



Statistique asymptotique dans des modèles à variables latentes

Catherine Matias

Habilitation à diriger des recherches

Soutenue le 17 octobre 2008 à l'Université d'Évry-Val d'Essonne

Laboratoire Statistique et Génome (UMR CNRS 8071),
Tour Évry 2, 523 pl. des Terrasses de l'Agora,
91000 Évry, France
e-mail: catherine.matias@genopole.cnrs.fr
url: stat.genopole.cnrs.fr/~cmatias

Table des matières

Remerciements	4
Notations	5
Contexte	6
Présentation générale	6
I Séquences : modélisation de la composition et des processus d'évolution	8
1 Les chaînes de Markov à régimes Markoviens	8
1.1 Contexte	8
1.2 Représentations des CMRM	10
1.3 Définition et estimation de l'ordre d'une CMRM	14
1.4 Données simulées	17
2 Les chaînes semi-Markov cachées	17
2.1 Contexte	17
2.2 Perspectives	19
3 Les modèles pair-Markov cachés pour modéliser l'évolution des séquences	19
3.1 Contexte	19
3.2 Description du modèle	21
3.3 Vraisemblance	22
3.4 Résultats	24
3.5 Commentaires et conclusions	25
4 Modèles d'évolution de séquences dépendants du contexte	26
4.1 Contexte	28
4.2 Modèle	30
4.3 Conclusions et perspectives	31
II Modèles semi paramétriques de signaux bruités	32
5 Le modèle de convolution	32
5.1 Convolution avec échelle du bruit inconnue	35
5.2 Convolution avec régularité du bruit inconnue	38
5.3 Approche générale de l'étude des estimateurs à noyau construits par « plug-in »	41
5.4 Tests d'adéquation en convolution semi ou non paramétrique	44
6 Fonctions périodiques bruitées, de période inconnue	49
III Graphes aléatoires	53
7 Les motifs dans les réseaux biologiques	53
8 Un modèle de mélange pour graphes	54

8.1	Quelques modèles de mélange de graphes identifiables	55
9	Identifiabilité des modèles de mélange pour application aux modèles de mélanges de graphes . .	59
10	Inférence de réseaux d'interaction	60
A	Identifiability of latent class models with many observed variables	63
A.1	Introduction	63
A.2	The discrete latent class model	66
A.3	Kruskal's theorem and its consequences	67
A.4	Finite mixtures of discrete multivariate distributions	69
A.5	Proofs	70
	Références	73
	Liste des travaux	80
	Liste des co-auteurs	82

Remerciements

Mes remerciements vont tout d'abord à Elisabeth Gassiat, qui a eu la bienveillance de suivre mon travail avec suffisamment de distance pour me laisser m'épanouir, mais sans jamais me fermer sa porte lorsque j'en ai eu besoin. Merci beaucoup Elisabeth.

Je remercie ensuite Fabienne Comte, Sylvie Huet et Aad van der Vaart, qui ont accepté d'écrire un rapport sur ce travail. C'est un honneur pour moi que vous ayez accepté cette tâche ingrate et je vous en suis reconnaissante. Je remercie également Laurent Cavalier et Stéphane Robin d'avoir accepté de faire partie des membres du jury de cette soutenance.

J'ai eu le plaisir de travailler avec de nombreuses personnes durant ces années de recherche, et je tiens à remercier ici tous mes co-auteurs. Sans eux, ce mémoire ne serait pas ce qu'il est, et j'ai eu avec chacun d'eux beaucoup de plaisir à travailler.

Mes remerciements vont ensuite à Bernard Prum, qui a su créer à Évry un laboratoire où l'on se sent tout simplement bien. Enfin, mes remerciements vont aux membres passés et présents du laboratoire Statistique et Génome, qui ont contribué à cette ambiance si particulière que je viens d'évoquer, ainsi qu'à tous les membres du groupe SSB.

Notations

Pour plus de lisibilité, je dresse ci-dessous la liste des notations et conventions utilisées dans ce manuscrit.

$X_{1:n}$ ou X_1^n désigne la suite de variables X_1, \dots, X_n .

\mathbb{N}^* désigne l'ensemble des entiers strictement positifs.

Si \mathcal{A} est un alphabet fini, $|\mathcal{A}|$ désigne son cardinal et \mathcal{A}^* l'ensemble des suites finies de \mathcal{A} .

Pour tout $s \in \mathcal{A}^*$, la longueur de s est notée $|s|$.

Pour tous $s, t \in \mathcal{A}^*$, la suite st est obtenue par concaténation de s et de t .

$1\{A\}$ est la fonction indicatrice de l'ensemble A .

\mathbb{P}, \mathbb{E} désignent une mesure de probabilité et l'espérance associée ; $\mathbb{P}_\theta, \mathbb{E}_\theta$ ou $\mathbb{P}_f, \mathbb{E}_f$ des mesures de probabilité et les espérances associées, dépendant d'un paramètre θ ou f ; $\mathbb{P}_0, \mathbb{E}_0$ les mêmes quantités pour le paramètre θ_0 ou f_0 .

$X \text{ II } Y$ signifie que les variables X et Y sont indépendantes.

Contexte

Après une thèse en statistique mathématique dans laquelle j'avais en particulier étudié des modèles de Markov cachés, j'ai été recrutée au laboratoire Statistique et Génome en octobre 2002 et me suis tournée vers une recherche principalement motivée par des applications en génomique ou post-génomique. Mon domaine de recherche est assez vaste, mais le dénominateur commun de mes travaux est la présence de variables latentes (non observées) dans les modèles étudiés. Mes préoccupations sont majoritairement théoriques : études asymptotiques, convergence des estimateurs, vitesses, identifiabilité . . . Les modèles considérés peuvent être aussi bien paramétriques que semi ou non paramétriques, et les outils statistiques utilisés sont donc relativement variés. Ma recherche a longtemps porté sur des séquences de variables aléatoires (processus « temporels ») et s'oriente à présent vers des observations organisées sous forme de *graphe*.

Je présente dans ce manuscrit les travaux effectués depuis la thèse. Les références du type $[Mi]$ où i est un numéro, renvoient à la liste de ma production scientifique. Ma présentation s'organise en trois grandes thématiques : les travaux portant sur des séquences, notamment sur la modélisation de leur distribution et des processus d'évolution sous-jacents ; les travaux de statistique semi ou non paramétrique portant sur des signaux observés avec du bruit ; et enfin les travaux (en partie en cours) portant sur les graphes aléatoires.

Présentation générale

Dans la première partie de ce manuscrit, je présente mes travaux liés à l'analyse statistique de la composition et de l'évolution des séquences biologiques. Les modèles de Markov cachés (que j'avais étudiés dans ma thèse mais dans un cadre différent puisque la chaîne cachée prenait des valeurs continues), y jouent un rôle prépondérant. Je présente tout d'abord des travaux sur l'estimation du nombre d'états cachés et sur la mémoire d'une chaîne de Markov à régimes Markoviens, ce modèle étant une variante des chaînes de Markov cachées. J'introduis également dans cette section des considérations sur l'identifiabilité de ces modèles qui ne font pas partie de la version publiée de ce travail. Je présente ensuite des considérations sur les chaînes semi-Markov cachées, cadre dans lequel je n'ai pas apporté de contribution personnelle. Les modèles pair-Markov cachés qui sont présentés ensuite sont destinés à la comparaison, par alignement, de séquences biologiques, dans un cadre évolutif. Enfin, je présente le cadre d'une thèse que je co-encadre sur les modèles d'évolution de séquences qui tiennent compte de dépendances locales.

Dans la seconde partie de ce manuscrit, je présente mes travaux portant sur l'analyse de signaux bruités. La majeure partie de ce travail se situe dans la lignée de mon second chapitre de thèse et porte sur l'étude de modèles de convolution semi paramétriques pour lesquels la distribution de la densité du bruit n'est connue qu'à paramètre près. Il s'agit de problèmes d'estimation de paramètres puis de densités dans un cadre minimax. Je présente également dans ce contexte des tests d'adéquation non paramétriques adaptatifs et minimax. Enfin, je présente un problème d'estimation de fonction périodique bruitée, lorsque

la période du bruit est inconnue. Là encore, il s'agit d'un problème d'estimation dans un cadre adaptatif minimax. Dans cette partie, je présente également une approche générale pour l'étude des estimateurs construits par « plug-in », avec un erratum et des considérations non publiées.

La troisième partie porte sur des données d'un type différent, puisqu'il s'agit de graphes. Je présente tout d'abord une étude de la moyenne et de la variance du nombre d'occurrences de motifs topologiques dans un modèle de graphe dont les degrés (nombre de connexions) des noeuds s'ajustent à ceux d'un graphe observé. Il s'agit là d'une première approche pour chercher à détecter des motifs sur ou sous représentés dans un graphe aléatoire. Je présente ensuite un modèle de mélange pour graphes aléatoires qui permet une modélisation relativement réaliste des réseaux réels observés. Un des problèmes de ce modèle réside dans l'identifiabilité des paramètres. Je fournis un certains nombres de cas particuliers de ce modèles pour lesquels on peut prouver l'identifiabilité des paramètres par des techniques simples (considérations de moments ou de lois marginales). Je présente ensuite deux travaux encore en cours. Le premier porte sur une notion d'identifiabilité générique, que nous étudions dans le but d'obtenir un résultat général d'identifiabilité dans le modèle de mélange pour graphes aléatoires mentionné ci-dessus. Le second concerne une procédure d'inférence de graphes de corrélation à partir de vecteurs Gaussiens dans un espace de très grande dimension. L'approche proposée ici mêle la régression pénalisée (LASSO) et les modèles de mélange de graphes, afin d'inférer des graphes ayant une structure cachée de groupes.

Première partie

Séquences : modélisation de la composition et des processus d'évolution

Cette partie regroupe les travaux [M7,M8], le sujet de thèse de mon étudiante, Audrey Finkler, ainsi que des considérations sur les chaînes semi-Markov cachées et sur la représentation des processus étudiés dans [M8], qui n'ont pas donné lieu à publication.

Les séquences biologiques (séquences d'ADN, de protéines, . . .) ont été produites en quantités phénoménales ces dernières années, grâce au développement des techniques de séquençage. La bio-informatique a pour tâche un traitement à grande échelle de ces données, en s'appuyant sur des modèles probabilistes simples mais pertinents. Les modèles de Markov, ou de Markov cachés, jouent un rôle prépondérant dans la modélisation et l'analyse de la composition des séquences biologiques ainsi que des processus d'évolution de ces séquences.

Mon approche a été de partir des modèles tels qu'ils étaient utilisés par les bio-informaticiens et d'étudier leurs propriétés statistiques, afin de donner un fondement à leur utilisation, ou des guides quant au choix de certains paramètres. Le travers de cette démarche est le suivant : les modèles utilisés n'ont pas été choisis pour leurs propriétés statistiques, et leur étude peut s'avérer compliquée. Il importe de savoir rester modeste quant aux résultats qui peuvent être établis sur de tels modèles.

1. Les chaînes de Markov à régimes Markoviens

1.1. Contexte

Les séquences biologiques sont des suites de variables aléatoires à valeurs dans un alphabet fini que nous noterons \mathcal{A} . Cette succession de lettres forme un texte, qui *prend du sens*, au moins dans les régions codantes de l'ADN. Une modélisation de ces séquences par un processus de variables indépendantes et identiquement distribuées (i.i.d.) est donc très peu adéquate. Les chaînes de Markov (CM) ont été quant à elles plus largement utilisées pour cette modélisation. Qu'elles soient d'ordre fixe, d'ordre variable (i.e. dépendant de la lettre à prédire) ou encore définies à partir d'un arbre de contexte, leur utilisation se limite à des séquences relativement courtes, que l'on peut considérer comme *homogènes*.

La modélisation de données hétérogènes peut se faire de diverses manières. L'approche la plus simple consiste à introduire, pour chaque variable observée X_i , une variable latente (non observée) et discrète, Z_i , qui indique le *type* ou *régime* de la variable observée. L'introduction d'un nombre fini (connu ou pas) de régimes différents permet donc de modéliser l'hétérogénéité des séquences à travers un nombre fini de groupes qui sont eux homogènes. Dans le cas des séquences biologiques, les régimes mis en évidence par une telle modélisation peuvent être par exemple des régions codantes/non codantes, des introns/exons

au sein des gènes, des morceaux de séquence provenant de transferts horizontaux (i.e. provenant d'autres organismes), etc ... Dans le cas le plus simple, les variables latentes sont i.i.d. et les observations sont indépendantes conditionnellement à la donnée des variables latentes, la loi de chaque X_i ne dépendant que de Z_i . Nous sommes dans le cadre d'un modèle de mélange, et les observations résultantes sont encore globalement i.i.d.

Afin d'introduire de la dépendance dans la succession des variables observées, il peut être intéressant de supposer que la succession des régimes suit en fait un processus Markovien à temps discret. Dans le cas des chaînes de Markov cachées (CMC), les observations sont encore supposées indépendantes, conditionnellement à la donnée des variables latentes, la loi de chaque X_i ne dépendant que de Z_i . On pourra se référer à [19, 42] pour plus de détails sur les chaînes de Markov cachées. La loi du temps de séjour d'une chaîne de Markov dans un état étant géométrique, ce modèle est adapté à des séquences dans lesquelles des *zones* ou *plages* de distribution homogène se succèdent le long de la séquence. Il est à noter que dans un tel modèle, la distribution résultante sur les observations n'est pas Markovienne et présente des phénomènes de dépendance entre des variables arbitrairement *éloignées* l'une de l'autre. Cependant, conditionnellement à la donnée des régimes, chaque plage homogène contient des variables qui sont i.i.d. Dans le cas d'une plage correspondant à une région codante de l'ADN par exemple, cette hypothèse peut sembler trop restrictive puisque l'on sait bien que l'information est structurée en codons (succession de trois nucléotides). C'est pour cela que la bio-informatique s'est rapidement tournée vers des modèles plus complexes, à savoir des chaînes de Markov à régimes Markoviens (CMRM). Dans cette classe de modèles, les variables latentes forment toujours une chaîne de Markov, mais conditionnellement à la donnée des régimes, les observations forment également une chaîne de Markov dont les transitions dépendent du régime considéré.

Ce modèle est largement répandu de nos jours, notamment grâce à l'algorithme EM [36] qui permet d'approcher l'estimateur du maximum de vraisemblance des paramètres (cet estimateur n'étant pas calculable analytiquement). Également appelé modèle auto-régressif à régimes Markoviens, il fut introduit à l'origine en économétrie [68] et est utilisé en particulier en finance (modèles à volatilité stochastique).

Cependant, son utilisation pratique pose des problèmes pour lesquels une réponse apportée par la statistique est souhaitable. En particulier, le choix du nombre d'états cachés et de l'ordre de la chaîne de Markov conditionnelle est un problème crucial, puisque l'on sait par exemple que les estimateurs du maximum de vraisemblance ne sont pas nécessairement convergents si l'ordre du modèle dans lequel ils sont calculés diffère de l'ordre du modèle qui a servi à générer les observations. Il s'agit essentiellement d'un problème d'identifiabilité des paramètres, voir par exemple à ce sujet [30] sur les modèles de mélange de familles exponentielles. Ce choix peut, dans certains cas, être dicté par le problème sous-jacent : s'il s'agit de détecter des gènes dans une séquence d'ADN, un nombre de régimes égal à 2 et un ordre de la chaîne conditionnelle égal à 2 également suffit certainement. Cependant, il est des cas où un tel choix est beaucoup moins évident et où le statisticien aimerait utiliser l'information contenue dans les observations pour orienter sa décision.

Un des problèmes majeurs qui se posent dans ce cadre vient du fait que les modèles considérés ne sont pas emboîtés. En effet, notons $\Pi^{k,m}$ l'ensemble des distributions sur \mathcal{A} , de type CMRM, avec k états cachés et une distribution conditionnelle des observations qui est une chaîne de Markov d'ordre m . Dans la suite, $k \geq 1$ et $m \geq 0$ et nous parlerons de couple d'ordre de la CMRM. Pour deux couples d'entiers quelconques (k, m) et (k', m') , il n'y a (en général) pas d'inclusion des espaces $\Pi^{k,m}$ et $\Pi^{k',m'}$, alors que pour m et k fixés, les suites $\{\Pi^{k,m'}\}_{m'}$ et $\{\Pi^{k',m}\}_{k'}$ sont naturellement croissantes au sens de l'inclusion. Dès lors, s'il existe deux représentations $\mathbb{P} \in \Pi^{k,m}$ et $\mathbb{P}' \in \Pi^{k',m'}$, laquelle sélectionner (en dehors du cas trivial où $k \leq k'$ et $m \leq m'$)? Mais d'abord, une telle situation peut-elle réellement se produire? Un premier exemple d'une telle situation est le suivant : toute CMRM d'ordre (k, m) peut aussi s'écrire comme une CMRM d'ordre $(ka^m, 0)$. Il suffit en effet de considérer les observations comme une fonction (déterministe) de la chaîne de Markov complète $\{(Z_i, X_{i-m+1}^i)\}_{i \geq m}$. [Cet exemple qui peut avoir l'air trivial est en fait à la base d'une littérature importante sur les CMC (ou plus généralement les CMRM) vues comme fonctions déterministes d'une chaîne de Markov.] Existe-t-il des situations « plus complexes » où deux ordres (k, m) et (k', m') décrivent la même CMRM? Ce problème en soulève un plus délicat encore, celui de l'identification d'un processus : comment caractériser la classe des processus décrits par les CMRM d'ordre (k, m) fixé?

Ci-dessous, je présente quelques résultats (non publiés), inspirés de la littérature des CMC.

1.2. Représentations des CMRM

Nous commençons par exprimer la distribution d'une CMRM en utilisant des produits de matrices bien choisies. Puis, suivant le travail de [59] et [51], nous utilisons de l'algèbre simple pour caractériser les CMRM.

Introduisons tout d'abord quelques notations. Dans la suite, \mathcal{Z} sera l'espace d'états de la chaîne de Markov latente. Nous noterons $\mathcal{M}(\mathcal{Z} \times \mathcal{A}^m)$ l'ensemble des mesures de probabilité sur $\mathcal{Z} \times \mathcal{A}^m$. L'ensemble $\Pi^{k,m}$ est naturellement paramétré par $\mathcal{M}(\mathcal{Z} \times \mathcal{A}^m) \times \Theta^{k,m}$, où

$$\Theta^{k,m} = \left\{ \theta = (A, B) : A = (a(i, j))_{1 \leq i, j \leq k}, a(i, j) \geq 0, \sum_{j=1}^k a(i, j) = 1 \text{ et} \right.$$

$$\left. B = (b(x|x_{1:m}; z))_{x \in \mathcal{A}, x_{1:m} \in \mathcal{A}^m, z \in \mathcal{Z}; b(x|x_{1:m}; z) \geq 0, \sum_{x=1}^a b(x|x_{1:m}; z) = 1 \right\}.$$

Ainsi, $\Pi^{k,m} = \{\mathbb{P} = \mathbb{P}_{\mu, \theta} : (\mu, \theta) \in \mathcal{M}(\mathcal{Z} \times \mathcal{A}^m) \times \Theta^{k,m}\}$. Dans la suite, nous considérons uniquement des processus stationnaires. Pour chaque paramètre $\theta \in \Theta^{k,m}$ (donnant lieu à un processus ergodique), nous notons π_θ la mesure stationnaire associée sur $\mathcal{Z} \times \mathcal{A}^m$, et $\mathbb{P}_\theta = \mathbb{P}_{\pi_\theta, \theta}$ est la CMRM stationnaire induite dans $\Pi^{k,m}$.

La matrice $A = (a(i, j))_{1 \leq i, j \leq k}$ est donc la matrice de transition du processus caché (taille $k \times k$) et pour tous $x \in \mathcal{A}$ et $x_1^m \in \mathcal{A}^m$, la matrice diagonale $B(x|x_1^m) = \text{diag}(b(x|x_1^m; i))_{1 \leq i \leq k}$ (de taille $k \times k$)

contient les probabilités d'émission du processus observé. De plus, pour tous $x \in \mathcal{A}$ et $x_1^m \in \mathcal{A}^m$, notons

$$M(x|x_1^m) = A \times B(x|x_1^m) = (m_{ij}(x|x_1^m))_{1 \leq i, j \leq k}$$

où $m_{ij}(x|x_1^m) = a(i, j)b(x|x_1^m; j) = \mathbb{P}_\theta(X_{t+1} = x, Z_{t+1} = j | Z_t = i, X_{t-m+1}^t = x_1^m)$. Pour tous $x_1^m \in \mathcal{A}^m$, notons $\pi_\theta(x_1^m)$ le vecteur ligne de la distribution initiale, $\pi_\theta(x_1^m) = (\pi_\theta(i, x_1^m))_{1 \leq i \leq k}$. Il est facile de voir que pour toute suite d'observations $x_1^n \in \mathcal{A}^n$, on a

$$(1.1) \quad \mathbb{P}_\theta(X_1^n = x_1^n) = \pi_\theta(x_1^m) \times A^m \times M(x_{m+1}|x_1^m) \times \dots \times M(x_n|x_{n-m}^{n-1}) \times e_{(k)},$$

où $e_{(k)}$ est le vecteur colonne de taille k qui ne contient que des 1. Ainsi, la distribution d'une CMRM est entièrement spécifiée par l'ensemble de paramètres $\mathcal{M} := \{k, m, A, \{M(x|x_1^m)\}\}$. Un tel ensemble est appelé une *représentation* de la CMRM.

Dans la proposition suivante, nous donnons une condition suffisante pour que deux représentations définissent la même CMRM.

Proposition 1.1. *Soient $\mathcal{M}_{(1)} = \{k_1, m_1, A_1, \{M_1(x|x_1^{m_1})\}\}$ et $\mathcal{M}_{(2)} = \{k_2, m_2, A_2, \{M_2(x|x_1^{m_2})\}\}$ telles qu'il existe deux matrices P, Q de tailles respectives $k_1 \times k_2$ et $k_2 \times k_1$ vérifiant*

- i) $PQ = I_{k_1}$ (la matrice identité de taille $k_1 \times k_1$) et $e_{(k_2)} = Qe_{(k_1)}$,*
- ii) Si $m_1 \geq m_2$, on suppose que $\forall x \in \mathcal{A}, x_1^{m_1} \in \mathcal{A}^{m_1}$, on a $M_2(x|x_{m_1-m_2+1}^{m_1}) = Q \times M_1(x|x_1^{m_1}) \times P$ et $\pi_{\theta_2}(x_1^{m_2})A_2^{m_2}M_2(x_{m_2+1}|x_1^{m_2}) \dots M_2(x_{m_1}|x_{m_1-m_2}^{m_1-1}) = \pi_{\theta_1}(x_1^{m_1})A_1^{m_1}P$.*
- iii) Si $m_2 \geq m_1$, on suppose que $\forall x \in \mathcal{A}, x_1^{m_2} \in \mathcal{A}^{m_2}$, on a $M_2(x|x_1^{m_2}) = Q \times M_1(x|x_{m_2-m_1+1}^{m_2}) \times P$ et $\pi_{\theta_2}(x_1^{m_2})A_2^{m_2} = \pi_{\theta_1}(x_1^{m_1})A_1^{m_1}M_1(x_{m_1+1}|x_1^{m_1}) \dots M_1(x_{m_2}|x_{m_2-m_1}^{m_1-1})P$.*

Alors les deux représentations $\mathcal{M}_{(1)}$ et $\mathcal{M}_{(2)}$ définissent la même CMRM.

La preuve de ce résultat découle très facilement de (1.1).

Remarque 1.1. *Puisque $PQ = I_{k_1}$, les rangs des matrices vérifient $\text{rang}(P) = \text{rang}(Q) = k_1$ d'où $k_1 \leq k_2$. Par conséquent, le cas $m_1 \geq m_2$ est le cas intéressant tandis que $m_2 \geq m_1$ correspond à des représentations $\mathcal{M}_{(2)}$ qui sont sur-paramétrées.*

Il importe à présent de chercher des conditions nécessaires pour que deux représentations définissent la même CMRM. Dans le cas des CMC, une voie intéressante a consisté en l'étude du *rang de la distribution* de la CMC.

Dans la littérature des processus de Markov et des CMC, il existe différentes notions de rang : la première, introduite par Gilbert [59] dans le cas des CMC repose sur la description d'une distribution sur un espace d'états finis à travers les probabilités d'occurrence des mots finis. Cette notion est intimement reliée à la définition d'une matrice de Hankel généralisée, et cette idée est en particulier exploitée dans [3, 124] pour le problème de la réalisation de CMC. Enfin, Heller [70] puis Holland [71] ont associé, à toute distribution sur l'ensemble fini \mathcal{A} , un \mathcal{A} -module dont la dimension (après une transformation) caractérise les chaînes de Markov d'ordre m . Notons que ces approches sont très liées à la notion d'identifiabilité générique introduite à la Section 9.

Nous nous limiterons ici à présenter la notion de rang introduite par Gilbert [59], qui repose sur la vision des CMC (ou ici CMRM) comme des fonctions déterministes d'une chaîne de Markov (latente). On note \mathcal{A}^* l'ensemble des mots finis de l'alphabet \mathcal{A} . Si s, t sont deux mots dans \mathcal{A}^* , alors st est le mot obtenu par concaténation de s et t ; et $|s|$ désigne la longueur du mot s .

Soit \mathbb{P} une distribution sur l'ensemble fini \mathcal{A} et $x \in \mathcal{A}$. Pour tout entier n et tout ensemble de mots $s_{1:n} = s_1, \dots, s_n$ et $t_{1:n} = t_1, \dots, t_n$ dans \mathcal{A}^* (i.e. $s_i \in \mathcal{A}^*$ et $t_j \in \mathcal{A}^*$), on considère la matrice composée des séquences $C_x(s_{1:n}, t_{1:n})$ de taille $n \times n$, définie par

$$\forall 1 \leq i, j \leq n, \quad \left(C_x(s_{1:n}, t_{1:n}) \right)_{ij} = \mathbb{P} \left(X_1^{|s_i x t_j|} = s_i x t_j \right).$$

Pour tout $x \in \mathcal{A}$, définissons le rang de la distribution \mathbb{P} au point x ,

$$r(x) = \max \left\{ \text{rang} \left(C_x(s_{1:n}, t_{1:n}) \right); n \geq 1, s_i, t_j \in \mathcal{A}^* \right\}.$$

Le rang d'une distribution \mathbb{P} est défini par $R = \sum_x r(x)$.

Le rang d'une distribution est un concept général qui n'est pas adapté uniquement aux CMC. En particulier, une chaîne de Markov d'ordre 1 satisfait $r(x) = 1$, pour tout x . En effet, pour toutes séquences s_1, s_2, t_1, t_2 on a $\mathbb{P}(s_1 x t_1) / \mathbb{P}(s_2 x t_1) = \mathbb{P}(s_1 x) / \mathbb{P}(s_2 x) = \mathbb{P}(s_1 x t_2) / \mathbb{P}(s_2 x t_2)$ ce qui donne $\det(C_x(s_{1:2}, t_{1:2})) = 0$. Ainsi, une $CM(m)$ sur \mathcal{A} a un rang qui vérifie $R \leq a^m$.

