter result has been generalized by Blanchard and Fleuret (2007), by showing that there is a family of step-up procedures (depending on the choice of a prior distribution) that still control the FDR under unspecified dependencies between the p-values.

All these procedures, which are built in order to control the FDR at a level α, do finally have a FDR smaller than $\pi_0\alpha$, where π_0 is the proportion of true null hypotheses. Therefore, when most of the hypotheses are false, these procedures are inevitably conservative. Therefore, the challenge of *adaptive control* of the FDR (see *e.g.* Benjamini and Hochberg (2000) and Black (2004)) is to integrate an estimation of the unknown proportion π_0 in the threshold of the previous procedures and to prove that the FDR is still rigorously controlled by α.

Recently, under independence, Benjamini *et al.* (2006) have shown that the Storey estimator (proposed by Storey (2002)) can be used to build an adaptive procedure that controls the FDR. They also give a new adaptive procedure (denoted here by "BKY06") that controls the FDR under independence and that seems robust to positive correlations. This adaptive procedure is said "two-stage" because it consists of two different steps: 1. Estimate π_0. 2. Use this estimate in a new threshold to build a new multiple testing procedure.

In this chapter, we present:

1. A simple step-up procedure more powerful in general than the LSU procedure that controls the FDR under independence. This procedure is said "one-stage" adaptive.

2. A new two-stage adaptive procedure more powerful in general than the "BKY06" procedure that controls the FDR under independence and that seems robust to positive correlations on simulations (using the above adaptive one-stage procedure as a first step).

3. A new two-stage adaptive version of the LSU procedure that control the FDR under positive dependencies (PRDS), resulting in an improvement of the power in "a certain regime".

4. New two-stage adaptive versions of all the procedures of Blanchard and Fleuret (2007) that control the FDR under unspecified dependencies, resulting in an improvement of the power in "a certain regime".

In the two last points "a certain regime" means that the number of rejected null hypotheses has to be large (typically more than 60%) in order to expect an improvement over the standalone non-adaptive procedures.

In this work, the results are proved using the probabilistic lemmas of Section 10.6. Similarly to Chapter 10, this provides synthetic proofs of FDR controls.

This chapter is organized as follows: in Section 11.2, we present the existing non-adaptive results in FDR control. Section 11.3 states the existing and new adaptive results in the independence context, and compares them in a simulations study. The case where the p-values have positive dependencies or unspecified dependencies is examined in Section 11.4. The proofs of the new results are given in Section 11.6.

11.2 Some existing non-adaptive step-up procedures that control the FDR

We consider the multiple testing framework of Section 9.2.2 (Chapter 9) where it is given a set of *p-values* $\mathbf{p} = (p_h, h \in \mathcal{H})$ for a set of null hypotheses $\mathcal{H}$. Remember that, for a multiple testing procedure R, the *false discovery rate*, is defined as the average proportion of true null hypotheses in the set of all the rejected hypotheses:

$$\mathrm{FDR}(R) = \mathbb{E}\left(\frac{|R \cap \mathcal{H}_0|}{|R|}\mathbf{1}\{|R| > 0\}\right).$$

Let us order the p-values $p_{(1)} \leq \cdots \leq p_{(m)}$ and put $p_{(0)} = 0$.

Definition 11.1 (Step-up procedure) *Let be $\alpha \in (0,1)$ and a shape function $\beta : \mathbb{R}^+ \mapsto \mathbb{R}^+$, that is, a non-decreasing function. The step-up procedure of shape function β (and at level α) is defined as*

$$R_\beta := \{h \in \mathcal{H} \mid p_h \leq p_{(k)}\}, \quad \textit{where} \;\; k = \max\{i \mid p_{(i)} \leq \alpha\beta(i)/m\}.$$

The function $\alpha\beta(\cdot)/m$ is called the threshold collection of the procedure. In the particular case where the shape function β is the identity function on $\mathbb{R}^+$, the procedure is called the linear step-up procedure (at level α).

Remark 11.2 *In our setting, the "linear step-up procedure" should rather be called the "identity step-up procedure". However, we choose here to keep the usual name.*

When the p-values are independent, the following theorem holds (the first part was proved by Benjamini and Hochberg (1995) whereas the second part was proved by Finner and Roters (2001)):

Theorem 11.3 *Suppose that the p-values of $\mathbf{p} = (p_h, h \in \mathcal{H})$ are independent. Then the linear step-up procedure has a FDR less than or equal to $\pi_0\alpha$, where $\pi_0 = m_0/m$ is the proportion of true null hypotheses. Moreover, if the p-values associated to true null hypotheses are exactly distributed like a uniform distribution, the linear step-up procedure has a FDR equal to $\pi_0\alpha$.*

Benjamini and Yekutieli (2001) extended the previous FDR control to the PRDS case (see Definition 10.8 of Chapter 10).

Theorem 11.4 *Suppose that the p-values of $\mathbf{p} = (p_h, h \in \mathcal{H})$ are PRDS on $\mathcal{H}_0$. Then the linear step-up procedure has a FDR less than or equal to $\pi_0\alpha$.*

When no particular assumptions are made on the dependencies between the p-values, Blanchard and Fleuret (2007) proved (extending a result of Benjamini and Yekutieli (2001)) that there is a class of step-up procedures that control the FDR:

Theorem 11.5 *Under unspecified dependencies between the p-values of $\mathbf{p} = (p_h, h \in \mathcal{H})$, consider β a shape function of the form:*

$$\beta(r) = \int_0^r u d\nu(u), \tag{11.1}$$

where ν is some probability distribution on $(0, \infty)$. Then the step-up procedure R_β has a FDR less than or equal to $\alpha\pi_0$.

For proofs of Theorems 11.3, 11.4 and 11.5, we refer the reader to Chapter 10. A direct corollary of these theorems is that the step-up procedure $R_{\beta^\star}$ with $\beta^\star = \beta/\pi_0$ has a FDR less than or equal to α in either of the following situations:

- $\beta(i) = i$ when the p-values are independent or PRDS,

- the shape function β is of the form (11.1) when the p-values have unspecified dependencies.

Moreover, since $\pi_0 \leq 1$, the procedure $R_{\beta^\star}$ is always less conservative than R_β (especially when π_0 is small). However, since π_0 is unknown, the procedure $R_{\beta^\star}$ cannot be only derived from the observations. Therefore, the procedure $R_{\beta^\star}$ is called the *Oracle step-up procedure* of shape function β (and at level α).

The role of the adaptive step-up procedures is to mimic the latter Oracle. They are defined as $R_{\beta G}$, where G is an estimator of π_0^{-1}.

Definition 11.6 (Two-stage adaptive step-up procedure) *Given a level $\alpha \in (0, 1)$, a shape function β and a measurable function $G : [0, 1]^{\mathcal{H}} \to (0, \infty)$. The (two-stage) adaptive step-up procedure of shape function β and using estimator G (at level α), is defined as*

$$R_{\beta G} = \{h \in \mathcal{H} \mid p_h \leq p_{(k)}\}, \quad \textit{where} \;\; k = \max\{i \mid p_{(i)} \leq \alpha\beta(i)G(\mathbf{p})/m\}.$$

The (data-dependent) function $\Delta(i) = \alpha\beta(i)G(\mathbf{p})/m$ is called the threshold collection of the adaptive procedure. In the particular case where the shape function β is the identity function on $\mathbb{R}^+$, the procedure is called the adaptive linear step-up procedure using estimator G (and at level α).

Following the previous definition, an adaptive procedure is composed of two different steps:

1. Estimate π_0^{-1} with an estimator G.

2. Take the step-up procedure of shape function βG.

The main theoretical task is to ensure that an adaptive procedure of this type still correctly controls the FDR. The mathematical difficulty obviously comes from the additional variations of the estimator G in the procedure.

11.3 Adaptive step-up procedures that control the FDR under independence

We suppose in this section that the p-values of $(p_h, h \in \mathcal{H})$ are independent. We introduce the following notations: for each $h \in \mathcal{H}$, we denote by $\mathbf{p}_{-h}$ the collection of p-values $(p_{h'}, h' \neq h)$ and by $\mathbf{p}_{0,h} = (\mathbf{p}_{-h}, 0)$ the collection $\mathbf{p}$ where p_h has been replaced by 0.

11.3.1 General theorem and some previously known procedures

The theorem is strongly inspired from techniques developed by Benjamini *et al.* (2006). It gives general conditions on the estimator to provide the FDR control of the corresponding adaptive procedure.

Theorem 11.7 *Suppose that the p-values of* $\mathbf{p} = (p_h, h \in \mathcal{H})$ *are independent and consider a measurable function* $G : [0,1]^{\mathcal{H}} \to (0,\infty)$ *coordinate-wise non-increasing, such that for each* $h \in \mathcal{H}_0$,

$$\mathbb{E}G(\mathbf{p}_{0,h}) \leq \pi_0^{-1}. \tag{11.2}$$

Then, the adaptive linear step-up procedure R *of threshold collection* $\Delta(i) = \alpha i G(\mathbf{p})/m$ *has a FDR less than or equal to* α.

Remark 11.8 *If* $G(\cdot)$ *is moreover supposed coordinate-wise left-continuous, we can prove that Theorem 11.7 still holds when the condition (11.2) is replaced by the slightly weaker condition:*

$$\mathbb{E}G(\widetilde{\mathbf{p}}_h) \leq \pi_0^{-1}, \tag{11.3}$$

where for each $h \in \mathcal{H}_0$, $\widetilde{\mathbf{p}}_h = (\mathbf{p}_{-h}, \widetilde{p}_h(\mathbf{p}_{-h}))$ *is the collection of p-values* $\mathbf{p}$ *where* p_h *has been replaced by* $\widetilde{p}_h(\mathbf{p}_{-h}) = \max\left\{p \in [0,1] \;\middle|\; p \leq \alpha\pi(h)|R(\mathbf{p}_{-h}, p)|G(\mathbf{p}_{-h}, p)\right\}$.

Following Benjamini *et al.* (2006), we can propose the following choices for G:

Corollary 11.9 (Essentially proved by Benjamini *et al.* (2006)) *Assume that the p-values of* $\mathbf{p} = (p_h, h \in \mathcal{H})$ *are independent. The adaptive linear step-up procedure at level* α *has a FDR less than or equal to* α *for one of the following choices for the estimator* G:

- $G_1(\mathbf{p}) = \frac{(1-\lambda)m}{\sum_{h\in\mathcal{H}} \mathbf{1}\{p_h > \lambda\} + 1}$, $\lambda \in [0,1[$.
- $G_2(\mathbf{p}) = \frac{1}{1+\alpha}\frac{m}{m - |R_0(\mathbf{p})| + 1}$, *where* R_0 *is the (non adaptive) linear step-up procedure at level* $\alpha/(1+\alpha)$.

Remark 11.10 *More precisely, the result proved by Benjamini* et al. *(2006) use a slightly better version of* G_2 *without the "+1" in the denominator (this could be derived here from Remark 11.8). We forget about this refinement here, noting that it results only in a very slight improvement.*

Remark 11.11 *The estimator* $\frac{\sum_{h\in\mathcal{H}} \mathbf{1}\{p_h > \lambda\} + 1}{(1-\lambda)m}$ *of* π_0 *is called the modified Storey's estimator and was initially introduced by Storey (2002) and Storey* et al. *(2004) (initially without the "+1" in the numerator, hence the name "modified"). Note that* G_1 *is not necessarily larger than* 1.

11.3.2 New adaptive one-stage step-up procedure

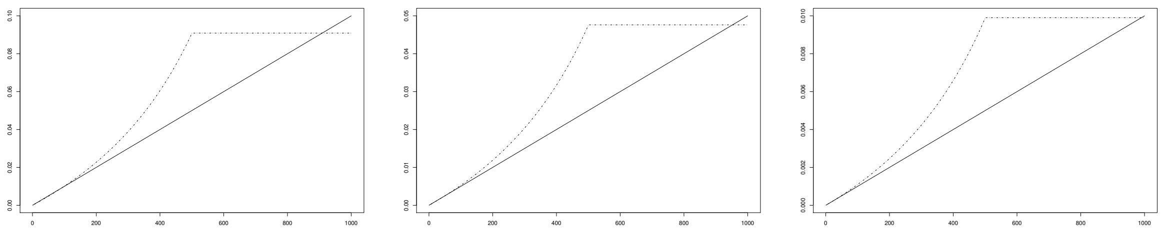
We now introduce our main first contribution, a "one-stage" adaptive procedure. This means that estimation step is directly included in the shape function β , and so does fall in the framework of Definition 11.1 (and not in the one of Definition 11.6).

Theorem 11.12 *Suppose that the p-values of* $\mathbf{p} = (p_h, h \in \mathcal{H})$ *are independent. The step-up procedure with the threshold collection*

$$\Delta(i) = \frac{\alpha}{1+\alpha} \min\left(\frac{i}{m-i+1}, 1\right),$$

has a FDR less than or equal to α.

Remark 11.13 *As Figure 11.1 illustrates, the procedure of Theorem 11.12 is generally less conservative than the (non-adaptive) linear step-up procedure (LSU). Precisely, the new procedure can be more conservative than the LSU procedure only in the marginal cases where the proportion of null hypotheses rejected by the LSU procedure is more than* $(1+\alpha)^{-1}$ *, or less than* $1/m + \alpha/(1+\alpha)$ *.*



$\alpha = 0.1$ $\alpha = 0.05$ $\alpha = 0.01$

Figure 11.1: For $m = 1000$ null hypotheses. These graphs represent the threshold collection of the new adaptive one-stage procedure of Theorem 11.12 (dashed-dotted line) and the threshold collection of the linear procedure $\Delta(i) = \alpha i/m$ (solid line). The left (resp. center, right) graph represents the case $\alpha = 0.1$ (resp. $\alpha = 0.05$, $\alpha = 0.01$).

11.3.3 New adaptive two-stage procedure

We can now use the previous one-stage procedure to estimate π_0^{-1} and build a two-stage procedure exactly following the philosophy that led to propose G_2. That is, we can use the same function G_2 as proposed earlier, except that we replace the first step using the standard step-up linear procedure by the adaptive procedure of Theorem 11.12. We obtain the following result:

Theorem 11.14 *Assume that the p-values of* $\mathbf{p} = (p_h, h \in \mathcal{H})$ *are independent and note* R_0' *the new one-stage adaptive procedure of Theorem 11.12. Then the adaptive linear step-up procedure with the threshold collection* $\Delta(i) = \alpha \frac{i}{m} G_3(\mathbf{p})$*, where*

$$G_3(\mathbf{p}) = \frac{1}{1+\alpha} \frac{m}{m - |R_0'| + 1},$$

has a FDR less than or equal to α *.*

11.3.4 Simulation study

How can we compare the different adaptive procedures defined above? Choosing $\lambda = \alpha/(1+\alpha)$, we have pointwise $G_1 \geq G_3 \geq G_2$ which shows that the adaptive procedure obtained using G_1 is always less conservative than the one derived from G_3, itself less conservative that the one using G_2 (except in the marginal cases where the one-stage adaptive procedure is more conservative than the standard step-up procedure, delineated earlier). It would therefore appear that one should always choose G_1 and disregard the other ones. Nevertheless, a point made by Benjamini *et al.* (2006) for introducing G_2 as a better alternative to the (already known earlier) G_1 was that on simulations with positively dependent test statistics, the FDR of the adaptive procedure using G_1 with $\lambda = 1/2$ resulted in very bad control of the FDR, which was not the case for G_2. While the positively dependent case is not covered by the theory, it is important to ensure that a multiple testing procedure is sufficiently robust in practice so that the FDR does not vary too much in this situation.

Therefore, in order to assess the quality of our new procedures, we here propose to evaluate the different methods on a simulation study following the setting used by Benjamini *et al.* (2006): Let $X_i = \mu_i + \varepsilon_i$, for $i, 1 \leq i \leq m$, where ε is a $\mathbb{R}^m$-valued centred Gaussian random vector such that $\mathbb{E}(\varepsilon_i^2) = 1$ and for $i \neq j$, $\mathbb{E}(\varepsilon_i \varepsilon_j) = \rho$, where $\rho \in [0,1]$ is a correlation parameter. Consequently, when $\rho = 0$ the X_i's are independent whereas when $\rho > 0$ the X_i's are positively correlated (with a constant correlation). For instance, the ε_i's can be constructed by taking $\varepsilon_i := \sqrt{\rho} U + \sqrt{1-\rho} Z_i$, where Z_i, $1 \leq i \leq m$ and U are all i.i.d $\sim \mathcal{N}(0,1)$.

Considering the one-sided null hypotheses h_i : "$\mu_i \leq 0$" against the alternatives "$\mu_i > 0$" for $1 \leq i \leq m$, we define the p-values $p_i = \overline{\Phi}(X_i)$, for $1 \leq i \leq m$, where $\overline{\Phi}$ is the standard Gaussian distribution tail. For $i, 1 \leq i \leq m_0$, $\mu_i = 0$ and for $i, m_0 + 1 \leq i \leq m$, $\mu_i = 3$, providing that the p-values corresponding to the null mean follow exactly a uniform distribution.

We perform the following step-up multiple testing procedures:

- [*LSU*] the (non-adaptive) linear procedure as defined in Definition 11.1 i.e. with the threshold collection $\Delta(i) = \alpha i/m$.

- [*LSU Oracle*] the procedure with the threshold collection $\Delta(i) = \alpha i/m_0$.

- [*Storey-λ*] the two-stage procedures corresponding to G_1 in Corollary 11.9. A classical choice for λ is 1/2. We try here also $\lambda = \alpha/(1+\alpha)$.

- [*BKY06*] The two-stage procedure corresponding to G_2 in Corollary 11.9.

- [*BR07-1S*] The new one-stage adaptive procedure of Theorem 11.12.

- [*BR07-2S*] The new two-stage adaptive procedure of Theorem 11.14.

Under independence ($\rho = 0$)

Remember that under independence, the *LSU* procedure has a FDR equal to $\alpha\pi_0$ and that the *LSU Oracle* procedure has a FDR equal to α (provided that $\alpha \leq \pi_0$). The other procedures have their FDR bounded by α. We can then define the *relative power of a procedure* as the mean of the number of true rejections of the procedure divided by the number of true rejections of the *LSU Oracle* procedure.

Figure 11.2 (page 180) represents the FDR and the relative power of these procedures in function of the proportion of true null hypotheses π_0 (estimated with 10 000 simulations). This experiment show that the procedures can be ordered in terms of (relative) power :

$$\textit{Storey-}1/2 \gg \textit{Storey-}\alpha/(1+\alpha) \gg \textit{BR07-2S} \gg \textit{BKY06},$$

the symbol "$\gg$" meaning "is (π_0-uniformly) more powerful than". The procedure *BR07-1S* is "between" *BKY06* and *BR07-2S*. We see here that the choice $\lambda = 1/2$ seems to be better than $\lambda = \alpha/(1+\alpha)$. However, the following remark gives drawbacks for the procedure *Storey*-1/2.

Remark 11.15 *Since the estimation of π_0 in Storey-$1/2$ is made using few p-values (the p-values larger than $1/2$), this procedure is very sensitive to small variations:*

- *Under independence, when we consider a least favorable case where μ_i takes negative values for some $i, 1 \le i \le m_0$, the procedure Storey-$1/2$ can be too conservative.*
- *As noticed by Benjamini* et al. *(2006) and as we will see in the next paragraph, when the p-values have positive correlations ($\rho > 0$), the procedure Storey-$1/2$ does not control the FDR anymore.*

Under positive dependencies ($\rho > 0$)

Under positive dependencies, the FDR of the procedure *LSU* (resp. *LSU Oracle*) is still bounded by $\alpha\pi_0$ (resp. α), but without equality. We do not know if the other procedures have a FDR smaller than α, so that they cannot be compared in terms of power.

Figure 11.3 (page 181) shows that the FDR control is no more provided for the procedure *Storey*-1/2. The maximum FDR for *BR07-2S* is smaller than the one of *Storey*-$\alpha/(1+\alpha)$. Thus our new two-stage procedure seems more robust to positive correlations than *Storey*-$\alpha/(1+\alpha)$ (for $\rho = 0.5$, the maximum FDR for *BR07-2S* is 0.0508 whereas the one of *Storey*-$\alpha/(1+\alpha)$ is 0.0539). An explanation is that the procedure *BR07-2S* is more conservative than *Storey*-$\alpha/(1+\alpha)$ in its estimation of π_0. When the p-value are very positively correlated $\rho = 0.9$, both procedures control the FDR. A reason is that both procedures are based on the linear step-up procedure which is very conservative in this case.

11.4 New adaptive step-up procedures that control the FDR under dependence

When the p-values may have some dependencies, we here propose to use Markov's inequality to estimate π_0^{-1}. Since Markov's inequality is general but not extremely precise, the resulting procedures are obviously quite conservative and are arguably of a limited practical interest. However, we will show that they still provide an improvement, in a certain regime, with respect to (non-adaptive) LSU procedure in the PRDS case and with respect to the family of (non-adaptive) procedures proposed in Theorem 11.5 when the p-values have unspecified dependencies.

For a fixed constant $\kappa \ge 2$, define the following function: for $x \in [0,1]$,

$$F_\kappa(x) = \begin{cases} 1 & \text{if } x \le \kappa^{-1} \\ \frac{2\kappa^{-1}}{1-\sqrt{1-4(1-x)\kappa^{-1}}} & \text{otherwise} \end{cases} . \tag{11.4}$$

We can prove the following general theorem:

Theorem 11.16 *Consider a shape function β and fix α_0 and α_1 in $(0,1)$ such that $\alpha_0 \leq \alpha_1$. Denote by R_0 the step-up procedure with threshold collection $\alpha_0 c/m$ and by R the adaptive step-up procedure with threshold collection $\alpha_1\beta(\cdot)F_\kappa(|R_0|/m)/m$. Suppose moreover that $FDR(R_0) \leq \alpha_0\pi_0$ and that for each $h \in \mathcal{H}_0$ and any constant $c > 0$,*

$$\mathbb{E}\left(\frac{\mathbf{1}\{p_h \leq c\beta(|R|)\}}{|R|}\mathbf{1}\{|R| > 0\}\right) \leq c\,. \tag{11.5}$$

Then R has a FDR less than or equal to $\alpha_1 + \kappa\alpha_0$.

Combining Theorem 11.16 with Theorems 11.4 and 11.5, we obtain the following corollary:

Corollary 11.17 *Consider a shape function β and fix α_0 and α_1 in $(0,1)$ such that $\alpha_0 \leq \alpha_1$. Denote by R_0 the step-up procedure with threshold collection $\alpha_0\beta(\cdot)/m$. Then the adaptive step-up procedure R with threshold collection $\alpha_1\beta(\cdot)F_\kappa(|R_0|/m)/m$ has a FDR less than or equal to $\alpha_1 + \kappa\alpha_0$ in either of the following dependence situations:*

- *the p-values $(p_h, h \in \mathcal{H})$ are PRDS on $\mathcal{H}_0$ and the shape function is the identity function.*
- *the p-values have unspecified dependencies and β is a shape function of the form (11.1).*

Remark 11.18 *If we choose $\kappa = 2$, $\alpha_0 = \alpha/4$ and $\alpha_1 = \alpha/2$, the adaptive procedure R defined in Corollary 11.17 (with either $\beta(i) = i$ in the PRDS case or β of the form (11.1) when the p-values have unspecified dependencies) has a FDR less than or equal to α. In this case, we note that R is less conservative than the non-adaptive step-up procedure with threshold collection $\alpha\beta(\cdot)/m$ if $F_2(|R_0|\,/\,|\mathcal{H}|) \geq 2$ or equivalently when R_0 rejects more than $F_2^{-1}(2) = 62,5\%$ of the null hypotheses. Conversely, R is more conservative otherwise, and we can lose up to a factor 2 in the threshold collection with respect to the standard one-stage version. Therefore, this adaptive procedure is only useful in the cases where it is expected that a "large" proportion of null hypotheses can easily be rejected.*
In particular, when we use Corollary 11.17 under general dependence, it is relevant to choose the shape function β from a prior distribution ν concentrated on the large numbers of $\{1,\dots,m\}$.