On peut montrer la propriété suivante : pour toute représentation \mathcal{M} d'une CMC, le rang de la distribution est toujours inférieur au nombre d'états cachés k de la représentation [59, Lemma 1]. La preuve de ce résultat est très simple et repose sur la factorisation de la probabilité $\mathbb{P}(X_1^n = x_1^n)$ et donc des matrices $C_x(s_{1:n}, t_{1:n})$ en produits de matrices. Les représentations \mathcal{M} qui utilisent un nombre d'états cachés égal au rang de la CMC sont dites *régulières* (mais elles n'existent pas toujours), et les représentations non régulières sont de mesure de Lebesgue nulle dans l'espace des paramètres de la représentation. Les représentations régulières d'une CMC ne sont pas sur-paramétrées : il n'existe pas d'autre représentation de la même CMC utilisant moins d'états cachés. Lorsqu'il existe une représentation régulière d'une CMC, on peut montrer une réciproque à la Proposition 1.1 (dans sa formulation pour les CMC), voir [51, Lemma 1.3.2].

Dans toute la suite, les CMRM sont vues comme des fonctions déterministes d'une chaîne de Markov sur un espace d'états de la forme $\mathcal{Z} \times \mathcal{A}^m$. En particulier, si $\{X_i\}$ est une CMRM, alors d'après ce qui précède, son rang vérifie $R \leq ka^m$.

Définition 1.1. Soit $\{X_i\}$ une CMRM admettant la représentation $\mathcal{M} := \{k, m, A, \{M(x|x_1^m)\}\}$. Cette représentation est dite *régulière* si le rang R de la distribution vérifie $R = ka^m$.

Nous avons vu que les représentations $\mathcal{M} = \{k, m, A, \{M(x|x_1^m)\}\}$ sont naturellement paramétrées par l'espace $\Theta^{k,m}$ qui est un sous ensemble de $\mathbb{R}^{ka^m(a-1)+k(k-1)}$. Les représentations $\{k, m, A, \{M(x|x_1^m)\}\}$ qui ne sont pas régulières sont de mesure de Lebesgue nulle dans $\mathbb{R}^{ka^m(a-1)+k(k-1)}$. En effet, l'ensemble des représentations qui ne sont pas régulières correspond à l'union (finie) des ensembles $\mathcal{I}_x, x \in \mathcal{A}$ définis de la façon suivante. Chaque ensemble \mathcal{I}_x contient les représentations telles que pour tout entier n et

pour toutes suites de mots $s_1, \dots, s_n, t_1, \dots, t_n \in \mathcal{A}^*$, le déterminant de la matrice $C_x(s_{1:n}, t_{1:n})$ est nul. Ceci correspond à un ensemble de mesure de Lebesgue nulle dans l'espace $\mathbb{R}^{ka^m(a-1)+k(k-1)}$. [L'argument utilisé ici est le suivant : $\det(C_x(s_{1:n}, t_{1:n}))$ est un polynôme non nul en les paramètres de la représentation, dont l'annulation définit une variété de dimension strictement inférieure à celle de l'espace des paramètres.] La question est maintenant de savoir si une réciproque à la Proposition 1.1 est possible. Comme nous allons le voir, il semble que la condition suffisante formulée dans la Proposition 1.1 soit trop forte pour être nécessaire.

En effet, introduisons les matrices $\tilde{M}(x)$ qui permettent de visualiser la CMRM comme une CMC. Pour tout $x \in \mathcal{A}$, on note $\tilde{M}(x)$ la matrice de taille $ka^m \times ka^m$ dont les coefficients sont

$$\tilde{M}(x)_{(i, u_1^m); (j, v_1^m)} = 1\{v_1^m = u_2^m x\} M(x|u_1^m)_{i,j}.$$

Soit $\{X_i\}$ une CMRM qui admet les représentations $\mathcal{M}_{(1)} = \{k_1, m_1, A_1, \{M_1(x|x_1^{m_1})\}\}$ et $\mathcal{M}_{(2)} = \{k_2, m_2, A_2, \{M_2(x|x_1^{m_2})\}\}$, avec $\mathcal{M}_{(1)}$ régulière. D'après [51, Lemma 1.3.2], on sait que $k_2 a^{m_2} \geq k_1 a^{m_1}$. Concentrons nous sur le cas intéressant d'égalité : $k_2 a^{m_2} = k_1 a^{m_1}$. Supposons également pour fixer les idées que $k_1 \leq k_2$ et $m_1 \geq m_2$. Toujours d'après [51, Lemma 1.3.2], il existe \tilde{P} et \tilde{Q} de tailles respectives $k_1 a^{m_1} \times k_2 a^{m_2}$ et $k_2 a^{m_2} \times k_1 a^{m_1}$ telles que $PQ = I_{k_1 a^{m_1}}$ (la matrice identité de taille $k_1 a^{m_1}$) et telles que pour tout $x \in \mathcal{A}$,

$$(1.2) \quad \tilde{M}_1(x) = \tilde{P} \tilde{M}_2(x) \tilde{Q}.$$

On décide de ranger les éléments des ensembles $\mathcal{Z} \times \mathcal{A}^{m_i}$ dans l'ordre suivant : on considère les a^{m_i} blocs successifs de longueur k_i et de la forme $\{(j, u_1^{m_i}), 1 \leq j \leq k_i\}$ où $u_1^{m_i} \in \mathcal{A}^{m_i}$ est fixé. Cette manipulation correspond à des changements de base et ne modifie pas ce qui précède. La matrice \tilde{P} (resp. \tilde{Q}) se décompose en blocs de taille $k_1 \times k_2$ (resp. de taille $k_2 \times k_1$) notés $P_{u_1^{m_1}, v_1^{m_2}}$ pour $u_1^{m_1} \in \mathcal{A}^{m_1}, v_1^{m_2} \in \mathcal{A}^{m_2}$ (resp. $Q_{u_1^{m_2}, v_1^{m_1}}$ pour $u_1^{m_2} \in \mathcal{A}^{m_2}, v_1^{m_1} \in \mathcal{A}^{m_1}$).

Les notations deviennent rapidement confuses. Je choisis d'illustrer ce qui se passe pour $a^{m_1} = 3$ et $a^{m_2} = 2$ (et donc je rappelle que $3k_1 = 2k_2$). On peut alors écrire (1.2) sous la forme

$$\left(\begin{array}{c|c|c} \tilde{M}_1^1(x) & 0 & 0 \\ \hline 0 & \tilde{M}_1^2(x) & 0 \\ \hline 0 & 0 & \tilde{M}_1^3(x) \end{array} \right) = \left(\begin{array}{c|c} \tilde{P}_{11} & \tilde{P}_{12} \\ \hline \tilde{P}_{21} & \tilde{P}_{22} \\ \hline \tilde{P}_{31} & \tilde{P}_{32} \end{array} \right) \times \left(\begin{array}{c|c|c} \tilde{M}_2^1(x) & 0 & 0 \\ \hline 0 & \tilde{M}_2^2(x) & 0 \\ \hline 0 & 0 & \tilde{M}_2^3(x) \end{array} \right) \times \left(\begin{array}{c|c|c} \tilde{Q}_{11} & \tilde{Q}_{12} & \tilde{Q}_{13} \\ \hline \tilde{Q}_{21} & \tilde{Q}_{22} & \tilde{Q}_{23} \end{array} \right)$$

où les matrices \tilde{P}_{ij} et \tilde{Q}_{ij} sont de tailles respectives $k_1 \times k_2$ et $k_2 \times k_1$, et pour $i = 1, 2$, chacune des matrices $\tilde{M}_i^j(x)$ est une matrice $M_i(x|j)$ pour un certain mot $j \in \mathcal{A}^{m_i}$. Dans notre exemple, notons $\mathcal{A}^{m_1} = \{1, 2, 3\}$ et $\mathcal{A}^{m_2} = \{1, 2\}$. On obtient alors les relations suivantes

$$(1.3) \quad \begin{aligned} \tilde{M}_1^1(x) &= M_1(x|1) = \tilde{P}_{11} M_2(x|1) \tilde{Q}_{11} + \tilde{P}_{12} M_2(x|2) \tilde{Q}_{21} \\ \tilde{M}_1^2(x) &= M_1(x|2) = \tilde{P}_{21} M_2(x|1) \tilde{Q}_{12} + \tilde{P}_{22} M_2(x|2) \tilde{Q}_{22} \\ \tilde{M}_1^3(x) &= M_1(x|3) = \tilde{P}_{31} M_2(x|1) \tilde{Q}_{13} + \tilde{P}_{32} M_2(x|2) \tilde{Q}_{23}. \end{aligned}$$

La relation obtenue est donc bien plus complexe que l'hypothèse suffisante formulée dans la Proposition 1.1 et il semble bien que s'il existe une réciproque à cette proposition, elle prenne une forme du type

(1.3).

En conclusion, il importe de savoir si pour une valeur de a fixée (le cardinal de l'espace des observations), on a beaucoup de couples (k, m) et (k', m') tels que $ka^m = k'a^{m'}$. Pour $a = 4$, on peut constater par exemple que (en excluant la valeur $k = 1$, voir le paragraphe *Données simulées*), les choix $(k, m) = (8, 1)$ et $(2, 2)$ donnent la même valeur de $ka^m = 32$. De plus, dans le cas d'existence de deux couples (k, m) et (k', m') avec $ka^m = k'a^{m'}$, on peut écrire un système de relations comme ci-dessus qui lie les probabilités d'émission dans chacune des deux représentations. Dans la section suivante, nous verrons qu'un critère qui peut être utilisé pour hiérarchiser les modèles est la dimension de l'espace de paramètres, qui vaut $N(k, m) = ka^m(a - 1) + k(k - 1)$. Au sens de ce critère, si deux couples (k, m) et (k', m') satisfont $ka^m = k'a^{m'}$, on préférera la représentation qui a le moins d'états cachés, car elle aura également moins de paramètres.

1.3. Définition et estimation de l'ordre d'une CMRM

Revenons à présent au problème de la définition de l'ordre d'une CMRM. Je décris ci-dessous un travail [M8] en collaboration avec Antoine Chambaz (Université René Descartes Paris 5). Il est important de noter que dans toute la suite, nous ne supposons jamais qu'il existe une borne a priori sur le nombre d'états cachés et sur la mémoire du processus.

Puisque les modèles CMRM ne sont pas emboîtés et puisqu'a priori, plusieurs représentations non comparables d'une CMRM sont possibles, il convient de choisir un critère qui permette de définir une *bonne* représentation (pour le statisticien), d'une CMRM. Nous avons choisi un critère de *parcimonie* : entre deux représentations d'une CMRM nous choisirons celle qui donne lieu à un minimum de paramètres pour décrire ce modèle. Nous hiérarchisons ainsi les modèles $\{\Pi^{k,m}\}$ en utilisant la dimension $N(k, m)$ de l'espace des paramètres associé. Plus précisément, une distribution stationnaire de l'ensemble $\Pi^{k,m}$ nécessite pour sa description $N(k, m) = ka^m(a - 1) + k(k - 1)$ paramètres. Cette quantité nous permet de définir une relation d'ordre total sur $\mathbb{N}^* \times \mathbb{N}$ de la façon suivante : pour tous $(k_1, m_1), (k_2, m_2) \in \mathbb{N}^* \times \mathbb{N}$,

$$(k_1, m_1) \prec (k_2, m_2) \quad \text{ssi} \quad \{N(k_1, m_1) < N(k_2, m_2)\} \text{ ou } \{N(k_1, m_1) = N(k_2, m_2) \text{ et } k_1 < k_2\}.$$

La définition fait de plus intervenir une dissymétrie des rôles de k et m afin d'obtenir un ordre total, mais ce choix n'a aucune conséquence sur la suite. Nous pouvons alors définir le couple d'ordre (k_0, m_0) d'une CMRM de distribution \mathbb{P} dans $\cup_{k \geq 1, m \geq 0} \Pi^{k,m}$ de la façon suivante

$$(k_0, m_0) = \min \{(k, m) \in (\mathbb{N}^* \times \mathbb{N}, \prec) : \mathbb{P} \in \Pi^{k,m}\}.$$

L'estimation de l'ordre est un problème statistique ancien, de nature très différente de l'estimation paramétrique dans un modèle de dimension fixée a priori. L'approche par pénalisation d'un critère empirique remonte à [2, 88, 109, 113]. Notre approche est basée sur la théorie de l'information et propose de pénaliser un critère obtenu via une mesure de codage sur l'espace des observations. Lorsque cette mesure

de codage est le maximum de vraisemblance, la procédure donne un estimateur du maximum de vraisemblance pénalisée. Il existe d'autres lois de codages classiques, comme la loi de Krichevsky-Trofimov [79] ou le maximum de vraisemblance renormalisé. Plus précisément, notre estimateur prend la forme suivante

$$(1.4) \quad (\widehat{k}, \widehat{m})_n = \underset{(k,m) \in (\mathbb{N}^* \times \mathbb{N}, \prec)}{\operatorname{argmin}} \left(-\log \mathbb{Q}_{k,m}(X_{1:n}) + \operatorname{pen}(n, k, m) \right),$$

où $\mathbb{Q}_{k,m}$ est une mesure (dite de codage) sur $\mathcal{A}^{\mathbb{N}}$ et $\operatorname{pen}(n, k, m)$ est un terme de pénalité à choisir. Nous utiliserons trois mesures de codage différentes, notées respectivement, $\text{KT}_{k,m}$, $\text{NML}_{k,m}$ et $\text{ML}_{k,m}$ et définies ci-dessous. Mais il nous faut pour cela introduire tout d'abord quelques notations.

Je rappelle que dans toute la suite, on ne considère que des paramètres θ générant une distribution ergodique. Notons $\nu_{k,m}$ la densité de probabilité définie sur $\Theta^{k,m}$, par

$$\nu_{k,m}(\theta) = \prod_{i=1}^k \frac{\Gamma(k/2)\Gamma(a/2)}{\Gamma(1/2)^k \Gamma(1/2)^a} \left(\prod_{j=1}^k \frac{1}{a(i,j)^{1/2}} \right) \left(\prod_{t_1^m \in \mathcal{A}^m} \prod_{t=1}^a \frac{1}{b(t|t_1^m; i)^{1/2}} \right),$$

où $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$.

La loi de mélange de Krichevsky-Trofimov est la distribution $\text{KT}_{k,m}$ sur $(\mathcal{Z} \times \mathcal{A})^{\mathbb{N}^*}$ dont les marginales ont la densité

$$(z_{1:n}, x_{1:n}) \mapsto \int_{\theta \in \Theta^{k,m}} \mathbb{P}_{\bar{\mu}^Z \otimes \bar{\mu}^{X,m}, \theta}(z_{1:n}, x_{1:n}) \nu_{k,m}(\theta) d\theta,$$

où $\bar{\mu}^Z$ et $\bar{\mu}^{X,m}$ sont les distributions uniformes respectivement sur \mathcal{Z} et \mathcal{A}^m .

La mesure du maximum de vraisemblance $\text{ML}_{k,m}$ et celle du maximum de vraisemblance renormalisée $\text{NML}_{k,m}$ sont définies très simplement

$$\text{ML}_{k,m}(x_{1:n}) = \sup_{\theta \in \Theta^{k,m}} \mathbb{P}_\theta(x_{1:n}),$$

et en notant $\mathcal{C} = \sum_{x_{1:n} \in \mathcal{A}^n} \sup_{\theta \in \Theta^{k,m}} \mathbb{P}_\theta(x_{1:n})$, alors

$$\text{NML}_{k,m}(x_{1:n}) = \sup_{\theta \in \Theta^{k,m}} \frac{\mathbb{P}_\theta(x_{1:n})}{\mathcal{C}} = \frac{\text{ML}_{k,m}(x_{1:n})}{\mathcal{C}}.$$

Il est à noter que $\text{KT}_{k,m}$ et $\text{NML}_{k,m}$ sont des mesures de probabilité, ce qui n'est pas le cas de $\text{ML}_{k,m}$. Bien que pertinents d'un point de vue théorique, les estimateurs qui résultent de l'utilisation de $\text{KT}_{k,m}$ et $\text{NML}_{k,m}$ ne peuvent pas être mis en pratique dans le cas de l'étude des CMRM (pas plus que pour l'étude des chaînes de Markov cachées d'ailleurs), car leur calcul effectif ne pourrait se faire que pour des tailles de séquences très modestes. Leur comportement est cependant fortement lié à celui de l'estimateur du maximum de vraisemblance pénalisée.

En nous appuyant sur des travaux existants [57], portant sur l'estimation de l'ordre d'une chaîne de Markov cachée (en l'occurrence, il s'agit du nombre d'états cachés de la chaîne), nous avons proposé une

procédure d'estimation du couple d'ordre d'une CMRM. Cette procédure est originale puisqu'il n'existait pas à notre connaissance, d'estimation d'ordre de type bidimensionnel validée théoriquement. Nous avons établi la convergence de nos estimateurs et avons étudié leur vitesse de sur-estimation.

Théorème 1.1. *Soit \mathbb{P}_0 une distribution stationnaire, ergodique appartenant à $\cup_{k \geq 1, m \geq 0} \Pi^{k, m}$ et d'ordre inconnu (k_0, m_0) . Soit $\{X_j\}_{1 \leq j \leq n}$ un processus stationnaire d'observations de loi \mathbb{P}_0 sur $\mathcal{A}^{\mathbb{N}}$.*

Soit φ une fonction croissante de $(\mathbb{N}^ \times \mathbb{N}, \preccurlyeq)$ dans \mathbb{N} . Fixons $\alpha > 1$ et définissons, pour tous $n \in \mathbb{N}^*$, $k \geq 1$ et $m \geq 0$,*

$$\tau(n, k, m) = \max \left(0, \log k + m \log a - k \log \frac{\Gamma(k/2)}{\Gamma(1/2)} - ka^m \log \frac{\Gamma(a/2)}{\Gamma(1/2)} + \frac{k^2(k-1)}{4n} + \frac{ka^{m+1}(a-1)}{4n} + \frac{5k}{24n}(1+a^m) \right).$$

On considère l'estimateur $(\widehat{k, m})_n$ défini par (1.4), avec $\mathbb{Q}_{k, m} = \text{ML}_{k, m}$ et

$$(1.5) \quad \text{pen}(n, k, m) = \sum_{(k', m') \preccurlyeq (k, m)} \left(\frac{1}{2} N(k', m') \log n + \tau(n, k', m') \right) + \alpha \varphi(k, m) \log n.$$

Alors, \mathbb{P}_0 -presque sûrement, $(\widehat{k, m})_n = (k_0, m_0)$, pour n assez grand.

Le choix naturel (car donnant la pénalité la plus petite possible) pour la fonction φ consiste à prendre $\varphi(k, m) = |\{(k', m') \in \mathbb{N}^* \times \mathbb{N} : (k', m') \preccurlyeq (k, m)\}|$.

Notre pénalité apparaît clairement comme une somme cumulée de termes de type pénalités BIC (c'est-à-dire de la forme $N(k, m) \log n/2$). En ce sens, cette pénalité est trop grande, et il serait souhaitable d'établir la consistance du critère BIC pour les CMRM. Cependant, sans supposer que le nombre de modèles possibles est fini et connu (ce qui revient à supposer qu'il existe une borne a priori sur le nombre d'états cachés et sur la mémoire du processus), un tel résultat n'est pas encore à notre portée. En effet, ce résultat n'est pas établi dans le cas beaucoup plus simple (unidimensionnel) des chaînes de Markov cachées. Les preuves de convergence de ce critère pour un nombre non fini de modèles à sélectionner, sont relativement peu nombreuses en dehors du cadre i.i.d.. On citera le cas des chaînes de Markov [32] et celui des modèles à arbres de contexte [33]. La preuve de [32] est assez difficile, et repose en particulier sur l'expression analytique (via de simples comptages) de l'estimateur du maximum de vraisemblance dans ce modèle. Dans [33], il est également fait usage de la forme explicite du maximum de vraisemblance. Or de telles expressions ne sont pas disponibles dans le cadre des chaînes de Markov cachées. Notons aussi que ces preuves utilisent un résultat délicat à obtenir sur les suites typiques d'un processus Markovien.

Pour revenir à notre pénalité, elle est donc certainement trop grande (par rapport à la pénalité BIC) mais son expression est en fait inspirée d'une étude similaire [57] dans le cas de l'estimation de l'ordre d'une chaîne de Markov cachée (i.e. du nombre d'états cachés). Nous reviendrons sur les performances de la pénalité BIC dans une étude de simulations.

La preuve de notre théorème passe par l'étude de deux événements très différents : la sous- et la sur-estimation du paramètre. Dans le cas de la sur-estimation, nous pouvons de plus donner un résultat de

vitesse de convergence.

Proposition 1.2. *Sous les hypothèses du Théorème 1.1, \mathbb{P}_0 -presque sûrement, $(\widehat{k}, \widehat{m})_n \preceq (k_0, m_0)$ pour n assez grand. De plus,*

$$\mathbb{P}_0 \left\{ (\widehat{k}, \widehat{m})_n \succ (k_0, m_0) \right\} = O(n^{-\alpha}),$$

où $\alpha > 1$ est choisi dans le Théorème 1.1.

1.4. Données simulées

Nous avons complété notre étude théorique par une série de simulations, en nous appuyant sur l'algorithme EM pour approcher la vraisemblance des observations. Le détail de ces simulations n'est pas redonné ici, et nous renvoyons le lecteur à [M8] pour de plus amples détails. Mentionnons simplement que notre procédure de sélection de l'ordre donne de très bons résultats lorsque le nombre d'observations est très grand ($n = 50000$) et fonctionne moins bien pour des tailles d'échantillon plus faibles ($n = 25000$). Il est à noter que cette très grande taille d'échantillon n'est pas nécessairement un facteur limitant dans le cadre de l'analyse de séquences biologiques. Nous avons également testé les performances du critère BIC, pour lequel le régime asymptotique est atteint beaucoup plus tôt (au moins à $n = 25000$ observations).

Toutes nos simulations ont été menées en éliminant le cas $k = 1$ (un seul régime). En effet, lorsque $k = 1$, la CMRM est réduite à une chaîne de Markov d'ordre m , c'est à dire à un processus homogène à mémoire finie. Ce type de processus est très différent en pratique des CMRM obtenues lorsque $k \geq 2$. Même si cette distinction n'est jamais apparue dans notre travail théorique, elle s'est avérée importante dans l'étude de simulations (les performances de la méthode lorsqu'on autorise $k = 1$ sont dégradées).

Remarquons également qu'une procédure Bayésienne de sélection de l'ordre (k, m) d'une CMRM a été proposée dans [13], procédure basée sur des méthodes de Monte Carlo par chaînes de Markov (MCMC) à sauts réversibles. Cette procédure est appliquée directement sur des données réelles, sans que sa pertinence ne soit justifiée par une étude de simulations. La convergence des méthodes MCMC à sauts réversibles est un problème délicat en général, d'autant plus sensible ici que l'espace des paramètres à explorer, de par sa double dimension, est gigantesque.

2. Les chaînes semi-Markov cachées

2.1. Contexte

L'utilisation d'une chaîne de Markov dans la modélisation du processus caché présente une caractéristique bien particulière : la longueur des plages homogènes qui sont ainsi modélisées suit une distribution géométrique. Or, les données empiriques obtenues sur des séquences réelles s'ajustent souvent très mal à cette contrainte. Par exemple, dans le cas de la reconnaissance de parole, cette distribution géométrique est peu adaptée à la longueur des segments de temps de parole. Il en est de même dans le cas de séquences biologiques, pour la longueur de segments spécifiques, tels les régions riches en $C + G$ ou encore les exons (pour lesquels une distribution Binomiale négative s'ajuste bien mieux aux observations).

Il a alors semblé assez naturel de se tourner vers d'autres types de modélisation, et les chaînes semi-Markoviennes sont apparues comme une alternative prometteuse. En effet, les chaînes semi-Markov sont une généralisation des chaînes de Markov pour lesquelles la distribution du temps de séjour dans un état n'est plus nécessairement géométrique, mais peut suivre n'importe quelle distribution fixée.

Les processus semi-Markoviens furent introduits simultanément par Lévy [84] et Smith [115]. Ces processus sont induits par des processus de renouvellement Markoviens, tout comme les processus de comptage sont induits par des processus de renouvellement simples. Une définition imprécise mais descriptive peut être donnée de la façon suivante : il s'agit d'un processus stochastique qui saute parmi un nombre fini d'états, la succession des états visités formant une chaîne de Markov, et les temps de séjour dans chacun des états suivent une distribution fixée, qui peut dépendre de cet état (la distribution du temps de séjour pourrait également dépendre de l'état suivant, mais ce cas de figure n'est en fait pas plus général). Un processus semi-Markovien peut donc être vu comme une chaîne de Markov dont l'indice temporel a subi un changement d'échelle aléatoire (de la même façon que les processus de comptage sont des processus i.i.d. dont l'échelle de temps a été modifiée de façon aléatoire). On pourra trouver plus de détails sur les processus Markoviens de renouvellement dans [29, 107, 108].

Les chaînes semi-Markov cachées (CSMC, parfois appelées *explicit duration HMM* en anglais) furent tout d'abord introduites dans le domaine de la reconnaissance de la parole. Ferguson [50] puis Russel [111] ont considéré des CSMC pour lesquelles les temps de séjour sont n'importe quelle distribution sur un ensemble fini $\{1, \dots, D_{\max}\}$, où D_{\max} représente le temps de séjour maximal dans un état. Ce cas est appelé le cas *non paramétrique* dans la littérature, car le nombre de paramètres à estimer, bien que fini, peut être très grand. Cependant, dans un cadre asymptotique où le nombre d'observations devient arbitrairement grand (et D_{\max} reste fini), le modèle est paramétrique. Plus tard, et afin de réduire la complexité du modèle, Levinson [83] introduisit le cadre appelé *paramétrique*, et qui correspond à des distributions de temps de séjour décrites par un très petit nombre de paramètres (lois Gaussiennes restreintes à \mathbb{R}^+ ou lois Gamma, pour lesquelles un ou deux paramètres décrivent entièrement la distribution).

Les algorithmes classiques des CMC, tels le « forward-backward », EM ou même Viterbi, ont été généralisés au cas des CSMC. La difficulté majeure de ces généralisations réside dans la complexité des algorithmes. Mentionnons qu'il existe une approche qui consiste à modéliser une CSMC via une CMC simple dans laquelle on crée artificiellement des macro-états : la mise en parallèle ou en série d'états identiques permet de modifier la distribution du temps de séjour [31]. Ainsi, la mise en série d'états identiques génère une distribution de temps de séjour qui est une convoluée de la distribution géométrique, tandis que la mise en parallèle des états correspond à un mélange de lois géométriques.

Une question délicate concerne la prise en compte des censures droite et gauche. L'approche classique qui ne tient pas compte de ces censures impose les hypothèses (non réalistes) que le processus observé est entré dans un nouvel état au premier temps d'observation (censure gauche) et qu'il en sort au dernier temps de l'observation (censure droite). La censure gauche peut être gérée en utilisant un processus de renouvellement *retardé*, c'est-à-dire que la toute première transition peut être différente des autres. Guédon [64] a proposé des équations de « forward-backward » qui prennent en compte le phénomène de

la censure droite.