In the PRDS case, the procedure R of Corollary 11.17 with $\kappa = 2$, $\alpha_0 = \alpha/4$ and $\alpha_1 = \alpha/2$, is the adaptive linear step-up procedure at level $\alpha/2$ with the estimator

$$\frac{1}{1 - \sqrt{(2|R_0|/m - 1)_+}}\,,$$

where $|R_0|$ is the number of rejections of the LSU procedure at level $\alpha/4$ and $(\cdot)_+$ denotes the positive part. This procedure was performed in the simulation setting of Section 11.3.4 with $\rho = 0.1$, $m_0 = 5$ and $m = 100$ (see Figure 11.4). The common value μ of the positive means is taken in the range $[2,5]$, so that large values of μ correspond to large rejection cases. We can notice that there is a regime where the adaptive procedure outperforms the regular one.

11.5 Conclusion

We proposed several adaptive multiple testing procedures that control the FDR. First, we introduced the procedures *BR07-1S* and *BR07-2S* and we proved that they have theoretical validity when the p-values are independent. The procedure *BR07-2S* is less conservative in general than the adaptive procedure proposed by Benjamini *et al.* (2006). Moreover, these new procedures have the advantage of appearing to be robustly controlling the FDR even in a positive dependence situation as shown in the simulations. This is an advantage with respect to the Storey procedure, which is less conservative but less robust. Second, we presented adaptive multiple testing procedures when the p-values are PRDS and when they have unspecified dependencies. Although their interest is mainly theoretical, it shows in principle that adaptivity can improve performance in a theoretically rigorous way even without the independence assumption.

11.6 Proofs of the results

Proof of Theorem 11.7. Denoting R the procedure of Theorem 11.7 and using Definition 11.1, R satisfies the following "self-consistency condition":

$$R \subset \{h \in \mathcal{H} \mid p_h \leq \alpha |R| G(\mathbf{p})/m\}. \tag{11.6}$$

Therefore,

$$\mathrm{FDR}(R) = \mathbb{E}\left(\frac{|R \cap \mathcal{H}_0|}{|R|}\mathbf{1}\{|R|>0\}\right) \leq \sum_{h\in\mathcal{H}_0} \mathbb{E}\left(\frac{\mathbf{1}\{p_h \leq \alpha |R(\mathbf{p})| G(\mathbf{p})/m\}}{|R(\mathbf{p})|}\right).$$

Since G is non-increasing, we get:

$$\begin{aligned}
\mathrm{FDR}(R) &\leq \sum_{h\in\mathcal{H}_0} \mathbb{E}\left(\frac{\mathbf{1}\{p_h \leq \alpha |R(\mathbf{p})| G(\mathbf{p}_{0,h})/m\}}{|R(\mathbf{p})|}\right) \\
&= \sum_{h\in\mathcal{H}_0} \mathbb{E}\left(\mathbb{E}\left(\frac{\mathbf{1}\{p_h \leq \alpha |R(\mathbf{p})| G(\mathbf{p}_{0,h})/m\}}{|R(\mathbf{p})|}\middle| \mathbf{p}_{-h}\right)\right) \leq \frac{\alpha}{m}\sum_{h\in\mathcal{H}_0} \mathbb{E}G(\mathbf{p}_{0,h}).
\end{aligned}$$

The last step is obtained with Lemma 10.17 of Chapter 10 with $U = p_h$, $g(U) = |R(\mathbf{p}_{-h}, U)|$ and $c = \alpha G(\mathbf{p}_{0,h})/m$, because the distribution of p_h conditionnally to $\mathbf{p}_{-h}$ is stochastically lower bounded by a uniform distribution, $|R|$ is coordinate-wise non-increasing and $\mathbf{p}_{0,h}$ depends only on the p-values of $\mathbf{p}_{-h}$. We apply then (11.2) to conclude. ■

Proof of Corollary 11.9. By Theorem 11.7, it is sufficient to prove that the condition (11.2) holds for G_1 and G_2. The bound for G_1 is obtained using Lemma 11.19 (see below) with $k = m_0$ and $q = 1 - \lambda$: for all $h \in \mathcal{H}_0$,

$$\mathbb{E}[G_1(\mathbf{p}_{0,h})] \leq m(1-\lambda)\mathbb{E}\left[\sum_{h'\in\mathcal{H}_0\backslash\{h\}} \mathbf{1}\{p_{h'} > \lambda\} + 1\right]^{-1} \leq \pi_0^{-1}.$$

The proof for G_2 is deduced from the one of G_1 with $\lambda = \alpha/(1+\alpha)$ because in this case $G_2 \leq G_1$ pointwise. ■

Proof of Theorem 11.12. We denote by R the corresponding procedure. Using Definition 11.1, R satisfies the "self-consistency condition" $R \subset \left\{h \in \mathcal{H} \mid p_h \leq \frac{\alpha}{1+\alpha} \min\left(\frac{|R|}{m-|R|+1}, 1\right)\right\}$. Therefore, we have

$$\begin{aligned}
\mathrm{FDR}(R) &\leq \sum_{h \in \mathcal{H}_0} \mathbb{E}\left(\frac{\mathbf{1}\{p_h \leq \frac{\alpha}{1+\alpha}\frac{|R(\mathbf{p})|}{m-|R(\mathbf{p})|+1}\}}{|R(\mathbf{p})|}\right) \\
&\leq \sum_{h \in \mathcal{H}_0} \mathbb{E}\left(\frac{\mathbf{1}\{p_h \leq \frac{\alpha}{1+\alpha}\frac{|R(\mathbf{p})|}{m-|R(\mathbf{p}_{0,h})|+1}\}}{|R(\mathbf{p})|}\right) \\
&= \sum_{h \in \mathcal{H}_0} \mathbb{E}\left(\mathbb{E}\left(\frac{\mathbf{1}\{p_h \leq \frac{\alpha}{1+\alpha}\frac{|R(\mathbf{p})|}{m-|R(\mathbf{p}_{0,h})|+1}\}}{|R(\mathbf{p})|}\middle|\mathbf{p}_{-h}\right)\right) \\
&\leq \frac{\alpha}{1+\alpha}\sum_{h \in \mathcal{H}_0} \mathbb{E}\left(\frac{1}{m-|R(\mathbf{p}_{0,h})|+1}\right),
\end{aligned}$$

The last step is obtained with Lemma 10.17 of Chapter 10 with $U = p_h$, $g(U) = |R(\mathbf{p}_{-h}, U)|$ and $c = \frac{\alpha}{1+\alpha}\frac{1}{m-|R(\mathbf{p}_{0,h})|+1}$, because the distribution of p_h conditionnally to $\mathbf{p}_{-h}$ is stochastically lower bounded by a uniform distribution and because $\mathbf{p}_{0,h}$ depends only on the p-values of $\mathbf{p}_{-h}$. Finally, since the threshold collection of R is less than or equal to $\alpha/(1+\alpha)$, we get

$$\frac{1}{1+\alpha}\mathbb{E}\left(m/(m-|R(\mathbf{p}_{0,h})|+1)\right) \leq \mathbb{E}G_1(\mathbf{p}_{0,h}),$$

where G_1 is the Storey estimator with $\lambda = \alpha/(1+\alpha)$. We then use $\mathbb{E}G_1(\mathbf{p}_{0,h}) \leq \pi_0^{-1}$ (see proof of Corollary 11.9) to conclude. ■

Proof of Theorem 11.14. From Theorem 11.7, it is sufficient to prove that $\mathbb{E}G_3(\mathbf{p}_{0,h}) \leq \pi_0^{-1}$, and this is the case because the procedure R_0' has a threshold collection less than or equal to $\alpha/(1+\alpha)$. ■

Proof of Theorem 11.16. Assume $\pi_0 > 0$ (otherwise the result is trivial). Note first that R satisfies the "self-consistency condition" $R \subset \{h \in \mathcal{H} \mid p_h \leq \alpha_1\beta(|R|)F_\kappa(|R_0|/m)/m\}$. Let us decompose the final output of the two-stage procedure R in the following way:

$$\begin{aligned}
\mathrm{FDR}(R) &\leq \sum_{h \in \mathcal{H}_0} \mathbb{E}\left[\frac{\mathbf{1}\{p_h \leq \beta(|R|)\alpha_1/m_0\}}{|R|}\mathbf{1}\{|R| > 0\}\right] \\
&+ \sum_{h \in \mathcal{H}_0} \mathbb{E}\left[\frac{\mathbf{1}\{\alpha_1\beta(|R|)/m_0 < p_h \leq \alpha_1\beta(|R|)F_\kappa(|R_0|/m)/m\}}{|R|}\mathbf{1}\{|R| > 0\}\right] \\
&\leq \alpha_1 + m_0\mathbb{E}\left[\frac{\mathbf{1}\{F_\kappa(|R_0|/m) > \pi_0^{-1}\}}{|R_0|}\right]
\end{aligned}$$

For the last inequality, we have used (11.5) with $c = \alpha_1/m_0$ for the first term. For the second term, we have used the two following facts:

(i) $F_\kappa(|R_0|/m) > \pi_0^{-1}$ implies $|R_0| > 0$,

(ii) because of the assumption $\alpha_0 \leq \alpha_1$ and $F_\kappa \geq 1$, the output of the second step is necessarily a set containing at least the output of the first step. Hence $|R| \geq |R_0|$.

Let us now concentrate on further bounding this second term. For this, first consider the generalized inverse of F_κ, $F_\kappa^{-1}(t) = \inf\{x \mid F_\kappa(x) > t\}$. Since F_κ is a non-decreasing left-continuous function, we have $F_\kappa(x) > t \Leftrightarrow x > F_\kappa^{-1}(t)$. Furthermore, the expression of F_κ^{-1} is given by: $\forall t \in [1, +\infty), F_\kappa^{-1}(t) = \kappa^{-1}t^{-2} - t^{-1} + 1$ (providing in particular that $F_\kappa^{-1}(\pi_0^{-1}) > 1 - \pi_0$). Hence

$$\begin{aligned} m_0\mathbb{E}\left[\frac{\mathbf{1}\{F_\kappa(|R_0|/m) > \pi_0^{-1}\}}{|R_0|}\right] &\leq m_0\mathbb{E}\left[\frac{\mathbf{1}\{|R_0|/m > F_\kappa^{-1}(\pi_0^{-1})\}}{|R_0|}\right] \\ &\leq \frac{\pi_0}{F_\kappa^{-1}(\pi_0^{-1})}\mathbb{P}\left[|R_0|/m \geq F_\kappa^{-1}(\pi_0^{-1})\right] . \end{aligned} \tag{11.7}$$

Now, by assumption, the FDR of the first step R_0 is controlled at level $\pi_0\alpha_0$, so that

$$\begin{aligned} \pi_0\alpha_0 &\geq \mathbb{E}\left[\frac{|R_0 \cap \mathcal{H}_0|}{|R_0|}\mathbf{1}\{|R_0| > 0\}\right] \\ &\geq \mathbb{E}\left[\frac{|R_0| + m_0 - m}{|R_0|}\mathbf{1}\{|R_0| > 0\}\right] \\ &= \mathbb{E}\left[[1 + (\pi_0 - 1)Z^{-1}]\mathbf{1}\{Z > 0\}\right] , \end{aligned}$$

where we denoted by Z the random variable $|R_0|/m$. Hence by Markov's inequality, for all $t > 1 - \pi_0$,

$$\mathbb{P}\left[Z \geq t\right] \leq \mathbb{P}\bigg([1 + (\pi_0 - 1)Z^{-1}]\mathbf{1}\{Z > 0\} \geq 1 + (\pi_0 - 1)t^{-1}\bigg) \leq \frac{\pi_0\alpha_0}{1 + (\pi_0 - 1)t^{-1}} ;$$

choosing $t = F_\kappa^{-1}(\pi_0^{-1})$ and using this into (11.7), we obtain

$$m_0\mathbb{E}\left[\frac{\mathbf{1}\{F_\kappa(|R_0|/m) > \pi_0^{-1}\}}{|R_0|}\right] \leq \alpha_0\frac{\pi_0^2}{F_\kappa^{-1}(\pi_0^{-1}) - 1 + \pi_0} .$$

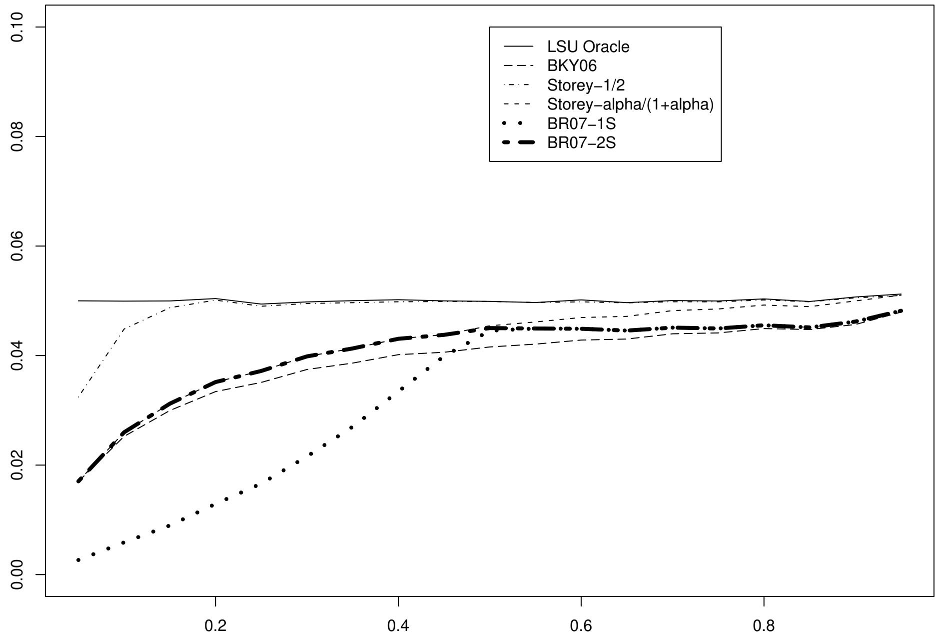
If we want this last quantity to be less than $\kappa\alpha_0$, this yields the condition $F_\kappa^{-1}(\pi_0^{-1}) \geq \kappa^{-1}\pi_0^2 - \pi_0 + 1$, and this is true from the expression of F_κ^{-1} (note that this is how the formula for F_κ was determined in the first place). ■

Proof of Corollary 11.17. We just have to prove that (11.5) is true for any fixed $h \in \mathcal{H}_0$. When the p-values have unspecified dependencies, this is a direct consequence of Lemma 10.19 of Chapter 10 with $U = p_h$ and $V = \beta(|R|)$. For the PRDS case, we note that since $|R(\mathbf{p})|$ is coordinate-wise non-increasing in each p-value, for any $v > 0$, $\{\mathbf{z} \in [0,1]^{\mathcal{H}} \mid |R(\mathbf{z})| < v\}$ is a non-decreasing set, so that the PRDS property implies that $u \mapsto \mathbb{P}(|R| < v \mid p_h \leq u)$ is non-decreasing. We can then apply Lemma 10.18 of Chapter 10 with $U = p_h$ and $V = |R|$. ■

The following lemma was proposed by Benjamini *et al.* (2006). It is a major point when we estimate π_0^{-1} in the independent case:

Lemma 11.19 *For all $k \geq 2$, $q \in]0, 1]$ and any random variable Y with a Binomial $(k-1, q)$ distribution, we have*

$$\mathbb{E}[1/(1+Y)] \leq 1/kq.$$



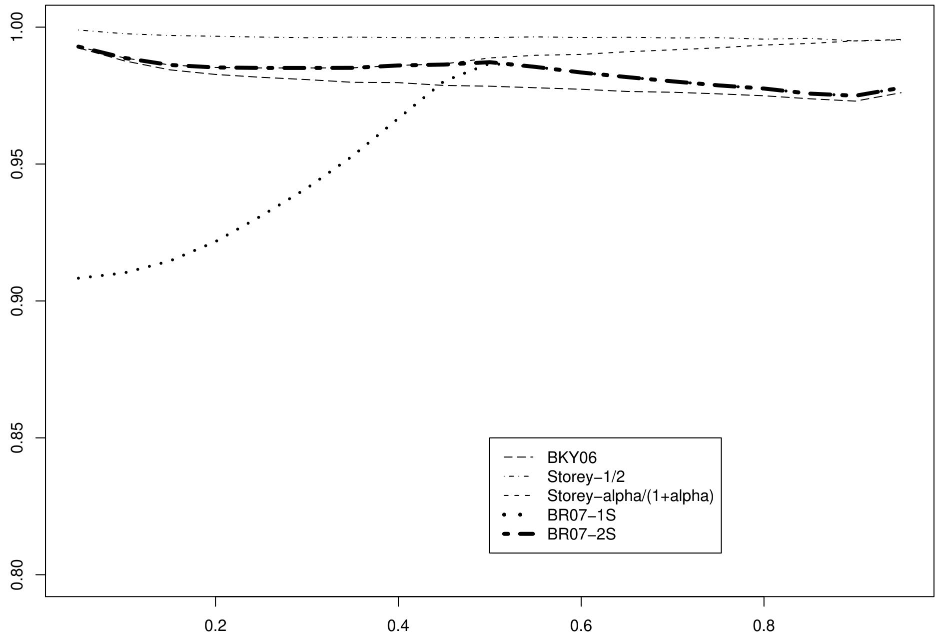
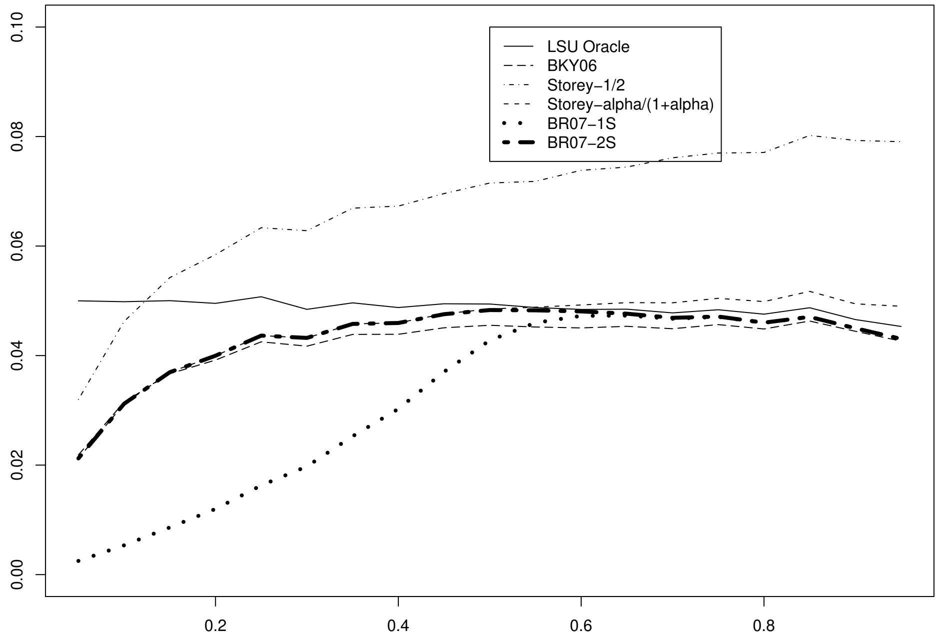


Figure 11.2: Top graph : estimated FDR in function of π_0. Bottom graph : estimated power (relative to the Oracle procedure) in function of π_0. Independent case ($\rho = 0$), 100 null hypotheses ($m = 100$), 10 000 simulations.



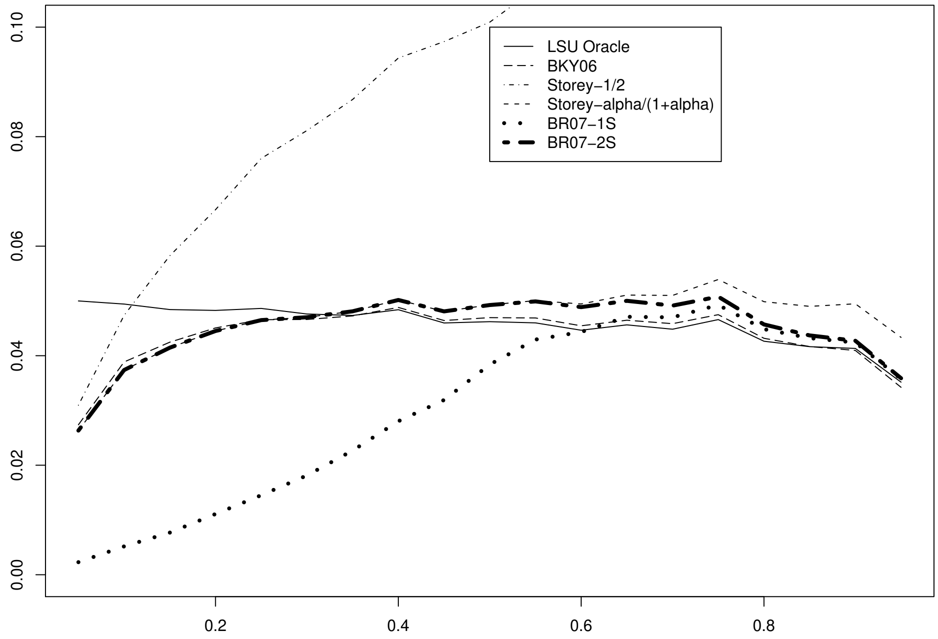


Figure 11.3: Both graphs : estimated FDR in function of π_0. Case of positive dependencies (top graph: $\rho = 0.2$, bottom graph: $\rho = 0.5$), 100 null hypotheses ($m = 100$), 10 000 simulations.

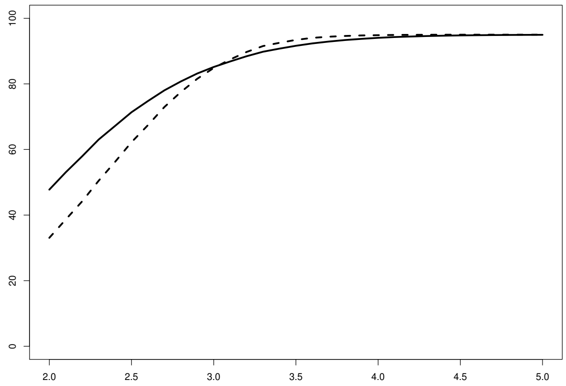


Figure 11.4: Y-axis: estimated expected number of correct rejections of the different procedures, X-axis: common value of all the positive means. The solid line corresponds to the LSU procedure. The dashed line corresponds to the two-stage adaptive procedure of Corollary 11.17 (PRDS case with $\kappa = 2$, $\alpha_0 = \alpha/4$ and $\alpha_1 = \alpha/2$). Case of positive correlations with $\rho = 0.1$, $m = 100$, $m_0 = 5$, 10 000 simulations.

Chapter 12

Resampling-based confidence regions and multiple tests for a correlated random vector

This chapter is a joint work with Sylvain Arlot[1] and Gilles Blanchard[2]. It corresponds to a long version of a paper published in the proceedings of COLT (see Arlot *et al.* (2007)).

We study generalized bootstrapped confidence regions for the mean of a random vector whose coordinates have an unknown dependence structure, with a non-asymptotic control of the confidence level. The random vector is supposed to be either Gaussian or to have a symmetric bounded distribution. We consider two approaches, the first based on a concentration principle and the second on a direct boostrapped quantile. The first one allows us to deal with a very large class of resampling weights while our results for the second are restricted to Rademacher weights. These results are applied in the one-sided and two-sided multiple testing problem, in which we derive several resampling-based step-down procedures providing a non-asymptotic FWER control. We compare our different procedures in a simulation study, and we show that they can outperform Bonferroni's or Holm's procedures as soon as the observed vector has sufficiently correlated coordinates.

12.1 Introduction

12.1.1 Goals and motivations

In this chapter, we assume that we observe a sample $\mathbf{Y} := (\mathbf{Y}^1, \dots, \mathbf{Y}^n)$ of $n \geq 2$ i.i.d. observations of an integrable random vector $\mathbf{Y}^i \in \mathbb{R}^K$ with a dimension K possibly much larger than n. Let $\mu \in \mathbb{R}^K$ denote the common mean of the $\mathbf{Y}^i$; our main goal is to find a non-asymptotic $(1-\alpha)$-confidence region for μ, of the form:

$$\left\{ x \in \mathbb{R}^K \mid \phi\left(\overline{\mathbf{Y}} - x\right) \leq t_\alpha(\mathbf{Y}) \right\} \quad , \tag{12.1}$$

[1]Univ Paris-Sud, Laboratoire de Mathématiques d'Orsay.
[2]Fraunhofer FIRST.IDA, Berlin, Germany.

where $\phi : \mathbb{R}^K \to \mathbb{R}$ is a measurable function (measuring a kind of distance), $\alpha \in (0,1)$, $t_\alpha : \left(\mathbb{R}^K\right)^n \to \mathbb{R}$ is a measurable data-dependent threshold, and $\overline{\mathbf{Y}} = \frac{1}{n}\sum_{i=1}^n \mathbf{Y}^i$ is the empirical mean of the sample $\mathbf{Y}$.

The form of the confidence region (12.1) is motivated by the following multiple testing problem: when we test simultaneously for all $1 \leq k \leq K$ the null hypotheses H_k : "$\mu_k \leq 0$" against A_k : "$\mu_k > 0$", a classical procedure consists in rejecting the H_k corresponding to

$$\left\{1 \leq k \leq K \mid \overline{\mathbf{Y}}_k > t_\alpha(\mathbf{Y})\right\} \quad . \tag{12.2}$$

The error of such a multiple testing procedure can be measured by the family-wise error rate (FWER) defined by the probability that at least one hypothesis is wrongly rejected. Denoting by $\mathcal{H}_0 = \{k \mid \mu_k \leq 0\}$ the set of coordinates corresponding to the true null hypotheses, the FWER of the procedure defined in (12.2) can be controlled as follows:

$$\begin{aligned}\mathbb{P}\left(\exists k \mid \overline{\mathbf{Y}}_k > t_\alpha(\mathbf{Y}) \text{ and } \mu_k \leq 0\right) &\leq \mathbb{P}\left(\exists k \in \mathcal{H}_0 \mid \overline{\mathbf{Y}}_k - \mu_k > t_\alpha(\mathbf{Y})\right) \\ &= \mathbb{P}\left(\sup_{k \in \mathcal{H}_0}\left\{\overline{\mathbf{Y}}_k - \mu_k\right\} > t_\alpha(\mathbf{Y})\right) \quad .\end{aligned}$$

Since μ_k is unknown under H_k, controlling the above probability by a level α is equivalent to establish a $(1-\alpha)$-confidence region for μ of the form (12.1) with $\phi = \sup_{\mathcal{H}_0}(\cdot)$. Similarly, the same reasoning with $\phi = \sup_{\mathcal{H}_0}|\cdot|$ in (12.1) allows us to test H_k : "$\mu_k = 0$" against A_k : "$\mu_k \neq 0$", by choosing the rejection set $\left\{1 \leq k \leq K \mid \left|\overline{\mathbf{Y}}_k\right| > t_\alpha(\mathbf{Y})\right\}$.

In our framework, we emphasize that:

- we want a non-asymptotical result valid for any fixed K and n, with K possibly much larger than the number of observations n.
- we do not make any assumptions on the dependency structure of the coordinates of $\mathbf{Y}^i$ (although we will consider some specific assumptions over the distribution of $\mathbf{Y}$, for example that it is Gaussian).

This viewpoint is motivated by practical applications, especially neuroimaging (see Pantazis *et al.* (2005); Darvas *et al.* (2005); Jerbi *et al.* (2007)). In a typical magnetoencephalography (MEG) experiment, each observation $\mathbf{Y}^i$ is a two or three dimensional brain activity map[3] of 15 000 points (or a time series of length between 50 and 1 000 of such data). The dimensionality K thus goes from 10^4 to 10^7. Such observations are repeated $n = 15$ up to 4 000 times, but this upper bound is very hard to attain (see Waberski *et al.* (2003)). Typically, $n \leq 100 \ll K$. In such data, there are strong dependencies between locations (the 15 000 points are obtained by pre-processing data of 150 sensors) which are highly spatially non-uniform, as remarked by Pantazis *et al.* (2005). Moreover, there may be distant correlations, *e.g.* depending on neural connections inside the brain, so that we cannot make use of a simple parametric model. Finally, notice that the false discovery rate (FDR), defined as the average proportion of wrongly rejected hypotheses among all the rejected hypotheses, is not always relevant in neuroimaging. Indeed, the signal is often strong over some well-known large areas of the brain (*e.g.* the motor and visual cortex). Therefore, if for instance 95 percent of the detected locations belong to these

[3]actually, $\mathbf{Y}^i$ is the difference between brain activities with and without some stimulation. Then, non-zero means are locations at which the stimulation has a significant effect.

well-known areas, FDR control (at level 5%) does not provide evidence for any new discovery. On the contrary, FWER control is more conservative, but each detected location outside these well-known areas is a new discovery with high probability.

12.1.2 Our two approaches

The ideal threshold t_α in (12.1) is obviously the $1-\alpha$ quantile of the distribution of $\phi\left(\overline{\mathbf{Y}}-\mu\right)$. However, this quantity depends on the unknown dependency structure of the coordinates of $\mathbf{Y}^i$ and is therefore itself unknown.

We propose here to approach t_α by some resampling scheme: the heuristics of the resampling method (introduced by Efron (1979), generalized to exchangeable weighted bootstrap by Mason and Newton (1992) and Præstgaard and Wellner (1993)) is that the distribution of $\overline{\mathbf{Y}}-\mu$ is "close" to the one of

$$\overline{\mathbf{Y}}_{[W-\overline{W}]} := \frac{1}{n}\sum_{i=1}^{n}(W_i-\overline{W})\mathbf{Y}^i = \frac{1}{n}\sum_{i=1}^{n}W_i(\mathbf{Y}^i-\overline{\mathbf{Y}}) = \overline{\left(\mathbf{Y}-\overline{\mathbf{Y}}\right)}_{[W]} \ ,$$

conditionally to $\mathbf{Y}$, where $(W_i)_{1\le i\le n}$ are real random variables independent of $\mathbf{Y}$ called the *resampling weights*, and $\overline{W}=n^{-1}\sum_{i=1}^{n}W_i$. We emphasize that the family $(W_i)_{1\le i\le n}$ itself *need not be independent.*

Following this idea, we propose two different approaches to obtain non-asymptotic confidence regions:

- Approach 1 ("concentration approach"):
  The expectations of $\phi\left(\overline{\mathbf{Y}}-\mu\right)$ and $\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)$ can be precisely compared, and the processes $\phi\left(\overline{\mathbf{Y}}-\mu\right)$ and $\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\middle|\mathbf{Y}\right]$ concentrate well around their expectations.
- Approach 2 ("quantile approach"):
  The $1-\alpha$ quantile of the distribution of $\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)$ conditionally to $\mathbf{Y}$ is close to the one of $\phi\left(\overline{\mathbf{Y}}-\mu\right)$.

Approach 1 above is closely related to the Rademacher complexity method in learning theory, and our results in this direction are heavily inspired by the work of Fromont (2004), who studies general resampling schemes in a learning theoretical setting. It may also be seen as a generalization of cross-validation methods. For approach 2, we will restrict ourselves specifically to Rademacher weights in our analysis, because we use a symmetrization trick.

12.1.3 Relation to previous work

Using resampling to construct confidence regions (see *e.g.* Efron (1979); Hall (1992); Hall and Mammen (1994)) or multiple testing procedures (see *e.g.* Westfall and Young (1993); Yekutieli and Benjamini (1999); Pollard and van der Laan (2003); Ge *et al.* (2003); Romano and Wolf (2007)) is a vast field of study in statistics. Roughly speaking, we can mainly distinguish between two types of results:

- asymptotic results, which are based on the fact that the bootstrap process is asymptotically close to the original empirical process (see van der Vaart and Wellner (1996)).

- exact randomized tests (see *e.g.* Romano (1989, 1990); Romano and Wolf (2005)), which are based on an invariance of the null distribution under a given transformation ; the underlying idea can be traced back to Fisher's permutation test (see Fisher (1935)).

As we have remarked earlier, the asymptotic approach is not adapted to the goals we have fixed here since we are looking for non-asymptotic results. On the other hand, what we called our "quantile approach" in the previous section is strongly related to exact randomization tests. Namely, we will only consider symmetric distributions: this is a specific instance of an invariance with respect to a transformation and will allow us to make use of distribution-preserving randomization via sign-flipping. The main difference with traditional exact randomization tests is that, because our first goal is to derive a confidence region, the vector of the means is unknown and therefore, so is the exact invariant transformation. Our contribution to this point is essentially to show that the true vector of the means can be replaced by the empirical one in the randomization, for the price of additional terms of smaller order in the threshold thus obtained. To our knowledge, this gives the first non-asymptotic approximation result on resampled quantiles with an unknown distribution mean.

Finally, our "concentration approach" of the previous section is not directly related to either type of the above previous results, but, as already pointed out earlier, is strongly inspired by results coming from learning theory.

12.1.4 Notations

Let us now define a few notations that will be useful throughout this chapter.

- Vectors, such as data vectors $\mathbf{Y}^i = (\mathbf{Y}^i_k)_{1 \le k \le K}$, will always be column vectors. Thus, $\mathbf{Y}$ is a $K \times n$ data matrix.
- If $\mu \in \mathbb{R}^K$, $\mathbf{Y} - \mu$ is the matrix obtained by subtracting μ from each (column) vector of $\mathbf{Y}$. If $c \in \mathbb{R}$ and $W \in \mathbb{R}^n$, $W - c = (W_i - c)_{1 \le i \le n} \in \mathbb{R}^n$.
- If X is a random variable, $\mathcal{D}(X)$ is its distribution and $\mathrm{Var}(X)$ is its variance.
- The vector $\sigma = (\sigma_k)_{1 \le k \le K}$ is the vector of the standard deviations of the data: $\forall k, 1 \le k \le K$, $\sigma_k = \mathrm{Var}^{1/2}(\mathbf{Y}^1_k)$.
- $\overline{\Phi}$ is the standard Gaussian upper tail function: if $X \sim \mathcal{N}(0,1)$, $\forall x \in \mathbb{R}$, $\overline{\Phi}(x) = \mathbb{P}(X \ge x)$.

Several properties may be assumed for the function $\phi : \mathbb{R}^K \to \mathbb{R}$:

- Subadditivity: $\forall x, x' \in \mathbb{R}^K, \quad \phi(x + x') \le \phi(x) + \phi(x')\,.$
- Positive-homogeneity: $\forall x \in \mathbb{R}^K, \forall \lambda \in \mathbb{R}^+, \quad \phi(\lambda x) = \lambda \phi(x)\,.$
- Bounded by the p-norm, $p \in [1, \infty]$: $\forall x \in \mathbb{R}^K$, $|\phi(x)| \le \|x\|_p$, where $\|x\|_p$ is equal to $(\sum_{k=1}^K |x_k|^p)^{1/p}$ if $p < \infty$ and $\max_k\{|x_k|\}$ otherwise.

Finally, we define the following possible assumptions on the generating distribution of $\mathbf{Y}$:

(GA) The Gaussian assumption: the $\mathbf{Y}^i$ are Gaussian vectors.

(SA) The symmetric assumption: the $\mathbf{Y}^i$ are symmetric with respect to μ i.e. $\mathbf{Y}^i - \mu \sim \mu - \mathbf{Y}^i$.

(BA)(p, M) The bounded assumption: $\left\|\mathbf{Y}^i - \mu\right\|_p \le M$ a.s.

In this chapter, our primary focus is on the Gaussian framework (GA), because the corresponding results will be more accurate. In addition, we will always assume that we know some upper bound on a p-norm of σ for some $p > 0$.

The chapter is organized as follows. We first build confidence regions with two different techniques : Section 12.2 deals with the concentration method with general weights, and Section 12.3 with a quantile approach with Rademacher weights. We then focus on the multiple testing problem in Section 12.4, where we deduce step-down procedures from our previous confidence regions. Finally, Section 12.5 illustrates our results on both confidence regions and multiple testing with a simulation study. All the proofs are given in Section 12.7.

12.2 Confidence region using concentration

We consider here a general *resampling weight vector* W, that is, a $\mathbb{R}^n$-valued random vector $W = (W_i)_{1\le i\le n}$ independent of $\mathbf{Y}$ satisfying the following properties : for all $i \in \{1, \dots, n\}$ $\mathbb{E}\left[W_i^2\right] < \infty$ and $n^{-1}\sum_{i=1}^n \mathbb{E}\left|W_i - \overline{W}\right| > 0$.

We will mainly consider in this section an *exchangeable resampling weight vector*, that is, a resampling weight vector W such that $(W_i)_{1\le i\le n}$ has an exchangeable distribution (*i.e.* invariant under any permutation of the indices). Several examples of exchangeable resampling weight vectors are given in Section 12.2.3, where we also tackle the question of choosing a resampling. Non-exchangeable weight vectors are studied in Section 12.2.4.

Four constants that depend only on the distribution of W appear in the results below (the fourth one is defined only for a particular class of weights). They are defined as follows and computed for classical resamplings in Tab. 12.1:

$$A_W := \mathbb{E}\left|W_1 - \overline{W}\right| \tag{12.3}$$

$$B_W := \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n \left(W_i - \overline{W}\right)^2\right)^{\frac{1}{2}}\right] \tag{12.4}$$

$$C_W := \left(\frac{n}{n-1}\mathbb{E}\left[\left(W_1 - \overline{W}\right)^2\right]\right)^{\frac{1}{2}} \tag{12.5}$$

$$D_W := a + \mathbb{E}\left|\overline{W} - x_0\right| \quad \text{if } \forall i,\ |W_i - x_0| = a \text{ a.s. (with } a > 0, x_0 \in \mathbb{R}) . \tag{12.6}$$

Note that these quantities are positive for an exchangeable resampling weight vector W:

$$0 < A_W \le B_W \le C_W\sqrt{1 - 1/n}.$$

Moreover, if the weights are i.i.d., we have $C_W = \operatorname{Var}(W_1)^{\frac{1}{2}}$. We can now state the main result of this section:

Theorem 12.1 *Fix $\alpha \in (0, 1)$ and $p \in [1, \infty]$. Let $\phi : \mathbb{R}^K \to \mathbb{R}$ be any function subadditive, positive-homogeneous and bounded by the p-norm, and let W be an exchangeable resampling weight vector.*

1. *If* $\mathbf{Y}$ *satisfies (GA), then*

$$\phi\left(\overline{\mathbf{Y}}-\mu\right) < \frac{\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)|\mathbf{Y}\right]}{B_W} + \|\sigma\|_p \overline{\Phi}^{-1}(\alpha/2)\left[\frac{C_W}{nB_W}+\frac{1}{\sqrt{n}}\right] \tag{12.7}$$

holds with probability at least $1-\alpha$*. The same bound holds for the lower deviations, i.e. with inequality* (12.7) *reversed and the additive term replaced by its opposite.*

2. *If* $\mathbf{Y}$ *satisfies (BA)*(p,M) *and (SA), then*

$$\phi\left(\overline{\mathbf{Y}}-\mu\right) < \frac{\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)|\mathbf{Y}\right]}{A_W} + \frac{2M}{\sqrt{n}}\sqrt{\log(1/\alpha)} \tag{12.8}$$

holds with probability at least $1-\alpha$ *. If moreover the weights satisfy the assumption of* (12.6)*, then*

$$\phi\left(\overline{\mathbf{Y}}-\mu\right) > \frac{\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)|\mathbf{Y}\right]}{D_W} - \frac{M}{\sqrt{n}}\sqrt{1+\frac{A_W^2}{D_W^2}}\sqrt{2\log(1/\alpha)} \tag{12.9}$$

holds with probability at least $1-\alpha$ *.*

Inequalities (12.7), (12.8) and (12.9) give thresholds such that the corresponding regions of the form (12.1) are confidence regions of level at least $1-\alpha$.

Additionally, if there exists a deterministic threshold t_α such that $\mathbb{P}(\phi\left(\overline{\mathbf{Y}}-\mu\right) > t_\alpha) \leq \alpha$ and in the Gaussian case, the following corollary establishes that we can combine the concentration threshold corresponding to (12.7) with t_α to obtain a threshold that is very close to the minimum of the two.

Corollary 12.2 *Fix* $\alpha,\delta\in(0,1)$*,* $p\in[1,\infty]$ *and take* ϕ *and* W *as in Theorem 12.1. Suppose that* $\mathbf{Y}$ *satisfies (GA) and that* $t_{\alpha(1-\delta)}$ *is a real number such that* $\mathbb{P}\left(\phi\left(\overline{\mathbf{Y}}-\mu\right) > t_{\alpha(1-\delta)}\right) \leq \alpha(1-\delta)$*. Then with probability at least* $1-\alpha$*,* $\phi\left(\overline{\mathbf{Y}}-\mu\right)$ *is upper bounded by the minimum between* $t_{\alpha(1-\delta)}$ *and*

$$\frac{\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)|\mathbf{Y}\right]}{B_W} + \frac{\|\sigma\|_p}{\sqrt{n}}\overline{\Phi}^{-1}\left(\frac{\alpha(1-\delta)}{2}\right) + \frac{\|\sigma\|_p C_W}{nB_W}\overline{\Phi}^{-1}\left(\frac{\alpha\delta}{2}\right). \tag{12.10}$$

Remark 12.3 1. *Corollary 12.2 is more precisely a consequence of Proposition 12.8 (ii).*

2. *Since the last term of* (12.10) *becomes negligible with respect to the rest when* n *grows large, if we use Corollary 12.2 with a small* δ *(for instance* $\delta = 1/n$*), we will obtain a threshold close to the minimum between* t_α *and the threshold corresponding to* (12.7)*.*

3. *For instance, if* $\phi=\sup(\cdot)$ *(resp.* $\sup|\cdot|$*), Corollary 12.2 may be applied with* t_α *equal to the classical Bonferroni threshold (obtained using a simple union bound over coordinates)*

$$t_{Bonf,\alpha} := \frac{1}{\sqrt{n}}\|\sigma\|_\infty \overline{\Phi}^{-1}\left(\frac{\alpha}{K}\right)\left(\textit{resp. } t'_{Bonf,\alpha} := \frac{1}{\sqrt{n}}\|\sigma\|_\infty \overline{\Phi}^{-1}\left(\frac{\alpha}{2K}\right)\right). \tag{12.11}$$

We thus obtain a confidence region almost equal to Bonferroni's for small correlations and better than Bonferroni's for strong correlations (see simulations in Section 12.5).

The proof of Theorem 12.1 involves results which are of self interest: the comparison between the expectations of the two processes $\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right) \middle| \mathbf{Y}\right]$ and $\phi\left(\overline{\mathbf{Y}}-\mu\right)$ and the concentration of these processes around their means. This is examined in the two following subsections. Then, we give some elements for a wise choice of resampling weight vectors among several classical examples. The last subsection tackles the practical issue of computation time.

12.2.1 Comparison in expectation

In this section, we compare $\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\right]$ and $\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}-\mu\right)\right]$. We note that these expectations exist in the Gaussian and the bounded cases provided that ϕ is measurable and bounded by a p-norm. Otherwise, in particular in Propositions 12.4 and 12.6, we assume that these expectations exist. In the Gaussian case, these quantities are equal up to a factor that depends only on the distribution of W:

Proposition 12.4 *Let* $\mathbf{Y}$ *be a sample satisfying (GA) and let* W *be a resampling weight vector. Then, for any measurable positive-homogeneous function* $\phi : \mathbb{R}^K \to \mathbb{R}$*, we have the following equality:*

$$B_W \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}-\mu\right)\right] = \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\right] \quad . \tag{12.12}$$

Remark 12.5

1. *In general, we can compute the value of* B_W *by simulation. For some classical weights, we give bounds or exact expressions (see Tab. 12.1 and Section 12.7.4).*
2. *In a non-Gaussian framework, the constant* B_W *is still relevant, at least asymptotically: in their Theorem 3.6.13, van der Vaart and Wellner (1996) use the limit of* B_W *when* n *goes to infinity as a normalizing constant.*
3. *If the weights satisfy* $\sum_{i=1}^{n}(W_i - \overline{W})^2 = n$ *a.s., then* (12.12) *holds for any function* ϕ *(and* $B_W = 1$*).*

When the sample is only symmetric we obtain the following inequalities :

Proposition 12.6 *Let be* $\mathbf{Y}$ *a sample satisfying (SA),* W *an exchangeable resampling weight vector and* $\phi : \mathbb{R}^K \to \mathbb{R}$ *any subadditive, positive-homogeneous function.*

(i) We have the general following lower bound:

$$A_W \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}-\mu\right)\right] \leq \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\right] \quad . \tag{12.13}$$

(ii) Moreover, if the weights satisfy the assumption of (12.6)*, we have the following upper bound:*

$$D_W \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}-\mu\right)\right] \geq \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\right] \quad . \tag{12.14}$$

Remark 12.7

1. *The bounds* (12.13) *and* (12.14) *are tight for Rademacher and Random hold-out (n/2) weights, but far less optimal in some other cases like Leave-one-out (see Section 12.2.3 for details).*

2. *When* $\mathbf{Y}$ *is not assumed to have a symmetric distribution and* $\overline{W} = 1$ *a.s., Proposition 2 of Fromont (2004) shows that* (12.13) *holds with* $\mathbb{E}(W_1 - \overline{W})_+$ *instead of* A_W*. Therefore, assumption (SA) allows us to get a tighter result (for instance twice sharper with Efron or Random hold-out (q) weights).*

12.2.2 Concentration around the expectation

In this section we present concentration results for the two processes $\phi\left(\overline{\mathbf{Y}} - \mu\right)$ and $\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right) \left|\mathbf{Y}\right.\right]$ in the Gaussian framework.