Enfin, des approches hybrides combinant les CSMC et les CMC ont été proposées pour modéliser des processus dans lesquels certains temps de séjour sont géométriques alors que d'autres suivent des distributions différentes [65, 66].

2.2. Perspectives

L'étude des propriétés asymptotiques du maximum de vraisemblance dans le cadre des CSMC est encore incomplète. Il n'existait aucun résultat dans ce domaine lorsque, avec Florence Muri (Université René Descartes, Paris) et Anne-Sophie Tocquet (Université d'Évry Val d'Essonne), nous avons essayé de généraliser ce qui se passait dans le cadre des CMC. Nous avons rapidement constaté que le cas des CSMC dont la distribution des temps de séjour est un espace d'états fini (du type $\{1, \dots, D_{\max}\}$) se résout sans aucune difficulté supplémentaire par rapport au cas des CMC. Notons que ce cadre a depuis été étudié dans [7]. Le cas où la distribution des temps de séjour est l'ensemble \mathbb{N} tout entier reste quant à lui plus difficile à traiter car il se ramène au cas d'un espace d'états quelconque pour la CMC. Les récents travaux de Fuh [55] pour les CMC à espace d'états quelconques sont peut-être une perspective prometteuse pour analyser les propriétés du maximum de vraisemblance dans le cadre de CSMC à temps de séjour non bornés.

3. Les modèles pair-Markov cachés pour modéliser l'évolution des séquences

3.1. Contexte

Les modèles pair-Markov cachés (ou pair-HMM pour pair-hidden Markov models) permettent de faire de l'alignement de séquences dans un contexte évolutif et en utilisant des techniques de vraisemblance (on parle d'alignement *probabiliste*). L'alignement est un outil de comparaison des séquences très répandu. Pour une revue appliquée des outils statistiques communément utilisés en bio-informatique pour la génomique comparative, on pourra consulter [97]. La méthode classique d'alignement (que j'appellerai *déterministe* par la suite) consiste à choisir une fonction de score puis à chercher l'alignement qui maximise cette fonction. Cette étape est réalisée en pratique grâce aux algorithmes de programmation dynamique (Needleman et Wunsch pour l'alignement global et Smith et Waterman pour l'alignement local, voir [37]). L'alignement par fonction de score est une approche biaisée par le problème du choix des paramètres de la fonction de score, qui sont liés au processus d'évolution sous-jacent et donc à l'alignement que l'on souhaite obtenir. L'approche pair-Markov caché propose une alternative à ce problème en maximisant la vraisemblance des séquences observées sous un modèle particulier d'évolution entre ces séquences. L'alignement recherché correspond alors à une suite de variables cachées qui traduit les événements d'insertion/délétion ou de mutation du processus d'évolution sous-jacent. Dans un tel contexte, les paramètres du modèle d'évolution (qui correspondent aux paramètres d'une fonction de score sous-jacente) sont directement estimés, par maximum de vraisemblance et non fixés arbitrairement.

Les algorithmes d'estimation des paramètres dans les modèles pair-Markov cachés sont formalisés depuis une dizaine d'années et sont des généralisations simples des algorithmes existants pour des modèles de Markov cachés. La différence fondamentale réside dans l'émission de deux séquences (au lieu d'une) pour chaque processus caché.

D'un point de vue théorique, les deux modèles (pair-Markov caché et Markov caché) sont pourtant très différents et rien ne garantit a priori la convergence des algorithmes, ni du maximum de vraisemblance dans le cadre pair-Markov caché.

En collaboration avec Ana Arribas-Gil (Universidad Carlos 3, Madrid) et Elisabeth Gassiat (Université Paris Sud Orsay), nous avons fourni dans [M7] un cadre formel qui permet l'étude des modèles pair-Markov cachés, ce qui n'existait pas jusqu'alors. Le processus caché est une marche aléatoire sur $\mathbb{N} \times \mathbb{N}$ astreinte à la croissance. Les pas élémentaires sont du type $(1, 0)$, qui correspond à une insertion dans la première séquence (ou une délétion dans la deuxième); $(0, 1)$ qui représente le phénomène inverse; et $(1, 1)$ qui correspond à un *match* entre les deux séquences (voir Figure 1).

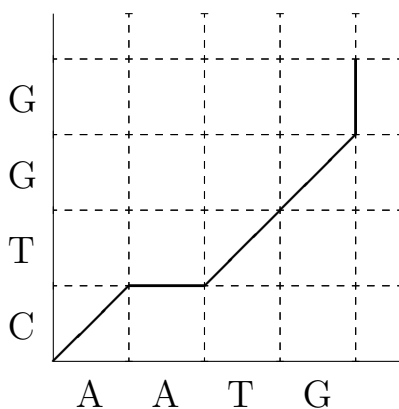


FIG. 1. Représentation graphique d'un alignement entre les deux séquences $X = AATG$ et $Y = CTGG$. L'alignement représenté correspond à $\begin{matrix} A & A & T & G & - \\ C & - & T & G & G \end{matrix}$.

Ce processus conditionne le tirage des deux séquences observées X et Y , avec un phénomène de distorsion aléatoire du temps : a priori, la variable aléatoire X_i n'a pas été tirée selon une loi qui dépend de la valeur Z_i du processus caché au temps i , et elle n'est pas tirée en même temps que Y_i . C'est là toute la différence avec les modèles de Markov caché où la distribution de l'observation au temps i dépend uniquement de la variable cachée à ce même temps i . Ce phénomène rend l'étude des modèles pair-Markov cachés délicate.

Les premiers modèles pair-Markov cachés sont apparus dans les travaux de Thorne Kishino et Felsenstein (TKF dans la suite) [119, 120] (et étaient déjà en germe dans [11]). Le modèle TKF est un modèle d'évolution de séquences, dans lequel chaque site évolue avec un taux de mutation constant, peut mourir avec un taux également constant, et de nouveaux sites peuvent venir s'insérer à la droite d'un site existant

depuis le début de l'évolution, avec un taux également constant. Ce modèle particulier d'évolution génère le modèle pair-Markov caché avec des paramètres contraints. Pour de plus amples détails sur les liens entre les modèles TKF et pair-Markov caché nous renvoyons à la thèse d'Ana Arribas-Gil [5]. Notons simplement que nous partons ici d'un modèle pair-Markov caché général, qui n'est pas forcément issu du modèle d'évolution TKF ou de ses variantes (i.e. nous n'imposons aucune contrainte sur les paramètres). Dans tous les cas, un alignement probabiliste obtenu sous un modèle pair-Markov caché (en maximisant la probabilité a posteriori du processus caché conditionnellement aux observations) correspond exactement à un alignement déterministe par score, pour certaines valeurs de la fonction de score (voir [37]). Nous reviendrons à une paramétrisation contrainte par un modèle d'évolution de séquences dans l'énoncé des derniers résultats (Théorèmes 3.2 et 3.3).

3.2. Description du modèle

Soit $\{\varepsilon_t\}_{t \geq 1}$, une chaîne de Markov stationnaire ergodique sur l'espace d'états $\mathcal{E} = \{(1, 0); (0, 1); (1, 1)\}$, de matrice de transition π et de loi stationnaire $\mu = (p, q, r)$. Cette chaîne induit une marche aléatoire $\{Z_t\}_{t \geq 0}$ à valeurs dans la grille $\mathbb{N} \times \mathbb{N}$, définie par $Z_0 = (0, 0)$ et $Z_t = \sum_{1 \leq s \leq t} \varepsilon_s$. Les coordonnées (aléatoires) de Z_t à l'instant t sont notées (N_t, M_t) (i.e. $Z_t = (N_t, M_t)$). Nous utiliserons tantôt la notation $\pi(\varepsilon_s, \varepsilon_{s+1})$ pour les probabilités de transition associées à π , tantôt des symboles explicites comme π_{HV} pour indiquer une transition de l'état *horizontal* $H = (1, 0)$ vers l'état *vertical* $V = (0, 1)$ (de la même façon avec $D = (1, 1)$ l'état *diagonal*).

Conditionnellement à cette marche aléatoire latente, les observations sont distribuées de la façon suivante. À l'instant t , si $\varepsilon_t = (1, 0)$ alors on tire une variable aléatoire X suivant une distribution f sur \mathcal{A} , si $\varepsilon_t = (0, 1)$, on tire une variable aléatoire Y suivant une distribution g sur \mathcal{A} et enfin si $\varepsilon_t = (1, 1)$, un couple de variables (X, Y) est tiré selon la distribution h sur $\mathcal{A} \times \mathcal{A}$. Conditionnellement au processus de Markov caché $\{\varepsilon_t\}_{t \geq 1}$, toutes les variables sont tirées indépendamment. Ce modèle est paramétré par $\theta = (\pi, f, g, h) \in \Theta$. La distribution conditionnelle des observations peut donc s'écrire de la façon suivante

$$(3.1) \quad \mathbb{P}(X_{1:N_t}, Y_{1:M_t} | \varepsilon_{1:t}, \{\varepsilon_s\}_{s > t}, \{X_i, Y_j\}_{i \neq N_s, j \neq M_s, 0 \leq s \leq t}) = \mathbb{P}(X_{1:N_t}, Y_{1:M_t} | \varepsilon_{1:t}) \\ = \prod_{s=1}^t f(X_{N_s})^{1_{\{\varepsilon_s=(1,0)\}}} g(Y_{M_s})^{1_{\{\varepsilon_s=(0,1)\}}} h(X_{N_s}, Y_{M_s})^{1_{\{\varepsilon_s=(1,1)\}}}.$$

De plus, la distribution complète \mathbb{P} s'écrit

$$\mathbb{P}(\varepsilon_{1:t}, X_{1:N_t}, Y_{1:M_t}) = \mu(\varepsilon_1) \left\{ \prod_{s=2}^t \pi(\varepsilon_{s-1}, \varepsilon_s) \right\} \mathbb{P}(X_{1:N_t}, Y_{1:M_t} | \varepsilon_{1:t}).$$

Cette distribution est naturellement paramétrée par $\theta = (\pi, f, g, h)$.

Notons qu'une condition nécessaire pour que le paramètre soit identifiable s'exprime par le fait que la probabilité d'occurrence de deux lettres alignées doit différer du produit des probabilités des occurrences de ces deux lettres, si elles étaient non-alignées. Autrement dit

Hypothèse 3.1.

$$\exists x, y \in \mathcal{A}, \text{ tels que } h(x, y) \neq f(x)g(y).$$

En effet, si $h = fg$, alors l'équation (3.1) nous donne

$$\mathbb{P}(X_{1:N_t}, Y_{1:M_t} | \varepsilon_{1:t}) = \left\{ \prod_{i=1}^{N_t} f(X_i) \right\} \left\{ \prod_{j=1}^{M_t} g(Y_j) \right\} = \mathbb{P}(X_{1:N_t}, Y_{1:M_t}).$$

Dans ce cas, la distribution des observations ne dépend pas du processus latent et le paramètre π ne peut pas être identifié. Dans la suite, nous supposons toujours que l'Hypothèse 3.1 est vérifiée.

3.3. Vraisemblance

Une des premières difficultés réside dans la définition de la notion de vraisemblance. En effet, lorsque l'on observe deux séquences $X_{1:n}$ et $Y_{1:m}$ dans le cas de figure le plus général, le point (n, m) n'est pas nécessairement un point qui appartient à la trajectoire de la marche Z_t non observée. Il convient donc, pour obtenir la distribution marginale de $X_{1:n}, Y_{1:m}$ de sommer sur tous les processus cachés. Conditionnellement à un alignement Z_t fixé, la distribution de $X_{1:n}, Y_{1:m}$ dépend des variables Z_t pour tous les temps t tels que $N_t \leq n$ ou $M_t \leq m$, et donc fait potentiellement intervenir d'autres variables $X_j, j > n$ ou $Y_j, j > m$. Il faut également noter que la longueur du processus caché peut être infinie. Ces considérations sont illustrées dans la Figure 2.

Introduisons tout d'abord quelques notations. L'ensemble \mathcal{E}_∞ désigne toutes les trajectoires possibles du processus caché; $\mathcal{E}_{n,m}$ celles qui passent par le point (n, m) et $\mathcal{E}_{n,m}^{-H}$ (resp. $\mathcal{E}_{n,m}^{-V}$) la restriction de l'ensemble $\mathcal{E}_{n,m}$ aux trajectoires dont le dernier état n'est pas horizontal (resp. vertical). Ainsi,

$$\begin{aligned} \mathcal{E}_\infty &= \{(0, 1); (1, 0); (1, 1)\}^{\mathbb{N}} = \{e = (e_1, e_2, \dots)\} = \mathcal{E}^{\mathbb{N}}, \\ \mathcal{E}_{n,m} &= \{e \in \{(0, 1); (1, 0); (1, 1)\}^l; n \vee m \leq l \leq n + m; \sum_{i=1}^l e_i = (n, m)\} \\ \mathcal{E}_{n,m}^{-H} &= \{e = (e_1, \dots, e_{|e|}) \in \mathcal{E}_{n,m}; e_{|e|} \neq (1, 0)\}, \quad \mathcal{E}_{n,m}^{-V} = \{e \in \mathcal{E}_{n,m}; e_{|e|} \neq (0, 1)\}. \end{aligned}$$

La vraisemblance des observations dans le modèle est donnée par

$$\begin{aligned} \mathbb{P}(X_{1:n}, Y_{1:m}) &= \sum_{e \in \mathcal{E}_\infty} \mathbb{P}(\varepsilon_{1:\infty} = e_{1:\infty}, X_{1:n}, Y_{1:m}) \\ &= \sum_{e \in \mathcal{E}_{n,m}} \mathbb{P}(\varepsilon_{1:|e|} = e, X_{1:n}, Y_{1:m}) + \sum_{l > n} \sum_{e \in \mathcal{E}_{l,m}^{-H}} \sum_{x_{n+1:l}} \mathbb{P}(\varepsilon_{1:|e|} = e, X_{1:n}, X_{n+1:l} = x_{n+1:l}, Y_{1:m}) \\ &\quad + \sum_{l > m} \sum_{e \in \mathcal{E}_{n,l}^{-V}} \sum_{y_{m+1:l}} \mathbb{P}(\varepsilon_{1:|e|} = e, X_{1:n}, Y_{1:m}, Y_{m+1:l} = y_{m+1:l}). \end{aligned}$$

Sous cette forme, une telle quantité n'est pas calculable. Nous avons donné dans [M7] des formules de récurrence qui permettent de la calculer. Cependant, ce n'est pas cette quantité qui est considérée par

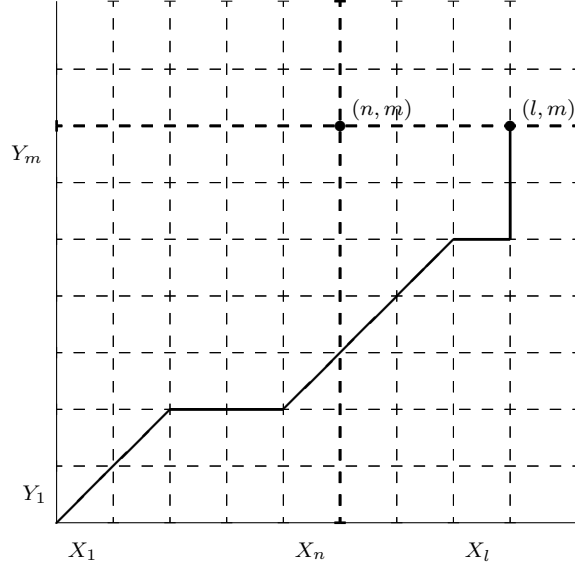


FIG. 2. Représentation graphique d'un alignement des séquences $X_{1:n}$ et $Y_{1:m}$ pour lequel le processus caché ne passe pas par le point (n, m) .

les algorithmes d'alignement probabiliste, et qui donc nous intéresse. Ceux-ci supposent en effet que l'alignement non observé est une trajectoire qui passe par le point (n, m) [37]. Nous introduisons donc

$$w_t(\theta) = \log \mathbb{Q}(X_{1:N_t}, Y_{1:M_t}), \quad t \geq 1$$

où, pour tous entiers n et m ,

$$\mathbb{Q}(X_{1:n}, Y_{1:m}) = \mathbb{P}(\exists s \geq 1, Z_s = (n, m); X_{1:n}, Y_{1:m}).$$

Ainsi, \mathbb{Q} est la probabilité d'observer les deux séquences, en supposant que le processus caché $\{\varepsilon_t\}_{t \geq 1}$ passe par le point (n, m) . Il faut noter que la longueur de l'alignement est inconnue lorsqu'on calcule \mathbb{Q} . Ainsi,

$$\mathbb{Q}(X_{1:n}, Y_{1:m}) = \sum_{e \in \mathcal{E}_{n,m}} \mathbb{P}(\varepsilon_{1:|e|} = e, X_{1:n}, Y_{1:m}).$$

Nous définissons donc

$$w_t(\theta) = \log \mathbb{P}(\exists s \geq 1, Z_s = (N_t, M_t); X_{1:N_t}, Y_{1:M_t}), \quad t \geq 1,$$

et la longueur de la trajectoire cachée n'est pas nécessairement t . Notons que \mathbb{Q} est la quantité qui est calculée par l'algorithme *forward* pour pair-Markov caché (voir [37]) et qui va jouer le rôle de la

vraisemblance pour le statisticien. De nombreux articles de bio-informatique proposent de maximiser cette quantité par rapport au paramètre θ puis d'utiliser l'algorithme de Viterbi ou la loi a posteriori du processus caché pour obtenir l'alignement des séquences [69, 75, 95, 96]. Il nous importait donc de comprendre le comportement asymptotique des estimateurs obtenus par cette méthode.

3.4. Résultats

Nous souhaitons prouver la consistance asymptotique de l'estimateur obtenu par maximisation de \mathbb{Q} . Il convient pour cela d'étudier la limite renormalisée de w_t lorsque le nombre d'observations tend vers $+\infty$, ce qui est équivalent ici à dire que t tend vers $+\infty$. Nous noterons θ_0 la valeur du vrai paramètre et $\mathbb{P}_0, \mathbb{E}_0$ la distribution et l'espérance associés. Nous avons pu identifier un certain nombre de cas pour lesquels le critère limite identifie le vrai paramètre. Malheureusement, ces cas ne sont pas exhaustifs.

Afin de présenter nos résultats, nous introduisons les espaces de paramètres suivants.

$$\begin{aligned}\Theta_0 &= \{\theta \in \Theta \mid \pi(i, j) > 0, f(x) > 0, g(y) > 0, h(x, y) > 0, \forall i, j \in \mathcal{E}, \forall x, y \in \mathcal{A}\}, \\ \Theta_{esp} &= \{\theta \in \Theta_0 : \forall \lambda > 0, \mathbb{E}(\varepsilon_1) \neq \lambda \mathbb{E}_0(\varepsilon_1)\} \\ \Theta_{marg} &= \{\theta \in \Theta_0 : h_X = f, h_Y = g\},\end{aligned}$$

où h_X (resp. h_Y) est la première (resp. seconde) marginale de h . Dans la suite, nous supposons toujours que $\theta_0 \in \Theta_0$. L'ensemble Θ_{esp} contient les paramètres tels que l'espérance de ε_1 sous le paramètre θ et sous le vrai paramètre θ_0 (inconnu) sont des points non alignés avec $(0, 0)$. Cette condition signifie que la marche aléatoire cachée Z_t prend des directions différentes sous les paramètres θ et θ_0 .

Théorème 3.1. *Pour tout $\theta \in \Theta_0 : t^{-1}w_t(\theta)$ converge \mathbb{P}_0 -presque sûrement et dans \mathbb{L}_1 , lorsque t tend vers $+\infty$ vers*

$$w(\theta) = \lim_{t \rightarrow \infty} \frac{1}{t} \mathbb{E}_0(\log \mathbb{Q}(X_{1:N_t}, Y_{1:M_t})) = \sup_t \frac{1}{t} \mathbb{E}_0(\log \mathbb{Q}(X_{1:N_t}, Y_{1:M_t})).$$

Définissons $D(\theta|\theta_0) = w(\theta_0) - w(\theta)$. On obtient alors

- Pour tout $\theta \in \Theta_0$, $D(\theta|\theta_0) \geq 0$.
- Pour tout $\theta \in \Theta_{esp}$, $\theta \neq \theta_0$, on a $D(\theta|\theta_0) > 0$.
- Si θ_0 et θ sont dans Θ_{marg} , alors $D(\theta|\theta_0) > 0$ dès que $f \neq f_0$ ou $g \neq g_0$.

Il faut remarquer que dans le cas où $p = q$ (les probabilités stationnaires d'apparition d'une insertion ou d'une délétion pour le processus caché), les espérances de ε_1 sous les paramètres θ et θ_0 sont alignées avec $(0, 0)$. Dans ce cas, nous ne savons pas montrer que si $h \neq h_0$, alors le critère limite identifie le vrai paramètre (i.e. $D(\theta|\theta_0) > 0$).

L'existence de la limite $w(\theta)$ est obtenue par un critère (presque classique) de sous-additivité de Kingman. La positivité de $D(\theta|\theta_0)$ découle simplement de l'écriture comme la limite d'une divergence de Kullback-Leibler. Par contre, l'identification de θ_0 par le critère limite se fait à chaque fois par une preuve originale.

Nous revenons à présent au problème de départ : le modèle pair-Markov caché provient d'un modèle d'évolution de séquences (type TKF ou autre) et les paramètres sont donc contraints par le modèle d'évolution initial. Notons $\beta \mapsto \theta(\beta)$ une paramétrisation continue d'un ensemble B vers Θ . Pour tout $\delta > 0$, nous définissons Θ_δ l'ensemble des paramètres dont toutes les coordonnées sont minorées par δ et $B_\delta = \theta^{-1}(\Theta_\delta)$. Nous supposons qu'il existe un $\delta > 0$ tel que $\beta_0 = \theta^{-1}(\theta_0)$ appartienne à B_δ . Notre estimateur est défini par

$$\hat{\beta}_t = \underset{\beta \in B_\delta}{\text{Argmax}} w_t(\theta(\beta)).$$

Alors on peut montrer le théorème suivant.

Théorème 3.2. *Si l'ensemble des maxima de $w(\theta(\beta))$ sur B_δ est réduit à $\{\beta_0\}$, alors $\hat{\beta}_t$ converge \mathbb{P}_0 -presque sûrement vers β_0 .*

On se place ensuite dans un cadre Bayésien où ν est une loi a priori sur l'ensemble B_δ . On s'intéresse à la loi a posteriori du paramètre sachant les observations, définie plus précisément de la façon suivante :

$$\nu_{|X_{1:N_t}, Y_{1:M_t}}(d\beta) = \frac{Q_{\theta(\beta)}(X_{1:N_t}, Y_{1:M_t})\nu(d\beta)}{\int_{B_\delta} Q_{\theta(\beta')} (X_{1:N_t}, Y_{1:M_t})\nu(d\beta')}.$$

On obtient le résultat suivant.

Théorème 3.3. *Si l'ensemble des maxima de $w(\theta(\beta))$ sur B_δ est réduit à $\{\beta_0\}$, et si ν charge β_0 , alors la suite de mesures a posteriori $\nu_{|X_{1:N_t}, Y_{1:M_t}}$ converge en loi \mathbb{P}_0 -presque sûrement, vers la masse de Dirac en β_0 .*

Évidemment, les deux théorèmes précédents reposent sur l'hypothèse très forte que le critère limite w identifie le vrai paramètre β_0 , ce que nous n'avons obtenu que partiellement dans le Théorème 3.1. Ces conditions ne sont pas aisément vérifiables même dans le modèle relativement simple d'évolution TKF (pour plus de détails sur la paramétrisation induite par TKF, voir [5]). Nous avons effectué des simulations afin de compléter cette approche théorique. Dans des cas où les paramètres ne vérifient pas les contraintes imposées par les théorèmes (comme par exemple $p = q$ et $h_X = h_Y = f = g$ fixés) la procédure semble cependant donner également de bons résultats.

3.5. Commentaires et conclusions

La généralisation du modèle pair-Markov caché à plus de deux séquences n'est pas immédiate et a été traitée dans [5, Chapitre 4] pour le modèle d'évolution TKF91, avec un arbre phylogénique fixé. En effet, le processus Markovien latent qui doit être introduit pour 3 séquences par exemple, n'est pas simplement une chaîne de Markov dans l'espace d'états $\{0, 1\}^3 \setminus \{(0, 0, 0)\}$, car un tel processus ne correspond pas au modèle d'évolution. Le processus latent est ici une chaîne de Markov indexée par les L sites de la séquence commune ancestrale située à la racine de l'arbre, et non plus par les sites de l'alignement. Pour chacun de ces sites, le processus latent indique les séquences pour lesquelles la position a été conservée (avec une éventuelle mutation), et pour les autres séquences, la taille éventuelle du segment inséré juste à droite de ce site.

Notons simplement que cette approche est différente de ce qu'on nomme en bio-informatique les chaînes de Markov cachées profils [38, 80], et qui permettent également de faire de l'alignement probabiliste multiple. Les chaînes de Markov cachées profils (*profile HMM* en anglais) sont constituées de L états *match*, L états de *délétion* et $L + 1$ états d'*insertion*, où L est encore la longueur de la séquence ancestrale (i.e. du nombre de colonnes homologues de l'alignement). Les séquences sont alors supposées indépendantes conditionnellement à la structure cachée, ce qui est la différence principale avec l'alignement multiple évoqué ci-dessus. La valeur de L est souvent choisie comme la longueur moyenne des séquences à aligner et les paramètres du modèle sont estimés par l'algorithme EM. L'alignement par chaînes de Markov cachées profils est souvent présenté comme un alignement par *score spécifique à chaque position*, puisque les paramètres d'émission des observations, conditionnellement à la chaîne cachée, sont différents suivants les positions de l'alignement.

Nous souhaitons également attirer l'attention du lecteur sur le fait que le modèle pair-Markov caché donne un alignement global de séquences. L'alignement global est utilisé par exemple sur des séquences de longueur similaire, en vue d'inférer la phylogénie de ces séquences. Mais il n'est pas utilisé pour inférer des relations d'homologie entre des séquences, car le comportement de la queue de distribution du score d'alignement global n'est pas connue (contrairement au score d'alignement local pour lequel il existe des comportements approchés). Dans un travail (en cours) en collaboration avec Ana Arribas-Gil et Elisabeth Gassiat, nous essayons d'utiliser le modèle pair-Markov caché pour mettre en place un test d'homologie des séquences. L'idée est de tester, via un test du rapport de vraisemblance, l'hypothèse H_0 : « les deux séquences sont indépendantes et i.i.d. », contre H_1 : « les séquences ont une distribution pair-Markov cachée ».

Enfin, notons qu'une version « hybride » de l'alignement probabiliste a été proposée par Yu et ses co-auteurs [129, 130] pour permettre un alignement local des séquences.