Proposition 12.8 *Let* $p \in [1, \infty]$, $\mathbf{Y}$ *a sample satisfying (GA) and* $\phi : \mathbb{R}^K \to \mathbb{R}$ *be any subadditive function, bounded by the p-norm.*

(i) For all $\alpha \in (0,1)$*, with probability at least* $1 - \alpha$ *the following holds:*

$$\phi\left(\overline{\mathbf{Y}} - \mu\right) < \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}} - \mu\right)\right] + \frac{\left\|\sigma\right\|_p \overline{\Phi}^{-1}(\alpha/2)}{\sqrt{n}} \quad , \tag{12.15}$$

and the same bound holds for the corresponding lower deviations.

(ii) Let W *be an exchangeable resampling weight vector. Then, for all* $\alpha \in (0,1)$*, with probability at least* $1 - \alpha$ *the following holds:*

$$\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right) \left|\mathbf{Y}\right.\right] < \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\right] + \frac{\left\|\sigma\right\|_p C_W \overline{\Phi}^{-1}(\alpha/2)}{n} \quad , \tag{12.16}$$

and the same bound holds for the corresponding lower deviations.

The bound (12.15) with a remainder in $n^{-1/2}$ is classical. The bound (12.16) is much more interesting because it illustrates one of the key properties of resampling: the "stabilization effect". Indeed, the resampling quantity $\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right) \left|\mathbf{Y}\right.\right]$ concentrates around its expectation at the rate $C_W n^{-1} = o\left(n^{-1/2}\right)$ for most of the weights (see Section 12.2.3 and Tab. 12.1 for more details). Thus, compared to the original process, it is "almost deterministic" and equal to $B_W \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}} - \mu\right)\right]$. In an asymptotic viewpoint, this may be understood through Edgeworth expansions. Indeed, it is well-known (see for instance Hall (1992)) that when ϕ is smooth enough, the first non-zero term in the Edgeworth expansion of $\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right) \left|\mathbf{Y}\right.\right] - \mathbb{E}\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)$ is at least of order n^{-1}.

Remark 12.9 *Combining expression* (12.12) *and Proposition 12.8 (ii), we derive that for a Gaussian sample* $\mathbf{Y}$ *and any* $p \in [1, \infty]$*, the following upper bound holds with probability at least* $1 - \alpha$*:*

$$\mathbb{E}\left\|\overline{\mathbf{Y}} - \mu\right\|_p < \frac{\mathbb{E}\left[\left\|\overline{\mathbf{Y}}_{[W-\overline{W}]}\right\|_p \left|\mathbf{Y}\right.\right]}{B_W} + \frac{\left\|\sigma\right\|_p C_W}{n B_W} \overline{\Phi}^{-1}(\alpha/2) \quad , \tag{12.17}$$

and a similar lower bound holds. This gives a control with high probability of the L^p*-risk of the estimator* $\overline{\mathbf{Y}}$ *of the mean* $\mu \in \mathbb{R}^K$ *at the rate* $C_W B_W^{-1} n^{-1}$*.*

Efron	$2\left(1-\frac{1}{n}\right)^n = A_W \le B_W \le \sqrt{\frac{n-1}{n}} \quad C_W = 1$
Efr., $n \to +\infty$	$\frac{2}{e} = A_W \le B_W \le 1 = C_W$
Rademacher	$1-\frac{1}{\sqrt{n}} \le A_W \le B_W \le \sqrt{1-\frac{1}{n}} \quad C_W = 1 \le D_W \le 1+\frac{1}{\sqrt{n}}$
Rad., $n \to +\infty$	$A_W = B_W = C_W = D_W = 1$
R. h.-o. (q)	$A_W = 2\left(1-\frac{q}{n}\right) \quad B_W = \sqrt{\frac{n}{q}-1}$ $C_W = \sqrt{\frac{n}{n-1}}\sqrt{\frac{n}{q}-1} \quad D_W = \frac{n}{2q} + \left\lvert 1-\frac{n}{2q}\right\rvert$
R. h.-o. $(n/2)$	$A_W = B_W = D_W = 1 \quad C_W = \sqrt{\frac{n}{n-1}}$
Leave-one-out	$\frac{2}{n} = A_W \le B_W = \frac{1}{\sqrt{n-1}} \quad C_W = \frac{\sqrt{n}}{n-1} \quad D_W = 1$

Table 12.1: Resampling constants for classical resampling weight vector.

12.2.3 Resampling weight vectors

In this section, we consider the question of choosing some appropriate exchangeable resampling weight vector W when using Theorem 12.1 or Corollary 12.2. We define the following classical resampling weight vectors:

1. **Rademacher**: W_i i.i.d. Rademacher variables, *i.e.* $W_i \in \{-1, 1\}$ with equal probabilities.
2. **Efron** (Efron's bootstrap weights): W has a multinomial distribution with parameters $(n; n^{-1}, \dots, n^{-1})$.
3. **Random hold-out** (q) (R. h.-o.), $q \in \{1, \dots, n\}$: $W_i = \frac{n}{q}\mathbb{1}_{i\in I}$, where I is uniformly distributed on subsets of $\{1, \dots, n\}$ of cardinality q. These weights may also be called cross validation weights, or leave-$(n-q)$-out weights. A classical choice is $q = n/2$ (when n is even). When $q = n-1$, these weights are called **leave-one-out** weights.

Note that in the resampling literature, the Random hold-out approach is also called "subsampling" (see Politis *et al.* (1999)). For these classical weights, exact or approximate values for the quantities A_W, B_W, C_W and D_W (defined by equations (12.3) to (12.6)) can be easily derived (see Tab. 12.1). Proofs are given in Section 12.7.4, where several other weights are considered. Now, to use Theorem 12.1 or Corollary 12.2, we have to choose a particular resampling weight vector. In the Gaussian case, we propose the following accuracy and complexity criteria: first, relation (12.7) suggests that the quantity $C_W B_W^{-1}$ can be proposed as *accuracy* index for W. Secondly, an upper bound on the computational burden to compute exactly the resampling quantity is given by the cardinality of the support of $\mathcal{D}(W)$, thus providing a *complexity* index. These two criteria are estimated in Tab. 12.2 for classical weights. For any exchangeable weight vector W, we have $C_W B_W^{-1} \ge [n/(n-1)]^{1/2}$ and the cardinality of the support of $\mathcal{D}(W)$ is larger than n. Therefore, the *leave-one-out weights* satisfy the best accuracy-complexity trade-off among exchangeable weights.

Remark 12.10 (Link to leave-one-out prediction risk estimation) *Consider using* $\overline{\mathbf{Y}}$ *for* predicting *a new data point* $\mathbf{Y}^{n+1} \sim \mathbf{Y}^1$ *(independent of* $\mathbf{Y} = (\mathbf{Y}^1, \dots, \mathbf{Y}^n)$*). The corresponding* L^p*-prediction risk is given by* $\mathbb{E}\left\lVert \overline{\mathbf{Y}} - \mathbf{Y}^{n+1}\right\rVert_p$*. For Gaussians, this prediction risk is proportional to the* L^p*-risk:* $\mathbb{E}\left\lVert \overline{\mathbf{Y}} - \mu\right\rVert_p = (n+1)^{\frac{1}{2}}\mathbb{E}\left\lVert \overline{\mathbf{Y}} - \mathbf{Y}^{n+1}\right\rVert_p$*, so that the estimator of the*

Resampling	$C_W B_W^{-1}$ (accuracy)	Card (supp $\mathcal{L}(W)$) (complexity)
Efron	$\leq \frac{1}{2}\left(1-\frac{1}{n}\right)^{-n} \xrightarrow[n\to\infty]{} \frac{e}{2}$	$\binom{2n-1}{n-1} = \Omega(n^{-\frac{1}{2}}4^n)$
Rademacher	$\leq \left(1-n^{-1/2}\right)^{-1} \xrightarrow[n\to\infty]{} 1$	2^n
R. h.-o. $(n/2)$	$= \sqrt{\frac{n}{n-1}} \xrightarrow[n\to\infty]{} 1$	$\binom{n}{n/2} = \Omega(n^{-1/2}2^n)$
Leave-one-out	$= \sqrt{\frac{n}{n-1}} \xrightarrow[n\to\infty]{} 1$	n

Table 12.2: Choice of the resampling weight vectors : accuracy-complexity trade-off.

L^p-risk proposed in Remark 12.9 leads to an estimator of the prediction risk. In particular, using leave-one-out weights and noting $\overline{\mathbf{Y}}^{(-i)}$ the mean of the $(\mathbf{Y}^j, j \neq i, 1 \leq j \leq n)$, we have then established that the leave-one-out estimator

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\overline{\mathbf{Y}}^{(-i)} - \mathbf{Y}^i\right\|_p$$

correctly estimates the prediction risk (up to the factor $(1-1/n^2)^{\frac{1}{2}} \sim 1$).

12.2.4 Practical computation of the thresholds

The exact computation of the resampling quantity $\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)|\mathbf{Y}\right]$ can be too complex for the weights define above. To adress this issue, we consider here to possible ways: first, we can use non-exchangeable weights with a lower complexity index and for which the exact computation is tractable. Alternatively, we propose to make a Monte-Carlo approximation. In both cases, the thresholds have to be made slightly larger in order to keep the level larger than $1-\alpha$. This is detailed in the two paragraphs below.

V-fold cross-validation weights

In order to reduce the computation complexity, we can use "piece-wise exchangeable" weights: consider a regular partition $(B_j)_{1\leq j\leq V}$ of $\{1,\ldots,n\}$ (where $V \in \{2,\ldots,n\}$ and $V|n$), and define the weights $W_i = \frac{V}{V-1}\mathbb{1}_{i\notin B_J}$ with J uniformly distributed on $\{1,\ldots,V\}$. These weights are called the **(regular) V-fold cross validation** weights (V-f. c.v.).

By applying our results to the process $(\widetilde{\mathbf{Y}}^j)_{1\leq j\leq K}$ where $\widetilde{\mathbf{Y}}^j = \frac{V}{n}\sum_{i\in B_j}\mathbf{Y}^i$ is the empirical mean of $\mathbf{Y}$ on block B_j, we can show that Theorem 12.1 can be extended to (regular) V-fold cross validation weights with the following resampling constants [4]:

$$A_W = \frac{2}{V} \quad B_W = \frac{1}{\sqrt{V-1}} \quad C_W = \frac{\sqrt{n}}{V-1} \quad D_W = 1 \ .$$

With V-f. c.v. weights, the complexity index is only V, but we lose a factor $[(n-1)/(V-1)]^{1/2}$ in the accuracy index. The most accurate weights are leave-one-out ones ($V=n$), whereas the

[4]When V does not divide n and the blocks are no longer regular, Theorem 12.1 can also be generalized, but the constants have more complex expressions. See Section 12.7.5.

2-fold ones are the best from the computational viewpoint. The choice of V is thus a trade-off between these two terms and depends on the particular constraints of each problem.

More general non-exchangeable weights are studied in Section 12.7.5. In this section, we focussed on regular V-fold cross-validation weights because they are both simple and efficient.

Monte-Carlo approximation

When we use a Monte-Carlo approximation in order to evaluate $\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)|\mathbf{Y}\right]$, we draw randomly a small number B of i.i.d. weight vectors $W^1, \ldots, W^B$ and compute

$$\frac{1}{B}\sum_{k=1}^{B}\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{\left[W^k-\overline{W^k}\right]}\right)|\mathbf{Y}\right].$$

This method is quite standard in the bootstrap literature and can be improved in several ways (see for instance Hall (1992), appendix II). In Proposition 12.11 below, we propose an explicit correction of the concentration thresholds that takes into account B weight vectors, for bounded weights.

Proposition 12.11 *Let $B \geq 1$ and $W^1, \ldots, W^B$ be i.i.d. exchangeable resampling weight vectors such that $W_1^1 - \overline{W^1} \in [c_1, c_2]$ a.s. Let $p \in [1, \infty]$, $\phi : \mathbb{R}^K \to \mathbb{R}$ be any subadditive function, bounded by the p-norm.*

If $\mathbf{Y}$ is a fixed sample and for every $k \in \{1, \ldots, K\}$, M_k is a median of $\left(\mathbf{Y}_k^i\right)_{1\leq i\leq n}$, then, for every $\beta \in (0,1)$,

$$\begin{aligned}\frac{1}{B}\sum_{k=1}^{B}\phi\left(\overline{\mathbf{Y}}_{\left[W^k-\overline{W^k}\right]}\right) \geq \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)|\mathbf{Y}\right]& \\ -(c_2-c_1)\sqrt{\frac{\ln(\beta^{-1})}{2B}}\left\|\frac{1}{n}\left(\sum_{i=1}^{n}\left|\mathbf{Y}_k^i - M_k\right|\right)_k\right\|_p&\end{aligned} \tag{12.18}$$

holds with probability at least $1-\beta$.

If $\mathbf{Y}$ is generated according to a distribution satisfying (GA), then, for every $\beta \in (0,1)$ and any deterministic $\nu \in \mathbb{R}^K$,

$$\left\|\frac{1}{n}\left(\sum_{i=1}^{n}\left|\mathbf{Y}_k^i - M_k\right|\right)_k\right\|_p \leq \mathbb{E}\left\|\frac{1}{n}\left(\sum_{i=1}^{n}\left|\mathbf{Y}_k^i - \nu_k\right|\right)_k\right\|_p + \frac{\|\sigma\|_p\,\overline{\Phi}^{-1}(\beta/2)}{\sqrt{n}} \tag{12.19}$$

holds with probability at least $1-\beta$.

For instance, with Rademacher weights, we can use (12.18) with $c_2 - c_1 = 2$ and $\beta = \delta\alpha$ ($\delta \in (0,1)$). Then, in the thresholds built upon Theorem 12.1 and Corollary 12.2, one can replace $\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)|\mathbf{Y}\right]$ by its Monte-Carlo approximation at the price of changing α into $(1-\delta)\alpha$, and adding

$$\frac{2}{B_W}\sqrt{\frac{\ln(1/(\delta\alpha))}{2B}}\left\|\frac{1}{n}\left(\sum_{i=1}^{n}\left|\mathbf{Y}_k^i - M_k\right|\right)_k\right\|_p \tag{12.20}$$

to the threshold.

Note that (12.18) holds conditionally to the observed sample, so that B can be chosen in function of $\mathbf{Y}$ in (12.20). Therefore, we can choose B with the following strategy: first, compute a rough estimate $t_{\mathrm{est},\alpha}$ of the final threshold (*e.g.* if $\phi = \|\cdot\|_\infty$ and $\mathbf{Y}$ is gaussian, take the Bonferroni threshold $\|\sigma\|_\infty n^{-1/2}\overline{\Phi}^{-1}(\alpha/(2K))$ or the single test threshold $\|\sigma\|_\infty n^{-1/2}\overline{\Phi}^{-1}(\alpha/2)$). Second, choose B such that (12.20) is much smaller than $t_{\mathrm{est},\alpha}$.

Remark 12.12 *In the Gaussian case,* (12.19) *gives a theoretical upper bound on the additive term (if one can bound the expectation term). This is only useful to ensure that the correction* (12.20) *is negligible for reasonable values of B.*

12.3 Confidence region using resampled quantiles

In this section, we consider a different approach to construct confidence regions, directly based on the estimation of the quantile via resampling. Remember that our setting is non-asymptotic, so that the standard asymptotic approaches cannot be applied here. For this reason, we based our approach on ideas coming from exact randomized tests and consider here the case where $\mathbf{Y}^1$ has a symmetric distribution and where W is an i.i.d Rademacher weight vector, that is, W_i i.i.d. with $W_1 \in \{-1, 1\}$ with equal probabilities.

The idea here is to approximate the quantiles of the distribution $\mathcal{D}\left(\phi\left(\overline{\mathbf{Y}} - \mu\right)\right)$ by the quantiles of the corresponding resampling-based distribution:

$$\mathcal{D}\left(\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right) | \mathbf{Y}\right). \tag{12.21}$$

For this, we take advantage of the symmetry of each $\mathbf{Y}^i$ around its mean. Let us define for a function ϕ the resampled empirical quantile by:

$$q_\alpha(\phi, \mathbf{Y}) := \inf\left\{x \in \mathbb{R} \mid \mathbb{P}_W\left[\phi(\overline{\mathbf{Y}}_{[W]}) > x\right] \le \alpha\right\} \quad .$$

We have the following lemma:

Lemma 12.13 *Let $\mathbf{Y}$ be a data sample satisfying assumption (SA). Then the following holds:*

$$\mathbb{P}\left[\phi(\overline{\mathbf{Y}} - \mu) > q_\alpha\left(\phi, \mathbf{Y} - \mu\right)\right] \le \alpha. \tag{12.22}$$

Of course, since $q_\alpha\left(\phi, \mathbf{Y} - \mu\right)$ still depends on the unknown μ, we cannot use this threshold to get a confidence region of the form (12.1). Therefore, following the general philosophy of resampling, we propose to replace μ by $\overline{\mathbf{Y}}$ in $q_\alpha\left(\phi, \mathbf{Y} - \mu\right)$. The main technical result of this section quantifies the price to pay to perform this operation:

Proposition 12.14 *Fix $\delta, \alpha \in (0,1)$. Let $\mathbf{Y}$ be a data sample satisfying assumption (SA). Let $f : \left(\mathbb{R}^K\right)^n \to [0, \infty)$ be a nonnegative (measurable) function on the set of data samples. Let ϕ be a* nonnegative*, subadditive, positive-homogeneous function. Denote $\widetilde{\phi}(x) = \max\left(\phi(x), \phi(-x)\right)$. Finally, for $\eta \in (0,1)$, denote*

$$\overline{\mathcal{B}}(n, \eta) = \min\left\{k \in \{0, \ldots, n\} \mid 2^{-n}\sum_{i=k+1}^{n}\binom{n}{i} < \eta\right\} \quad ,$$

the upper quantile function of a Binomial $(n, \frac{1}{2})$ *variable. Then we have:*

$$\mathbb{P}\left[\phi(\overline{\mathbf{Y}}-\mu) > q_{\alpha(1-\delta)}\left(\phi, \mathbf{Y}-\overline{\mathbf{Y}}\right) + f(\mathbf{Y})\right] \;\leq\; \alpha + \mathbb{P}\left[\widetilde{\phi}(\overline{\mathbf{Y}}-\mu) > \frac{n}{2\overline{\mathcal{B}}\left(n, \frac{\alpha\delta}{2}\right) - n} f(\mathbf{Y})\right]. \tag{12.23}$$

Remark 12.15 *Note that from Hoeffding's inequality, we have*

$$\frac{n}{2\overline{\mathcal{B}}\left(n, \frac{\alpha\delta}{2}\right) - n} \geq \left(\frac{n}{2\ln\left(\frac{2}{\alpha\delta}\right)}\right)^{1/2}.$$

We can use this in (12.23) *to derive a more explicit (but slightly less accurate) inequality.*

By iteration of Proposition 12.14, we obtain the following corollary:

Corollary 12.16 *Fix* J *a positive integer,* $(\alpha_i)_{i=0,\ldots,J-1}$ *a finite sequence in* $(0,1)$ *and* $\beta, \delta \in (0,1)$. *Let* $\mathbf{Y}$ *be a data sample satisfying assumption (SA). Let* $\phi : \mathbb{R}^K \to \mathbb{R}$ *be a nonnegative, subadditive, positive-homogeneous function and* $f : \left(\mathbb{R}^K\right)^n \to [0,\infty)$ *be a nonnegative function on the set of data samples. Then the following holds:*

$$\mathbb{P}\left[\phi(\overline{\mathbf{Y}}-\mu) > q_{(1-\delta)\alpha_0}(\phi, \mathbf{Y}-\overline{\mathbf{Y}}) + \sum_{i=1}^{J-1} \gamma_i q_{(1-\delta)\alpha_i}(\widetilde{\phi}, \mathbf{Y}-\overline{\mathbf{Y}}) + \gamma_J f(\mathbf{Y})\right]$$
$$\leq \sum_{i=0}^{J-1} \alpha_i + \mathbb{P}\left[\widetilde{\phi}(\overline{\mathbf{Y}}-\mu) > f(\mathbf{Y})\right], \tag{12.24}$$

where, for $k \geq 1$, $\gamma_k = n^{-k} \prod_{i=0}^{k-1} \left(2\overline{\mathcal{B}}\left(n, \frac{\alpha_i \delta}{2}\right) - n\right)$.

The rationale behind this result is that the sum appearing inside the probability in (12.24) should be interpreted as a series of corrective terms of decreasing order of magnitude, since we expect the sequence γ_k to be sharply decreasing. Looking at Hoeffding's bound, this will be the case if the levels are such that $\alpha_i \gg \exp(-n)$.

Looking at (12.24), we still have to deal with the trailing term on the right-hand-side to obtain a useful result. We did not succeed in obtaining a self-contained result based on the symmetry assumption (SA) alone. However, to upper-bound the trailing term, we can assume some additional regularity assumption on the distribution of the data. For example, if the data are Gaussian or bounded, we can apply the results of the previous section (or apply some other device like Bonferroni's bound (12.11)). Explicit formulas for the resulting thresholds are given in Section 12.4 and 12.5 (with $J = 1$). We want to emphasize that the bound used in this last step does not have to be particularly sharp: since we expect (in favorable cases) γ_J to be very small, the trailing probability term on the right-hand side as well as the contribution of $\gamma_J f(\mathbf{Y})$ to the left-hand side should be very minor. Therefore, even a coarse bound on this last term should suffice.

Finally, we note as in the previous section that, for computational reasons, it might be relevant to consider a block-wise Rademacher resampling scheme. For this, let $(B_j)_{1\leq j\leq V}$ be a regular partition of $\{1,\ldots,n\}$ and for all $i \in B_j$, $W_i = W_j^B$, where $(W_j^B)_{1\leq j\leq V}$ are i.i.d. Rademacher. This is equivalent to applying the previous method to the block-averaged sample $(\widetilde{Y}_1,\ldots,\widetilde{Y}_V)$, where $\widetilde{Y}_k$ is the average of the $(Y_i)_{i\in B_k}$. Because the $\widetilde{Y}_i$ are i.i.d. variables, all of the previous results carry over when replacing n by V.

12.4 Application to multiple testing

In this section, we describe how the results of Section 12.2 and 12.3 can be used to derive multiple testing procedures. We focus on the two following multiple testing problems:

- *One-sided problem*: test simultaneously the null hypotheses H_k : "$\mu_k \leq 0$" against A_k : "$\mu_k > 0$", for $1 \leq k \leq K$.
- *Two-sided problem*: test simultaneously the null hypotheses H_k : "$\mu_k = 0$" against A_k : "$\mu_k \neq 0$", for $1 \leq k \leq K$.

In this context, we precise the link between confidence regions and multiple testing, and explain how to improve our resampling-based thresholds. We first introduce a few more notations:

- Put $\mathcal{H} := \{1,\ldots,K\}$, $\mathcal{H}_0 := \{1 \leq k \leq K \mid H_k \text{ is true}\}$ and $\mathcal{H}_1$ its complementary in $\mathcal{H}$.
- For any $x \in \mathbb{R}$, the bracket $[x]$ denotes either x in the one-sided context or $|x|$ in the two-sided context.
- Reordering the coordinates of $\overline{\mathbf{Y}}$

  $$\left[\overline{\mathbf{Y}}_{\sigma(1)}\right] \geq \left[\overline{\mathbf{Y}}_{\sigma(2)}\right] \geq \cdots \geq \left[\overline{\mathbf{Y}}_{\sigma(K)}\right] \quad ,$$

  with a permutation σ of $\{1,\ldots,K\}$, we define for every $i \in \{1,\ldots,K\}$, $\mathcal{C}_i(\mathbf{Y}) := \{\sigma(j) \mid j \geq i\}$ the set which contains the $K-i+1$ smaller coordinates of $\left[\,\overline{\mathbf{Y}}\,\right]$. In particular, $\mathcal{C}_1 = \mathcal{H}$.
- For any $\mathcal{C} \subset \mathcal{H}$,

  $$T(\mathcal{C}) := \sup_{k\in\mathcal{C}} \left[\overline{\mathbf{Y}}_k - \mu_k\right] \qquad T'(\mathcal{C}) := \sup_{k\in\mathcal{C}} \left[\overline{\mathbf{Y}}_k\right]$$

  We remark that $T(\mathcal{H}) \geq T(\mathcal{H}_0) \geq T'(\mathcal{H}_0)$ in general and $T(\mathcal{H}_0) = T'(\mathcal{H}_0)$ in the two-sided context.

12.4.1 Multiple testing and connection with confidence regions

A multiple testing procedure is a (measurable) function

$$R(\mathbf{Y}) \subset \mathcal{H} \ ,$$

that rejects the null hypotheses H_k with $k \in R(\mathbf{Y})$. For such a multiple testing procedure R, a type I error arises as soon as R rejects at least one hypothesis which is in fact true. The family-wise error rate of R is then the probability that at least one type I error occurs:

$$\mathrm{FWER}(R) := \mathbb{P}\left(|R(\mathbf{Y}) \cap \mathcal{H}_0| > 0\right) \ .$$

Given a level $\alpha \in (0,1)$, our goal is to build a multiple testing procedure R with

$$\mathrm{FWER}(R) \leq \alpha. \tag{12.25}$$

Of course, choosing the procedure $R = \emptyset$ (i.e. the procedure which rejects no null hypothesis) satisfies trivialy this property. Therefore, provided that (12.25) holds, we want the average number of rejected false null hypotheses, that is

$$\mathbb{E}|R(\mathbf{Y}) \cap \mathcal{H}_1| \ , \tag{12.26}$$

to be as large as possible.

A common way to build a multiple testing procedure is to reject the null hypotheses H_k corresponding to

$$R(\mathbf{Y}) = \left\{1 \leq k \leq K \mid \left[\overline{\mathbf{Y}}_k\right] > t\right\} \ , \tag{12.27}$$

where t is a (possibly data-dependent) threshold. From now on, we will restrict our attention to multiple testing procedures of the previous form. In this case, the deterministic threshold that maximises (12.26) provided that (12.25) holds is obviously the $1-\alpha$ quantile of the distribution of $T'(\mathcal{H}_0)$.

This should be compared to the confidence region context, where the smallest deterministic threshold for which (12.1) holds with $\phi = \sup[\cdot]$ is the $(1-\alpha)$ quantile of the distribution of $T(\mathcal{H})$. Since $T(\mathcal{H}) \geq T'(\mathcal{H}_0)$, we observe following:

1. The thresholds that give confidence regions of the form (12.1) with $\phi = \sup[\cdot]$ also give multiple testing procedures with a FWER less than or equal to α (following the thresholding procedure (12.27)). Therefore, we can directly derived from Sections 12.2 and 12.3 resampling-based multiple testing procedures that control the FWER.

2. One might expect to be able to find better (i.e. smaller) thresholds in the multiple testing framework than in the confidence region framework. Therefore, when $\mathcal{H}_1$ is "large", $T(\mathcal{H})$ is "significantly larger" than $T'(\mathcal{H}_0)$ and then procedures based on upper bounding $T(\mathcal{H})$ are conservative. A method commonly used to adress this issue is to consider step-down procedures. This is examinated in the following section.

12.4.2 Background on step-down procedures

We review in this section known facts on step-down procedure (see Romano and Wolf (2005)). We consider here thresholds $\mathbf{t}$ of the following general form:

$$\mathbf{t} : \mathcal{C} \subset \mathcal{H} \mapsto \mathbf{t}(\mathcal{C}) \in \mathbb{R} \ .$$

We call such a threshold a *subset-based threshold* since it gives a value to each subset of $\mathcal{H}$. A subset-based threshold is said to be *non-decreasing* if for all subsets $\mathcal{C}$ and $\mathcal{C}'$, we have

$$\mathcal{C} \subset \mathcal{C}' \quad \Rightarrow \quad \mathbf{t}(\mathcal{C}) \leq \mathbf{t}(\mathcal{C}') \ .$$

In our setting, a non-decreasing subset-based thresholds is easily obtained by taking a supremum over a subset $\mathcal{C}$ of coordinates. In particular, the thresholds derived from Section 12.2 (resp. Section 12.3) define non-decreasing subset-based thresholds, by taking $\phi = \sup_{\mathcal{C}}[\cdot]$ (resp. $\phi = 0 \vee \sup_{\mathcal{C}}[\cdot]$).

Definition 12.17 (Step-down procedure with subset-based threshold) *Let* $\mathbf{t}$ *be a non-decreasing subset-based threshold and note for all* i, $t_i = \mathbf{t}(\mathcal{C}_i)$. *The step-down procedure with threshold* $\mathbf{t}$ *rejects*

$$\left\{1 \leq k \leq K \mid \left[\overline{\mathbf{Y}}_k\right] \geq t_{\hat{\ell}}\right\}$$

where $\hat{\ell} = \max\left\{1 \leq i \leq K \mid \forall j \leq i, \left[\overline{\mathbf{Y}}_{\sigma(j)}\right] \geq t_j\right\}$ *when the latter maximum exists, and the procedure rejects no null hypothesis otherwise.*

A step-down procedure of the above form can be computed using the following iterative algorithm:

Algorithm 12.18

1. *Init: define* $R_0 := \emptyset$, $\mathcal{E}_0 := \mathcal{H}$.
2. *Iteration* $i \geq 1$: *put* $\mathcal{E}_i := \mathcal{E}_{i-1} \backslash R_{i-1}$ *and* $R_i = \left\{k \mid \left[\overline{\mathbf{Y}}_k\right] \geq \mathbf{t}(\mathcal{E}_i)\right\} \backslash R_{i-1}$.
   If $R_i = \emptyset$, *stop and reject the null hypotheses corresponding to:*

$$R(\mathbf{Y}) := \left\{\sigma(k), k \in \cup_{j \leq i-1} R_j\right\} \ .$$

   Otherwise, go to iteration $i+1$ *(if* $\cup_{j \leq i} R_j = \mathcal{H}$ *stop and reject all the null hypotheses).*

We recall here Theorem 1 of Romano and Wolf (2005), adapted to our setting:

Theorem 12.19 (Romano and Wolf, 2005) *Let* $\mathbf{t}$ *be a non-decreasing subset-based threshold. Then the step-down procedure* R *of threshold* $\mathbf{t}$ *satisfies,*

$$FWER(R) \leq \mathbb{P}(T(\mathcal{H}_0) \geq \mathbf{t}(\mathcal{H}_0)). \tag{12.28}$$

As a consequence, Algorithm 12.18 with any threshold derived from Section 12.2 (resp. Section 12.3) with $\phi = \sup_{\mathcal{H}_0}[\cdot]$ (resp. $\phi = 0 \vee \sup_{\mathcal{H}_0}[\cdot]$) gives a multiple testing procedure with control of the FWER. We detail this in the following section.

12.4.3 Using our confidence regions to build step-down procedures

Using Theorem 12.19 and Corollary 12.2 with the Bonferroni threshold, we derive:

Corollary 12.20 *Fix* $\alpha, \delta \in (0,1)$. *Let* W *be an exchangeable resampling weight vector and suppose that* $\mathbf{Y}$ *satisfies (GA). Then, in the one-sided context, the step-down procedure with the following subset-based threshold controls the FWER at level* α:

$$\mathcal{C} \mapsto \min\left(\frac{\|\sigma\|_\infty}{\sqrt{n}}\overline{\Phi}^{-1}\left(\frac{\alpha(1-\delta)}{|\mathcal{C}|}\right), \frac{\mathbb{E}\left[\sup_{k\in\mathcal{C}}\left\{\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)_k\right\}|\mathbf{Y}\right]}{B_W} + \varepsilon(\alpha,\delta,n)\right)$$

where $\varepsilon(\alpha,\delta,n) = \frac{\|\sigma\|_\infty}{\sqrt{n}}\overline{\Phi}^{-1}\left(\frac{\alpha(1-\delta)}{2}\right) + \frac{\|\sigma\|_\infty C_W}{nB_W}\overline{\Phi}^{-1}\left(\frac{\alpha\delta}{2}\right)$.

Using Theorem 12.19 and Proposition 12.14, we derive:

Corollary 12.21 *Fix $\alpha, \gamma, \delta \in (0,1)$. Let W be a Rademacher weight vector and suppose that $\mathbf{Y}$ satisfies (GA). Then, in the one-sided context, the step-down procedure with the following subset-based threshold controls the FWER at level α:*

$$\mathcal{C} \mapsto q_{\alpha(1-\delta)(1-\gamma)}\left(0 \vee \sup_{\mathcal{C}}(\cdot), \mathbf{Y} - \overline{\mathbf{Y}}\right) + \varepsilon'(\alpha, \delta, \gamma, n, |\mathcal{C}|)$$

where $\varepsilon'(\alpha, \delta, \gamma, n, k) = \frac{2\overline{\mathcal{B}}(n, \alpha(1-\gamma)\delta/2) - n}{n} \frac{\|\sigma\|_\infty}{\sqrt{n}} \overline{\Phi}^{-1}\left(\frac{\alpha\gamma}{2k}\right)$.

Of course, analogues of Corollary 12.20 and 12.21 can also be derived for the two-sided problem.

Remark 12.22 *1. Note that the above (data-dependent) subset-based thresholds are translation-invariant because $\mathbf{Y} - \overline{\mathbf{Y}}$ is. Therefore, large values of non-zero means μ_k will not enlarge these thresholds.*

2. Both subset-based thresholds of Corollary 12.20 and 12.21 are built in order to improve "Bonferroni's subset-based threshold"

$$\mathcal{C} \mapsto \frac{\|\sigma\|_\infty}{\sqrt{n}} \overline{\Phi}^{-1}\left(\frac{\alpha}{|\mathcal{C}|}\right).$$

Therefore, the corresponding step-down procedures are expected to perform better than Holm's procedure (i.e. the step-down version of Bonferroni's procedure, see Holm (1979)).

12.4.4 Uncentered quantile approach for two-sided testing

We focus here on the two-sided multiple testing problem. According to (12.28), we only need a weak[5] control of $T'(\mathcal{C}) = \sup_{k\in\mathcal{C}} |\overline{\mathbf{Y}}_k|$ to obtain a step-down procedure with a strong[6] control of the FWER. Then, similarly to Lemma 12.13, an exact quantile approach is possible.

Corollary 12.23 *Let W be a Rademacher weight vector and suppose that $\mathbf{Y}$ satisfies (SA). Then for two-sided testing, the step down procedure with the subset-based threshold*

$$\mathcal{C} \mapsto q_\alpha\left(\sup_{\mathcal{C}} |\cdot|, \mathbf{Y}\right)$$

controls the FWER at level α.

The main difference with our approach is that the data $\mathbf{Y}$ is not recentered here. In the following, we will call the threshold $q_\alpha(\sup_{\mathcal{C}} |\cdot|, \mathbf{Y})$ the "uncentered quantile". When $\mathcal{H}_0 = \mathcal{H}$, this procedure is very accurate since it achieves the exact level (up to 2^{-n}). It certainly performs better than our "empirically centered" procedures based on Proposition 12.14, because of the second-order term (see the simulation study of Section 12.5). However, large values of the non-zero means μ_k will enlarge this threshold, so that the procedure of Corollary 12.23 may need more steps. In order to fix this drawback, we propose to mix the step-down uncentered and empirically centered quantiles. Up to some small loss in the level, the following algorithm should be faster than the one corresponding to Corollary 12.23.

[5] *i.e.* when $\mathcal{C} = \mathcal{H}_0$.
[6] *i.e.* for every $\mu \in \mathbb{R}^K$.

Algorithm 12.24

1. *Reject the null hypotheses corresponding to:*

$$R_0 := \left\{k \mid \left|\overline{\mathbf{Y}}_k\right| \geq q_{\alpha(1-\delta)(1-\gamma)}(\|\cdot\|_\infty, \mathbf{Y} - \overline{\mathbf{Y}}) + \varepsilon'(\alpha, \delta, \gamma, n, K)\right\}$$

2. *If $R_0 = \mathcal{H}$ then stop.*
Otherwise, consider the set of the remaining coordinates $\mathcal{H} \backslash R_0$ and apply on it the step-down algorithm 12.18 with the subset-based threshold

$$\mathcal{C} \mapsto q_{\alpha(1-\gamma)}\left(\sup_{\mathcal{C}} |\cdot|, \mathbf{Y}\right) \quad .$$

Proposition 12.25 *Fix $\alpha, \gamma, \delta \in (0,1)$. Let W be a Rademacher weight vector and suppose that* $\mathbf{Y}$ *satisfies (GA). In the two-sided context, the algorithm 12.24 gives a multiple testing procedure with a FWER less than or equal to α.*

12.5 Simulations

For simulations, we consider data of the form $\mathbf{Y}_t = \mu_t + G_t$, where t belongs to an $d \times d$ discretized 2D torus of $K = d^2$ "pixels", identified with $\mathbb{T}_d^2 = (\mathbb{Z}/d\mathbb{Z})^2$, and G is a centered Gaussian vector obtained by 2D discrete convolution of an i.i.d. standard Gaussian field ("white noise") on $\mathbb{T}_d^2$ with a function $F : \mathbb{T}_d^2 \to \mathbb{R}$ such that $\sum_{t \in \mathbb{T}_d^2} F^2(t) = 1$. This ensures that G is a stationary Gaussian process on the discrete torus, it is in particular isotropic with $\mathbb{E}\left[G_t^2\right] = 1$ for all $t \in \mathbb{T}_d^2$.

In the simulations below we consider for the function F a "pseudo Gaussian" convolution filter of bandwidth b on the torus:

$$F_b(t) = C_b \exp\left(-d(0,t)^2/b^2\right) \quad ,$$

where $d(t,t')$ is the standard distance on the torus and C_b is a normalizing constant. Note that for actual simulations it is more convenient to work in the Fourier domain and to apply the inverse DFT which can be computed efficiently. We then compare the different thresholds obtained by the methods proposed in this work for varying values of b. Remember that the only information available to our algorithms is the bound on the marginal variance; the form of the function F_b itself is of course unknown.

12.5.1 Confidence balls

On Figure 12.1 we compare the thresholds obtained when $\phi = \sup |\cdot|$, which corresponds to L^∞ confidence balls. Remember that these thresholds can be also directly used in the two-sided multiple testing situation (see Section 12.4). We use the different approaches proposed in this work, with the following parameters: the dimension is $K = 128^2 = 16384$, the number of data points per sample is $n = 1000$ (much smaller than K, so that we really are in a non-asymptotic framework), the width b takes even values in the range $[0, 40]$, the overall level is $\alpha = 0.05$.

Recall that the Bonferroni threshold is

$$t'_{\mathrm{Bonf},\alpha} := \frac{1}{\sqrt{n}} \|\sigma\|_\infty \overline{\Phi}^{-1}\left(\frac{\alpha}{2K}\right) \quad .$$

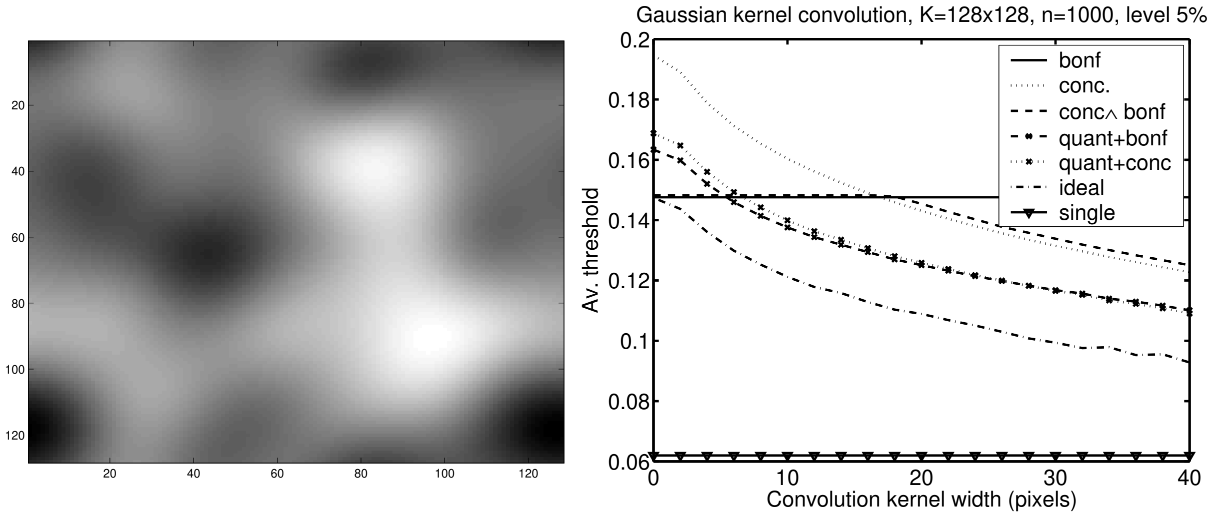


Figure 12.1: Left: example of a 128x128 pixel image obtained by convolution of Gaussian white noise with a (toroidal) Gaussian filter with width $b = 18$ pixels. Right: average thresholds obtained for the different approaches, see text.

For the concentration threshold (12.7)

$$t_{\mathrm{conc},\alpha}(\mathbf{Y}) := \frac{\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\big|\mathbf{Y}\right]}{B_W} + \|\sigma\|_p\,\overline{\Phi}^{-1}(\alpha/2)\left[\frac{C_W}{nB_W}+\frac{1}{\sqrt{n}}\right] \quad ,$$

we used Rademacher weights. For the "compound" threshold of Corollary 12.2 (with the Bonferroni threshold as deterministic reference threshold)

$$t_{\mathrm{conc}\wedge\mathrm{bonf},\alpha}(\mathbf{Y}) := \min\left\{t'_{\mathrm{bonf},\alpha}, \frac{\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\big|\mathbf{Y}\right]}{B_W} + \frac{\|\sigma\|_p}{\sqrt{n}}\overline{\Phi}^{-1}\left(\frac{\alpha(1-\delta)}{2}\right) + \frac{\|\sigma\|_p\, C_W}{nB_W}\overline{\Phi}^{-1}\left(\frac{\alpha\delta}{2}\right)\right\} \quad ,$$

we used $\delta = 0.1$. For the quantile approach (12.24)

$$t_{\mathrm{quant+bonf},\alpha}(\mathbf{Y}) := q_{\alpha_0(1-\delta)}\left(\phi,\mathbf{Y}-\overline{\mathbf{Y}}\right) + \frac{2\overline{\mathcal{B}}\left(n,\frac{\alpha_0\delta}{2}\right)-n}{n}t'_{\mathrm{Bonf},\alpha-\alpha_0}$$

$$t_{\mathrm{quant+conc},\alpha}(\mathbf{Y}) := q_{\alpha_0(1-\delta)}\left(\phi,\mathbf{Y}-\overline{\mathbf{Y}}\right) + \frac{2\overline{\mathcal{B}}\left(n,\frac{\alpha_0\delta}{2}\right)-n}{n}t_{\mathrm{conc},\alpha-\alpha_0}(\mathbf{Y}) \quad ,$$

we used $J = 1$, $\alpha_0 = 0.9\alpha$, $\delta = 0.1$ and took f either equal to the Bonferroni or the concentration threshold, respectively. Finally, for comparison, we included in the figure the threshold corresponding to $K = 1$ (estimation of a single coordinate mean)

$$t_{\mathrm{single},\alpha} := \frac{1}{\sqrt{n}}\|\sigma\|_\infty\,\overline{\Phi}^{-1}\left(\frac{\alpha}{2}\right).$$

We also included an estimation of the true quantile (actually, an empirical quantile over 1000 samples), *i.e.* $t_{\mathrm{ideal},\alpha}$ the $1-\alpha$ quantile of the distribution of $\phi\left(\overline{\mathbf{Y}}-\mu\right)$.

Each point represents an average over 50 experiments (except of course for $t'_{\mathrm{Bonf},\alpha}$ and $t_{\mathrm{single},\alpha}$). The quantiles or expectations with respect to Rademacher weights were estimated by Monte-Carlo with 1000 draws (without the additional term introduced in Section 12.2.4). On the figure we did not include standard deviations: they are quite low, of the order of 10^{-3}, although it is worth noting that the quantile threshold has a standard deviation roughly twice as large as the concentration threshold (we did not investigate at this point what part of this variation is due to the MC approximation).

We also computed the quantile threshold $q_\alpha(\phi, \mathbf{Y} - \overline{\mathbf{Y}})$ without second-order term: it is so close to $t_{\mathrm{ideal},\alpha}$ that they would be almost indistinguishable on Figure 12.1.

The overall conclusion of this first preliminary experiment is that the different thresholds proposed in this work are relevant in the sense that they are smaller than the Bonferroni threshold provided the vector has strong enough correlations. As expected, the quantile approach appears to lead to tighter thresholds. (However, this might not be always the case for smaller sample sizes.) One advantage of the concentration approach is that the 'compound' threshold can "fall back" on the Bonferroni threshold when needed, at the price of a minimal threshold increase.

12.5.2 Multiple testing

We now focus on the multiple testing problem. We present here only the two-sided case because the one-sided case gives similar results, except that we can not use the "uncentered quantile" method of Corollary 12.23.

We consider the experiment of the previous section, with the following choice for the vector of means:

$$\forall (i,j) \in \{0, \dots, 127\}^2, \qquad \mu_{(i,j)} = \frac{(64-j)_+}{64} \times 20 t'_{\mathrm{Bonf},\alpha} \ . \tag{12.29}$$

In this situation, note that the half of the null hypotheses are true while the non-zero means are increasing linearly from $(5/16)t'_{\mathrm{Bonf},\alpha}$ to $20t'_{\mathrm{Bonf},\alpha}$. The thresholds obtained are given on Figure 12.2 (100 simulations). The ideal threshold $t_{\mathrm{ideal},\alpha}$ is now derived from the $1-\alpha$ quantile of the distribution of $T'(\mathcal{H}_0) = \sup_{\mathcal{H}_0} |\overline{\mathbf{Y}}|$. We did not report $t_{\mathrm{conc},\alpha}$ and $t_{\mathrm{conc}\wedge\mathrm{bonf},\alpha}$ in order to simplify Figure 12.2 (their values are unchanged, since these thresholds are translation invariant). In addition to the previous thresholds, we considered:

- the uncentered quantile defined by:

$$t_{\mathrm{quant.uncent.},\alpha}(\mathbf{Y}) := q_\alpha\left(||\cdot||_\infty, \mathbf{Y}\right),$$

  and its step down version $t_{\mathrm{s.d.quant.uncent.},\alpha}(\mathbf{Y})$ (see Corollary 12.23).