4. Modèles d'évolution de séquences dépendants du contexte

Depuis septembre 2006, je co-dirige avec Léonid Galtchouk (Université Louis Pasteur à Strasbourg) la thèse d'Audrey Finkler. Audrey est une étudiante de Strasbourg qui souhaitait orienter son doctorat de statistique vers des applications à la génomique. Elle a contacté notre laboratoire au printemps 2006 et je lui ai proposé de faire un mémoire de Master(2) sur les modèles d'évolution des séquences, avec l'idée de l'orienter par la suite vers des modèles d'évolution « dépendants du contexte ». C'est dans cette voie qu'elle a commencé sa thèse (à Strasbourg, avec visites mensuelles à Évry). J'expose ici le contexte et les enjeux de cette thématique, ainsi que les premières voies explorées.

Dans toute la suite on ne considère que des processus de mutation (substitution) sur les séquences. Tous les autres phénomènes évolutifs, comme par exemple les insertions, délétions, translocations, inversions, ... ne sont pas pris en compte par ce modèle.

Les modèles de substitution de séquences sont des processus Markoviens à temps continu qui décrivent les mutations temporelles d'un site i le long d'une séquence (biologique). Les observations sont ici composées des colonnes d'un alignement de deux ou plusieurs séquences, de même longueur. Ces séquences

sont reliées par une phylogénie (i.e. un arbre enraciné dont les feuilles sont les séquences observées). La probabilité d'observer un ensemble de p séquences $X^{(1)}, \dots, X^{(p)}$ conditionnellement à la phylogénie τ (un arbre et des longueurs de branche) s'obtient facilement à partir de l'algorithme de Felsenstein [48]. Dans le cas simple de deux séquences X, Y , provenant de la racine avec des temps d'évolution respectifs depuis la racine t_1 et t_2 , cette probabilité s'écrit

$$\mathbb{P}(X, Y, \tau) = \sum_R \pi(R) \mathbb{P}(X|R, t_1) \mathbb{P}(Y|R, t_2),$$

où la somme porte sur toutes les séquences possibles à la racine de l'arbre, π est la loi stationnaire (on suppose qu'elle existe) du processus de Markov, et $\mathbb{P}(X|R, t)$ est la probabilité de transition de la séquence R à la séquence X en un temps t . Lorsque le processus Markovien est *réversible*, cette probabilité se simplifie et devient

$$\mathbb{P}(X, Y, \tau) = \pi(Y) \mathbb{P}(X|Y, t) = \pi(X) \mathbb{P}(Y|X, t),$$

où $t = t_1 + t_2$ est la distance phylogénique entre les deux séquences. Dans toute la suite, nous supposons toujours que la phylogénie des séquences (et en particulier le temps d'évolution entre deux séquences) est fixé et connu.

La réversibilité du processus d'évolution est une hypothèse en contradiction avec la biologie, mais elle peut simplifier l'analyse du problème. Dans un modèle de substitution réversible, la racine de l'arbre ne peut pas être positionnée sans l'observation d'un *outgroup*, i.e. d'une séquence dont on sait a priori qu'elle est en dehors du groupe monophylétique observé. L'hypothèse de loi stationnaire (aussi bien à la racine qu'aux feuilles de l'arbre) est raisonnable étant donné l'échelle de temps de l'évolution et le fait que l'ancêtre commun le plus récent d'une famille de séquences (i.e. la racine de l'arbre) est proche des séquences observées relativement à cette échelle.

La vaste majorité des modèles de substitution fait l'hypothèse d'indépendance et de distribution identique pour chacun des sites (colonnes de l'alignement). La vraisemblance d'un ensemble de séquences se factorise alors sur les sites, et sur chaque site agit un processus de Markov à temps continu. Si Q est la matrice des taux de transition du processus, alors

$$\mathbb{P}(Y|X, t) = [\exp(Qt)]_{X,Y}.$$

Si X et Y sont des nucléotides, la matrice Q est de taille 4×4 (ou 20×20 pour des acides aminés) et le calcul de la matrice $\exp(Qt)$ se fait facilement, par diagonalisation.

Depuis une dizaine d'années environ, l'hypothèse de distribution identique de ces sites a pu être relâchée, via l'introduction de modèles de mélange, mais en préservant l'hypothèse d'indépendance des sites [49]. Cependant, pour les biologistes, cette hypothèse d'indépendance des sites est trop simplificatrice, et ne permet pas de rendre compte de nombreux phénomènes observés, tels les forts taux de mutation à partir des paires *CpG*. Pour plus de détails sur les évidences biologiques de ces phénomènes, je renvoie à la très bonne (et très détaillée) introduction de [114]. Il importe donc de fournir à la bio-informatique des modèles d'évolution qui prennent en compte le contexte d'un site (i.e. ses voisins) dans la modélisation de cette évolution. D'un point de vue statistique, la relaxe de cette hypothèse reste un véritable enjeu.

En effet, lorsque les sites ne sont pas indépendants, il faut considérer la séquence dans son ensemble. La matrice Q est alors de taille $4^n \times 4^n$ (pour des séquences de longueur n et l'alphabet des nucléotides) et le calcul exact de $\exp(Qt)$ est numériquement trop coûteux pour des séquences de longueur raisonnable. De plus, l'existence de la loi stationnaire d'un tel processus Markovien, et la convergence vers cette loi sont loin d'être des questions triviales. Dans la plupart des articles de bio-informatique, cette question est balayée. De façon générale, il importe de restreindre les relations de dépendance entre sites afin d'obtenir des modèles dans lesquels l'estimation des paramètres reste possible. Dans la pratique, le nombre de paramètres à inférer doit également rester raisonnable et on utilise des considérations biologiques pour réduire le nombre de paramètres distincts.

4.1. Contexte

La bibliographie dans ce domaine, bien que récente, est déjà très vaste et je ne prétends nullement ici à l'exhaustivité. Les premières tentatives de prise en compte de la dépendance sont des modèles pour lesquels on ne considère pas la suite des nucléotides, mais la suite des couples de nucléotides (modèle « doublet » de Schöniger et von Haeseler [112, 125]) ou de codons [61, 98]. Les « doublets » ou les codons sont alors supposés i.i.d et la dépendance ne se fait qu'au sein des nouvelles unités, qui ne se chevauchent pas. Ainsi, le processus est indépendant sur une sous-séquence de la séquence initiale. Ces approches étaient déjà en germe dans [85]. Les modèles de « doublet » sont en particulier utilisés pour modéliser des dépendances entre paires de sites contraints par la structure secondaire d'un ARN (les doublets ne sont pas a priori des nucléotides successifs le long de la séquence). Les modèles de codon sont utilisés pour modéliser des séquences de gènes. Ils présentent l'intérêt de prendre en compte des taux de substitution différents pour les substitutions synonymes (qui donnent le même acide aminé) et les substitutions non synonymes. Cependant, un seul cadre de lecture est alors pris en compte (puisque la suite des codons est fixée a priori avant l'analyse).

Un modèle de codon qui autorise la dépendance aux codons voisins a ensuite été proposé par Pedersen et Jensen [74, 105]. Plus précisément, le modèle proposé permet de tenir compte de l'existence de deux cadres de lecture dans une séquence codante. Le modèle de substitution est le suivant : le taux instantané de mutation en un site est proportionnel à la fréquence stationnaire du nucléotide qui apparaît. De plus, les transitions (i.e. les substitutions qui respectent les classes *purines* = $\{A, G\}$ et *pyrimidines* = $\{C, T\}$, par opposition aux *transversions*) sont multipliées par un facteur K . Enfin, lorsque la substitution modifie le codon dans le premier cadre de lecture (resp. le second, ou les deux) le taux de mutation instantanée est encore multiplié par un facteur f_I (resp. f_{II} et $f_{I,II}$). Le taux de mutation est nul lorsque la substitution entraîne l'apparition d'un codon *stop* dans l'un au moins des deux cadres de lecture.

Ainsi, les nucléotides de chaque codon suivent un processus de Markov dont la matrice des taux de substitution dépend des deux voisins de gauche et de droite de chaque nucléotide. Mais les taux de substitution sont suffisamment simples pour que le modèle soit réversible, avec une loi stationnaire qui se factorise sur les codons.

Dans ce modèle, la vraisemblance des séquences alignées ne peut pas être obtenue comme un produit de facteurs sur tous les sites (i.e. les codons dans le cas présent). Les auteurs proposent donc d'utiliser

une méthode MCMC pour simuler le ratio $\mathbb{P}_{\theta_1}(X, Y, t_1)/\mathbb{P}_{\theta_2}(X, Y, t_2)$ des probabilités de transition, sous deux paramètres différents θ_1 et θ_2 , d'une séquence X à une séquence Y (et avec des temps d'évolution t_1, t_2 différents). L'estimation de ce ratio permet d'obtenir un estimateur du maximum de vraisemblance (EMV) approchée, ainsi que de faire des tests du rapport de vraisemblance. Pour le calcul de l'EMV, le paramètre θ_2 est fixé tandis que le ratio est maximisé par rapport à θ_1 . Au cours de cette maximisation, il convient de réajuster la valeur de θ_2 (et donc l'estimation du ratio) afin qu'une trop grande distance entre θ_1 et θ_2 ne viennent augmenter la variance de l'estimation du ratio. La méthode est donc extrêmement lourde en temps de calcul.

D'autres approches de ce problème ont été considérées. Arndt Burge et Hwa [4] étudient un modèle autorisant des substitutions de dinucléotides. La mesure stationnaire du processus est difficilement calculable (la fréquence d'un nucléotide dépend de celle des dinucléotides, qui elle-même dépend de la fréquence des tri-nucléotides, etc ...). Les auteurs proposent de l'approcher par une chaîne de Markov d'ordre 1 (*two-clusters approximation*) et arguent des bons résultats de cette approximation comparée à une simulation par méthode de Monte Carlo. Pour estimer les paramètres du modèle d'évolution, les auteurs proposent d'utiliser uniquement la distribution stationnaire (approchée) d'une séquence observée. On n'utilise pas ici de séquences alignées (d'où une perte d'information sur le processus de substitution). Les auteurs modélisent une séquence observée par une chaîne de Markov d'ordre 1, puis par des méthodes numériques (sur un espace de paramètres d'évolution restreint), ils obtiennent les paramètres du modèle d'évolution qui correspondent aux comptages observés.

Lunter et Hein [87] ont proposé un modèle de substitution de dinucléotides, irréversible (et donc permettant d'inférer la position de la racine de l'arbre). Ils utilisent une approximation du calcul de $\exp(Qt)$ où Q est la matrice de taille $4^n \times 4^n$ des taux de transition entre les séquences, approximation basée sur les premiers termes du développement de l'exponentielle (et en négligeant les termes d'ordre supérieur). La loi stationnaire est approchée par une chaîne de Markov d'ordre 2 (il s'agit d'une *3-cluster approximation*) avec une correction aux bords. Une approche de ré-échantillonnage Bayésien par MCMC est utilisée pour obtenir la loi a posteriori des paramètres, sachant deux séquences observées.

Dans [126], Whelan et Goldman utilisent une modélisation du processus d'évolution par un mélange de trois composantes : un processus de substitution sur les nucléotides, un autre qui agit sur les dinucléotides (en changeant les deux bases simultanément) et un troisième agissant sur les codons. Grâce à une approximation du type champ moyen ils maintiennent l'hypothèse de sites (codons) indépendants ce qui simplifie leur analyse, tout en prenant en compte des événements qui affectent jusqu'à trois nucléotides à la fois.

Christensen et co-auteurs [27, 28] ont proposé une approche basée sur une technique de pseudo-vraisemblance, inspirée des champs de Markov [8]. La pseudo-vraisemblance des observations est le produit sur tous les sites, de la vraisemblance du site conditionnelle à ses voisins. L'état des voisins n'étant pas connu tout au long du processus d'évolution, on suppose que pour un site i , si l'état initial ($t = 0$) de la séquence X_i diffère de l'état final ($t = 1$) de la séquence Y_i , alors il y a eu une substitution de X_i vers Y_i au temps $t = 1/2$. Un algorithme EM est proposé pour maximiser cette pseudo-vraisemblance par rapport aux paramètres. Ce travail est étendu à un modèle non réversible et non stationnaire dans [27].

Yang [128] propose un modèle très intéressant où la suite des taux de transition en un site suit un processus de Markov non observé et conditionnellement à la donnée des taux de transition, les sites sont indépendants et suivent un processus Markovien à temps continu. Plus précisément, les matrices de taux de substitution utilisées sont obtenues à partir d'une seule matrice commune, via un facteur multiplicatif. Ce facteur multiplicatif est une variable aléatoire (non observée) distribuée selon une chaîne de Markov à espace d'états fini. Les équations de forward-backward sont utilisées pour le calcul de la vraisemblance, mais l'auteur préfère une stratégie d'optimisation de type quasi-Newton à un algorithme EM.

Notons que ces approches sont souvent limitées au cas de deux séquences, et (sauf pour [87]) font l'hypothèse d'un modèle réversible, et n'estiment donc pas la phylogénie des séquences. Dans [114], Siepel et Haussler proposent une nouvelle approche, nommée phyloHMM, qui permet de traiter plusieurs séquences ainsi que d'estimer la phylogénie de ces séquences. À partir d'un alignement de plusieurs séquences, Siepel et Haussler combinent le modèle de Yang [128] à taux de mutation dépendants avec la donnée d'un arbre phylogénique (éventuellement différent en chaque site). Dans les phyloHMM, la probabilité d'observer une colonne (site) de l'alignement est obtenue par l'algorithme de Felsenstein [48]. Cependant, l'approche de Siepel et Haussler n'est pas basée sur un modèle probabiliste, mais plutôt sur une approche empirique de modélisation de N -uplets qui peuvent se chevaucher.

Dans [17], Bérard, Gouéré et Piau s'attachent à décrire un modèle d'évolution de séquences où un seul site peut muter à chaque instant de temps, mais où on superpose des taux de mutation simples (i.e. qui ne dépendent que du site considéré), avec des taux de mutation doubles (i.e. qui dépendent d'un site et de son voisin de gauche ou de droite). En imposant certaines contraintes d'égalité sur ces taux de mutation, les auteurs se ramènent à un modèle dans lequel toute sous-séquence (de nucléotides successifs) de taille finie évolue indépendamment du reste de la séquence. En particulier, la loi stationnaire du processus existe et ses marginales s'obtiennent en résolvant des systèmes non linéaires. Dans [18], Bérard et Piau explorent la robustesse de leurs résultats lorsque le modèle de départ est perturbé (i.e. les taux de mutation sont légèrement perturbés et ne vérifient donc plus les contraintes fortes du modèle initial).

4.2. Modèle

Audrey Finkler travaille en ce moment sur un modèle de dépendance au voisin de gauche. Plus précisément, le processus d'évolution reste un processus Markovien (à temps discrétisé), et nous supposons que l'évolution d'un site dépend de l'état de ce site et du site voisin gauche, au temps précédent. En considérant l'évolution temporelle de chaque site comme un vecteur colonne, nous obtenons un processus Markovien spatial : la distribution de chaque colonne (état d'un site sur un intervalle de temps fixé) ne dépend que de l'état de la colonne précédente. Ce modèle d'évolution est en fait un cas particulier d'une chaîne de Markov cachée, et nous pouvons utiliser l'algorithme EM afin d'estimer les paramètres.

La perspective naturelle de ce travail est de prendre en compte les dépendances aux deux voisins (gauche et droite) en introduisant un modèle de champ Markovien.

4.3. Conclusions et perspectives

Le but de ces modèles d'évolution dépendants du contexte est de pouvoir faire, à terme, des inférences de phylogénie améliorées par rapport au cadre des sites indépendants. C'est déjà ce que font les phyloHMM, mais sans la référence à un modèle probabiliste.

Récemment, des algorithmes d'alignement (déterministe) de séquences ont été proposés, avec des fonctions de score qui prennent en compte le contexte de la position à aligner (voir [56] et également des approches plus anciennes [73, 127]). D'une part, le comportement de la queue de distribution des scores obtenus par de telles méthodes n'est pour le moment pas du tout connu, et empêche donc de tester la significativité d'un tel alignement. D'autre part, un tel alignement par fonction de score pose encore le difficile problème du choix des paramètres de la fonction de score, paramètres qui sont fortement liés à l'alignement que l'on souhaite obtenir. À terme, il serait intéressant de pouvoir proposer un modèle d'évolution dépendant du contexte, qui permettrait de faire un alignement probabiliste comme c'est le cas avec les modèles pair-Markov cachés. Une idée simple est de s'intéresser à des généralisations du modèle pair-Markov caché dans lesquelles, conditionnellement à l'alignement non observé, les séquences sont émises suivant une chaîne de Markov. Malheureusement, un tel modèle n'est pas a priori issu d'un modèle d'évolution sur les séquences.

Deuxième partie

Modèles semi paramétriques de signaux bruités

Dans cette partie, j'ai regroupé mes travaux concernant le modèle de convolution [M4,M9,M10] avec des travaux sur les fonctions périodiques bruitées [M6]. Hormise une partie de la section 5.4, le cadre général de ces études est semi paramétrique, et nous nous intéressons à l'impact de la partie fini-dimensionnelle sur l'estimation non paramétrique du paramètre infini dimensionnel. Ces travaux n'ont pas été motivés par des problématiques issues de la bio-informatique, mais pourraient trouver des applications dans ce domaine. Par exemple, les modèles de convolution sont utilisés pour la normalisation des données de biopuces ; les estimateurs à noyau non paramétrique sont utilisés pour des problématiques d'estimation du *local False Discovery Rate* [25], etc ...

5. Le modèle de convolution

Cette série d'articles ([M4,M9,M10]) reprend un modèle déjà étudié dans le second chapitre de ma thèse, et entreprend de fournir des résultats d'estimation mais aussi des tests d'adéquation non paramétriques sur la densité d'un signal convolué dans un modèle semi ou non paramétrique.

Dans un modèle de convolution, les observations sont constituées d'un échantillon de variables qui résultent de la somme indépendante d'un signal X et d'un bruit ε .

$$Y_k = X_k + \sigma\varepsilon_k, \quad 1 \leq k \leq n, \quad X_k \text{ i.i.d.}, \quad \varepsilon_k \text{ i.i.d.}, \quad \{X_k\} \perp \{\varepsilon_k\}.$$

En général, la distribution du bruit est supposée entièrement connue et on cherche à obtenir des résultats de type non paramétriques sur la densité du signal non observé. Cette hypothèse de la connaissance complète de la distribution du bruit est très forte et peut sembler peu réaliste. En effet, un paramètre d'échelle comme la variance du bruit fait souvent l'objet d'une estimation préliminaire et est ensuite supposé connu. Notre travail est motivé par l'étude de l'impact de cet éventuelle estimation préliminaire.

Les articles dans lesquels la distribution n'est pas supposée connue font parfois l'hypothèse que le statisticien dispose d'un second échantillon indépendant du premier, pour lequel seul le bruit est observé. C'est par exemple le cas de [99] mais ce n'est pas l'approche que nous suivrons ici. Koltchinskii [76] a proposé un estimateur de la matrice de covariance d'un bruit Gaussien dans un modèle de convolution multidimensionnel. Meister a étudié différents contextes de convolution où la loi du bruit n'est pas entièrement spécifiée [91–94].

Nous avons entamé une série de travaux portant sur des modèles semi paramétriques de convolution, dans lesquels la densité du bruit est supposée connue uniquement à paramètre (d'échelle ou de régularité) près. Une première partie de l'étude consiste à mettre en évidence l'existence de familles de modèles de convolution semi paramétriques qui soient *identifiables*.

Dans toute la suite, la densité des X_i sur \mathbb{R} est notée f , supposée dans $\mathbb{L}_2(\mathbb{R})$, et sa transformée de Fourier ($u \mapsto \int e^{iux} f(x) dx$) est notée Φ . Cette densité f sera toujours supposée inconnue. Les variables ε_i ont quant à elles une distribution de régularité dite **s-exponentielle** (sauf dans une partie de la Section 5.4), c'est-à-dire que leur densité f^ε a une transformée de Fourier notée Φ^ε vérifiant, pour tout $|u|$ assez grand,

$$be^{-|u|^s} \leq |\Phi^\varepsilon(u)| \leq Be^{-|u|^s},$$

où les constantes $b, B > 0$ sont supposées connues. Nous considérons successivement les deux cas suivants.

A) Cas d'échelle inconnue : le paramètre d'échelle $\sigma > 0$ est supposé inconnu, alors que la régularité $s > 0$ est supposée connue.

B) Cas de régularité inconnue : le paramètre d'échelle $\sigma > 0$ est supposé connu (et fixé sans perte de généralité à la valeur 1), alors que la régularité $s > 0$ est supposée inconnue.

Dans tous les cas, le paramètre de régularité s appartient à l'intervalle $(0; 2]$ (au-delà de la valeur 2, les fonctions obtenues ne sont plus des densités). Dans chacun des cas, il faudra faire des hypothèses supplémentaires sur la densité f du signal afin d'obtenir un modèle identifiable. Nous notons $f^Y = f \star f^\varepsilon$ la densité des observations (où \star désigne le produit de convolution) et Φ^Y sa transformée de Fourier. Notons que $\Phi^Y(u) = \Phi(u)\Phi^\varepsilon(\sigma u)$, pour tout $u \in \mathbb{R}$. Nous utiliserons également l'estimateur empirique de la fonction caractéristique Φ^Y de la distribution des observations, défini par

$$\hat{\Phi}_n^Y(u) = \frac{1}{n} \sum_{k=1}^n e^{iuY_k}, \quad \forall u \in \mathbb{R}.$$

Dans le modèle de convolution (avec densité du bruit connue), les estimateurs à noyau de la densité f sont définis de la façon suivante. À partir d'un noyau quelconque (ici nous prendrons $J(x) = \sin(x)/(\pi x)$ dont la transformée de Fourier est simplement $\Phi^J(u) = 1_{\{|u| \leq 1\}}$), nous définissons le noyau de déconvolution en utilisant la transformée de Fourier de la densité du bruit. Ainsi, nous introduisons le noyau k_n défini par sa transformée de Fourier

$$\Phi^{k_n}(u) = \left\{ \Phi^\varepsilon(\sigma u/h_n) \right\}^{-1} 1_{|\sigma u| \leq 1},$$

où la fenêtre h_n est positive et tend vers 0. Lorsque la densité du bruit est connue à paramètre près, il suffit alors de remplacer, dans cette expression, le paramètre d'échelle σ ou de régularité s par un estimateur préliminaire (méthode de « plug-in »). Le noyau de déconvolution devient aléatoire. Il est noté \hat{k}_n et défini par sa transformée de Fourier $\Phi^{\hat{k}_n}$. Ainsi, dans le cas où σ est inconnu et estimé par $\hat{\sigma}_n$, on définit

$$\Phi^{\hat{k}_n}(u) = \left\{ \Phi^\varepsilon(\hat{\sigma}_n u/h_n) \right\}^{-1} 1_{|\hat{\sigma}_n u| \leq 1},$$

et lorsque c'est la régularité s qui est inconnue et estimée par \hat{s}_n , on suppose pour simplifier les notations que $\sigma = 1$ et $\Phi^\varepsilon(u) = \exp(-|u|^s)$ et on utilise

$$(5.1) \quad \Phi^{\hat{k}_n}(u) = \exp(|u/\hat{h}_n|^{\hat{s}_n}) 1_{|u| \leq 1}.$$

[Ici, la fenêtre h_n dépend également de l'estimateur \hat{s}_n et devient aléatoire.] Ce nouveau noyau de déconvolution permet de définir l'estimateur à noyau de la densité f de la façon habituelle

$$(5.2) \quad \hat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^n \hat{k}_n \left(\frac{Y_i - x}{h_n} \right).$$

Suivant les cas, la densité f appartient soit à un ensemble $\mathcal{S}(\alpha, r, L)$ de fonctions analytiques sur une bande du plan complexe, soit à un espace de Sobolev $W(\beta, L)$. Dans le premier cas, f est dite *super régulière* ou encore de *régularité exponentielle*. Dans le second, f est *simplement régulière* ou encore de *régularité polynomiale*. Nous introduisons donc les ensembles de fonctions

$$\begin{aligned} \mathcal{S}(\alpha, r, L) &= \left\{ f; \text{densité telle que } \int |\Phi(u)|^2 e^{2\alpha|u|^r} du \leq L^2 \right\}, \\ \text{et } W(\beta, L) &= \left\{ f; \text{densité telle que } \int |\Phi(u)|^2 (1 + |u|^{2\beta}) du \leq L^2 \right\}, \end{aligned}$$

avec $\alpha, r, L > 0$ et $\beta > 1/2$.

Dans les parties suivantes, j'expose tout d'abord les résultats obtenus pour l'estimation du paramètre inconnu et de la densité f . On observe deux phénomènes très différents. Dans le cas **A**) (échelle inconnue) les vitesses d'estimation du paramètre d'échelle σ sont très lentes. Il en résulte une dégradation des vitesses d'estimation non paramétriques de la densité f , par rapport au cas où la distribution du bruit est entièrement connue (ces dernières ont été calculées dans [15, 16, 39]). De plus, nos vitesses sont optimales au sens du risque minimax : le paramètre σ joue donc un vrai rôle de *nuisance* dans ce modèle semi paramétrique. Dans le cas **B**) (régularité inconnue) au contraire, il n'y a aucune perte de vitesse de l'estimateur par *plug-in*, par rapport à l'estimateur à noyau classique, construit lorsque la densité du bruit est entièrement spécifiée. Cependant, nous estimons le paramètre de régularité s uniquement sur une grille, et pas dans un intervalle de valeurs continues. Ce paramètre ne joue pas le rôle d'une nuisance pour l'estimation de la densité f .

Les résultats de borne inférieure qui sont établis pour la vitesse de convergence des estimateurs reposent sur une variante du lemme de Fano [9, 10, 47] qui n'utilise que deux points [16, Lemma 8]. Le principe général est de construire deux jeux de paramètres les plus éloignés possibles, et tels que les distributions des observations associées à ces paramètres soient proches pour la distance du χ^2 . Nous ne reviendrons pas davantage ici sur ces techniques.

Enfin, je présenterai des travaux concernant des tests d'adéquation adaptatifs dans le modèle de convolution. Ils se situent dans la lignée de [14] et proposent une version adaptative (en la régularité de f) de ces résultats dans le cas d'un bruit de régularité polynomiale, mais aussi une version semi paramétrique et adaptative (en la régularité de f et de celle de f^ε) dans le cas d'un bruit de régularité exponentielle.

5.1. Convolution avec échelle du bruit inconnue

En collaboration avec Cristina Butucea (Université des sciences et technologies de Lille 1), nous avons proposé [M4] un estimateur du paramètre d'échelle du bruit dans le cas où celui-ci est inconnu (cas **A**) ci-dessus). Rappelons qu'alors la régularité s de la densité du bruit est supposée connue. Nous faisons alors l'une ou l'autre des hypothèses suivantes sur la densité inconnue f du signal. Soit

Hypothèse 5.1. *La densité f appartient à la classe de fonctions $\mathcal{A}(\alpha, r)$ dont la transformée de Fourier ne décroît pas (en l'infini) plus rapidement qu'une exponentielle, i.e.*

$$|\Phi(u)| \geq ce^{-\alpha|u|^r}, \quad |u| \text{ assez grand},$$

où les paramètres $\alpha > 0$ et $r \in (0, s)$ sont connus et $c > 0$ est une constante arbitraire.

ou bien,

Hypothèse 5.2. *La densité f appartient à la classe de fonctions $\mathcal{B}(\beta)$ dont la transformée de Fourier ne décroît pas (en l'infini) plus rapidement qu'un polynôme, i.e.*

$$|\Phi(u)| \geq c|u|^{-\beta}, \quad |u| \text{ assez grand},$$

où le paramètre $\beta > 1$ est connu et $c > 0$ est une constante arbitraire.