- the step down version $t_{\mathrm{s.d.quant+bonf},\alpha}(\mathbf{Y})$ of $t_{\mathrm{quant+bonf},\alpha}(\mathbf{Y})$.

- Holm's threshold $t_{\mathrm{Holm},\alpha}(\mathbf{Y})$ (*i.e.* the step-down version of the Bonferroni procedure).

On the right-hand-side of Figure 12.2, we evaluated the powers of the different thresholds $t_\alpha(\mathbf{Y})$, defined as follows :

$$\mathrm{Power}(t_\alpha(\mathbf{Y})) := \mathbb{E}\left(\frac{|\{1 \le k \le K \mid \mu_k \neq 0 \text{ and } |\mathbf{Y}_k| > t_\alpha(\mathbf{Y})\}|}{|\{1 \le k \le K \mid \mu_k \neq 0\}|} \right) \ . \tag{12.30}$$

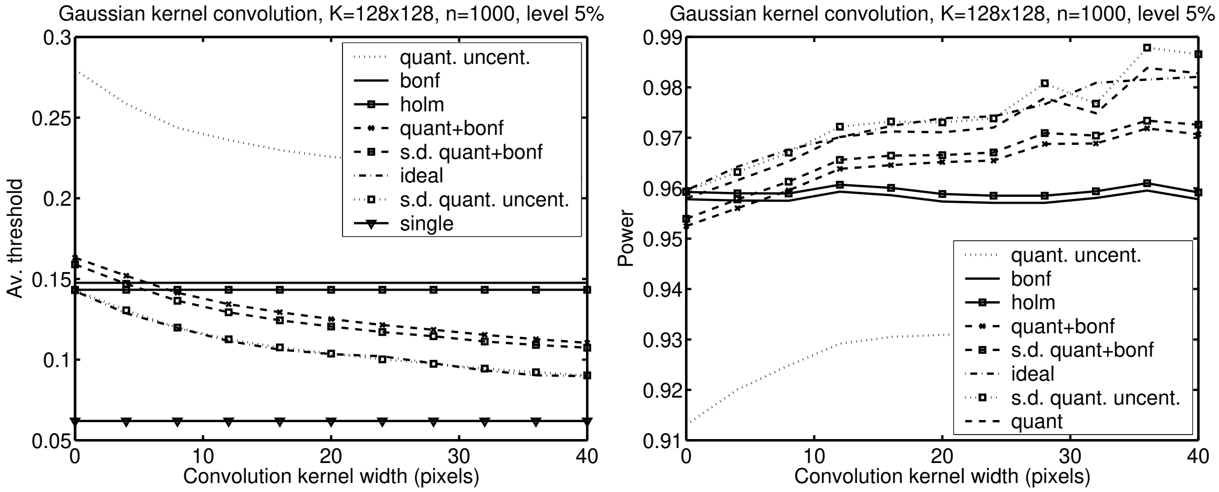


Figure 12.2: Multiple testing problem with μ defined by (12.29) for different approaches, see text. Left: average thresholds. Right: power, defined by (12.30).

This experiment shows that:

1. for single-step resampling-based procedures :

   - the single-step procedure based on our quantile approach ("quant+bonf") can outperform Holm's procedure as soon as the the coordinates of the vector are sufficiently correlated.
   - the single-step procedure based on the uncentered quantile ("quant. uncent") has bad performance.

2. for step-down resampling-based procedures :

   - the step-down procedure based on our quantile approach ("s.d. quant+bonf") can outperform Holm's procedure as soon as the the coordinates of the vector are sufficiently correlated (obvious from the point 1).
   - the step-down procedure based on the uncentered quantile ("s.d. quant+bonf") seems to be the most efficient thresholds of the step-down procedures considered here.

However, when K and n are large, each iteration of the step-down algorithm for the uncentered quantiles may be quite long to compute (typically one day in the neuroimaging framework). Therefore, while the procedure "s.d. quant+bonf" seems to be the more accurate, our quantile approach ("quant+bonf") provides in only one step a quite good accuracy . In this direction, a speed-accuracy trade-off can be made with the algorithm 12.24 (called here "mixed approach"), which uses our quantile approach ("quant+bonf") at first step and the uncentered quantile ("s.d. quant. uncent.") in the remaining steps (at slightly more conservative level of confidence).

We illustrate this with a preliminary study: consider the same simulation framework as above unless that the bandwidth b is fixed at 30, the size of the sample is $n = 100$ and the means are given by: $\forall (i,j) \in \{0,\dots,127\}^2$, $\mu_{(i,j)} = f(i+128j)$, where

$$\forall k \in \{0,\dots,8192\}, \quad f(k) = 50 t'_{\mathrm{Bonf},\alpha} \times \exp\left(-\frac{(8192-k)_+}{8192}\log(100)\right), \tag{12.31}$$

and $f(k) = 0$ for the other values of k. In this situation, the non-zero means are increasing log-linearly from $0.5\, t'_{\mathrm{Bonf},\alpha}$ to $50\, t'_{\mathrm{Bonf},\alpha}$. With 100 simulations, we computed in Table 12.3 the average number of iterations in the step-down algorithm 12.18 for the above step-down procedures. Moreover, on Figure 12.3, the power is given in function of the number of iterations in the step-down algorithm for the different approaches. We can see that the "mixed approach" outperforms method "s.d. quant. uncent." for iterations 1 and 2 and performs almost as well in the following iterations. Moreover, the "mixed approach" is faster because it needs less iterations.

Therefore, this mixed approach can be an interesting alternative of the uncentered quantile approach when several long iterations in the step-down algorithm are expected. This situation arises typically when the signal (non-zero means) has a very wide dynamic range, which was the case in our above simulation where the signal-to-noise ratio for non-true null hypotheses varies between 0.25 and 25.

Holm's procedure	" s.d. quant+bonf"	"s.d. quant. uncent."	"mixed approach"
3.25	3.13	4.92	3.94

Table 12.3: Multiple testing problem with μ corresponding to (12.31) for different step-down approaches. Average number of iterations in the step-down algorithm.

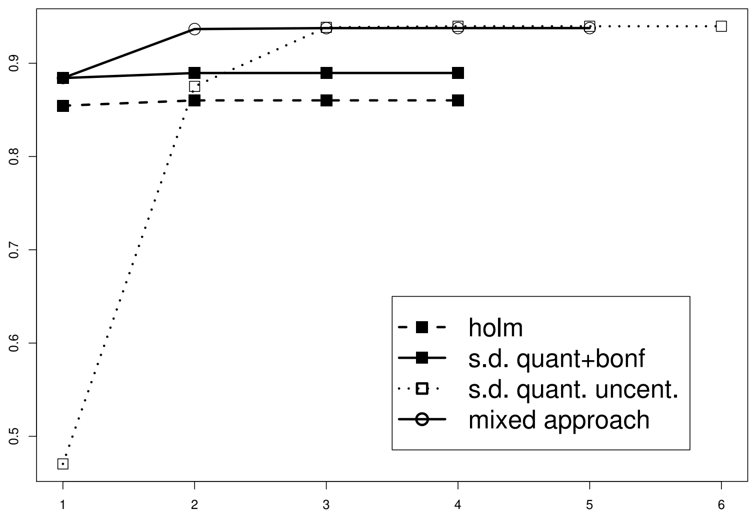


Figure 12.3: Multiple testing problem with μ corresponding to (12.31) for different step-down approaches. Power in function of the number of iterations in the step-down algorithm.

12.6 Conclusion

In this chapter, we proposed two approaches to build non-asymptotic resampling-based confidence regions for a correlated random vector:

- The first one is strongly inspired by results coming from learning theory and is based on a concentration argument. An advantage of this method is that it allows to use a very large class of resampling weights. However, these concentration-based thresholds have relatively conservative deviation terms and they are better than the Bonferroni threshold only if there is very strong correlations in the data. Therefore, using this method when we do not have any prior knowledge on the correlations can be too risky. To adress this issue, we proposed under the Gaussian assumption to combine the corresponding concentration threshold with the Bonferroni threshold to obtain a threshold very closed to the best of the two (using the so-called "stabilization property" of the resampling).

- The second method is more close in spirit to randomization tests: it estimates directly the quantile of $\phi(\overline{\mathbf{Y}} - \mu)$ using a symmetrization argument (it is therefore restricted to Rademacher weights). The point is that an exact approach is not possible because we have to replace the unknown parameter μ by the empirical mean $\overline{\mathbf{Y}}$. Therefore, the derived thresholds have a remainder term, but it is quite small when n is sufficiently large (typically $n \geq 1\,000$).

Our simulations have shown that the confidence regions obtained with the second method are often better than the Bonferroni ones. Moreover, it seems that the quantile threshold without the remainder term is very close to the ideal quantile, so that we may conjecture that the additional term is unnecessary (or at least too large).

Finally, we have used the two previous methods to derive step-down multiple testing procedures that control the FWER when testing simultaneously the means of a (Gaussian) random vector (in the one-sided or two-sided context). Because these procedures use translation-invariant thresholds, the number of iterations in the step-down algorithm is generally small. Moreover, they can outperform Holm's procedure when the coordinates of the observed vector has strong enough correlations. However, these procedures are quite conservative because of the remaining terms (coming from our "non-asymptotic and non-exact" framework).

In the two-sided context, an exact step-down procedure is valid and is more accurate than the above methods (because it has no remainder term). However, this exact method needs generally more iterations in the step-down algorithm. Therefore, we proposed to combine our quantile approach with the latter exact method to get a faster procedure.

Again, we may conjecture that the step-down procedure using the recentred quantile without the additional term (or at least with a smaller term) still controls the FWER for a fixed n. This would give an accurate procedure in both two-sided and one-sided contexts, and the latter would be faster than the exact step-down procedure in the two-sided context. This is an interesting direction for future work.

12.7 Proofs

12.7.1 Confidence regions using concentration

In this section, we prove all the statements of Section 12.2 except computations of resampling weight constants (made in Section 12.7.4) and statements with non-exchangeable resampling weights (made in Section 12.7.5).

Comparison in expectation

Proof of Proposition 12.4. Denoting by $\mathbf{\Sigma}$ the common covariance matrix of the $\mathbf{Y}^i$, we have $\mathcal{D}(\overline{\mathbf{Y}}_{[W-\overline{W}]}|W) = \mathcal{N}\big(0, (n^{-1}\sum_{i=1}^n (W_i - \overline{W})^2) n^{-1}\mathbf{\Sigma}\big)$, and the result follows because $\mathcal{D}(\overline{\mathbf{Y}} - \mu) = \mathcal{N}(0, n^{-1}\mathbf{\Sigma})$ and ϕ is positive-homogeneous. ■

Proof of Proposition 12.6. (i). By independence between W and $\mathbf{Y}$, exchangeability of W and the positive homogeneity of ϕ, for every realization of $\mathbf{Y}$ we have:

$$A_W \phi\left(\overline{\mathbf{Y}} - \mu\right) = \phi\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \left|W_i - \overline{W}\right| \left(\mathbf{Y}^i - \mu\right) \middle| \mathbf{Y}\right]\right) \ .$$

Then, by convexity of ϕ,

$$A_W \phi\left(\overline{\mathbf{Y}} - \mu\right) \le \mathbb{E}\left[\phi\left(\frac{1}{n}\sum_{i=1}^n \left|W_i - \overline{W}\right| \left(\mathbf{Y}^i - \mu\right)\right) \middle| \mathbf{Y}\right] \ .$$

We integrate with respect to $\mathbf{Y}$, and use the symmetry of the $\mathbf{Y}^i$ with respect to μ and again the independence between W and $\mathbf{Y}$ to show finally that

$$\begin{aligned} A_W \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}} - \mu\right)\right] &\le \mathbb{E}\left[\phi\left(\frac{1}{n}\sum_{i=1}^n \left|W_i - \overline{W}\right| \left(\mathbf{Y}^i - \mu\right)\right)\right] \\ &= \mathbb{E}\left[\phi\left(\frac{1}{n}\sum_{i=1}^n \left(W_i - \overline{W}\right) \left(\mathbf{Y}^i - \mu\right)\right)\right] = \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\right] \ . \end{aligned}$$

The point (ii) comes from :

$$\begin{aligned} \mathbb{E}\phi(\overline{\mathbf{Y}}_{W-\overline{W}}) &= \mathbb{E}\phi\Big(\frac{1}{n}\sum_{i=1}^n (W_i - \overline{W})(\mathbf{Y}^i - \mu)\Big) \\ &\le \mathbb{E}\phi\Big(\frac{1}{n}\sum_{i=1}^n (W_i - x_0)(\mathbf{Y}^i - \mu)\Big) + \mathbb{E}\phi\Big(\frac{1}{n}\sum_{i=1}^n (x_0 - \overline{W})(\mathbf{Y}^i - \mu)\Big). \end{aligned}$$

Then, by symmetry of the $\mathbf{Y}^i$ with respect to μ and independence between W and $\mathbf{Y}$, we get

$$\begin{aligned} \mathbb{E}\phi(\overline{\mathbf{Y}}_{W-\overline{W}}) &\le \mathbb{E}\phi\Big(\frac{1}{n}\sum_{i=1}^n |W_i - x_0|(\mathbf{Y}^i - \mu)\Big) + \mathbb{E}\phi\Big(\frac{1}{n}\sum_{i=1}^n |x_0 - \overline{W}|(\mathbf{Y}^i - \mu)\Big) \\ &\le (a + \mathbb{E}|\overline{W} - x_0|)\mathbb{E}\phi(\overline{\mathbf{Y}} - \mu). \end{aligned}$$

■

Concentration inequalities

Proof of Proposition 12.8. We denote by $\mathbf{A}$ a square root of the common covariance matrix of the $\mathbf{Y}^i$. If $\mathbf{G}$ is a $K \times n$ matrix with standard centered i.i.d. Gaussian entries, then $\mathbf{AG}$ has the same distribution as $\mathbf{Y} - \mu$. We let for all $\zeta \in \left(\mathbb{R}^K\right)^n$, $T_1(\zeta) := \phi\left(\frac{1}{n}\sum_{i=1}^n \mathbf{A}\zeta_i\right)$ and $T_2(\zeta) := \mathbb{E}\left[\phi\left(\frac{1}{n}\sum_{i=1}^n (W_i - \overline{W})\mathbf{A}\zeta_i\right)\right]$. From the Gaussian concentration theorem of Cirel'son, Ibragimov and Sudakov (see for instance Massart (2005), Theorem 3.8), we just need to prove that T_1 (resp. T_2) is a Lipschitz function with constant $\|\sigma\|_p / \sqrt{n}$ (resp. $\|\sigma\|_p C_W/n$) with respect to the Euclidean norm $\|\cdot\|_{2,Kn}$ on $\left(\mathbb{R}^K\right)^n$. Let $\zeta, \zeta' \in \left(\mathbb{R}^K\right)^n$ and denote by $(a_k)_{1\leq k\leq K}$ the rows of $\mathbf{A}$. Using that ϕ is 1-Lipschitz with respect to the p-norm (because it is subadditive and bounded by the p-norm), we get

$$\begin{aligned}
|T_1(\zeta) - T_1(\zeta')| &\leq \left\| \frac{1}{n}\sum_{i=1}^n \mathbf{A}(\zeta_i - \zeta_i') \right\|_p \\
&\leq \left\| \left(\left\langle a_k, \frac{1}{n}\sum_{i=1}^n (\zeta_i - \zeta_i') \right\rangle \right)_k \right\|_p .
\end{aligned}$$

For each coordinate k, by Cauchy-Schwartz's inequality and since $\|a_k\|_2 = \sigma_k$, we deduce

$$\left| \left\langle a_k, \frac{1}{n}\sum_{i=1}^n (\zeta_i - \zeta_i') \right\rangle \right| \leq \sigma_k \left\| \frac{1}{n}\sum_{i=1}^n (\zeta_i - \zeta_i') \right\|_2 .$$

Therefore, we get

$$\begin{aligned}
|T_1(\zeta) - T_1(\zeta')| &\leq ||\sigma||_p \left\| \frac{1}{n}\sum_{i=1}^n (\zeta_i - \zeta_i') \right\|_2 \\
&\leq \frac{||\sigma||_p}{\sqrt{n}} ||\zeta - \zeta'||_{2,Kn} ,
\end{aligned}$$

using the convexity of $x \in \mathbb{R}^K \mapsto \|x\|_2^2$, and we obtain (i). For T_2, we use the same method as for T_1 :

$$\begin{aligned}
\left|T_2(\zeta) - T_2(\zeta')\right| &\leq \|\sigma\|_p \mathbb{E}\left\| \frac{1}{n}\sum_{i=1}^n (W_i - \overline{W})(\zeta_i - \zeta_i') \right\|_2 \\
&\leq \frac{\|\sigma\|_p}{n} \sqrt{\mathbb{E}\left\| \sum_{i=1}^n (W_i - \overline{W})(\zeta_i - \zeta_i') \right\|_2^2} . \qquad (12.32)
\end{aligned}$$

Note that since $\left(\sum_{i=1}^n (W_i - \overline{W})\right)^2 = 0$, we have $\mathbb{E}(W_1 - \overline{W})(W_2 - \overline{W}) = -C_W^2/n$. We now develop $\left\|\sum_{i=1}^n (W_i - \overline{W})(\zeta_i - \zeta_i')\right\|_2^2$ in the Euclidean space $\mathbb{R}^K$:

$$\begin{aligned}
\mathbb{E}\left\| \sum_{i=1}^n (W_i - \overline{W})(\zeta_i - \zeta_i') \right\|_2^2 &= C_W^2\left(1 - n^{-1}\right)\sum_{i=1}^n \|\zeta_i - \zeta_i'\|_2^2 - \frac{C_W^2}{n}\sum_{i\neq j}\left\langle \zeta_i - \zeta_i', \zeta_j - \zeta_j' \right\rangle \\
&= C_W^2 \sum_{i=1}^n \|\zeta_i - \zeta_i'\|_2^2 - \frac{C_W^2}{n}\left\| \sum_{i=1}^n (\zeta_i - \zeta_i') \right\|_2^2 .
\end{aligned}$$

Consequently,

$$\mathbb{E}\left\| \sum_{i=1}^{n} (W_i - \overline{W})\left(\zeta_i - \zeta_i'\right) \right\|_2^2 \le C_W^2 \sum_{i=1}^{n} \left\|\zeta_i - \zeta_i'\right\|_2^2 = C_W^2 \left\|\zeta - \zeta'\right\|_{2,Kn}^2 . \tag{12.33}$$

Combining expression (12.32) and (12.33), we find that T_2 is $\|\sigma\|_p C_W/n$-Lipschitz. ■

Remark 12.26 *The proof of Proposition 12.8 is still valid under the weaker assumption (instead of exchangeability of W) that $\mathbb{E}\left[(W_i - \overline{W})(W_j - \overline{W})\right]$ can only take two possible values depending on whether or not $i = j$.*

Main results

Proof of Theorem 12.1. The case (BA)(p, M) and (SA) is obtained by combining Proposition 12.6 and McDiarmid's inequality (see for instance Fromont (2004)). The (GA) case is a straightforward consequence of Proposition 12.4 and the proof of Proposition 12.8. ■

Proof of Corollary 12.2. From Proposition 12.8 (i), with probability at least $1 - \alpha(1-\delta)$, $\phi\left(\overline{\mathbf{Y}} - \mu\right)$ is upper bounded by the minimum between $t_{\alpha(1-\delta)}$ and $\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}} - \mu\right)\right] + \frac{\|\sigma\|_p \overline{\Phi}^{-1}(\alpha(1-\delta)/2)}{\sqrt{n}}$ (because these thresholds are deterministic). In addition, Proposition 12.4 and Proposition 12.8 (ii) give that with probability at least $1 - \alpha\delta$, $\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}} - \mu\right)\right] \le \frac{\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}} - \mu\right) | \mathbf{Y}\right]}{B_W} + \frac{\|\sigma\|_p C_W}{B_W n} \overline{\Phi}^{-1}(\alpha\delta/2)$. The result follows by combining the two last expressions. ■

Monte-Carlo approximation

Proof of Proposition 12.11. The idea of the proof is to apply McDiarmid's inequality conditionally to $\mathbf{Y}$ (see McDiarmid (1989)). For any realizations W and W' of the resampling weight vector and any $\nu \in \mathbb{R}^k$,

$$\begin{aligned}\left|\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right) - \phi\left(\overline{\mathbf{Y}}_{[W'-\overline{W'}]}\right)\right| &\le \phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]} - \overline{\mathbf{Y}}_{[W'-\overline{W'}]}\right) \\ &\le \frac{c_2 - c_1}{n} \left\| \left(\sum_{i=1}^{n} \left|\mathbf{Y}_k^i - \nu_k\right| \right)_k \right\|_p\end{aligned}$$

since ϕ is sub-additive, bounded by the p-norm and $W_i - \overline{W} \in [c_1, c_2]$ a.s.

The sample $\mathbf{Y}$ being deterministic, we can take ν equal to the median M of the sample, which realizes the infimum. Since $W^1, \ldots, W^B$ are independent, McDiarmid's inequality gives (12.18).

When $\mathbf{Y}$ satisfies (GA), a proof very similar to the one of (12.15) in Proposition 12.8 can be applied to the remainder term with any deterministic ν. We then obtain (12.19). ■

12.7.2 Quantiles

Remember the following inequality coming from the definition of the quantile q_α : for any fixed $\mathbf{Y}$

$$\mathbb{P}_W\left[\phi\left(\overline{\mathbf{Y}}_{[W]}\right) > q_\alpha(\phi, \mathbf{Y})\right] \le \alpha \le \mathbb{P}_W\left[\phi\left(\overline{\mathbf{Y}}_{[W]}\right) \ge q_\alpha(\phi, \mathbf{Y})\right] . \tag{12.34}$$

Proof of Lemma 12.13. We have

$$\begin{aligned}\mathbb{P}_{\mathbf{Y}}\left[\phi(\overline{\mathbf{Y}}-\mu) > q_\alpha(\phi,\mathbf{Y}-\mu)\right] &= \mathbb{E}_W\left[\mathbb{P}_{\mathbf{Y}}\left[\phi(\overline{(\mathbf{Y}-\mu)_{[W]}}) > q_\alpha(\phi,(\mathbf{Y}-\mu)_{[W]})\right]\right]\\ &= \mathbb{E}_{\mathbf{Y}}\left[\mathbb{P}_W\left[\phi\left(\overline{(\mathbf{Y}-\mu)_{[W]}}\right) > q_\alpha(\phi,\mathbf{Y}-\mu)\right]\right]\\ &\leq \alpha\,. \end{aligned}\tag{12.35}$$

The first equality is due to the fact that the distribution of $\mathbf{Y}$ satisfies assumption (SA), hence the distribution of $(\mathbf{Y}-\mu)$ invariant by reweighting by (arbitrary) signs $W \in \{-1,1\}^n$. In the second equality we used Fubini's theorem and the fact that for any arbitrary signs W as above $q_\alpha(\phi,(\mathbf{Y}-\mu)_{[W]}) = q_\alpha(\phi,\mathbf{Y}-\mu)$; finally the last inequality comes from (12.34). ■

Proof of Proposition 12.14. Let us define the event

$$\Omega = \left\{\mathbf{Y} \mid q_\alpha(\phi,\mathbf{Y}-\mu) \leq q_{\alpha(1-\delta)}(\phi,\mathbf{Y}-\overline{\mathbf{Y}}) + f(\mathbf{Y})\right\}\;;$$

then we have using (12.35) :

$$\begin{aligned}\mathbb{P}&\left[\phi(\overline{\mathbf{Y}}-\mu) > q_{\alpha(1-\delta)}(\phi,\mathbf{Y}-\overline{\mathbf{Y}}) + f(\mathbf{Y})\right]\\ &\leq \mathbb{P}\left[\phi(\overline{\mathbf{Y}}-\mu) > q_\alpha(\phi,\mathbf{Y}-\mu)\right] + \mathbb{P}\left[\mathbf{Y}\in\Omega^c\right]\\ &\leq \alpha + \mathbb{P}\left[\mathbf{Y}\in\Omega^c\right]\;. \end{aligned}\tag{12.36}$$

We now concentrate on the event Ω^c. Using the subadditivity of ϕ, and the fact that $\overline{(\mathbf{Y}-\mu)_{[W]}} = \overline{(\mathbf{Y}-\overline{\mathbf{Y}})_{[W]}} + \overline{W}(\overline{\mathbf{Y}}-\mu)$, we have for any fixed $\mathbf{Y}\in\Omega^c$:

$$\begin{aligned}\alpha &\leq \mathbb{P}_W\left[\phi(\overline{(\mathbf{Y}-\mu)_{[W]}}) \geq q_\alpha(\phi,\mathbf{Y}-\mu)\right]\\ &\leq \mathbb{P}_W\left[\phi(\overline{(\mathbf{Y}-\mu)_{[W]}}) > q_{\alpha(1-\delta)}(\phi,\mathbf{Y}-\overline{\mathbf{Y}}) + f(\mathbf{Y})\right]\\ &\leq \mathbb{P}_W\left[\phi(\overline{(\mathbf{Y}-\overline{\mathbf{Y}})_{[W]}}) > q_{\alpha(1-\delta)}(\phi,\mathbf{Y}-\overline{\mathbf{Y}})\right] + \mathbb{P}_W\left[\phi(\overline{W}(\overline{\mathbf{Y}}-\mu)) > f(\mathbf{Y})\right]\\ &\leq \alpha(1-\delta) + \mathbb{P}_W\left[\phi(\overline{W}(\overline{\mathbf{Y}}-\mu)) > f(\mathbf{Y})\right]\;. \end{aligned}$$

For the first and last inequalities we have used (12.34), and for the second inequality the definition of Ω^c. From this we deduce that

$$\Omega^c \subset \left\{\mathbf{Y} \mid \mathbb{P}_W\left[\phi(\overline{W}(\overline{\mathbf{Y}}-\mu)) > f(\mathbf{Y})\right] \geq \alpha\delta\right\}\;.$$

Now using the homogeneity of ϕ, and the fact that both ϕ and f are nonnegative:

$$\begin{aligned}\mathbb{P}_W\left[\phi(\overline{W}(\overline{\mathbf{Y}}-\mu)) > f(\mathbf{Y})\right] &= \mathbb{P}_W\left[\left|\overline{W}\right| > \frac{f(\mathbf{Y})}{\phi(\operatorname{sign}(\overline{W})(\overline{\mathbf{Y}}-\mu))}\right]\\ &\leq \mathbb{P}_W\left[\left|\overline{W}\right| > \frac{f(\mathbf{Y})}{\widetilde{\phi}(\overline{\mathbf{Y}}-\mu)}\right]\\ &= 2\mathbb{P}\left[\frac{1}{n}(2B_{n,\frac{1}{2}}-n) > \frac{f(\mathbf{Y})}{\widetilde{\phi}(\overline{\mathbf{Y}}-\mu)}\middle|\mathbf{Y}\right]\,, \end{aligned}$$

where $B_{n,\frac{1}{2}}$ denotes a binomial $(n, \frac{1}{2})$ variable (independent of $\mathbf{Y}$). From the two last displays we conclude

$$\Omega^c \subset \left\{ \mathbf{Y} \mid \widetilde{\phi}(\overline{\mathbf{Y}} - \mu) > \frac{n}{2\overline{\mathcal{B}}\left(n, \frac{\alpha\delta}{2}\right) - n} f(\mathbf{Y}) \right\},$$

which, put back in (12.36), leads to the desired conclusion. ■

Proof of Corollary 12.16. Applying proposition 12.14 with the function

$$g(\mathbf{Y}) = \sum_{i=1}^{J-1} \gamma_i q_{(1-\delta)\alpha_i}(\widetilde{\phi}, \mathbf{Y} - \overline{\mathbf{Y}}) + \gamma_J f(\mathbf{Y}),$$

we get the following bound:

$$\begin{aligned} &\mathbb{P}_W \left[\phi(\overline{\mathbf{Y}} - \mu) > q_{(1-\delta)\alpha_0}(\phi, \mathbf{Y} - \overline{\mathbf{Y}}) + g(\mathbf{Y})\right] \\ &\leq \alpha_0 + \mathbb{P}_W \left[\phi(\overline{\mathbf{Y}} - \mu) > q_{(1-\delta)\alpha_1}(\widetilde{\phi}, \mathbf{Y} - \overline{\mathbf{Y}}) + \frac{n}{2\overline{\mathcal{B}}\left(n, \frac{\alpha_0\delta}{2}\right) - n} \left(g(\mathbf{Y}) - \gamma_1 q_{(1-\delta)\alpha_1}(\widetilde{\phi}, \mathbf{Y} - \overline{\mathbf{Y}})\right)\right]; \end{aligned} \tag{12.37}$$

note that the above left-hand side is the quantity of interest appearing in the conclusion of the theorem. Now applying repeatedly the proposition to the probabilities appearing on the right-hand side, we obtain

$$\mathbb{P}_W \left[\phi(\overline{\mathbf{Y}} - \mu) > q_{(1-\delta)\alpha_0}(\phi, \mathbf{Y} - \overline{\mathbf{Y}}) + g(\mathbf{Y})\right] \leq \sum_{i=0}^{J-1} \alpha_i + \mathbb{P}\left[\widetilde{\phi}(\overline{\mathbf{Y}} - \mu) > f(\mathbf{Y})\right],$$

as announced. ■

12.7.3 Multiple testing

Proof of Theorem 12.19. (from Romano and Wolf (2005)) We use the notations of Definition 12.17. If the procedure rejects at least one true null hypothesis, we may consider $j_0 = \min\{j \leq \hat{\ell} \mid H_{\sigma(j)}$ is true $\}$. By definition of a step-down procedure, we have $[\overline{\mathbf{Y}}_{\sigma(j_0)}] \geq t_{j_0}$. By definition of j_0, we have $\mathcal{H}_0 \subset \mathcal{C}_{j_0}$ so that, since $\mathbf{t}$ is non-decreasing, $\mathbf{t}(\mathcal{C}_{j_0}) \geq \mathbf{t}(\mathcal{H}_0)$. Finally, we can obtain (12.28) as follows:

$$\begin{aligned} \mathrm{FWER}(R) &\leq \mathbb{P}\left(\exists j_0 \mid H_{\sigma(j_0)} \text{ is true and } \left[\overline{\mathbf{Y}}_{\sigma(j_0)}\right] \geq \mathbf{t}(\mathcal{H}_0)\right) \\ &\leq \mathbb{P}\left(T'(H_0) \geq \mathbf{t}(\mathcal{H}_0)\right) \\ &\leq \mathbb{P}\left(T(H_0) \geq \mathbf{t}(\mathcal{H}_0)\right) . \end{aligned}$$

■

Proof of Proposition 12.25. First note that

$$q_{\alpha(1-\gamma)}\left(\sup_{\mathcal{H}_0} |\cdot|, \mathbf{Y}\right) \le q_{\alpha(1-\gamma)}(\|\cdot\|_\infty, \mathbf{Y}-\mu) \ .$$

Recall that from the proof of Proposition 12.14, with probability larger than $1-\alpha\gamma$ we have

$$q_{\alpha(1-\gamma)}(\|\cdot\|_\infty, \mathbf{Y}-\mu) \le q_{\alpha(1-\delta)(1-\gamma)}\left(\|\cdot\|_\infty, \mathbf{Y}-\overline{\mathbf{Y}}\right) + \varepsilon'(\alpha,\delta,\gamma,n,K) \ .$$

Take $\mathbf{Y}$ in the event where the above inequality holds. If the global procedure rejects at least one true null hypothesis, we denote j_0 the first time that this occurs ($j_0 = 0$ if it is in the first step). There are two cases:

- if $j_0 = 0$ then we have

$$T(\mathcal{H}_0) \ge q_{\alpha(1-\delta)(1-\gamma)}\left(\|\cdot\|_\infty, \mathbf{Y}-\overline{\mathbf{Y}}\right) + \varepsilon'(\alpha,\delta,\gamma,n,K) \ge q_{\alpha(1-\gamma)}\left(\sup_{\mathcal{H}_0}|\cdot|, \mathbf{Y}\right)$$

- if $j_0 \ge 1$, following the proof of Theorem 12.19, $T(\mathcal{H}_0) \ge q_{\alpha(1-\gamma)}\left(\sup_{\mathcal{H}_0}|\cdot|, \mathbf{Y}\right)$.

In both cases, $T(\mathcal{H}_0) \ge q_{\alpha(1-\gamma)}\left(\sup_{\mathcal{H}_0}|\cdot|, \mathbf{Y}\right)$, which occurs with probability smaller than $\alpha(1-\gamma)$. ■

12.7.4 Exchangeable resampling computations

In this section, we compute constants A_W, B_W, C_W and D_W (defined by (12.3) to (12.6)) for some exchangeable resamplings. This implies all the statements in Tab. 12.1. We first define several additional exchangeable resampling weights:

- **Bernoulli** (p), $p \in (0,1)$: pW_i i.i.d. with a Bernoulli distribution of parameter p. A classical choice is $p = \frac{1}{2}$.
- **Efron** (q), $q \in \{1\ldots,n\}$: $qn^{-1}W$ has a multinomial distribution with parameters $(q; n^{-1}, \ldots, n^{-1})$. A classical choice is $q = n$.
- **Poisson** (μ), $\mu \in (0,+\infty)$: μW_i i.i.d. with a Poisson distribution of parameter μ. A classical choice is $\mu = 1$.

Notice that $\overline{Y}_{[W-\overline{W}]}$ and all the resampling constants are invariant under translation of the weights, so that Bernoulli (1/2) weights are completely equivalent to Rademacher weights in this chapter.

Lemma 12.27 *1. Let W be Bernoulli (p) weights with $p \in (0,1)$. Then,*

$$2(1-p) - \sqrt{\frac{1-p}{pn}} \le A_W \le B_W \le \sqrt{\frac{1}{p}-1}\sqrt{1-\frac{1}{n}}$$

$$C_W = \sqrt{\frac{1}{p}-1} \qquad \text{and} \qquad D_W \le \frac{1}{2p} + \left|\frac{1}{2p}-1\right| + \sqrt{\frac{1-p}{np}} \ .$$

2. *Let W be Efron (q) weights with $q \in \{1, \dots, n\}$. Then,*

$$A_W \le B_W \le \sqrt{\frac{n-1}{n}} \qquad \text{and} \qquad C_W = 1 \ .$$

Moreover, if $q \le n$,

$$A_W = 2\left(1 - \frac{1}{n}\right)^q \ .$$

3. *Let W be Poisson (μ) weights with $\mu > 0$. Then,*

$$A_W \le B_W \le \frac{1}{\sqrt{\mu}}\sqrt{1 - \frac{1}{n}} \qquad \text{and} \qquad C_W = \frac{1}{\sqrt{\mu}} \ .$$

Moreover, if $\mu = 1$,

$$\frac{2}{e} - \frac{1}{\sqrt{n}} \le A_W \ .$$

4. *Let W be Random hold-out (q) weights with $q \in \{1, \dots, n\}$. Then,*

$$A_W = 2\left(1 - \frac{q}{n}\right) \qquad B_W = \sqrt{\frac{n}{q} - 1}$$

$$C_W = \sqrt{\frac{n}{n-1}}\sqrt{\frac{n}{q} - 1} \qquad \text{and} \qquad D_W = \frac{n}{2q} + \left|1 - \frac{n}{2q}\right| \ .$$

Proof of Lemma 12.27. We consider the following cases:

General case We first only assume that W is exchangeable. Then, from the concavity of $\sqrt{\cdot}$ and the triangular inequality, we have

$$\begin{aligned}\mathbb{E}\left|W_1 - \mathbb{E}[W_1]\right| - \sqrt{\mathbb{E}\left(\overline{W} - \mathbb{E}[W_1]\right)^2} &\le \mathbb{E}\left|W_1 - \mathbb{E}[W_1]\right| - \mathbb{E}\left|\overline{W} - \mathbb{E}[W_1]\right| \\ &\le A_W \le B_W \le \sqrt{\frac{n-1}{n}} C_W \ . \end{aligned} \tag{12.38}$$

Independent weights When we suppose that the W_i are i.i.d.,

$$\mathbb{E}\left|W_1 - \mathbb{E}[W_1]\right| - \frac{\sqrt{\mathrm{Var}(W_1)}}{\sqrt{n}} \le A_W \qquad \text{and} \qquad C_W = \sqrt{\mathrm{Var}(W_1)} \ . \tag{12.39}$$

Bernoulli These weights are i.i.d. with $\mathrm{Var}(W_1) = p^{-1} - 1$, $\mathbb{E}[W_1] = 1$ and

$$\mathbb{E}\left|W_1 - 1\right| = p\left(p^{-1} - 1\right) + (1 - p) = 2(1 - p) \ .$$

With (12.38) and (12.39), we obtain the bounds for A_W, B_W and C_W. Moreover, Bernoulli (p) weights satisfy the assumption of (12.6) with $x_0 = a = (2p)^{-1}$. Then,

$$D_W = \frac{1}{2p} + \mathbb{E}\left|\overline{W} - \frac{1}{2p}\right| \le \frac{1}{2p} + \left|1 - \frac{1}{2p}\right| + \mathbb{E}\left|\overline{W} - 1\right| \le \frac{1}{2p} + \frac{1}{p}\left|\frac{1}{2} - p\right| + \sqrt{\frac{1-p}{np}} \ .$$

Efron We have $\overline{W} = 1$ a.s. so that

$$C_W = \sqrt{\frac{n}{n-1}} \operatorname{Var}(W_1) = 1 \enspace .$$

If moreover $q \leq n$, then $W_i < 1$ implies $W_i = 0$ and

$$\begin{aligned} A_W &= \mathbb{E}\left|W_1 - 1\right| = \mathbb{E}\left[W_1 - 1 + 2\mathbf{1}\{W_1 = 0\}\right] \\ &= 2\mathbb{P}(W_1 = 0) = 2\left(1 - \frac{1}{n}\right)^q \enspace . \end{aligned}$$

The result follows from (12.38).

Poisson These weights are i.i.d. with $\operatorname{Var}(W_1) = \mu^{-1}$, $\mathbb{E}[W_1] = 1$. Moreover, if $\mu \leq 1$, $W_i < 1$ implies $W_i = 0$ and

$$\mathbb{E}\left|W_1 - 1\right| = 2\mathbb{P}(W_1 = 0) = 2e^{-\mu} \enspace .$$

With (12.38) and (12.39), the result follows.

Random hold-out These weights are such that $\{W_i\}_{1 \leq i \leq n}$ takes only two values, with $\overline{W} = 1$. Then, A_W, B_W and C_W can be directly computed. Moreover, they satisfy the assumption of (12.6) with $x_0 = a = n/(2q)$. The computation of D_W is straightforward. ■

12.7.5 Non-exchangeable weights

In Section 12.2.4, we considered non-exchangeable weights in order to reduce the complexity of computation of expectations w.r.t. the resampling randomness. Then, we are mainly interested in non-exchangeable weights with small support. This is why we focus on the two following cases:

1. deterministic weights

2. V-fold weights ($V \in \{2, \ldots, n\}$) : let $(B_j)_{1 \leq j \leq V}$ be a partition of $\{1, \ldots, n\}$ and $W^B \in \mathbb{R}^V$ an exchangeable resampling weight vector of size V. Then, for any $i \in \{1, \ldots, n\}$ with $i \in B_j$, define $W_i = W_j^B$.

We will often assume that the partition $(B_j)_{1 \leq j \leq V}$ is "regular", *i.e.* that V divides n and $\operatorname{Card}(B_j) = n/V$ for every $j \in \{1, \ldots, V\}$. When V does not divide n, the B_j can be chosen approximatively of the same size.

In the following, we make use of five constants that depend only on the resampling scheme: B_W and D_W stay unchanged (see definitions (12.4) and (12.6)), we modify the definitions of A_W and C_W (notice that we stay consistent with (12.3) and (12.5) when W is exchangeable), and we introduce a fifth constant E_W (which is equal to A_W in the exchangeable case):

$$A_W := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left|W_i - \overline{W}\right| \tag{12.40}$$

$$C_W := \sqrt{n} B_W \qquad \text{if } W \text{ is deterministic} \tag{12.41}$$

$$C_W := \sqrt{\max_j \operatorname{Card}(B_j) C_{W^B} + \sqrt{n}\mathbb{E}\left|\overline{W^B} - \overline{W}\right|} \qquad \text{if } W \text{ is } V\text{-fold} \tag{12.42}$$

$$E_W := \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(\mathbb{E}|W_i - \overline{W}|\right)^2} \ . \tag{12.43}$$

We can now state the main theorem of this section.

Theorem 12.28 *Let W be either a deterministic or V-fold resampling weight vector, and define the constants A_W, B_W, C_W, D_W and E_W by* (12.40), (12.4), (12.41), (12.42), (12.6) *and* (12.43). *Then, all the results of Theorem 12.1 and Corollary 12.2 hold, with only a slight modification in* (12.8)*:*

$$\phi\left(\overline{\mathbf{Y}} - \mu\right) < \frac{\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\middle|\mathbf{Y}\right]}{A_W} + \frac{M}{\sqrt{n}}\sqrt{1 + \frac{A_W^2}{E_W^2}}\sqrt{2\log(1/\alpha)} \ .$$

Proof of Theorem 12.28. In the gaussian case, we use the same proof as Theorem 12.1 and Corollary 12.2, but we replace the concentration result (12.16) by the one of Proposition 12.29.

In the bounded case, the proof is identical (it relies on Mc Diarmid inequality), but we no longer have $A_W = E_W$ because the weights are non-exchangeable. ■

When V divides n, we can compute the constants for regular V-fold weights:

$$A_W = E_W = A_{W^B} \qquad B_W = B_{W^B} \qquad C_W = \sqrt{\frac{n}{V}} C_{W^B} \ .$$

We now give two natural examples of non-exchangeable weights:

1. **Hold-out** (q) : $W_i = \frac{n}{q}\mathbf{1}\{i \in I\}$ for some deterministic subset $I \subset \{1, \ldots, n\}$ of cardinality q. A classical choice is $q = \lfloor n/2 \rfloor$.

2. **(Possibly non-regular) V-fold cross validation**, $V \in \{2, \ldots, n\}$: V-fold weights with W^B leave-one-out (which is often called cross-validation). More precisely, $W_i = \frac{V}{V-1}\mathbf{1}\{i \notin B_J\}$, J uniform on $\{1, \ldots, V\}$), $(B_j)_{1\leq j\leq V}$ partition of $\{1, \ldots, n\}$.

The terms "hold-out", "cross-validation" and "V-fold cross-validation" refer to slightly different procedures which inspired these weights. In those two cases, we can compute the resampling constants :

1. **Hold-out** (q) :

$$A_W = 2\left(1 - \frac{q}{n}\right) \qquad B_W = E_W = \sqrt{\frac{n}{q} - 1}$$

$$C_W = \sqrt{n\left(\frac{n}{q} - 1\right)} \qquad \text{and} \qquad D_W = \frac{n}{2q} + \left|1 - \frac{n}{2q}\right| \ .$$

2. **(Possibly non-regular) V-fold cross validation** :

$$A_W = \frac{2}{V-1} \sum_{j=1}^{V} \frac{\operatorname{Card}(B_j)}{n} \left(1 - \frac{\operatorname{Card}(B_j)}{n}\right)$$

$$B_W = \frac{1}{V-1} \sum_{j=1}^{V} \sqrt{\frac{\operatorname{Card}(B_j)}{n} \left(1 - \frac{\operatorname{Card}(B_j)}{n}\right)}$$

$$C_W = \sqrt{\max_j \operatorname{Card}(B_j)} \frac{\sqrt{V}}{V-1} + \frac{\sqrt{n}}{V-1} \sum_{j=1}^{V} \left| \frac{\operatorname{Card}(B_j)}{n} - \frac{1}{V} \right|$$

$$D_W = \frac{1}{V-1} \sum_{j=1}^{V} \left(\frac{1}{2} + \left| \frac{1}{2} - \frac{\operatorname{Card}(B_j)}{n} \right| \right)$$

$$E_W = \frac{2}{V-1} \sqrt{\sum_{j=1}^{V} \frac{\operatorname{Card}(B_j)}{n} \left(1 - \frac{\operatorname{Card}(B_j)}{n}\right)^2} \ .$$

When the partition $(B_j)_{1 \le j \le V}$ is almost regular, *i.e.* $\max_j \left|\operatorname{Card}(B_j) - nV^{-1}\right| \le 1$ and $n \gg V \ge 3$, then $C_W B_W^{-1} \le \sqrt{n/(V-1)}\,(1 + o(1))$ which is close to its value in the "regular" case. This means that the concentration thresholds behave as in the regular case provided that n is large enough.

The proofs of these results are given at the end of this section. Before this, we give analogues of the results of Section 12.2.1 and 12.2.2 in the non-exchangeable case.

Expectations

The Proposition 12.4 is valid with non-exchangeable weights. The proof of Proposition 12.6 remains unchanged with non-exchangeable weights, with A_W defined by (12.40).

Concentration inequalities

Whereas Proposition 12.8 deals only with exchangeable weights, we can derive a similar result for deterministic and V-fold exchangeable weights. This is the object of the following result.

Proposition 12.29 *Let $p \in [1, +\infty]$, $\mathbf{Y}$ a sample satisfying (GA) and $\phi : \mathbb{R}^K \to \mathbb{R}$ any subadditive function, bounded by the p-norm. Let W be some resampling weight vector among*

(i) Deterministic weights.

(ii) V-fold exchangeable resampling weight for some $V \in \{2, \dots, n\}$.

Then, for all $\alpha \in (0,1)$, (12.16) *and the corresponding lower bound hold with C_W defined by* (12.41) *(deterministic case)* (12.42) *(V-fold case).*

Proof of Proposition 12.29. Deterministic weights (i): we can use (12.15) and the corresponding lower bound with $B_W\sigma$ instead of σ because $\mathcal{D}\big(\overline{\mathbf{Y}}_{[W-\overline{W}]}\big) = \mathcal{D}(B_W(\overline{\mathbf{Y}} - \mu))$. The result follows with $C_W = \sqrt{n}B_W$.