Sous l'une ou l'autre de ces deux hypothèses, nous obtenons un modèle identifiable. Il suffit pour cela de considérer les transformées de Fourier afin d'obtenir

$$\begin{cases} -\alpha|u|^{r-s} \leq |u|^{-s} \log |\Phi(u)| \leq 0, & \text{pour } |u| \text{ grand, sous l'hypothèse 5.1} \\ |u|^{-s}(-\beta \log |u| + \log c) \leq |u|^{-s} \log |\Phi(u)| \leq 0, & \text{pour } |u| \text{ grand, sous l'hypothèse 5.2} \end{cases}$$

(rappelons que la régularité s est connue). Ainsi, $\lim_{|u| \rightarrow \infty} |u|^{-s} \log |\Phi(u)| = 0$, ce qui nous donne

$$\lim_{|u| \rightarrow \infty} \frac{\log |\Phi^Y(u)|}{|u|^s} = \lim_{|u| \rightarrow \infty} \frac{\log |\Phi^\varepsilon(\sigma u)|}{|u|^s} = -\sigma^s.$$

Ainsi, l'échelle du bruit σ , et par conséquent la densité inconnue f , sont donc bien définies de façon unique à partir de la distribution des observations.

Pour obtenir l'optimalité de nos vitesses de convergence, nous sommes amenées à préciser la forme de la densité du bruit. Nous choisissons de faire l'hypothèse que la distribution du bruit est une loi stable.

Hypothèse 5.3. *Les variables ε sont distribuées suivant une loi stable $S(1, s, \nu, \mu)$, de paramètre d'échelle fixé à 1, d'indice d'auto-similarité $s \in (0, 2]$, de paramètre de symétrie $\nu \in [-1, 1]$ et de paramètre de position $\mu \in \mathbb{R}$.*

Rappelons quelques propriétés des lois stables (pour plus de détails, nous renvoyons à [131]). Sous l'hypothèse précédente, la variable aléatoire $\sigma\varepsilon$ a également une loi stable dont la transformée de Fourier

est donnée par

$$\Phi^\varepsilon(\sigma u) = \begin{cases} \exp\{-\sigma^s |u|^s (1 - i\nu \operatorname{sgn}(u) \tan(\pi s/2)) + iu\sigma\mu\} & , s \neq 1 \\ \exp\{-\sigma |u| (1 + i\nu \operatorname{sgn}(u) 2/\pi \log |u|) + iu\sigma(\mu - \nu 2/\pi \log \sigma)\} & , s = 1. \end{cases}$$

On a donc exactement $|\Phi^\varepsilon(\sigma u)| = e^{-\sigma^s |u|^s}$. Signalons de plus que les sommes indépendantes de variables de loi stable avec le même indice d'auto-similarité s sont distribuées suivant une loi stable d'indice d'auto-similarité s . Ces distributions sont donc particulièrement adaptées à la modélisation de bruits additifs. Notons également que le modèle peut s'écrire sous la forme $Y = X + \sigma(\varepsilon_0 + \mu)$ où le bruit ε_0 est distribué selon une loi stable de paramètre de position nul. Puisque σ est inconnu, cette expression montre que le paramètre de position μ ne peut pas être négligé. En particulier, le paramètre \tilde{s} défini par

$$(5.3) \quad \tilde{s} = \begin{cases} s \vee 1 & \text{si } \mu \neq 0, \\ s & \text{si } \mu = 0. \end{cases},$$

joue un rôle dans les vitesses de convergence de nos estimateurs.

Nous introduisons à présent l'estimateur du paramètre d'échelle σ . Considérons, pour tous $\tau, u > 0$,

$$\widehat{F}_n(\tau, u) = \widehat{\Phi}_n^Y(u) e^{(\tau u)^s}.$$

L'estimateur de σ est alors défini par

$$\widehat{\sigma}_n = \widehat{\sigma}_n(Y_1, \dots, Y_n) = \left(\inf \left\{ \tau : \tau > 0, |\widehat{F}_n(\tau, u_n)| \geq 1 \right\} \wedge M \right),$$

pour une suite de réels $u_n \rightarrow \infty$ à choisir, et où $M > 0$ est une constante assez grande. Cette procédure est basée sur le comportement asymptotique de la fonction $u \mapsto |\Phi^Y(u)|e^{(\tau u)^s}$ qui est bornée tant que $\tau \leq \sigma$ et qui tend vers l'infini dès que $\tau > \sigma$ et $u \rightarrow +\infty$. Nous tronquons l'estimateur afin de garder une quantité toujours bornée. Nous obtenons alors les deux résultats suivants.

Théorème 5.1. *Sous l'Hypothèse 5.1, pour tout $\sigma_0 > 0$, pour un bon choix du paramètre u_n , et pour toute famille de voisinages $\mathcal{V}_\delta(\sigma_0) = (\sigma_0 - \delta, \sigma_0 + \delta)$ de σ_0 , on a*

$$\lim_{\delta \rightarrow 0} \limsup_{n \rightarrow \infty} \sup_{\sigma \in \mathcal{V}_\delta(\sigma_0)} \sup_{f \in \mathcal{A}(\alpha, r)} \varphi_n^{-2} \mathbb{E}(|\widehat{\sigma}_n - \sigma|^2) \leq 1,$$

avec la vitesse

$$\varphi_n = \frac{\alpha}{s\sigma_0^{r-1}} \left(\frac{\log n}{2} \right)^{r/s-1}.$$

Supposons de plus que l'Hypothèse 5.3 de loi stable pour le bruit est vérifiée, alors

$$\lim_{\delta \rightarrow 0} \liminf_{n \rightarrow \infty} \inf_{\sigma_n} \sup_{\sigma \in \mathcal{V}_\delta(\sigma_0)} \sup_{f \in \mathcal{A}(\alpha, r)} \varphi_n^{-2} \mathbb{E}(|\sigma_n - \sigma|^2) \geq 1,$$

où l'infimum est pris sur tous les estimateurs σ_n de σ .

Et également,

Théorème 5.2. *Sous l'Hypothèse 5.2, pour tout $\sigma_0 > 0$, pour un bon choix du paramètre u_n , et pour toute famille de voisinages $\mathcal{V}_\delta(\sigma_0) = (\sigma_0 - \delta, \sigma_0 + \delta)$ de σ_0 , on a*

$$\lim_{\delta \rightarrow 0} \limsup_{n \rightarrow \infty} \sup_{\sigma \in \mathcal{V}_\delta(\sigma_0)} \sup_{f \in \mathcal{B}(\beta)} \psi_n^{-2} \mathbb{E}(|\hat{\sigma}_n - \sigma|^2) \leq 1,$$

avec la vitesse

$$\psi_n = \frac{2\beta\sigma_0 \log \log n}{s^2 \log n}.$$

Supposons de plus que l'Hypothèse 5.3 de loi stable pour le bruit est vérifiée, alors

$$\lim_{\delta \rightarrow 0} \liminf_{n \rightarrow \infty} \inf_{\sigma_n} \sup_{\sigma \in \mathcal{V}_\delta(\sigma_0)} \sup_{f \in \mathcal{B}(\beta)} \psi_n^{-2} \mathbb{E}(|\sigma_n - \sigma|^2) \geq \left(1 - \frac{|s-1|}{2\beta}\right)^2,$$

où l'infimum est pris sur tous les estimateurs σ_n de σ .

Les résultats de borne supérieure sont basés sur un contrôle précis de la vitesse de convergence vers zéro des quantités $\mathbb{P}(|\hat{\sigma}_n - \sigma| \geq \varphi_n)$ dans le premier cas et $\mathbb{P}(|\hat{\sigma}_n - \sigma| \geq \psi_n)$ dans le second. Il faut noter que les vitesses de convergence sont très lentes, et en particulier pas du tout paramétriques. Mais nos résultats de borne inférieure montrent que ces vitesses ne peuvent pas être améliorées, au moins au sens du risque minimax. De plus, les vitesses sont meilleures quand le bruit est beaucoup plus régulier que le signal (Théorème 5.2), puisqu'on a alors plus d'information pour estimer le paramètre d'échelle σ . Enfin, pour les densités *super régulières*, nous obtenons un résultat minimax *exact* (i.e. mêmes constantes dans les bornes supérieures et inférieures).

Nous avons également montré que la vitesse de convergence du paramètre est ici un facteur limitant des vitesses de convergence de l'estimateur « plug-in » de la densité. Ces vitesses sont bien plus lentes que celles obtenues dans le cas d'une loi du bruit entièrement connu. De plus, cette dégradation des vitesses est *inévitable* : il n'existe pas d'estimateur de la densité qui puisse converger à une vitesse plus rapide que celle que nous avons exhibée, au sens du risque minimax.

Mais avant d'obtenir des vitesses d'estimation de la densité f , il faut supposer que cette densité appartient à une boule pour un certain espace de régularité. Nous considérons les cas où la densité f appartient soit à l'espace $\mathcal{S}(\alpha', R, L)$, soit à l'espace $W(\beta', L)$ avec $\alpha', R, L > 0$ et $\beta' > 1/2$. Trois cas différents peuvent alors se produire, suivant l'appartenance de f à l'un des trois espaces suivants :

- $\mathcal{A}(\alpha, r) \cap \mathcal{S}(\alpha', R, L)$, qui est non vide pour $R < r$ ou $\{R = r \text{ et } \alpha' < \alpha\}$;
- $\mathcal{A}(\alpha, r) \cap W(\beta', L)$;
- $\mathcal{B}(\beta) \cap W(\beta', L)$, qui est non vide pour $\beta > \beta' + 1/2$. On a alors automatiquement $\beta > 1$.

[L'intersection $\mathcal{B}(\beta) \cap \mathcal{S}(\alpha', R, L)$ est toujours vide.] Nous montrons alors trois théorèmes, qui établissent que les vitesses de convergence φ_n et ψ_n obtenues dans les Théorèmes 5.1 et 5.2 régissent la vitesse de convergence de \hat{f}_n . Nous ne reproduisons ici que le résultat concernant l'espace $\mathcal{A}(\alpha, r) \cap \mathcal{S}(\alpha', R, L)$.

Théorème 5.3. *Sous les hypothèses et les notations du Théorème 5.1 et l'Hypothèse 5.3 de loi stable*

pour le bruit, pour tout $\sigma_0 > 0$, et tout voisinage borné $\mathcal{V}(\sigma_0)$, en choisissant la fenêtre

$$h_n = \left\{ \frac{(\sigma_0 + \delta)^R}{\alpha'} \left(1 - \frac{r}{s}\right) \log \log n - \frac{(\sigma_0 + \delta)^R}{\alpha'} \frac{1-R}{2R} \log \log \log n \right\}^{-1/R},$$

et pour tout x dans \mathbb{R} , on obtient

$$\limsup_{n \rightarrow \infty} \sup_{\sigma \in \mathcal{V}(\sigma_0)} \sup_{f \in \mathcal{A}(\alpha, r) \cap \mathcal{S}(\alpha', R, L)} \varphi_n^{-2} \mathbb{E} \left(|\hat{f}_n(x) - f(x)|^2 \right) \leq C < \infty$$

et

$$\liminf_{n \rightarrow \infty} \inf_{f_n} \sup_{\sigma \in \mathcal{V}_s(\sigma_0)} \sup_{f \in \mathcal{A}(\alpha, r) \cap \mathcal{S}(\alpha', R, L)} \varphi_n^{-2} \mathbb{E} \left(|f_n(x) - f(x)|^2 \right) \geq c > 0,$$

où l'infimum est pris sur tous les estimateurs f_n de f .

Dans le cas d'une densité f qui appartient à une classe de Sobolev $W(\beta', L)$, les vitesses d'estimation de f lorsque $\beta' > \tilde{s} + 1/2$, où \tilde{s} est défini par (5.3), sont celles obtenues pour l'estimation de σ , i.e. φ_n si $f \in \mathcal{A}(\alpha, r) \cap W(\beta', L)$ et ψ_n si $f \in \mathcal{B}(\beta) \cap W(\beta', L)$. Ces vitesses sont même dégradées lorsque $\beta' \leq \tilde{s} + 1/2$ mais nous ne discutons pas ces résultats ici et renvoyons le lecteur intéressé vers [M4].

Les mêmes vitesses de convergence sont obtenues pour l'étude du risque en norme \mathbb{L}_2 , avec quelques dégradations supplémentaires dans certains cas particuliers (je renvoie encore à [M4] pour plus de détails).

5.2. Convolution avec régularité du bruit inconnue

Dans un travail [M10] en collaboration avec Cristina Butucea et Christophe Pouet (Université de Provence, Marseille), nous considérons cette fois que le paramètre de régularité s de la densité du bruit est inconnu. Nos procédures d'estimation vont donc s'adapter automatiquement à cette régularité s . De plus, ces procédures sont également adaptatives en la régularité du signal que nous supposons également inconnue. Pour simplifier les notations, nous supposons que l'on a exactement

Hypothèse 5.4. $\Phi^\varepsilon(u) = \exp(-|\sigma u|^s)$, $\sigma, s > 0$.

Tout d'abord, et afin d'obtenir un modèle identifiable, nous devons nous restreindre à des densités f de signal qui sont moins régulières que la densité f^ε du bruit. Nous supposons donc que f appartient à une classe de Sobolev $W(\beta, L)$ où $L > 0$ est une constante (connue) et le paramètre de régularité $\beta > 0$ est inconnu. Nous supposons que β appartient à un intervalle $[\underline{\beta}, \bar{\beta}] \subset (0, +\infty)$ qui est connu. Lorsque nous considérons le problème de l'estimation de f , nous nous restreignons à $[\underline{\beta}, \bar{\beta}] \subset (1/2, +\infty)$. De plus, nous supposons que f n'est pas *trop régulière*, au sens où sa transformée de Fourier ne décroît pas plus vite vers zéro que l'inverse d'un polynôme connu d'ordre β' .

Hypothèse 5.5. Il existe une constante $A > 0$ telle que pour tout $|u|$ assez grand, $|\Phi(u)| \geq A|u|^{-\beta'}$.

Lorsque f appartient à la classe de Sobolev $W(\beta, L)$ et vérifie l'Hypothèse 5.5, on a nécessairement $\beta' > \beta + 1/2$. Dans la suite, nous utilisons la notation $q_{\beta'}(u) = A|u|^{-\beta'}$. Sous les Hypothèses 5.4 et 5.5, on obtient facilement que le modèle est identifiable. En effet, supposons qu'il existe deux jeux de

paramètres (Φ_1, s) et (Φ_2, s') générant la même distribution des observations. Alors les transformées de Fourier vérifient $\Phi_1^Y = \Phi_2^Y$, où $\Phi_1^Y(u) = \Phi_1(u)e^{-|\sigma u|^s}$ et $\Phi_2^Y(u) = \Phi_2(u)e^{-|\sigma u|^{s'}}$. Supposons par exemple que $s \leq s'$. Alors on obtient

$$|\sigma u|^{-s} \log |\Phi_1(u)| - 1 = |\sigma u|^{-s} \log |\Phi_2(u)| - |\sigma u|^{s'-s}$$

et en prenant la limite lorsque $|u|$ tend vers $+\infty$, on obtient $s = s'$ et donc $\Phi_1 = \Phi_2$. Dans la suite, nous supposons (sans perte de généralité) que $\sigma = 1$.

Je présente maintenant notre procédure d'estimation du paramètre s . Mais tout d'abord, j'explique pourquoi une démarche similaire à celle suivie pour estimer le paramètre d'échelle σ ne fonctionne plus ici. En effet, introduisons pour tous $t, u > 0$,

$$G(t, u) = \frac{\log |\Phi^Y(u)|}{|u|^t}.$$

On observe que lorsque $|u|$ tend vers l'infini, cette fonction tend vers zéro, si $s < t$, vers 1 si $s = t$ et vers $+\infty$ si $s > t$. On pourrait donc vouloir estimer le paramètre s en détectant l'instant où la fonction $\hat{G}_n(t, u_n) = |u_n|^{-t} \log |\hat{\Phi}_n^Y(u_n)|$ devient grande, pour une suite u_n de points bien choisis et tendant vers l'infini. Malheureusement cette procédure ne fonctionne pas car lorsque $|u|$ devient grand, les fonctions $\Phi^Y(u)$ et $\hat{\Phi}_n^Y(u)$ tendent vers zéro et on ne peut pas garantir que $\log |\hat{\Phi}_n^Y(u)|$ reste proche de $\log |\Phi^Y(u)|$.

Afin d'estimer le paramètre s , nous supposons qu'il appartient à une grille

$$S_n = \{\underline{s} = s_1 < s_2 < \dots < s_N = \bar{s}\},$$

avec $0 < \underline{s} < \bar{s} \leq 2$ et où le nombre de points de la grille N peut croître vers l'infini avec le nombre d'observations n . Nous utilisons encore une fois le comportement asymptotique de la transformée de Fourier Φ^Y de la densité des observations. Ainsi, pour $|u|$ assez grand, nous avons

$$q_{\beta'}(u) \exp(-|u|^s) = A|u|^{-\beta'} \exp(-|u|^s) \leq |\Phi^Y(u)| \leq \exp(-|u|^s).$$

Notons $\Phi^{[k]}(u) = e^{-|u|^{s_k}}$ et $I_k(u)$ l'intervalle

$$I_k(u) = [(q_{\beta'} \Phi^{[k]})(u), \Phi^{[k]}(u)].$$

Nous utilisons également une suite de réels positifs $u_{n,k}$ pour $k = 1, \dots, N$, choisie ultérieurement. Notre procédure sélectionne les valeurs de k dans $1, \dots, N$ pour lesquelles l'intervalle $I_k(u_{n,k})$ est l'intervalle le plus proche de la valeur de $|\hat{\Phi}_n^Y(u_{n,k})|$ (il peut éventuellement contenir cette valeur, mais cela n'est pas nécessaire). L'estimateur \hat{s}_n est alors choisi comme la plus petite des valeurs précédentes, ou égal à s_1 si l'ensemble des valeurs précédentes est vide.

Autrement dit, soit $\hat{S}_n \subset S_n$ le sous-ensemble des points de la grille S_n défini de la façon suivante.

– $s_k \in \hat{S}_n$ si $2 \leq k \leq N - 1$ et

$$\frac{1}{2} \left\{ q_{\beta'} \Phi^{[k]} + \Phi^{[k+1]} \right\} (u_{n,k}) \leq |\hat{\Phi}_n^Y(u_{n,k})| < \frac{1}{2} \left\{ q_{\beta'} \Phi^{[k-1]} + \Phi^{[k]} \right\} (u_{n,k}),$$

- $s_1 \in \hat{S}_n$ si $|\hat{\Phi}_n^Y(u_{n,1})| \geq \frac{1}{2} \{q_{\beta'} \Phi^{[1]} + \Phi^{[2]}\}(u_{n,1})$,
- $s_N \in \hat{S}_n$ si $|\hat{\Phi}_n^Y(u_{n,N})| < \frac{1}{2} \{q_{\beta'} \Phi^{[N-1]} + \Phi^{[N]}\}(u_{n,N})$.

Si cet ensemble \hat{S}_n est vide, nous lui ajoutons la valeur s_1 . Notre estimateur est alors défini par

$$\hat{s}_n = \min \hat{S}_n.$$

Cette procédure assure qu'avec une grande probabilité, notre estimateur ne sur-estime pas la vraie valeur s . La sur-estimation de la régularité s est un phénomène plus préjudiciable dans ce cadre que la sous-estimation, car elle pourrait avoir pour conséquences un estimateur à noyau \hat{f}_n dont le risque n'est pas borné.

Nous prouvons la consistance de cette procédure. Pour plus de clarté, la loi \mathbb{P} est ici indiquée par les paramètres f, s .

Proposition 5.1. *Sous les Hypothèses 5.4 et 5.5, en choisissant*

$$u_{n,k} = \left(\frac{\log n}{2} - \frac{\delta}{s_k} \log \log n \right)^{1/s_k},$$

avec $\delta > \beta'$, et si la grille $\underline{s} = s_1 < s_2 < \dots < s_N = \bar{s}$ vérifie

$$|s_{k+1} - s_k| \geq d_n = \frac{c}{\log n}, \text{ avec } c > 2\beta', \quad N - 1 \leq (\bar{s} - \underline{s})/d_n,$$

alors, pour tout $k \in \{1, \dots, N\}$, on a

$$\mathbb{P}_{f, s_k}(\hat{s}_n \neq s_k) \leq \exp \left(-\frac{A^2}{4} 2^{2\beta'/\bar{s}} (\log n)^{2(\delta-\beta')/\bar{s}} (1 + o(1)) \right),$$

où A et β' sont définis dans l'Hypothèse 5.5.

Nous considérons alors l'estimateur à noyau (5.2) avec fenêtre aléatoire

$$\hat{h}_n = \left(\frac{\log n}{2} - \frac{\bar{\beta} - \hat{s}_n + 1/2}{\hat{s}_n} \log \log n \right)^{-1/\hat{s}_n},$$

et nous obtenons la vitesse de convergence de cet estimateur. Celle-ci est la même que dans le cas de la loi du bruit entièrement spécifiée [15, 16, 39], et nous obtenons automatiquement que cette vitesse est adaptative et optimale au sens du minimax.

Théorème 5.4. *Sous les hypothèses et notations de la Proposition 5.1, pour tous $\bar{\beta} > \underline{\beta} > 1/2$, et si $\delta > \beta' + \bar{s}^2/(2\underline{s})$, on obtient, pour tout $x \in \mathbb{R}$,*

$$\limsup_{n \rightarrow \infty} \sup_{s \in \hat{S}_n} \sup_{\beta \in [\underline{\beta}, \bar{\beta}]} \sup_{f \in W(\beta, L)} (\log n)^{(2\beta-1)/s} \mathbb{E}_{f, s_k} |\hat{f}_n(x) - f(x)|^2 < \infty.$$

De plus, cette vitesse est asymptotiquement adaptative optimale au sens du minimax.

Nous avons également mené une étude de simulations qui montre les bonnes performances de notre estimateur.

5.3. Approche générale de l'étude des estimateurs à noyau construits par « plug-in »

Dans cette partie, j'aimerais donner une vue d'ensemble de l'approche que nous avons suivie pour étudier les estimateurs construits avec un noyau aléatoire, lorsque le paramètre fini-dimensionnel est estimé dans un intervalle. Il s'agit de l'approche suivie dans [M2,M4]. Par contre, dans [M10], le paramètre s est estimé sur une grille et l'étude de l'estimateur à noyau ne nécessite donc pas une approche aussi sophistiquée. Je présente également un erratum (écrit en collaboration avec Cristina Butucea) pour la preuve que nous avons publiée dans [M4] ainsi qu'une discussion motivée par cette correction.

Pour simplifier les notations, je me place dans le cadre de l'estimation ponctuelle de $f(x)$. Je m'intéresse au risque quadratique de l'estimateur $\hat{f}_n(x)$. Cet estimateur est construit à partir d'un estimateur préliminaire $\hat{\tau}_n$ du paramètre inconnu τ (précédemment, σ ou s), que l'on suppose unidimensionnel. Je note $\hat{f}_{n,\hat{\tau}_n}$ l'estimateur en question, et $\hat{f}_{n,\tau}$ la version « classique » de cet estimateur, construit avec un noyau non aléatoire qui fait intervenir le paramètre (inconnu) τ . Notons que $\hat{f}_{n,\tau}$ n'est pas connu du statisticien et est utilisé ici uniquement comme un outil. Le principe général que nous avons utilisé consiste à découper le risque de l'estimateur en deux contributions, suivant l'appartenance de $\hat{\tau}_n$ à un voisinage du vrai paramètre τ , de rayon donné par la vitesse de convergence en probabilité φ_n de $\hat{\tau}_n$ vers τ .

$$\begin{aligned} \mathbb{E}\{(\hat{f}_{n,\hat{\tau}_n}(x) - f(x))^2\} &= \mathbb{E}\{(\hat{f}_{n,\hat{\tau}_n}(x) - f(x))^2 1_{|\hat{\tau}_n - \tau| \leq \varphi_n}\} + \mathbb{E}\{(\hat{f}_{n,\hat{\tau}_n}(x) - f(x))^2 1_{|\hat{\tau}_n - \tau| > \varphi_n}\} \\ &\leq \mathbb{E}\left\{ \sup_{|t - \tau| \leq \varphi_n} (\hat{f}_{n,t}(x) - f(x))^2 \right\} + C\mathbb{P}(|\hat{\tau}_n - \tau| > \varphi_n). \end{aligned}$$

L'inégalité précédente est vraie dès que l'on peut majorer (uniformément) les quantités $\hat{f}_{n,\hat{\tau}_n}(x)$ et $f(x)$ (ce qui est généralement vrai puisque l'on se place dans une boule d'un espace de régularité). Le rayon φ_n est choisi de sorte que le second terme de cette inégalité converge vers zéro. Le contrôle de ce terme résulte donc des propriétés de convergence de $\hat{\tau}_n$ vers τ . Le premier terme est plus délicat à contrôler. Nous avons supprimé l'aléa qui existait dans $\hat{\tau}_n$ mais au prix de l'introduction d'un supremum dans l'espérance. Ce terme se décompose en un terme de biais et un terme de « variance » comme suit.

$$\mathbb{E}\left\{ \sup_{|t - \tau| \leq \varphi_n} (\hat{f}_{n,t}(x) - f(x))^2 \right\} \leq 2 \sup_{|t - \tau| \leq \varphi_n} \left| \mathbb{E}\hat{f}_{n,t}(x) - f(x) \right|^2 + 2\mathbb{E}\left(\sup_{|t - \tau| \leq \varphi_n} |\hat{f}_{n,t}(x) - \mathbb{E}\hat{f}_{n,t}(x)|^2 \right).$$

Le contrôle du biais de l'estimateur « classique » (i.e. à loi du bruit entièrement connue) $\hat{f}_{n,t}$ doit donc « simplement » être uniforme pour les paramètres t dans l'intervalle $[\tau - \varphi_n; \tau + \varphi_n]$, ce qui ne pose pas en général de problème. Le contrôle du terme de « variance » est plus subtil. Pour majorer ce terme, nous avons recours à des inégalités de contrôle des moments de processus empiriques [123].

Introduisons $k_{n,t}$ le noyau de déconvolution déterministe obtenu lorsque la loi du bruit a pour paramètre $\tau = t$. Ainsi, on a

$$(5.4) \quad \hat{f}_{n,t}(x) = \frac{1}{nh_n} \sum_{i=1}^n k_{n,t} \left(\frac{Y_i - x}{h_n} \right),$$

et je note \mathbb{G}_n le processus empirique associé à la mesure \mathbb{P} , i.e. $\mathbb{G}_n(g) = n^{-1/2} \sum_{i=1}^n (g(Y_i) - \mathbb{P}g)$. Il y a au moins deux façons possibles de faire apparaître le processus empirique dans le terme de variance, les deux approches conduisant à des hypothèses très différentes sur le modèle. L'approche suivie dans [M2], puis également dans [M4] est décrite ci-dessous. Nous écrivons

$$\begin{aligned} |\hat{f}_{n,t}(x) - \mathbb{E}\hat{f}_{n,t}(x)|^2 &= \frac{1}{4\pi^2} \left| \int \Phi^{k_{n,t}}(uh_n) e^{iux} (\widehat{\Phi}_n^Y(u) - \Phi^Y(u)) du \right|^2 \\ &\leq \frac{1}{4\pi^2 n} \left(\sup_{u \in I_{n,t}} |\mathbb{G}_n f_u^{(1)}| \right)^2 \left(\int_{u \in I_{n,t}} |\Phi^{k_{n,t}}(uh_n)| du \right)^2 \end{aligned}$$

où $I_{n,t}$ est le support (compact) de $\Phi^{k_{n,t}}$ et $f_u^{(1)} : y \mapsto e^{iuy}$. On obtient donc

$$(5.5) \quad \mathbb{E} \left(\sup_{|t-\tau| \leq \varphi_n} |\hat{f}_{n,t}(x) - \mathbb{E}\hat{f}_{n,t}(x)|^2 \right) \leq \frac{C}{n} \mathbb{E} \left(\sup_{|t-\tau| \leq \varphi_n} \sup_{u \in I_{n,t}} |\mathbb{G}_n f_u^{(1)}| \right)^2 \left(\int_{u \in I_{n,t}} |\Phi^{k_{n,t}}(uh_n)| du \right)^2.$$

Mais il existe aussi une approche directe qui consiste à écrire

$$(5.6) \quad \mathbb{E} \left(\sup_{|t-\tau| \leq \varphi_n} |\hat{f}_{n,t}(x) - \mathbb{E}\hat{f}_{n,t}(x)|^2 \right) = \frac{1}{n} \mathbb{E} \left(\sup_{|t-\tau| \leq \varphi_n} |\mathbb{G}_n f_t^{(2)}| \right)^2$$

où

$$f_t^{(2)} : y \mapsto (h_n)^{-1} k_{n,t}[(y-x)/h_n].$$

Le moment d'ordre 2 du processus empirique \mathbb{G}_n dépend de l'entropie de la classe de fonctions sur laquelle il est considéré. Dans la première approche, nous sommes donc amenés à étudier l'entropie de la classe de fonctions $\mathcal{F}_n^1 = \{y \mapsto e^{iuy}; u \in \cup_{|t-\tau| \leq \varphi_n} I_{n,t}\}$ alors qu'avec la seconde, nous étudions l'entropie de la classe $\mathcal{F}_n^2 = \{y \mapsto (h_n)^{-1} k_{n,t}[(y-x)/h_n]; t \in [\tau - \varphi_n; \tau + \varphi_n]\}$. Notons que dans la première approche, la classe de fonctions considérée n'est pas du tout liée à la forme du noyau et celui-ci intervient alors dans le terme le plus à droite de l'équation (5.5). Au contraire, dans la seconde approche, la classe de fonctions dépend fortement du noyau de déconvolution, et donc de la transformée de Fourier de la densité du bruit.

Dans la suite nous décrivons la démarche générale de contrôle du moment du processus empirique sur la classe de fonctions $\mathcal{F}_n = \{f_t; t \in J_{n,\tau}\}$. Les notations utilisées dans la suite sont celles de [123]. Comme le paramètre qui indice cette classe de fonctions varie dans l'ensemble (c'est en fait un intervalle) $J_{n,\tau}$, il est intéressant de relier les taux d'accroissements des fonctions aux variations de leur indice. Pour cela, nous étudions des propriétés de Lipschitz des fonctions $t \mapsto f_t(y)$ où y est cette fois fixé. Si nous obtenons une propriété de la forme

$$|f_{t_1}(y) - f_{t_2}(y)| \leq |t_1 - t_2| F_n(y),$$

alors, en utilisant [123, Théorème 2.7.11], nous pouvons relier le *bracketing number* de la classe \mathcal{F}_n , i.e. le nombre minimal de « crochets » de taille ϵ nécessaires pour couvrir la classe de fonctions \mathcal{F}_n , au *covering number* de l'intervalle $J_{n,\tau}$, i.e. le nombre minimal de boules de rayon ϵ nécessaires pour le couvrir. Ainsi, on obtient

$$N_{[]} (2\epsilon \|F_n\|_{L_2(Q)}; \mathcal{F}_n; L_2(Q)) \leq N(\epsilon; J_{n,\tau}; |\cdot|),$$

où Q est n'importe quelle mesure de probabilité discrète telle que $\|F_n\|_{L_2(Q)} > 0$. Il est alors aisé de contrôler le nombre de couverture de l'intervalle $J_{n,\tau}$ en utilisant le diamètre $|J_{n,\tau}|$ de l'intervalle

$$N(\epsilon; J_{n,\tau}; |\cdot|) \leq \frac{|J_{n,\tau}|}{\epsilon}.$$

En utilisant que le nombre de couverture $N(\epsilon\|F_n\|_{L_2(Q)}; \mathcal{F}_n; L_2(Q))$ est toujours majoré par le nombre de crochet $N_{[]} (2\epsilon\|F_n\|_{L_2(Q)}; \mathcal{F}_n; L_2(Q))$, on obtient

$$(5.7) \quad N(\epsilon\|F_n\|_{L_2(Q)}; \mathcal{F}_n; L_2(Q)) \leq \frac{|J_{n,\tau}|}{\epsilon}.$$

Nous introduisons alors l'entropie de la classe \mathcal{F}_n , définie par

$$(5.8) \quad J(\mathcal{F}_n) = \sup_Q \int_0^1 \{1 + \log N(\epsilon\|F_n\|_{L_2(Q)}, \mathcal{F}_n, L_2(Q))\}^{1/2} d\epsilon,$$

où le supremum est pris sur toutes les mesures de probabilité discrètes Q . En appliquant [123, Théorème 2.14.1] à la classe de fonctions mesurables \mathcal{F}_n admettant l'enveloppe mesurable F_n , on obtient

$$(5.9) \quad \mathbb{E} \left\{ \left(\sup_{t \in J_{n,\tau}} |\mathbb{G}_n f_t| \right)^2 \right\} \leq c \|F_n\|_{L_2(\mathbb{P})}^2 J(\mathcal{F}_n)^2,$$

où c est une constante absolue.

Revenons maintenant plus précisément au cadre de l'article [M4]. Nous avons suivi la première approche et en utilisant le théorème des accroissements finis, nous obtenons une propriété de Lipschitz (non améliorable) des fonctions de la classe $\mathcal{F}_n^{(1)}$, par rapport au paramètre.

$$\forall t, s, y \in \mathbb{R}, \quad |f_t^{(1)}(y) - f_s^{(2)}(y)| = |e^{ity} - e^{isy}| = |y| \times |t - s|,$$

et donc $F_n^{(1)}(y) = |y|$. Or, ce choix de fonction qui s'était avérée convenir dans [M2] pose ici un problème. En effet, l'inégalité (5.9) fait intervenir la quantité $\|F_n^{(1)}\|_{L_2(\mathbb{P})}$ qui n'est finie que si les observations admettent un moment d'ordre 2, i.e. si le bruit considéré est Gaussien (ce qui était le cas dans [M2]). La preuve que nous avons publiée dans [M4] est donc incorrecte et il faut utiliser la seconde approche (beaucoup plus directe) pour obtenir le résultat annoncé.

En effet, montrons que la classe de fonctions \mathcal{F}_n^2 est Lipschitz par rapport au paramètre t dans $[\tau -$

$\varphi_n, \tau + \varphi_n] = [\sigma - \varphi_n, \sigma + \varphi_n]$. Soient $t_1 < t_2$ et $y \in \mathbb{R}$, on a

$$\begin{aligned}
& |f_{t_1}^{(2)}(y) - f_{t_2}^{(2)}(y)| = \frac{1}{2\pi} \left| \int e^{-iu(y+x)} (\Phi^{k_n}(h_n t_1 u) - \Phi^{k_n}(h_n t_2 u)) du \right| \\
&= \frac{1}{2\pi} \left| \int_{|u| \leq 1/(h_n t_2)} e^{-iu(y+x)} (\exp(t_1^s |u|^s) - \exp(t_2^s |u|^s)) du \right. \\
&\quad \left. + \int_{1/(h_n t_2) \leq |u| \leq 1/(h_n t_1)} e^{-iu(y+x)} \exp(t_1^s |u|^s) du \right| \\
&\leq |t_1 - t_2| \frac{s}{2\pi} (\sigma + \varphi_n)^{s-1} \int_{|u| \leq 1/(h_n t_2)} |u|^s \exp[(\sigma + \varphi_n)^s |u|^s] du \\
&\quad + \frac{|t_1 - t_2|}{2\pi h_n (\sigma - \varphi_n)^2} \exp \left[\left(\frac{\sigma + \varphi_n}{(\sigma - \varphi_n) h_n} \right)^s \right] \\
&\leq |t_1 - t_2| \exp \left[\left(\frac{\sigma + \varphi_n}{\sigma - \varphi_n} \right)^s h_n^{-s} \right] \left(\frac{s}{2\pi} (\sigma + \varphi_n)^{3s-2} h_n^{2s-1} (1 + o(1)) + \frac{1}{h_n 2\pi (\sigma - \varphi_n)^2} \right) \\
&\leq |t_1 - t_2| F_n^{(2)},
\end{aligned}$$

où $F_n^{(2)}$ est cette fois une fonction constante, définie par

$$F_n^{(2)} = \frac{h_n^{-1}}{2\pi (\sigma - \varphi_n)^2} \exp \left[\left(\frac{\sigma + \varphi_n}{\sigma - \varphi_n} \right)^s h_n^{-s} \right] (1 + o(1)).$$

En reprenant la définition (5.8) combinée à l'inégalité (5.7) on obtient aisément que l'entropie $J(\mathcal{F}_n^2)$ est bornée par une constante (uniforme en n). On obtient alors la borne suivante

$$\mathbb{E} \left\{ \left(\sup_{|t-\sigma| \leq \varphi_n} |\mathbb{G}_n f_t^{(2)}| \right)^2 \right\} \leq \kappa (F_n^{(2)})^2 = \kappa' h_n^{-2} \exp \left[2 \left(\frac{\sigma + \varphi_n}{\sigma - \varphi_n} \right)^s h_n^{-s} \right] (1 + o(1)),$$

où κ, κ' sont des constantes universelles. Le résultat global annoncé dans [M4] reste vrai, mais la borne obtenue sur le terme de variance est un peu plus grande que celle que nous avons annoncée. En conclusion, et sur cet exemple uniquement, la seconde approche permet de s'affranchir d'une hypothèse de moment d'ordre 2 sur les observations (qui semble être une hypothèse nécessaire pour la première méthode), au prix d'une borne un peu plus grande (au moins lorsque $s = 2$ et que les deux bornes sont vraies).

5.4. Tests d'adéquation en convolution semi ou non paramétrique

Dans cette partie en collaboration avec Cristina Butucea et Christophe Pouet (et qui regroupe des résultats de [M9] et de [M10]), nous nous intéressons à des tests d'adéquation sur la densité f du signal, non paramétriques et adaptatifs par rapport à la régularité inconnue de la densité f .

Précisons avant de commencer que nous nous intéressons à des tests non paramétriques sur la densité f , et non pas sur la densité des observations f^Y . Lorsque la distribution du bruit est entièrement spécifiée, les hypothèses nulles sur f (de la forme $f = f_0$ ou bien f appartient à un modèle paramétrique

$\mathcal{M} = \{f(\cdot, \theta), \theta \in \Theta\}$ se traduisent de façon unique en hypothèses nulles sur f^Y . En ce sens, il semble équivalent de faire un test sur la densité des observations et sur la densité du signal. Cependant, les alternatives sont différentes dans chacun des cas. En particulier, nous nous intéressons dans la suite à des alternatives H_1 de la forme $\psi_n^{-2} \|f - f_0\|_2^2 \geq \mathcal{C}$ qui ne sont pas en bijection avec des alternatives du même type sur f^Y . En effet, des densités f_1 et f_2 très éloignées (au sens de la norme \mathbb{L}_2) peuvent résulter en des densités d'observations f_1^Y et f_2^Y très proches, et réciproquement. Les tests d'adéquation dans le modèle de convolution ne peuvent donc pas se réduire, même dans le cas où la densité du bruit est spécifiée, aux tests d'adéquation dans le cas d'observations « directes ».

Les tests d'adéquation non paramétriques dans le modèle de convolution ont été étudiés dans [72] et [14]. L'approche de [72] concerne un test d'adéquation à une famille paramétrique dans un modèle de convolution où la densité du bruit (connue) est simplement régulière. Les auteurs étudient la distribution d'une statistique de test du type Bickel-Rosenblatt (i.e. $T_n = \int (\hat{f}_n - J_h \star f_0)^2$ où $J_h = J(\cdot/h)/h$ est un noyau), sous une hypothèse nulle simple $f = f_0$ ainsi que sous diverses alternatives locales ou fixes. L'approche de [14] qui est reprise ici se place dans un cadre minimax et repose sur l'estimation de la fonctionnelle $\int (f - f_0)^2$ (voir 5.12), d'où des alternatives exprimées en terme de distance pour la norme \mathbb{L}_2 . Les procédures de test proposées dans [14] atteignent les vitesses minimax de test dans les cas suivants : densité f dans un espace de Sobolev ou super régulière et bruit de régularité polynomiale ; densité f dans un espace de Sobolev et bruit de régularité exponentielle. Le cas difficile où les densités du signal et du bruit sont toutes les deux super régulières est également étudié, mais l'optimalité de la procédure n'est obtenue que lorsque la densité du signal est moins régulière que celle du bruit. Ces procédures ne sont cependant pas adaptatives, et nous allons ici en proposer des versions adaptatives, étudier les vitesses de test associées et leur optimalité.

Dans un premier temps, nous supposons que f^ε est entièrement connue et de régularité polynomiale. En effet, cette régularité de la densité du bruit ne nous permet pas de considérer un cadre semi paramétrique où les paramètres sont identifiables. Nous proposons un test d'adéquation de l'hypothèse $H_0 : f = f_0$ lorsque l'alternative H_1 est exprimée à partir de la norme \mathbb{L}_2 (i.e. de la forme $\psi_n^{-2} \|f - f_0\|_2^2 \geq \mathcal{C}$). Cette procédure est adaptative (par rapport à la régularité inconnue de f) et présente différentes vitesses de test (ψ_n) en fonction du type de régularité de f_0 (simple ou super régulière). L'adaptativité induit une perte sur la vitesse de test, perte qui est calculée grâce à un théorème de type Berry-Esseen non uniforme pour des U -statistiques dégénérées. Dans le cas d'une régularité simple pour f , nous prouvons que cette perte est inévitable et donc optimale.

Dans un second temps, nous nous plaçons dans le cadre plus large du modèle semi paramétrique étudié à la Section 5.2. Nous utilisons notre estimateur de la régularité s pour estimer la fonctionnelle $\int (f - f_0)^2$ et construire le test de l'hypothèse H_0 . Ces procédures sont adaptatives par rapport à s et à la régularité inconnue de f , et atteignent les vitesses optimales du cas où s et la régularité de f sont connus. Enfin, lorsque f^ε est connue et de régularité exponentielle, une conséquence de notre résultat est que cette procédure de test est adaptative lorsque f_0 appartient à un espace de Sobolev.

Je présente tout d'abord le cadre des tests d'adéquation adaptatifs non paramétriques qui nous occupe. Pour simplifier les notations, nous appellerons τ le paramètre de régularité de la densité inconnue f et $\mathcal{F}(\tau, L)$ l'ensemble auquel appartient f . Ainsi on a $\tau = (\alpha, r, \beta)$ et soit $r > 0$ ce qui signifie que f appartient à $\mathcal{S}(\alpha, r, L)$ ou bien $r = 0$ et alors f appartient à $W(\beta, L)$. Nous supposons que le paramètre de régularité τ est inconnu et qu'il varie dans un ensemble fermé connu \mathcal{T} .

Soit f_0 une densité fixée dans la classe $\mathcal{F}(\tau_0)$. Nous souhaitons construire à partir des observations Y_1, \dots, Y_n , un test de l'hypothèse

$$H_0 : f = f_0.$$

Nous calculons la famille de vitesses $\Psi_n = \{\psi_{n,\tau}\}_{\tau \in \mathcal{T}}$ qui sépare, pour la norme \mathbb{L}_2 , l'hypothèse nulle H_0 de l'alternative

$$H_1(\mathcal{C}, \Psi_n) : f \in \cup_{\tau \in \mathcal{T}} \{f \in \mathcal{F}(\tau, L) \text{ et } \psi_{n,\tau}^{-2} \|f - f_0\|_2^2 \geq \mathcal{C}\}$$

au sens suivant :

- 1) pour tout $0 < \epsilon < 1$, il existe une statistique de test $\Delta_n^* = 1_{\{\text{Rejet de } H_0\}}$ telle que pour tout $\mathcal{C}^0 > 0$, l'erreur du test (i.e. la somme des erreurs de première et de seconde espèce) vérifie

$$(5.10) \quad \limsup_{n \rightarrow \infty} \left\{ \mathbb{P}_0[\Delta_n^* = 1] + \sup_{f \in H_1(\mathcal{C}, \Psi_n)} \mathbb{P}_f[\Delta_n^* = 0] \right\} \leq \epsilon,$$

pour toute valeur $\mathcal{C} > \mathcal{C}^0$. Cette inégalité est appelée la « borne supérieure » du test.

- 2) Cette procédure est optimale, i.e.

$$(5.11) \quad \liminf_{n \rightarrow \infty} \inf_{\Delta_n} \left\{ \mathbb{P}_0[\Delta_n = 1] + \sup_{f \in H_1(\mathcal{C}, \Psi_n)} \mathbb{P}_f[\Delta_n = 0] \right\} \geq \epsilon,$$

pour une certaine constante $\mathcal{C}_0 > 0$ et pour tout $0 < \mathcal{C} < \mathcal{C}_0$, où l'infimum est pris sur toutes les statistiques de test Δ_n . Cette inégalité est appelée la « borne inférieure » du test.

Les vitesses $\Psi_n = \{\psi_{n,\tau}\}_{\tau \in \mathcal{T}}$ sont alors dites *vitesses adaptatives minimax de test*.

Afin de construire une telle procédure et d'obtenir les vitesses associées, nous introduisons notre statistique de test qui est en fait un estimateur de la fonctionnelle $\int (f - f_0)^2$. Ainsi

$$(5.12) \quad T_{n,h_n} = \frac{2}{n(n-1)} \sum_{1 \leq k < j \leq n} \left\langle h_n^{-1} k_n \left(\frac{\cdot - Y_k}{h_n} \right) - f_0, h_n^{-1} k_n \left(\frac{\cdot - Y_j}{h_n} \right) - f_0 \right\rangle.$$

où k_n est le noyau de déconvolution introduit dans l'introduction de la Section 5. Les termes de la somme pour $k = j$ ne sont pas pris en compte afin de réduire le biais de l'estimateur. L'estimateur $T_{n,h}$ n'est pas nécessairement positif, mais sa moyenne l'est.

L'adaptation du test est obtenue grâce à une grille $\mathcal{T}_N = \{\tau_i; 1 \leq i \leq N\}$ sur l'ensemble des paramètres \mathcal{T} . À chaque point τ_i de cette grille, nous associons un seuil $t_{n,i}^2$ ainsi qu'une fenêtre h_n^i et une statistique de test T_{n,h_n^i} comme ci-dessus.

Le test rejette alors l'hypothèse nulle dès qu'au moins un des tests sur la grille (basé sur la valeur τ_i du paramètre) est rejeté. Ainsi, le test est donné par

$$(5.13) \quad \Delta_n^* = \begin{cases} 1 & \text{si } \sup_{1 \leq i \leq N} |T_{n, h_n^i}| t_{n, i}^{-2} > C^* \\ 0 & \text{sinon,} \end{cases}$$

pour une constante $C^* > 0$ à choisir. Dans la pratique, $C^* > 0$ sera choisie par une procédure de simulation par méthode de Monte Carlo sous l'hypothèse nulle, i.e. telle que l'erreur de première espèce du test soit assez petite.

Cas du bruit de régularité polynomiale

Nous considérons tout d'abord le cas d'un bruit de régularité polynomiale.

Hypothèse 5.6.

$$(5.14) \quad |\Phi^\varepsilon(u)| \sim c_g |u|^{-\gamma}, \quad |u| \rightarrow \infty, \quad \gamma > 1;$$

Notons que dans ce cadre (entièrement non paramétrique) le noyau utilisé est le noyau classique de déconvolution (et n'est donc pas aléatoire).

Lorsque la densité f_0 sous l'hypothèse nulle appartient à la classe de Sobolev $W(\bar{\beta}, L)$ (de régularité maximale), la grille est définie de la façon suivante. Pour N fixé, on choisit $\mathcal{T}_N = \{\tau_i; 1 \leq i \leq N+1\}$ telle que

$$\begin{cases} \forall 1 \leq i \leq N, \tau_i = (0; 0; \beta_i) \text{ et } \beta_1 = \underline{\beta} < \beta_2 < \dots < \beta_N = \bar{\beta}, \\ \forall 1 \leq i \leq N-1, \beta_{i+1} - \beta_i = (\bar{\beta} - \underline{\beta})/(N-1), \\ \text{et } \tau_{N+1} = (\underline{\alpha}; \bar{r}; 0) \end{cases}$$

Ainsi, les N premiers points de la grille servent à l'adaptation de la procédure vis-à-vis du paramètre β de la fonction f à tester, lorsque f appartient à un espace de Sobolev, alors que le dernier point τ_{N+1} sert à l'adaptivité de la procédure vis-à-vis de la régularité r de la densité de f , lorsque celle-ci est super régulière.

Théorème 5.5. Soit $f_0 \in W(\bar{\beta}, L)$. La statistique de test Δ_n^* dont les paramètres sont

$$N = \lceil \log n \rceil; \quad \forall 1 \leq i \leq N : \begin{cases} h_n^i = \left(\frac{n}{\sqrt{\log \log n}} \right)^{-2/(4\beta_i + 4\gamma + 1)} \\ t_{n, i}^2 = \left(\frac{n}{\sqrt{\log \log n}} \right)^{-4\beta_i/(4\beta_i + 4\gamma + 1)} \end{cases}, \\ h_n^{N+1} = n^{-2/(4\bar{\beta} + 4\gamma + 1)}; \quad t_{n, N+1}^2 = n^{-4\bar{\beta}/(4\bar{\beta} + 4\gamma + 1)},$$

et une constante C^* assez grande, satisfait (5.10) pour tout $\epsilon \in (0, 1)$, avec la vitesse de test $\Psi_n = \{\psi_{n, \tau}\}_{\tau \in \mathcal{T}}$ donnée par

$$\psi_{n, \tau} = \left(\frac{n}{\sqrt{\log \log n}} \right)^{-2\beta/(4\beta + 4\gamma + 1)} \mathbf{1}_{r=0} + n^{-2\bar{\beta}/(4\bar{\beta} + 4\gamma + 1)} \mathbf{1}_{r>0}, \quad \forall \tau = (\alpha, r, \beta) \in \mathcal{T}.$$

De plus, si $f_0 \in W(\bar{\beta}, cL)$ pour une constante $0 < c < 1$ et si des hypothèses sur la queue de distribution de f_0 et sur les dérivées de Φ^ε sont satisfaites, alors le test est minimax adaptatif sur la famille de classes $\{W(\beta, L), \beta \in [\underline{\beta}, \bar{\beta}]\}$ (i.e. (5.11) est satisfaite).

Les hypothèses nécessaires à l'obtention d'un résultat minimax relèvent des résultats de [14] et ne sont pas reprises ici. Notre procédure de test atteint la vitesse polynomiale $n^{-2\bar{\beta}/(4\bar{\beta}+4\gamma+1)}$ sur l'union de toutes les classes de fonctions super-régulières (ou régulières), qui sont plus (ou autant) régulières que f_0 ; et pas seulement sur les fonctions des classes de Sobolev aussi régulières que f_0 . À notre connaissance, c'est le premier résultat de ce type dans la littérature. De plus, d'après les résultats de [14], cette vitesse est la vitesse minimax de test sur la classe de Sobolev $W(\bar{\beta}, L)$. En conséquence, nous prouvons que le terme en puissance de $\log \log n$ qui apparaît lorsque la densité f est moins régulière que f_0 est une perte inévitable due à l'adaptation. Nous avons établi un résultat de type Berry-Esseen pour des U -statistiques dégénérées afin d'obtenir ce résultat.

Lorsque la densité f_0 sous l'hypothèse nulle appartient à la classe de densités super régulières $\mathcal{S}(\bar{\alpha}, \bar{r}, L)$, la grille sur l'ensemble des paramètres est définie de la façon suivante. Soient N_1, N_2 et $\mathcal{T}_N = \{\tau_i; 1 \leq i \leq N = N_1 + N_2\}$ tels que

$$\begin{cases} \forall 1 \leq i \leq N_1, \tau_i = (0; 0; \beta_i) \text{ et } \beta_1 = \underline{\beta} < \beta_2 < \dots < \beta_{N_1} = \bar{\beta}, \\ \forall 1 \leq i \leq N_1 - 1, \beta_{i+1} - \beta_i = (\bar{\beta} - \underline{\beta}) / (N_1 - 1), \\ \text{et } \forall 1 \leq i \leq N_2, \tau_{N_1+i} = (\bar{\alpha}; r_i; \beta_0) \text{ et } r_1 = \underline{r} < r_2 < \dots < r_{N_2} = \bar{r}, \\ \forall 1 \leq i \leq N_2 - 1, r_{i+1} - r_i = (\bar{r} - \underline{r}) / (N_2 - 1). \end{cases}$$

Dans ce contexte, les N_1 premiers points de la grille servent à rendre la procédure adaptative par rapport à β lorsque f est dans une classe de Sobolev ($r = 0$), tandis que les N_2 points suivants sont utilisés pour rendre la procédure adaptative par rapport à r lorsque f est super régulière.