V-fold weights (ii): the proof is widely inspired from the one of Proposition 12.8. We have to compute the Lipschitz constant of T_2 defined by

$$T_2(\zeta) = \mathbb{E}\phi\left(\frac{1}{n}\sum_{i=1}^{n}\left(W_i - \overline{W}\right)A\zeta_i\right) .$$

For all $\zeta, \zeta' \in \mathbb{R}^K$, we use the triangular inequality and the same arguments as in the proof of Proposition 12.8:

$$\begin{aligned}
\left|T_2(\zeta) - T_2(\zeta')\right| &\leq \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(W_i - \overline{W}\right)A(\zeta_i - \zeta_i')\right\|_p \\
&\leq \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(W_i - \overline{W^B}\right)A(\zeta_i - \zeta_i')\right\|_p + \mathbb{E}\left|\overline{W^B} - \overline{W}\right|\left\|\frac{1}{n}\sum_{i=1}^{n}A(\zeta_i - \zeta_i')\right\|_p \\
&\leq \frac{\|\sigma\|_p}{n}\sqrt{\mathbb{E}\left\|\sum_{i=1}^{n}(W_i - \overline{W^B})(\zeta_i - \zeta_i')\right\|_2^2} + \mathbb{E}\left|\overline{W^B} - \overline{W}\right|\frac{\|\sigma\|_p}{\sqrt{n}}\left\|\zeta - \zeta'\right\|_{2,Kn}
\end{aligned}$$

Using the exchangeability of the W^B, we show that

$$\begin{aligned}
\mathbb{E}\left\|\sum_{i=1}^{n}(W_i - \overline{W^B})(\zeta_i - \zeta_i')\right\|_2^2 &= \mathbb{E}\left\|\sum_{j=1}^{V}\left(W_j^B - \overline{W^B}\right)\sum_{i\in B_j}(\zeta_i - \zeta_i')\right\|_2^2 \\
&\leq C_{W^B}^2\sum_{j=1}^{V}\left\|\sum_{i\in B_j}(\zeta_i - \zeta_i')\right\|_2^2 \\
&\leq C_{W^B}^2\sum_{j=1}^{V}\operatorname{Card}(B_j)\sum_{i\in B_j}\left\|\zeta_i - \zeta_i'\right\|_2^2
\end{aligned}$$

by convexity of $\|\cdot\|_2^2$. Finally, this implies that T_2 is Lipschitz of parameter

$$\frac{\|\sigma\|_p}{n}\sqrt{\max_j \operatorname{Card}(B_j)C_{W^B}} + \frac{\|\sigma\|_p}{\sqrt{n}}\mathbb{E}\left|\overline{W^B} - \overline{W}\right| .$$

■

Computation of the constants

We first remark that the following statements are straightforward:

- if W is deterministic, $B_W = E_W$.
- if W is regular V-fold exchangeable,

$$A_W = E_W = A_{W^B} \qquad B_W = B_{W^B} \qquad C_W = \sqrt{\frac{n}{V}} C_{W^B} .$$

In the hold-out (q) case, we compute A_W, B_W and D_W exactly as in the Random hold-out (q) case.

In the general V-fold cross-validation case, we use the following trick : conditionally to the index J of the removed block, W is a deterministic hold-out ($n - \mathrm{Card}(B_J)$) weight multiplied by a factor $c(J) = \frac{V(n-\mathrm{Card}(B_J))}{(V-1)n}$. This allows to compute A_W, B_W and D_W from the hold-out case: for instance,

$$\begin{aligned} A_W &= \frac{1}{V} \sum_{j=1}^{V} \left[2c(J) \left(1 - \frac{q}{n} \right) \right] \\ &= \frac{2}{V-1} \sum_{j=1}^{V} \frac{\mathrm{Card}(B_j)}{n} \left(1 - \frac{\mathrm{Card}(B_j)}{n} \right) \quad . \end{aligned}$$

This also shows

$$\mathbb{E} \left| \overline{W^B} - \overline{W} \right| = \frac{1}{V} \sum_{j=1}^{V} \left| \frac{V}{V-1} \frac{n - \mathrm{Card}(B_J)}{n} - 1 \right|$$

from which we obtain C_W. The computation of E_W is done directly by noting that

$$\mathbb{E} \left| W_j^B - \overline{W} \right| = \frac{V}{V-1} \mathbb{E} \left| \mathbf{1}\{j \neq J\} - 1 + \frac{\mathrm{Card}(B_J)}{n} \right| = \frac{2}{V-1} \left(1 - \frac{\mathrm{Card}(B_j)}{n} \right) \quad ,$$

$$\begin{aligned} E_W^2 &= \frac{1}{n} \sum_{i=1}^{n} (\mathbb{E} \left| W_i - \overline{W} \right|)^2 \\ &= \sum_{j=1}^{V} \frac{\mathrm{Card}(B_j)}{n} (\mathbb{E} \left| W_j^B - \overline{W} \right|)^2 \\ &= \left(\frac{2}{V-1} \right)^2 \sum_{j=1}^{V} \frac{\mathrm{Card}(B_j)}{n} \left(1 - \frac{\mathrm{Card}(B_j)}{n} \right)^2 \quad . \end{aligned}$$

We now prove the last statement about "almost regular" V-fold cross-validation: when $\max_j \mathrm{Card}(B_j) \leq nV^{-1} + 1$,

$$\begin{aligned} C_W &\leq \sqrt{\frac{n}{V} + 1} \frac{\sqrt{V}}{V-1} + \frac{V\sqrt{n}}{n(V-1)} \\ &\leq \frac{\sqrt{n}}{V-1} \left(1 + \sqrt{\frac{V}{n}} + \frac{V}{n} \right) \quad . \end{aligned}$$

If moreover $V^{-1} + n^{-1} \leq 1/2$, we have:

$$
\begin{aligned}
B_W &\geq \frac{V}{V-1}\sqrt{\left(\frac{1}{V}-\frac{1}{n}\right)\left(1-\frac{1}{V}+\frac{1}{n}\right)} \\
&= \frac{1}{\sqrt{V-1}}\sqrt{1+\frac{V^2}{(V-1)n}\left(\frac{2}{V}-1-\frac{1}{n}\right)} \\
&\geq \frac{1}{\sqrt{V-1}} - \frac{V}{(V-1)\sqrt{n}}\sqrt{\left(1+\frac{1}{n}-\frac{2}{V}\right)_+} \enspace .
\end{aligned}
$$

Conclusion générale

Cette thèse a présenté dans une première partie des outils pour la recherche de motifs exceptionnels dans une séquence d'ADN. Contrairement aux méthodes classiques, ces nouveaux outils prennent en compte une certaine hétérogénéité (connue a priori) dans la séquence, présentée sous la forme d'une segmentation qui peut être déterministe (connue) ou aléatoire (de loi markovienne connue). Nous avons proposé plusieurs approximations de Poisson composée pour la loi du comptage valables pour des motifs rares ; leur qualité dépend de conditions spécifiques concernant le nombre de ruptures ou la longueur des segments de la segmentation. Les paramètres des lois de Poisson composées sont explicites et l'erreur d'approximation (calculée en variation totale) est contrôlée en utilisant la méthode de Chen-Stein.

Dans le cas d'une segmentation déterministe (souvent le cas en pratique pour les séquences biologiques), les nouvelles approximations ont été implémentées dans une extension du logiciel R'MES[7]. Nous montrons sur plusieurs exemples que lorsque la séquence est suffisamment hétérogène, l'approche homogène n'est pas satisfaisante et l'hétérogénéité doit être prise en compte dans l'évaluation de l'exceptionnalité.

Par ailleurs, le fait de chercher simultanément les motifs exceptionnels parmi tous les motifs d'une longueur donnée pose un problème de multiplicité. Nous avons proposé de résoudre ce problème par une approche de tests multiples, en utilisant des procédures contrôlant la probabilité qu'il y ait au moins k motifs non-exceptionnels parmi les motifs sélectionnés (k-FWER). Cette démarche pour l'analyse des séquences d'ADN ouvre de nombreuses perspectives, comme la mise en place de procédures contrôlant le taux moyen de motifs non-exceptionnels parmi les motifs sélectionnés (FDR) (critère souvent préféré en pratique). Ceci m'a conduit à considérer le problème théorique du contrôle des taux d'erreurs pour les procédures de tests multiples.

La deuxième partie de cette thèse a présenté des contributions à la théorie et à la méthodologie des tests multiples. Nous avons proposé des preuves concises pour les résultats classiques du contrôle du FDR (entre autres pour les procédures de Benjamini and Hochberg (1995), Benjamini and Yekutieli (2001) et Benjamini *et al.* (2006)). En particulier, nous avons introduit une condition "d'auto-consistance" (satisfaite par les procédures *step-up*) qui dépend d'une fontion β (*shape function*) et nous avons montré que certains choix de β permettent le contrôle du FDR lorsque les p-values sont indépendantes, positivement dépendantes (PRDS) ou de structure de dépendance quelconque.
Nous avons également proposé de nouvelles procédures de tests multiples adaptatives à π_0 (la

[7]http ://genome.jouy.inra.fr/ssb/rmes

proportion d'hypothèses nulles qui sont vraies) qui contrôlent le FDR : lorsque les p-values sont indépendantes, nous avons amélioré la procédure de Benjamini *et al.* (2006). Lorsque les p-values peuvent avoir des dépendances, nous avons proposé des procédures adaptatives contrôlant rigoureusement le FDR. Cependant, ces dernières sont uniquement utiles dans des situations où une large proportion des hypothèses nulles est rejetée.

Finalement, afin de prendre en compte la structure de dépendance entre les p-values, nous avons proposé des procédures par rééchantillonnage. Ces procédures permettent d'effectuer simultanément des tests unilatéraux (ou bilatéraux) sur les moyennes d'un vecteur gaussien, ou alternativement d'un vecteur de loi symétrique borné, tout en garantissant un contrôle du FWER. L'originalité de cette étude est qu'elle a été réalisée dans un cadre de rééchantillonnage non-asymptotique "approché", ce qui a exigé un contrôle de termes d'erreurs. Dans la continuité de ce travail, une perspective intéressante serait d'utiliser des procédures de rééchantillonnage pour contrôler rigoureusement et non-asymptotiquement le FDR (ou d'autres quantités liées au taux de fausses découvertes).

Bibliographie

Arlot, S. (2007). *Rééchantillonnage et Sélection de modèles.* PhD thesis, Université Paris XI.

Arlot, S., Blanchard, G. and Roquain, E. (2007). Resampling-based confidence regions and multiple tests for a correlated random vector. In *Learning theory*, volume 4539 of *Lecture Notes in Comput. Sci.*, 127–141. Springer, Berlin.

Arratia, R., Goldstein, L. and Gordon, L. (1989). Two moments suffice for Poisson approximations : the Chen-Stein method. *Ann. Prob.* **17** 9–25.

Arratia, R., Goldstein, L. and Gordon, L. (1990). Poisson approximation and the Chen-Stein method. *Statistical Science.* **5** 403–434.

Baraud, Y., Huet, S. and Laurent, B. (2003). Adaptive tests of linear hypotheses by model selection. *Ann. Statist.* **31 (1)** 225–251.

Baraud, Y., Huet, S. and Laurent, B. (2005). Testing convex hypotheses on the mean of a Gaussian vector. Application to testing qualitative hypotheses on a regression function. *Ann. Statist.* **33 (1)** 214–257.

Barbour, A. D., Holst, L. and Janson, S. (1992). *Poisson approximation.* Oxford-University Press.

Benjamini, Y., Krieger, A. M. and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika.* **93 (3)** 491–507.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate : a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B.* **57 (1)** 289–300.

Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery fate in multiple testing with independent statistics. *J. Behav. Educ. Statist.* **25** 60–83.

Benjamini, Y. and Liu, W. (1999a), A distribution-free multiple test procedure that controls the false discovery rate. Technical report, Dept. of statistics, University of Tel-Aviv.

Benjamini, Y. and Liu, W. (1999b). A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *J. Statist. Plann. Inference.* **82 (1-2)** 163–170.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29 (4)** 1165–1188.

Black, M. A. (2004). A note on the adaptive control of false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66 (2)** 297–304.

Blanchard, G. and Fleuret, F. (2007). Occam's hammer : a link between randomized learning and multiple testing FDR control. ArXiV preprint math.ST/0608713.

Blom, G. and Thorburn, D. (1982). How many random digits are required until given sequences are obtained ? *J. Appl. Prob.* **19** 518–531.

Chrysaphinou, O. and Papastavridis, S. (1990). The occurrence of sequence patterns in repeated dependent experiments. *Teor. Veroyatnost. i Primenen.* **35 (1)** 167–173.

Chryssaphinou, O., Papastavridis, S. and Vaggelatou, E. (2001). Poisson approximation for the non-overlapping appearances of several words in Markov chains. *Combinatorics, Probability and Computing.* **10** 293–308.

Dacunha-Castelle, D. and Duflo, M. (1983). *Probabilités et statistiques. Tome 2.* Collection Mathématiques Appliquées pour la Maîtrise. [Collection of Applied Mathematics for the Master's Degree]. Masson, Paris, Problèmes à temps mobile. [Movable-time problems].

Darvas, F., Rautiainen, M., Pantazis, D., Baillet, S., Benali, H., Mosher, J., Garnero, L. and Leahy, R. (2005). Investigations of dipole localization accuracy in meg using the bootstrap. *NeuroImage.* **25** 355–368.

Dudoit, S., Shaffer, J. P. and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statist. Sci.* **18 (1)** 71–103.

Dudoit, S., van der Laan, M. J. and Pollard, K. S. (2004). Multiple testing. I. Single-step procedures for control of general type I error rates. *Stat. Appl. Genet. Mol. Biol.* **3** Art. 13, 71 pp. (electronic).

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998). *Biological sequences analysis.* Cambridge University Press.

Efron, B. (1979). Bootstrap methods : another look at the jackknife. *Ann. Statist.* **7 (1)** 1–26.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96 (456)** 1151–1160.

Farcomeni, A. (2007). Some results on the control of the false discovery rate under dependence. *Scandinavian Journal of Statistics.* **34 (2)** 275–297.

Finner, H. and Roters, M. (1998). Asymptotic comparison of step-down and step-up multiple test procedures based on exchangeable test statistics. *Ann. Statist.* **26 (2)** 505–524.

Finner, H. and Roters, M. (2001). On the false discovery rate and expected type I errors. *Biom. J.* **43 (8)** 985–1005.

Fisher, R. A. (1935). *The Design of Experiments.* Oliver and Boyd, Edinburgh.p.

Fromont, M. (2004). Model selection by bootstrap penalization for classification. In *Learning theory*, volume 3120 of *Lecture Notes in Comput. Sci.*, 285–299. Springer, Berlin.

Ge, Y., Dudoit, S. and Speed, T. P. (2003). Resampling-based multiple testing for microarray data analysis. *Test.* **12 (1)** 1–77. With comments and a rejoinder by the authors.

Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64 (3)** 499–517.

Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *Ann. Statist.* **32 (3)** 1035–1061.

Genovese, C. R., Roeder, K. and Wasserman, L. (2006). False discovery control with p-value weighting. *Biometrika.* **93 (3)** 509–524.

Godbole, A. P. (1991). Poisson approximations for runs and patterns of rare events. *Adv. in Appl. Probab.* **23 (4)** 851–865.

Hall, P. (1992). *The bootstrap and Edgeworth expansion.* Springer Series in Statistics. Springer-Verlag, New York.

Hall, P. and Mammen, E. (1994). On general resampling algorithms and their performance in distribution estimation. *Ann. Statist.* **22 (4)** 2011–2030.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika.* **75 (4)** 800–802.

Hoebeke, M. and Schbath, S. (2006). R'mes : Finding exceptional motifs, version 3. user guide. http ://genome.jouy.inra.fr/ssb/rmes.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6 (2)** 65–70.

Hommel, G. (1983). Tests of the overall hypothesis for arbitrary dependence structures. *Biometrical J.* **25 (5)** 423–430.

Jerbi, K., Lachaux, J.-P., N'Diaye, K., Pantazis, D., Leahy, R. M., Garnero, L. and Baillet, S. (2007). Coherent neural representation of hand speed in humans revealed by meg imaging. *PNAS.* **104 (18)** 7676–7681.

Karlin, S. and Rinott, Y. (1980). Classes of orderings of measures and related correlation inequalities. I. Multivariate totally positive distributions. *J. Multivariate Anal.* **10 (4)** 467–498.

Lehmann, E. L., Romano, J. P. and Schaffer, J. P. (2005). On optimality of stepdown and stepup multiple test procedures. *The Annals of Statistics.* **33** 1084–1108.

Lehmann, E. L. and Romano, J. P. (2005a). Generalizations of the familywise error rate. *The Annals of Statistics.* **33** 1138–1154.

Lehmann, E. L. and Romano, J. P. (2005b). *Testing statistical hypotheses.* Springer Texts in Statistics. Springer, New York, third edition.

LOTHAIRE, M. (2005). *Applied combinatorics on words.* Cambridge University Press.

MASON, D. M. and NEWTON, M. A. (1992). A rank statistics approach to the consistency of a general bootstrap. *Ann. Statist.* **20 (3)** 1611–1624.

MASSART, P. (2005). Concentration inequalities and model selection (lecture notes of the St-Flour probability summer school 2003). Available online at `http ://www.math.u-psud.fr/~massart/stf2003_massart.pdf`.

MCDIARMID, C. (1989). On the method of bounded differences. In *Surveys in combinatorics*, volume 141 of *London Mathematical Society Lecture Notes*, 148–188.

MELO DE LIMA, C. (2005). *Développement d'une approche markovienne pour l'analyse de l'organisation spatiale des génomes.* PhD thesis, Université Claude Bernard, Lyon I.

MEYN, S. P. and TWEEDIE, R. L. (1993). *Markov chains and stochastic stability.* Communications and Control Engineering Series. Springer-Verlag London Ltd., London.

MONFORT, A. (1997). *Cours de Statistique Mathematique.* Economica.

MURI, F. (1997). *Comparaison d'algorithmes d'identification de chaînes de Markov cachées et application à la détection de régions homogènes dans les séquences d'ADN.* PhD thesis, Université René Descartes, Paris V.

PANTAZIS, D., NICHOLS, T. E., BAILLET, S. and LEAHY, R. M. (2005). A comparison of random field theory and permutation methods for statistical analysis of meg data. *NeuroImage.* **25** 383–394.

PERONE PACIFICO, M., GENOVESE, C., VERDINELLI, I. and WASSERMAN, L. (2004). False discovery control for random fields. *J. Amer. Statist. Assoc.* **99 (468)** 1002–1014.

POLITIS, D. N., ROMANO, J. P. and WOLF, M. (1999). *Subsampling.* Springer Series in Statistics. Springer-Verlag, New York.

POLLARD, K. S. and VAN DER LAAN, M. J. (2003), Resampling-based multiple testing : Asymptotic control of type i error and applications to gene expression data. Working Paper Series Working Paper 121, U.C. Berkeley Division of Biostatistics. available at http ://www.bepress.com/ucbbiostat/paper121.

PRÆSTGAARD, J. and WELLNER, J. A. (1993). Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.* **21 (4)** 2053–2086.

PRUM, B., RODOLPHE, F. and TURCKHEIM, É. (1995). Finding words with unexpected frequencies in DNA sequences. *J. R. Statist. Soc. B.* **57** 205–220.

REINERT, G., SCHBATH, S. and WATERMAN, M. (2000). Probabilistic and statistical properties of words. *J. Comp. Biol.* **7** 1–46.

REINERT, G. and SCHBATH, S. (1998). Compound Poisson and Poisson process approximations for occurrences of multiple words in Markov chains. *J. Comp. Biol.* **5** 223–253.

Robin, S. and Daudin, J. J. (1999). Exact distribution of word occurrences in a random sequence of letters. *J. Appl. Probab.* **36 (1)** 179–193.

Robin, S., Rodolphe, F. and Schbath, S. (2003a). *ADN, mots et modèles.* BELIN.

Robin, S., Rodolphe, F. and Schbath, S. (2003b). *DNA, Words and Models.* Cambridge University Press.

Robin, S. and Schbath, S. (2001). Numerical comparison of several approximations of the word count distribution in random sequences. *J. Comp. Biol.* **8** 349–359.

Robin, S., Bar-Hen, A., Daudin, J.-J. and Pierre, L. (2007). A semi-parametric approach for mixture models : Application to local false discovery rate estimation. *Comput. Stat. Data Anal.* **51 (12)** 5483–5493.

Romano, J. P. (1989). Bootstrap and randomization tests of some nonparametric hypotheses. *Ann. Statist.* **17 (1)** 141–159.

Romano, J. P. (1990). On the behavior of randomization tests without a group invariance assumption. *J. Amer. Statist. Assoc.* **85 (411)** 686–692.

Romano, J. P. and Shaikh, A. M. (2006a). On stepdown control of the false discovery proportion. volume 49 of *Lecture Notes-Monograph Series*, 33–50.

Romano, J. P. and Shaikh, A. M. (2006b). Stepup procedures for control of generalizations of the familywise error rate. *Ann. Statist.* **34 (4)** 1850–1873.

Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *J. Amer. Statist. Assoc.* **100 (469)** 94–108.

Romano, J. P. and Wolf, M. (2007). Control of generalized error rates in multiple testing. *Ann. Statist.* **35 (4)** 1378–1408.

Roquain, E. and Schbath, S. (2007). Improved compound poisson approximation for the number of occurrences of any rare word family in stationnary markov chain. *Adv. Appl. Prob.* **39** 128–140.

Régnier, M. (2000). A unified approach to word occurrence probabilities. *Discrete Applied Mathematics.* **104** 259–280.

Sarkar, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Ann. Statist.* **30 (1)** 239–257.

Sarkar, S. K. (2006), Two-stage stepup procedures controlling fdr. Technical report.

Sarkar, S. K. and Guo, W. (2006), Procedures controlling generalized false discovery rate. Technical report.

Schbath, S. (1995a). Compound poisson approximation of word counts in DNA sequences. *ESAIM : Probability and Statistics.* **1** 1–16.

Schbath, S. (1995b). *Étude 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, Université René Descartes, Paris V.

Senoussi, R. (1990). Statistique asymptotique presque sûre de modèles statistiques convexes. *Ann. Inst. H. Poincaré Probab. Statist.* **26 (1)** 19–44.

Storey, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64 (3)** 479–498.

Storey, J. D. (2003). The positive false discovery rate : a Bayesian interpretation and the q-value. *Ann. Statist.* **31 (6)** 2013–2035.

Storey, J. D. (2005), The optimal discovery procedure : A new approach to simultaneous significance testing. UW Biostatistics Working Paper Series. Working Paper 259. Available at http ://www.bepress.com/uwbiostat/paper259.

Storey, J. D., Taylor, J. E. and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates : a unified approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66 (1)** 187–205.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes.* Springer Series in Statistics. Springer-Verlag, New York, With applications to statistics.

Waberski, T., Gobbele, R., Kawohl, W., Cordes, C. and Buchner, H. (2003). Immediate cortical reorganization after local anesthetic block of the thumb : source localization of somatosensory evoked potentials in human subjects. *Neurosci. Lett.* **347** 151–154.

Wasserman, L. and Roeder, K. (2006), Weighted hypothesis testing. Technical report, Dept. of statistics, Carnegie Mellon University.

Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing.* Wiley, Examples and Methods for P- Value Adjustment.

van de Wiel, M. A. and In Kim, K. (2007), Estimating the false discovery rate using nonparametric deconvolution. To appear in Biometrics. Available at http ://www.blackwell-synergy.com/doi/abs/10.1111/j.1541-0420.2006.00736.x.

Yekutieli, D. and Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J. Statist. Plann. Inference.* **82 (1-2)** 171–196. Multiple comparisons (Tel Aviv, 1996).