Théorème 5.6. *Supposons que $f_0 \in \mathcal{S}(\bar{\alpha}, \bar{r}, L)$. La statistique de test Δ_n^* de paramètres \mathcal{C}^* assez grand et*

$$N_1 = \lceil \log n \rceil; \quad \forall 1 \leq i \leq N_1 : \begin{cases} h_n^i = \left(\frac{n}{\sqrt{\log \log n}} \right)^{-2/(4\beta_i+4\gamma+1)} \\ t_{n,i}^2 = \left(\frac{n}{\sqrt{\log \log n}} \right)^{-4\beta_i/(4\beta_i+4\gamma+1)} \end{cases},$$

$$N_2 = \lceil \log \log n / (\bar{r} - \underline{r}) \rceil; \quad \forall 1 \leq i \leq N_2 : \begin{cases} h_n^{N_1+i} = \left(\frac{\log n}{2c} \right)^{-1/r_i}, c < \underline{\alpha} \exp\left(-\frac{1}{r}\right) \\ t_{n,N_1+i}^2 = \frac{(\log n)^{(4\gamma+1)/(2r_i)}}{n} \sqrt{\log \log \log n} \end{cases},$$

satisfait l'inégalité (5.10), avec la vitesse de test $\Psi_n = \{\psi_{n,\tau}\}_{\tau \in \mathcal{T}}$ suivante

$$\psi_{n,\tau} = \left(\frac{n}{\sqrt{\log \log n}} \right)^{-2\beta/(4\beta+4\gamma+1)} 1_{r=0} + \frac{(\log n)^{(4\gamma+1)/(4r)}}{\sqrt{n}} (\log \log \log n)^{1/4} 1_{r \in [\underline{r}, \bar{r}]}.$$

Sous les mêmes hypothèses que dans le théorème précédent, on peut montrer de façon analogue que la perte de vitesse en puissance $\log \log n$ est optimale pour les alternatives dans l'union $\bigcup_{\beta \in [\underline{\beta}, \bar{\beta}]} W(\beta, L)$.

Lorsque l'on considère une alternative qui contient des densités super régulières, qui sont moins régulières que f_0 , alors on obtient une perte de la forme $(\log \log \log n)^{1/4}$ par rapport à la vitesse minimax (non adaptative) obtenue par [14]. Nous ne prouvons pas que cette partie de la perte est optimale.

Cas du bruit de régularité exponentielle

Considérons à présent le cas d'un bruit de régularité exponentielle. Nous nous plaçons dans le cadre semi paramétrique de la Section 5.2 où la régularité s de la densité du bruit est inconnue et de forme donnée par l'Hypothèse 5.4. Dans la définition de la statistique de test T_{n,h_n} , le noyau k_n utilisé est le noyau aléatoire défini en (5.1) qui utilise la valeur de l'estimateur \hat{s}_n .

Corollaire 5.1. *Sous les hypothèses du Théorème 5.4, pour toute densité $f_0 \in W(\bar{\beta}, L)$, on considère la procédure de test qui utilise le seuil et la fenêtre aléatoire (légèrement modifiée) suivants*

$$\hat{t}_n^2 = \left(\frac{\log n}{2} \right)^{-2\bar{\beta}/\hat{s}_n} \quad ; \quad \hat{h}_n = \left(\frac{\log n}{2} - \frac{2\bar{\beta}}{\hat{s}_n} \log \log n \right)^{-1/\hat{s}_n}$$

et une constante C^* assez grande. Alors, cette procédure satisfait (5.10) pour tout $\epsilon \in (0, 1)$ avec la vitesse de test $\Psi_n = \{\psi_{n,\beta}\}_{\beta \in [\underline{\beta}, \bar{\beta}]}$ donnée par

$$\psi_{n,\beta} = \left(\frac{\log n}{2} \right)^{-\beta/s}.$$

De plus, si $f_0 \in W(\bar{\beta}, cL)$ pour un certain $0 < c < 1$ et sous des hypothèses supplémentaires sur la queue de distribution de f_0 , le test est minimax adaptatif sur la famille de classes $\{W(\beta, L), \beta \in [\underline{\beta}, \bar{\beta}]\}$ (i.e. (5.11) est satisfaite).

L'adaptivité de cette procédure découle directement des résultats de [14] puisqu'il n'y a pas ici de perte de vitesse par rapport au cas non adaptatif (à β et s connus). Notons également que dans le cas où s est connu, le résultat sur l'adaptivité de la procédure par rapport à β est un résultat nouveau.

6. Fonctions périodiques bruitées, de période inconnue

Dans ce travail [M6] en collaboration avec Ismaël Castillo (Vrije Universiteit, Amsterdam) et Céline Lévy-Leduc (CNRS-Telecom ParisTech), nous nous sommes intéressés au problème de l'observation d'un signal périodique bruité dans un cadre semi paramétrique. Le signal, ainsi que sa période sont supposés inconnus. Le cadre formel est celui du bruit blanc Gaussien. Plus précisément, nous observons le processus $\{X_t\}_{|t| \leq T/2}$ sur l'intervalle de temps $[-T/2; T/2]$, satisfaisant l'équation de diffusion suivante

$$(6.1) \quad dX_t = f(t/\theta)dt + dW_t,$$

où $\theta \in (0; +\infty)$ est la période inconnue du signal, la fonction f est supposée périodique, de période égale à 1 et $\{W_t\}$ est le mouvement Brownien standard. Nous supposons que f appartient à $\mathbb{L}_2([0; 1])$ et nous introduisons la série de ses coefficients de Fourier

$$\forall k \in \mathbb{Z}, \quad c_k = \int_0^1 f(x) e^{-2i\pi kx} dx.$$

Des estimateurs de la période, convergents à vitesse paramétrique, sont déjà connus dans ce contexte (voir par exemple [21, 58, 62]). Nous nous sommes donc penchés sur le problème de la reconstruction du signal (i.e. estimation de f), utilisant l'estimation de la période (estimateurs « plug-in »). A priori, l'estimation préliminaire de la période peut dégrader considérablement les performances d'un estimateur du signal par rapport au cas où la période est supposée connue (comme c'était le cas dans la Section 5.1). D'un point de vue pratique, la période du signal n'est pas connue a priori; il est donc important de connaître le comportement d'un estimateur du signal dans lequel on a injecté non pas la vraie valeur de la période, mais un estimateur de celle-ci.

Nous prouvons que notre estimateur du signal f (construit à partir de la méthode de Stein par blocs) réalise en fait les mêmes performances qu'un estimateur qui utilise la connaissance de la période. Plus précisément, cet estimateur est minimax et adaptatif sur une classe d'espaces de Sobolev.

Nous avons également proposé des simulations des performances de notre estimateur dans le modèle discrétisé correspondant au modèle de départ (6.1).

Je décris ci-dessous les hypothèses principales et les résultats obtenus. L'approche naturelle lorsque le paramètre θ est connu est de projeter le processus observé X_t sur la base de Fourier $\{t \rightarrow \exp(2i\pi kt/\theta)\}_{k \in \mathbb{Z}}$ de $\mathbb{L}_2([0, \theta])$, ramenant ainsi le problème à un modèle de suites Gaussiennes. Si nous disposons d'un estimateur consistant $\hat{\theta}_T$ de la période, il est donc naturel de vouloir projeter le processus observé X_t sur la famille de fonctions $\{t \rightarrow \exp(2i\pi kt/\hat{\theta}_T)\}_{k \in \mathbb{Z}}$. Tout d'abord, nous donnons un sens à la quantité

$$\int_{-T/2}^{T/2} e^{-2i\pi kt/\hat{\theta}_T} dX_t,$$

qui ne peut pas être définie au sens de l'intégrale d'Itô (car $\hat{\theta}_T$ n'est pas nécessairement adapté à la filtration du Brownien). Cette projection nous permet d'obtenir une version approchée du modèle de suites gaussiennes. Ainsi, nous nous ramenons à l'observation de

$$(6.2) \quad z_k = \frac{1}{T} \int_{-T/2}^{T/2} e^{-2i\pi kt/\hat{\theta}_T} dX_t = \gamma_k(\theta; \hat{\theta}_T) + \frac{1}{\sqrt{T}} \xi_k(\hat{\theta}_T), \quad k \in \mathbb{Z},$$

où

$$\begin{aligned} \gamma_k(\theta; \hat{\theta}_T) &= T^{-1} \int_{-T/2}^{T/2} f(t/\theta) e^{-2i\pi kt/\hat{\theta}_T} dt, \\ \xi_k(\hat{\theta}_T) &= T^{-1/2} \int_{-T/2}^{T/2} u_{\hat{\theta}_T}(t) dW_t. \end{aligned}$$

Ici, les variables $\xi_k(\hat{\theta}_T)$ ne sont plus Gaussiennes (comme c'est le cas lorsque θ est connu) et $\gamma_k(\theta; \hat{\theta}_T)$ est aléatoire (ce n'est plus le k ème coefficient de Fourier de f). Tout notre travail a consisté à montrer que le modèle approché de suites Gaussiennes (SGA) (6.2) se comporte comme un modèle de suites Gaussiennes (SG) classique, et que les mêmes résultats d'estimation y sont obtenus. Pour une bonne référence sur la version « classique » des résultats qui vont suivre, nous renvoyons à [122].

Nous supposons que la fonction périodique f appartient à un espace de Sobolev (périodisé)

$$W^{per}(\beta, L) = \{f \text{ 1-périodique, } f \in \mathbb{L}_2([0; 1]); \sum_{k \in \mathbb{Z}} |2\pi k|^{2\beta} |c_k|^2 \leq L\}.$$

Nous imposons également une condition sur les coefficients de Fourier c_k qui implique en particulier que θ est la période minimale du signal. Enfin, le paramètre θ varie dans un intervalle $[\alpha_T; \beta_T]$ qui couvre asymptotiquement $]0; +\infty[$.

Dans le modèle SG, les estimateurs linéaires, i.e. de la forme $(\lambda_k z_k)_{k \in \mathbb{Z}}$ sont des estimateurs consistants de la suite des coefficients de Fourier c_k (et donc de la fonction f). Pinsker [106] a proposé un choix optimal des poids λ_k , au sens où l'estimateur obtenu est asymptotiquement minimax *exact* (i.e. la même constante, dite de Pinsker, apparaît dans les bornes supérieures et inférieures du contrôle du risque quadratique), parmi toutes les procédures d'estimation de f dans l'espace $W^{per}(\beta, L)$. Nous montrons tout d'abord que si l'estimateur $\hat{\theta}_T$ vérifie certaines hypothèses (qui sont vérifiées par exemple par l'estimateur proposé dans [21], sous des hypothèses raisonnables sur le modèle), alors l'estimateur linéaire avec poids de Pinsker dans le modèle SGA est également asymptotiquement minimax exact parmi toutes les procédures d'estimation de f dans l'espace $W^{per}(\beta, L)$. Les poids de Pinsker sont définis de la façon suivante.

$$\forall k \in \mathbb{Z}, \quad q_k = (1 - w|2\pi k|^{2\beta})_+ ,$$

où w est solution de l'équation

$$\frac{1}{wT} \sum_{k \in \mathbb{Z}} |2\pi k|^{2\beta} (1 - w|2\pi k|^{2\beta})_+ = L .$$

En particulier, ces poids sont nuls à partir d'un certain rang (qui tend vers l'infini avec T). Notre estimateur linéaire avec poids de Pinsker dans le modèle SGA est donc défini simplement comme

$$\hat{f}_T(x) = \sum_{k \in \mathbb{Z}} q_k z_k e^{2i\pi kx} .$$

[Noter ici que l'estimateur $\hat{\theta}_T$ se trouve dans l'observation z .] On obtient alors le résultat suivant.

Théorème 6.1. *Soient $\beta \geq 2$ et $L > 0$. Sous de bonnes hypothèses sur le modèle et sur l'estimateur $\hat{\theta}_T$, l'estimateur \hat{f}_T satisfait*

$$\begin{aligned} & \lim_{T \rightarrow +\infty} \sup_{\theta \in [\alpha_T; \beta_T]} \sup_{f \in W^{per}(\beta, L)} T^{2\beta/(2\beta+1)} \mathbb{E}_{\theta, f} \|\hat{f}_T - f\|_2^2 \\ & = \lim_{T \rightarrow +\infty} \inf_{f_T} \sup_{\theta \in [\alpha_T; \beta_T]} \sup_{f \in W^{per}(\beta, L)} T^{2\beta/(2\beta+1)} \mathbb{E}_{\theta, f} \|f_T - f\|_2^2 = C^* , \end{aligned}$$

où l'infimum est pris sur tous les estimateurs f_T utilisant l'observation du processus $\{X_t\}_{|t| \leq T/2}$ et C^* est la constante de Pinsker, définie par

$$C^* = [L(2\beta + 1)]^{\frac{1}{2\beta+1}} \left(\frac{\beta}{\pi(\beta + 1)} \right)^{\frac{2\beta}{2\beta+1}} .$$

L'estimateur de Pinsker n'est pas adaptatif puisque le choix des poids requiert la connaissance de la régularité β ainsi que du rayon L de l'espace de Sobolev $W^{per}(\beta, L)$. Cavalier et Tsybakov [24] ont proposé, pour le modèle SG, une procédure linéaire adaptative, utilisant des poids aléatoires construits via la méthode de Stein par blocs. Cette procédure adaptative atteint la même vitesse et la même constante que la procédure de Pinsker, et est donc automatiquement asymptotiquement adaptative minimax exacte, pour des fonctions f dans $W^{per}(\beta, L)$. De plus, cette procédure satisfait une inégalité oracle exacte pour la classe des estimateurs linéaires à poids monotones. Nous montrons que la même procédure, dans le modèle SGA, obtient les mêmes performances (sous les mêmes hypothèses sur $\hat{\theta}_T$ que précédemment) et est donc également asymptotiquement adaptative minimax exacte dans le modèle SGA.

Je rappelle tout d'abord brièvement la construction de la procédure par blocs de Stein, telle qu'elle est présentée dans [24]. Nous fixons tout d'abord une valeur N_{max} maximale au-delà de laquelle les poids affectés à l'estimateur seront tous nuls. Ici, nous choisissons $N_{max} = T^{1/4}$. L'intervalle $[-N_{max}; N_{max}]$ est alors partitionné en blocs B_j pour $j = -J, \dots, J$. Le cardinal de chaque bloc B_j est noté T_j . Je ne donnerai pas plus de détails ici sur la construction de ces blocs. J'ajouterais simplement que nous avons choisi une construction où les blocs croissent de façon géométrique (voir par exemple [24] ou [122, Section 3.6]). Les poids de James-Stein sont alors définis par

$$\psi_k(z) = \sum_{j=-J}^J \left(1 - \frac{T_j}{T \|z\|_j^2} \right)_+ 1_{k \in B_j} , \quad \text{pour } k \in \mathbb{Z} ,$$

où $\|z\|_j = (\sum_{l \in B_j} |z_l|^2)^{1/2}$ est la norme ℓ_2 du vecteur z sur le bloc B_j . Les poids de James-Stein sont tels que les observations dont « l'énergie » $\|z\|_j^2$ sur le j ème bloc est inférieure au niveau attendu T_j/T sur ce bloc, ne sont pas prises en compte dans la procédure d'estimation. Ces poids sont constants sur chaque blocs et nuls à partir d'un certain rang ($N_{max} = T^{1/4}$). De plus, ces poids n'utilisent pas les valeurs de β et de L .

L'estimateur dans le modèle SGA utilisant les poids de Stein par blocs est alors simplement défini par

$$\tilde{f}_T(x) = \sum_{k \in \mathbb{Z}} \psi_k(z) z_k e^{2ik\pi x} .$$

On obtient alors le résultat suivant.

Théorème 6.2. *Sous de bonnes hypothèses sur le modèle et sur l'estimateur $\hat{\theta}_T$, l'estimateur \tilde{f}_T satisfait, pour tous $\beta \geq 2$ et tout $L > 0$*

$$\lim_{T \rightarrow \infty} \sup_{\theta \in [\alpha_T; \beta_T]} \sup_{f \in W^{per}(\beta, L)} T^{2\beta/(2\beta+1)} \mathbb{E}_{\theta, f} \|\tilde{f}_T - f\|_2^2 = C^* ,$$

où C^* est la constante de Pinsker.

Troisième partie

Graphes aléatoires

Cette partie regroupe les travaux [M5,M11,M12].

La bio-informatique s'est pendant longtemps consacré principalement à l'étude de séquences génomiques. De nos jours, l'apparition de nouvelles techniques à haut débit nous permet de disposer de données de type nouveau. On assiste en particulier à l'émergence de données relatives à des réseaux biologiques (réseaux d'interaction protéiques, voies métaboliques, réseaux de régulation...) qui nécessitent de nouveaux modèles et techniques d'analyse associées.

Les enjeux majeurs d'un point de vue statistique relativement à ce type de données concernent d'une part l'inférence, d'autre part l'analyse de ces réseaux. Concernant l'analyse, un des enjeux reste la modélisation probabiliste de ces réseaux et l'étude statistique qui découle de telles modélisations. En effet, le seul modèle mathématique de graphe aléatoire parfaitement défini (et très étudié) jusqu'à présent est le modèle d'Erdős-Rényi [43–46] dans lequel toutes les arêtes sont des variables aléatoires indépendantes qui suivent une loi de Bernoulli de paramètre p . Ce modèle s'ajuste très mal aux réseaux connus (aussi bien biologiques que les réseaux de communication ou autres) car la distribution du degré d'un noeud qui en découle est Binomiale (approximativement Poisson quand le nombre de noeuds du réseau devient grand), alors qu'on observe en pratique une loi de puissance pour cette distribution (de la forme $k \rightarrow c(\rho)k^{-(\rho+1)}$ avec ρ généralement dans l'intervalle $[1, 2]$). Il est donc important de fournir des modèles probabilistes de graphes qui décrivent mieux les réseaux réels observés. Quant au problème de l'inférence de ces réseaux, l'enjeu statistique réside dans la très grande dimension de l'espace des paramètres et du très petit nombre d'observations disponibles.

Cette thématique est une des plus récentes dans mes activités de recherche. Elle s'inscrit au sein d'un groupe de travail (SSBnet) constitué de membres du Laboratoire Statistique et Génome, de l'équipe MIG de l'INRA à Jouy-en-Josas, ainsi que de l'équipe Statistique et Génome du département OMIP de l'AgroParisTech.

7. Les motifs dans les réseaux biologiques

Cet article [M5] réalisé en collaboration avec des membres de SSBnet, s'inspire de travaux récents issus de revues de bio-informatique. Dans les travaux en question, les auteurs s'intéressent à la distribution des motifs dans des réseaux (biologiques). La recherche de motifs sur- ou sous-représentés (par rapport à un modèle qu'il s'agit de définir et d'ajuster sur les observations) est un domaine dans lequel il existe déjà une vaste littérature qui concerne les séquences. Dans les graphes, la motivation est la même (un motif trop peu ou trop souvent présent est susceptible de révéler un phénomène biologique particulier), mais les problèmes sont plus complexes (tout simplement parce qu'un graphe n'est pas une séquence linéaire). Les premiers travaux sur ce sujet proposent donc de considérer certains motifs (par exemple tous les motifs de taille 3 ou 4) et de compter leur nombre d'occurrences dans un réseau observé. Il faut alors

comparer ce nombre à celui attendu dans un modèle à spécifier (et éventuellement calculer sa variance dans un tel modèle). Ces auteurs ont alors recours à des simulations très coûteuses : ils génèrent un très grand nombre de graphes qui ont les mêmes degrés que le graphe observé. L'espérance et la variance du nombre d'occurrences d'un motif est alors approché par les espérances et variances empiriques obtenus par cette méthode. Nous avons proposé un modèle de graphe aléatoire dans lequel les arêtes sont des variables aléatoires indépendantes mais qui ont toutes une probabilité d'apparition différente, et qui est proportionnelle aux degrés des deux noeuds que cette arête connecte. Ce modèle s'ajuste donc sur les degrés observés des noeuds, comme ce qui est fait dans les simulations. Nous pouvons alors fournir des calculs exacts de l'espérance et de la variance du nombre d'occurrences d'un motif dans ce modèle, ce qui évite d'avoir recours aux simulations (coûteuses, et spécifiques à chaque réseau observé puisqu'il faut refaire les simulations lorsque les degrés changent).

Ce travail est un premier pas vers la description complète de la distribution du nombre d'occurrences d'un motif dans des modèles de graphes aléatoires qui s'ajustent à la réalité.

8. Un modèle de mélange pour graphes

Comme expliqué ci-dessus, il existe peu de modèles probabilistes de graphes aléatoires en dehors du modèle d'Erdős. Un modèle intéressant car relativement riche, mais suffisamment simple pour permettre la mise en place de méthodes d'estimation rapides des paramètres, est un modèle de mélange dans lequel les noeuds du graphe sont supposés appartenir à des groupes non-observés. Conditionnellement à la donnée des groupes des noeuds, les arêtes sont distribuées de façon indépendante, et de loi dépendant des groupes des noeuds qui sont reliés.

Dans la suite, nous considérons un modèle pour graphe non dirigé avec n noeuds notés $\{1, \dots, n\}$ et où la présence/absence d'une arête entre deux noeuds i et j est donnée par la variable indicatrice de présence d'arête X_{ij} . Soient $\{Z_i\}_{1 \leq i \leq n}$ des variables latentes i.i.d. à valeurs dans $\{1, \dots, r\}$ dont la distribution est donnée par $\pi = (\pi_1, \dots, \pi_r)$ représentant les groupes des noeuds. Conditionnellement à la donnée des groupes des noeuds $\{Z_i\}$, les indicateurs d'arêtes X_{ij} sont supposés indépendants suivant une loi de Bernoulli de paramètre $\rho_{Z_i Z_j}$ (loi notée $\mathcal{B}(\rho_{Z_i Z_j})$). Les paramètres de connectivité $\rho_{ij} \in [0, 1]$ satisfont la relation de symétrie $\rho_{ij} = \rho_{ji}, \forall 1 \leq i, j \leq r$.

Dans toute la suite, le nombre r (≥ 2) de classes non observées sera supposé connu. Son estimation relève d'un problème plus difficile que celui que nous cherchons à résoudre ici. L'espace des paramètres est défini de la façon suivante

$$\Theta = \{\theta = (\pi, \rho); \pi \in (0, 1)^r, \sum_{q=1}^r \pi_q = 1, \rho = (\rho_{ql})_{1 \leq q, l \leq r}, \rho_{ql} \in [0, 1], \rho_{ql} = \rho_{lq}\}.$$

L'intérêt de ces modèles de mélange pour graphes réside dans le fait que des noeuds différents peuvent avoir des connectivités différentes. Ainsi, le modèle peut permettre de décrire une classe de *hubs* qui sont

des noeuds à très forte connectivité. Pour plus d'exemples de types de graphes qui peuvent être modélisés par cette approche, on se référera à [34].

Contrairement à ce qui résulte d'un modèle de mélange classique, les variables observées X_{ij} ne sont pas ici indépendantes. Chaque variable cachée Z_i induit une certaine distribution sur l'ensemble des $(n - 1)$ indicateurs d'arête $\{X_{ik}\}_{k \neq i}$. En conséquence, la distribution des variables cachées $\{Z_i\}_i$ conditionnelle aux observations $\{X_{ij}\}_{i,j}$ n'est pas un produit de termes indépendant sur chacun des noeuds. Toutes ces dépendances compliquent énormément l'analyse statistique du modèle et en particulier, empêchent l'utilisation exacte de l'algorithme EM pour estimer les paramètres. Des approches de type EM variationnel ont été proposées pour pallier à ce problème [34].

Ce modèle a été redécouvert à de nombreuses reprises et dans différents domaines d'application. Une bibliographie non exhaustive inclut [34, 52, 101, 116, 117]. Cependant, la question de l'identifiabilité des paramètres du modèle (à savoir, l'unicité du paramétrage pour une distribution donnée) n'a (à ma connaissance) jamais été traitée. Mentionnons à ce propos les travaux de Franck et Harary [52], portant sur l'estimation des paramètres dans un modèle restreint appelé modèle α - β ou encore modèle d'affiliation. Dans ce cadre, on utilise uniquement deux paramètres pour traduire les connectivités intra et inter groupe, i.e. $\rho_{ii} = \alpha, 1 \leq i \leq r$ et $\rho_{ij} = \beta, 1 \leq i < j \leq r$. En utilisant les statistiques du nombre total d'arêtes présentes et du nombre de triangles, Franck et Harary [52] obtiennent dans de nombreux cas particuliers, des systèmes d'équations (non linéaires) pour des estimateurs des paramètres α, β et parfois du nombre de groupes r . Cependant, les auteurs n'abordent pas le problème de l'unicité des solutions de leurs systèmes, et l'identifiabilité du modèle d'affiliation n'est pas résolue de façon générale. Plus généralement, l'identifiabilité des paramètres n'est pas établie pour ces modèles de mélange pour graphes, et ceci pourrait conduire à des algorithmes d'inférence non convergents, ou fortement dépendants de leur initialisation. Il importe donc de bien comprendre quel type de structure peuvent être inférées ou non avec les données.

Je présente ci-dessous quelques premiers résultats non publiés portant sur des exemples de sous modèles identifiables.

8.1. Quelques modèles de mélange de graphes identifiables

Comme dans tous les modèles de mélange, lorsque le nombre r de classes est fixé, les paramètres ne peuvent être identifiés qu'à permutation près sur les étiquettes des classes. Cependant, la réduction de l'espace des paramètres à classe d'équivalence près n'est pas suffisante en général pour assurer l'identifiabilité du modèle. Cette non identifiabilité du modèle à permutation près a été mentionnée dans [116] mais le problème plus délicat de l'identifiabilité des paramètres à permutation près sur les étiquettes des classes n'est pas abordée. Nous introduisons tout d'abord la relation d'équivalence \sim sur les paramètres

qui diffèrent à permutation près des étiquettes des classes.

Pour tout $r \geq 2$ et $\theta = (\pi, \rho), \theta' = (\pi', \rho') \in \Theta$, nous écrivons $\theta \sim \theta'$

$$\text{ssi } \exists \sigma \in \mathfrak{S}_r, \text{ telle que } \forall 1 \leq q \leq l \leq r, \pi_q = \pi'_{\sigma(q)} \text{ et } \rho_{ql} = \rho'_{\sigma(q)\sigma(l)},$$

où \mathfrak{S}_r est l'ensemble des permutations sur $\{1, \dots, r\}$. L'espace quotienté $\tilde{\Theta} = \Theta/\sim$ contient alors les paramètres, définis à permutation près sur les étiquettes des classes.

Une autre restriction évidente qui doit être faite sur l'espace des paramètres est la suivante. La matrice des connectivités ρ ne peut pas avoir deux lignes (ou de façon équivalente, deux colonnes) identiques. En effet, si tel était le cas, les classes correspondantes auraient des proportions non uniquement déterminées. Supposons par exemple que θ est tel que $\rho_{1i} = \rho_{2i}$ pour tout $1 \leq i \leq r$. Il est aisé de voir que tout paramètre θ' tel que $\rho' = \rho$, $\pi'_i = \pi_i$ pour $i \geq 3$ et $\pi'_1 + \pi'_2 = \pi_1 + \pi_2$ avec $\pi'_1 \neq \pi_1$ et $\pi'_2 \neq \pi_2$, génère la même distribution sur les observations (i.e. $\mathbb{P}_\theta = \mathbb{P}_{\theta'}$) alors que les paramètres sont bien différents ($\theta \neq \theta'$). Nous introduisons donc un espace de paramètres respectant cette contrainte.

$$\tilde{\Theta}_{\text{res.}} = \{\theta \in \tilde{\Theta}; \forall 1 \leq q \neq l \leq r, (\rho_{q1}, \dots, \rho_{qr}) \neq (\rho_{l1}, \dots, \rho_{lr})\}.$$

Dans la suite, nous nous intéressons à des cas particuliers de ce modèle de mélange pour graphes, dans lesquels l'identifiabilité des paramètres est assurée. Notre approche est basée sur une caractérisation de la distribution du processus à travers ses moments.

Exemple 1 (Modèle d'affiliation avec groupes uniformes). Soit \mathcal{E}_1 l'ensemble défini par

$$\mathcal{E}_1 = \left\{ \theta \in \tilde{\Theta}; \exists \alpha \neq \beta \text{ tels que } \forall 1 \leq q \neq l \leq r, \pi_q = \frac{1}{r}, \rho_{qq} = \alpha \text{ et } \rho_{ql} = \beta \right\}.$$

L'identifiabilité du modèle est vérifiée sur \mathcal{E}_1 .

Démonstration. (La preuve s'inspire d'idées apparaissant dans [52]). Fixons $\theta \in \mathcal{E}_1$ et notons $\alpha = \rho_{qq}$ et $\beta = \rho_{ql}$ pour tout $1 \leq q \neq l \leq r$. Nous considérons les premiers moments des variables X_{12} et $X_{12}X_{13}X_{23}$ (le premier moment de la variable $X_{12}X_{13}$ n'est pas informatif ici car $\mathbb{E}_\theta(X_{12}X_{13}) = \mathbb{E}_\theta(X_{12})^2$). On obtient facilement

$$\begin{aligned} \mathbb{E}_\theta(X_{12}) &= \sum_{q=1}^r \sum_{l=1}^r \pi_q \pi_l \rho_{ql} = (\alpha - \beta)/r + \beta \\ \mathbb{E}_\theta(X_{12}X_{13}X_{23}) &= \frac{(\alpha - \beta)^3}{r^2} + \frac{3\beta(\alpha - \beta)^2}{r^2} + \frac{2\beta^2(\alpha - \beta)}{r^2} + \frac{\beta^2(\alpha - \beta)}{r} + \beta^3. \end{aligned}$$

En combinant les deux équations précédentes, on obtient un polynôme de degré 3 en β uniquement. Après quelques calculs simples (mais fastidieux), on peut montrer que ce polynôme a une unique solution dans $[0, +\infty[$. Les paramètres α et β sont donc définis de façon unique. \square

Le même type de preuve permet de montrer l'identifiabilité du modèle dans le cadre plus général où les proportions des classes sont connues (mais pas nécessairement égales). Un tel modèle peut être utile lorsque l'on dispose d'une connaissance a priori sur les proportions des groupes.

Exemple 2 (Modèle d'affiliation avec proportions fixes). Soient $\pi^0 = (\pi_q^0)_{1 \leq q \leq r} \in (0; 1)$ des proportions fixées telles que $\sum_{q=1}^r \pi_q^0 = 1$ et \mathcal{E}_2 l'ensemble défini par

$$\mathcal{E}_2 = \left\{ \theta = (\pi^0, \rho) \in \tilde{\Theta}; \exists \alpha \neq \beta \text{ tels que } \forall 1 \leq q \neq l \leq r, \rho_{qq} = \alpha \text{ and } \rho_{ql} = \beta \right\}.$$

L'identifiabilité du modèle est vérifiée sur \mathcal{E}_2 .

Je propose à présent de regarder certaines lois marginales du processus. Je souligne le fait qu'une approche basée sur l'étude des moments ou sur l'étude des lois marginales du processus ne peut pas permettre de caractériser la distribution des observations. En effet, la dépendance entre les variables empêche la réduction du problème à l'étude d'une loi discrète. En conséquence, ces approches ne peuvent donner que des conditions suffisantes (et jamais nécessaires) sur l'ensemble des paramètres qui satisfait à l'identifiabilité.

Remarquons tout d'abord que la distribution marginale d'une arête X_{ij} est un mélange de lois de Bernoulli ($\sum_q \sum_l \pi_q \pi_l \mathcal{B}(\rho_{ql})$) qui n'est pas une distribution dont les paramètres sont identifiables. Il est donc nécessaire de faire appel à la dépendance entre les variables et considérer des statistiques qui font intervenir plus d'une arête pour pouvoir caractériser les paramètres.

Lemme 8.1. Notons $X_{i+} = \sum_{j \neq i} X_{ij}$ la variable de degré (nombre d'arêtes) du noeud i . La variable X_{i+} est distribuée suivant un mélange de lois Binomiale

$$\sum_{q=1}^r \pi_q \mathcal{B}(n-1, \bar{\rho}_q), \quad \text{où } \bar{\rho}_q = \sum_{l=1}^r \pi_l \rho_{ql}.$$

Démonstration. (voir [34]) Conditionnellement à $Z_i = q$, les variables aléatoires $\{X_{ij}\}_{1 \leq j \leq n, j \neq i}$ sont i.i.d. Bernoulli de paramètre

$$\bar{\rho}_q = \mathbb{P}_\theta(X_{ij} = 1 | Z_i = q) = \sum_{l=1}^r \pi_l \rho_{ql}.$$

Ainsi, conditionnellement à $Z_i = q$, la variable aléatoire X_{i+} suit une loi Binomiale de paramètres $(n-1, \bar{\rho}_q)$. \square

En utilisant l'identifiabilité (sous certaines hypothèses) des mélanges de lois Binomiale, on peut alors exhiber des conditions suffisantes pour obtenir l'identifiabilité du modèle. Supposons par exemple que l'hypothèse suivante est vérifiée.

Hypothèse 8.1. Soit $\mathcal{E} \subset \tilde{\Theta}_{res.}$, tel que pour tous $\theta = (\pi, \rho)$ et $\theta' = (\pi', \rho')$ dans \mathcal{E} , si pour tout $1 \leq q \leq r$, on a $\bar{\rho}_q = \bar{\rho}'_q$ et $\pi_q = \pi'_q$ alors nécessairement $\rho = \rho'$.

L'Hypothèse 8.1 peut sembler assez restrictive, mais nous verrons qu'elle peut être vérifiée pour certains exemples. J'introduis également l'ensemble

$$\Theta_{marg} = \{ \theta = (\pi, \rho) \in \tilde{\Theta}; 1 \leq q \neq q' \leq r, \bar{\rho}_q \neq \bar{\rho}_{q'}, \bar{\rho}_q \in (0, 1) \}$$

Proposition 8.1. Si l'ensemble $\mathcal{E} \subset \tilde{\Theta}_{res.}$ satisfait l'Hypothèse 8.1, alors le modèle est identifiable sur $\Theta_{marg} \cap \mathcal{E}$.

Démonstration. Fixons $\theta = (\pi, \rho), \theta' = (\pi', \rho') \in \mathcal{E} \cap \Theta_{\text{marg}}$ et supposons que ces paramètres génèrent la même distribution sur les observations, i.e. pour tout $n \geq 0$ et tout $\{x_{ij}\}_{1 \leq i < j \leq n}$, on a

$$\mathbb{P}_\theta(\{X_{ij} = x_{ij}\}_{1 \leq i < j \leq n}) = \mathbb{P}_{\theta'}(\{X_{ij} = x_{ij}\}_{1 \leq i < j \leq n}).$$

Nous voulons montrer que $\theta = \theta'$. Comme $\mathbb{P}_\theta = \mathbb{P}_{\theta'}$ sur les observations, les distributions des degrés (qui ne dépendent que des observations) sont également les mêmes pour les deux paramètres. Ainsi, pour tout $n \geq 0$, et toute suite $\{k_i\}_{1 \leq i \leq n} \in \{0, \dots, n-1\}^n$, on obtient

$$\mathbb{P}_\theta(X_{i+} = k_i, 1 \leq i \leq n) = \mathbb{P}_{\theta'}(X_{i+} = k_i, 1 \leq i \leq n).$$

D'après le Lemme 8.1, ceci implique (pour tout $n \geq 0$) l'égalité des lois de mélange de Binomiale

$$(8.1) \quad \sum_{z=1}^r \pi_z \mathcal{B}(n-1, \bar{\rho}_z) = \sum_{z=1}^r \pi'_z \mathcal{B}(n-1, \bar{\rho}'_z).$$

Or, le mélange de r lois Binomiale $\{\mathcal{B}(m, \gamma_i)\}_{1 \leq i \leq r}$ avec les poids $\{\pi_i\}_{1 \leq i \leq r}$ est identifiable (à permutation près des étiquettes des groupes) dès que $m \geq 2r-1$, les poids π_i appartiennent à $(0, 1)$ et les paramètres γ_i sont tous distincts et appartiennent à $(0, 1)$ (voir [12]). Ici, les paramètres θ, θ' appartiennent à l'ensemble Θ_{marg} et l'égalité (8.1) est vérifiée pour tout n (et donc en particulier pour $n-1 \geq 2r-1$). On obtient ainsi

$$\pi = \pi' \text{ et } \bar{\rho} = \bar{\rho}'.$$

(Ces égalités s'entendent dans $\tilde{\Theta}$, c'est à dire à permutation près sur les étiquettes des groupes). Le résultat final découle alors directement de l'Hypothèse 8.1. \square

Ces résultats nous permettent d'exhiber deux exemples supplémentaires de modèles identifiables.

Exemple 3 (Modèle d'affiliation avec proportions distinctes). Soit \mathcal{E}_3 l'ensemble défini par

$$\mathcal{E}_3 = \{\theta \in \tilde{\Theta}; \exists \alpha, \beta \in (0, 1), \alpha \neq \beta \text{ tels que } \forall 1 \leq q \neq l \leq r, \pi_q \neq \pi_l, \rho_{qq} = \alpha \text{ et } \rho_{ql} = \beta\}.$$

L'identifiabilité du modèle est vérifiée sur \mathcal{E}_3 .

Démonstration. Fixons θ dans \mathcal{E}_3 et notons α (resp. β) la valeur commune des paramètres ρ_{qq} (resp. $\rho_{ql}, q \neq l$). Il est facile de voir que

$$\forall 1 \leq q \leq r, \bar{\rho}_q = \pi_q(\alpha - \beta) + \beta.$$

Montrons alors que l'Hypothèse 8.1 est satisfaite sur l'ensemble \mathcal{E}_3 . En effet, supposons que $\theta, \theta' \in \mathcal{E}_3$ sont tels que $\bar{\rho}_q = \bar{\rho}'_q$ et $\pi_q = \pi'_q$ pour tout $1 \leq q \leq r$. Alors on obtient $\pi_q(\alpha - \beta) + \beta = \pi_q(\alpha' - \beta') + \beta'$. Puisque $r \geq 2$ et $\pi_1 \neq \pi_2$, ces équations définissent de façon unique les paramètres et on obtient $\alpha = \alpha'$ et $\beta = \beta'$. De plus, en supposant que tous les π_i sont distincts dans $(0, 1)$ et que $\alpha, \beta \in (0, 1)$ avec $\alpha \neq \beta$, on obtient que l'ensemble \mathcal{E}_3 est inclus dans Θ_{marg} . La Proposition 8.1 permet alors de conclure. \square

Remarquons que les exemples précédents ne permettent pas de conclure que le modèle d'affiliation général (avec $\alpha \neq \beta$) est identifiable.

Exemple 4 (Connectivité multiplicative). Soit \mathcal{E}_4 l'ensemble défini par

$$\mathcal{E}_4 = \{\theta \in \tilde{\Theta}; \exists(\eta_q)_{1 \leq q \leq r} \in (0, 1)^r, \forall 1 \leq q, l \leq r, \rho_{ql} = \eta_q \eta_l, \text{ et } \eta_q \neq \eta_l \text{ dès que } q \neq l\}.$$

L'identifiabilité du modèle est vérifiée sur \mathcal{E}_4 .

Démonstration. Pour tout $\theta \in \mathcal{E}_4$, on a $\bar{\rho}_q = \eta_q \bar{\eta}$, où $\bar{\eta} = \sum_{l=1}^r \pi_l \eta_l$. Montrons que l'Hypothèse 8.1 est vérifiée. Supposons que $\theta, \theta' \in \mathcal{E}_4$ sont tels que $\pi_q = \pi'_q$ et $\bar{\rho}_q = \bar{\rho}'_q$, pour tous $1 \leq q \leq r$. Alors on obtient $\eta_q \bar{\eta} = \eta'_q \bar{\eta}'$ pour tous $1 \leq q \leq r$. En multipliant les deux termes de cette égalité par $\pi_q = \pi'_q$ et en sommant sur les différentes valeurs de q , il découle $(\bar{\eta})^2 = (\bar{\eta}')^2$, et donc $\bar{\eta} = \bar{\eta}'$ (ces quantités sont positives). En réinjectant ce résultat dans l'égalité $\eta_q \bar{\eta} = \eta'_q \bar{\eta}'$, il vient $\eta_q = \eta'_q$ pour tous $1 \leq q \leq r$. Ainsi, la Proposition 8.1 et la restriction $\eta_q \neq \eta_l, \forall q \neq l$ permettent de conclure. \square

Une généralisation naturelle du Lemme 8.1 est donnée ci-dessous.

Lemme 8.2. Considérons $D_{1,2} = \sum_{i \geq 3} X_{i1} X_{i2}$ (le nombre de paires d'arêtes reliées aux noeuds $\{1, 2\}$). La variable aléatoire $D_{1,2}$ est distribuée suivant une loi mélange de Binomiale

$$\sum_{q=1}^r \sum_{l=1}^r \pi_q \pi_l \mathcal{B}(n-2, \bar{\rho}_{ql}),$$

$$\text{où } \bar{\rho}_{ql} = \sum_{s=1}^r \pi_s \rho_{qs} \rho_{ls}.$$

Démonstration. Conditionnellement à $Z_1 = q, Z_2 = l$, les variables aléatoires $\{X_{i1} X_{i2}\}_{i \geq 3}$ sont i.i.d. de loi $\mathcal{B}(\bar{\rho}_{ql})$. \square

Plus généralement, on peut facilement obtenir la distribution de comptages d'ordre supérieur, comme par exemple $T_{1,2,3} = \sum_{i \geq 4} X_{i1} X_{i2} X_{i3}$, qui suit un mélange de lois Binomiale de paramètres $n-3$ et $\bar{\rho}_{qls} = \sum_t \pi_t \rho_{qt} \rho_{lt} \rho_{st}$. Cependant, il semble difficile d'utiliser ces variables pour établir l'identifiabilité du modèle de mélange pour graphes. En effet, les différents paramètres $\bar{\rho}_{ql}, \bar{\rho}_{qls}, \dots$ etc contiennent une information de moins en moins précise sur le paramètre θ .

Dans la section suivante, nous développons des résultats sur l'identifiabilité générique de modèles multivariés discrets à variables latentes. Une collaboration est en cours (avec Elizabeth Allman et John Rhodes) pour appliquer ces résultats à l'identifiabilité des modèles de mélange pour graphes.

9. Identifiabilité des modèles de mélange pour application aux modèles de mélanges de graphes

Très récemment, des techniques d'algèbre ou de géométrie algébrique ont été importées en statistique et semblent être une voie prometteuse pour la résolution de certains problèmes statistiques [104]. En effet, les modèles statistiques usuels en bio-informatique (comme les modèles de Markov, Markov caché ou de mélange) s'appuient sur des vraisemblances qui sont des polynômes en les paramètres du modèle. De plus, les distributions obtenues pour des modèles de variables aléatoires discrètes sont alors entièrement déterminées par un ensemble fini de fonctions polynomiales en les paramètres du modèle. Les questions

d'identifiabilité s'expriment alors de façon simple dans le vocabulaire de l'algèbre : il s'agit de caractériser des variétés, i.e. l'ensemble des zéros d'une famille de polynômes.

Dans une collaboration [M11] avec Elizabeth Allman et John Rhodes (University of Fairbanks, Alaska), nous utilisons un résultat d'algèbre dû à Kruskal [81], qui nous permet d'obtenir des résultats d'identifiabilité générique pour des modèles à variables latentes. À l'origine de cette collaboration, je cherchais à utiliser des méthodes algébriques pour résoudre la question de l'identifiabilité des modèles de mélange de graphes. Les techniques que nous développons ne sont cependant pas du tout limitées à ce cadre d'application. En particulier, nous obtenons des résultats très intéressants sur des modèles de mélange non paramétriques, qui font l'objet d'un intérêt très récent.

Ces travaux sont encore en cours, et non disponibles sous forme d'article. J'ai choisi d'en présenter une version préliminaire en anglais dans l'Annexe A.

10. Inférence de réseaux d'interaction

L'inférence de réseaux d'interaction de gènes à partir de données de bio-puces est un des grands enjeux de la bio-informatique actuelle. D'un point de vue statistique, le problème est difficile, puisqu'il s'agit d'inférer un nombre de paramètres gigantesque avec très peu d'observations. L'hypothèse sur laquelle se basent ces approches est celle d'un nombre relativement peu élevé d'interactions réelles entre les gènes (hypothèse en accord avec la biologie). En particulier, des méthodes de pénalisation ℓ_1 (de type LASSO) ont été récemment appliquées avec succès à ces données [6, 54, 90].

La modélisation des réseaux par modèles de mélange (voir Section 8) permet de prendre en compte une structure de groupe sur les noeuds (gènes) du graphe. En collaboration avec Christophe Ambroise et Julien Chiquet (Université d'Évry Val d'Essonne), nous proposons [M12] d'utiliser ces modèles de mélange pour inférer, via des méthodes de pénalisation ℓ_1 , des graphes structurés. En pénalisant de façon différente les arêtes entre des noeuds identifiés comme étant dans un même groupe, ou dans des groupes différents, nous souhaitons favoriser l'inférence de graphes qui respectent une structure sous-jacente (non-observée) des gènes.

Nous nous plaçons ici dans le cadre des modèles graphiques Gaussiens. L'estimation de la matrice de précision (l'inverse de la matrice de covariance) d'un vecteur Gaussien de grande dimension, sous l'hypothèse que cette matrice est *creuse* (i.e. contient peu de valeurs non nulles), est un problème qui a fait l'objet de nombreux travaux ces dernières années. Les modèles graphiques Gaussiens forment un cadre particulièrement adapté pour modéliser les dépendances entre variables. Dans ce contexte, un graphe non dirigé $G = (V, E)$ est associé à un vecteur Gaussien de la façon suivante : chaque noeud correspond à une coordonnée du vecteur, et deux noeuds sont reliés si et seulement si, les variables aléatoires correspondantes sont indépendantes, conditionnellement aux variables restantes. Or, l'indépendance conditionnelle de deux coordonnées d'un vecteur Gaussien correspond exactement à la présence d'une entrée nulle dans la matrice de précision. Ainsi, la détection des entrées non nulles dans la matrice de précision permet la reconstruction du graphe du modèle graphique Gaussien. C'est ce graphe qui est candidat à être le réseau d'interaction entre les gènes.

Dans la suite, nous nous intéressons donc au problème de la sélection des coefficients non nuls de la matrice de précision d'un vecteur Gaussien. Notre approche privilégie l'aspect sélection sur l'aspect estimation. Les domaines d'application sont variés, incluant, en plus de l'inférence de réseaux de régulation biologiques, la spectroscopie ou encore les données climatiques.

L'idée de la sélection de covariance est apparue tout d'abord dans les travaux de Dempster [35]. Dans le contexte de grande dimension (i.e. quand le nombre de variables est grand devant le nombre d'observations), la sélection de covariance est primordiale puisque la matrice de covariance empirique des observations n'est plus régulière.

Tibshirani et co-auteurs [121] ont introduit la méthode LASSO dans le cadre de la régression linéaire. La méthode LASSO est une technique de régularisation qui sélectionne les variables et estime les coefficients de la régression à la fois. Cette approche est aussi appelée *basis pursuit* en traitement du signal [26]. Il s'agit de pénaliser les moindres carrés ordinaires en y ajoutant ρ (paramètre de pénalisation) fois la norme ℓ_1 du vecteur des paramètres de la régression. Le choix de la norme 1 permet d'obtenir à la fois un problème d'optimisation convexe et une solution creuse. Plusieurs algorithmes ont alors été proposés pour la résolution effective du problème d'optimisation du LASSO. Un des algorithmes les plus célèbres est le LARS [40], qui résout le problème LASSO pour toutes les valeurs du paramètre de pénalisation ρ . En utilisant des résultats d'optimisation convexe, on peut reformuler le problème LASSO comme un problème primal et chercher à résoudre le problème dual associé [103]. Cette approche aboutit en particulier à un algorithme itératif *la méthode par homotopie* [102]. Il existe également une méthode itérative très simple de *descente par coordonnées* [53] et que nous utilisons ici.

Revenons à présent au problème de l'inférence dans les modèles graphiques Gaussien (MGG). Meinshausen et ses co-auteurs [90] ont appliqué le LASSO au problème de l'inférence de la matrice de covariance dans un MGG. Pour cela, les auteurs considèrent successivement la régression de chaque gène sur tous les autres et résolvent autant de problèmes LASSO qu'il y a de gènes p observés. Le défaut d'une telle approche vient du fait qu'il faut ensuite rendre ces p résultats cohérents, puisque si le coefficient de régression de la i ème variable sur la j ème est estimé nul, il n'en va pas nécessairement de même de celui de la j ème variable sur la i ème. Les auteurs proposent alors deux alternatives : une règle *ET* ou une règle *OU*, entre lesquelles il est difficile de choisir. Banerjee et ses co-auteurs [6] formulent quant à eux le problème d'estimation de la matrice de précision sous forme de maximum de vraisemblance pénalisée, et constatent que ce problème se ramène alors à un seul problème de type LASSO. Leur algorithme de résolution du problème d'optimisation repose sur la méthode de Nesterov et présente des problèmes d'efficacité en grande dimension. Friedman et ses co-auteurs [54], en faisant appel aux techniques de *descente par coordonnées* [53], reprennent la formulation de Banerjee et al. et proposent une solution numériquement plus performante.

Dans notre approche, nous introduisons une structure cachée (voir Section 8), afin d'orienter la pénalisation des coefficients de la matrice de précision. Cette approche s'apparente un peu au LASSO adaptatif [132] qui a pour but de débiaiser les grands coefficients de la régression. En utilisant un es-

estimateur préliminaire des coefficients de la régression, le LASSO adaptatif consiste à pénaliser le j ème coefficient en utilisant un poids égal à l'inverse de la j ème coordonnée de l'estimateur préliminaire. La procédure LASSO peut se révéler inconsistante dans certains cas, alors que la procédure de LASSO adaptative possède de bonnes propriétés oracles [132]. Plus précisément, dans un cadre asymptotique (le nombre d'observations n devient grand alors que le nombre de variables p reste fixé), l'estimateur de l'ensemble des coefficients non nuls de la régression est consistant et les estimateurs des paramètres de cette régression sont asymptotiquement normaux, avec variance asymptotique minimale.

Nous proposons un algorithme itératif qui permet d'inférer successivement la matrice de concentration du modèle, avec des pénalités différentes sur les arêtes en fonction de la classe des noeuds correspondants (méthode de type LASSO), et la classe des noeuds pour un graphe (i.e. matrice de corrélation) fixé (algorithme variationnel EM développé dans le cadre de la Section 8). Les premiers résultats obtenus semblent très prometteurs pour estimer des graphes présentant une structure d'affiliation non observée.

Annexe A: Identifiability of latent class models with many observed variables

Joint work with Elizabeth S. Allman and John A. Rhodes.

A.1. Introduction

This paper is devoted to the study of identifiability of many different latent-class statistical models with discrete observations. Existence of a latent (unobserved) structure is widely used and helps reflecting some heterogeneity in the data. Latent-class models form a very large class of models, including for instance finite univariate or multivariate mixtures [89], hidden Markov models [19, 42] and nonparametric mixtures [86].

The general formulations of the identification problems were made by several authors (mainly in econometrics and by staff members of the Cowles Commission for Research in Economics). One can cite among many others [78] and the collection of papers in [77]. Identification literature is concerned with the problem of whether it is possible to have access to some characteristics (a parameter) of the probability distribution of the observed variables. Lack of identification reflects the fact that a random variable may have the same distribution for different values of a parameter. The study of identifiability proceeds from a hypothetical exact knowledge of the distribution of observed variables rather than from a finite sample of observations drawn from this distribution. Thus, identification problems are not problems of statistical inference in a strict sense. However, it is clear that unidentified parameters cannot be consistently estimated. Identification thus appears as a prerequisite of statistical inference.

In the following, we are interested in models defined by a family $\mathcal{M}(\Theta) = \{\mathbb{P}_\theta, \theta \in \Theta\}$ of probability distributions on some space Ω . Here the parameter θ is some characteristic of the distribution the statistician is interested in. It may be for instance the mean of the distribution, or a functional of its density. The classical definition of an identifiable model requires that for any two different values $\theta \neq \theta'$ in Θ , the corresponding probability distributions \mathbb{P}_θ and $\mathbb{P}_{\theta'}$ be different. This is exactly to require injectivity of the parameterization map Ψ for this model, which is defined by $\Psi(\theta) = \mathbb{P}_\theta$.

In many cases, the above map will not be strictly injective while the model remains useful. For instance, it is well known that in latent class models (such as finite mixtures or hidden Markov models), the latent classes can be freely relabeled without changing the distribution of the observations. The phenomenon is known as 'label swapping'. In this sense, the above map is always at least $r!$ -to-one, where r is the number of classes in the model. However, this does not prevent the statistician from using these models. Indeed, a result stating that a latent class model is identifiable, up to a permutation on the class labels is largely enough for practical use, at least in a maximum likelihood setting. Note that the label swapping issue may cause major problems in a Bayesian framework, see for instance [89, Section 4.9].

Practitioners might be more interested in the concept of *local identifiability* which only requires the parameter to be unique in small neighborhoods in the parameter space. This corresponds to local injecti-

vity of the parameterization map. Under some regularity conditions and for parametric models, there is an equivalence between local identification of the parameter and nonsingularity of the information matrix [110]. When using an algorithmic procedure to infer an estimator of the parameter, different initializations can help to detect multiple solutions of the estimation procedure. This often corresponds to the existence of multiple parameter values giving rise to the same distributions. However, the validity of such procedures relies on the insurance that the parameterization map is at most finite-to-one and a precise characterization of the value of k such that it is a k -to-one map would be most useful.

Thus, ensuring that the parameterization map is finite-to-one might be too weak a result from a statistical perspective on identifiability. Moreover, we will argue in the following that, surprisingly, infinite-to-one maps might not be a problem as long as they are *generic* finite-to-one maps.

Indeed, while the above mentioned approaches focus on one-to-one or k -to-one parameterization maps and are well-suited for most of the classical models encountered in the literature, they are non-adapted in some important cases. For instance, it is well-known that finite mixtures of multivariate Bernoulli distributions are not identifiable [67], even up to a relabelling of latent classes. However, these distributions are widely used to model data when many binary variables are observed from individuals belonging to different unknown populations. For instance, these models may be used in numerical identification of bacteria (see [67] and the references therein). Statisticians are aware of this apparent contradiction and we refer to the article [20] whose title, *Practical identifiability of finite mixtures of multivariate Bernoulli distributions* is an attempt to reconcile non-identifiability and practical use of such models. This clearly indicates that the above notion of identifiability is not useful in this specific context. Thus, one has to rely on a weaker notion of identifiability, that would explain this *practical identifiability* that statisticians are looking for. Here, we explain that finite mixtures of multivariate Bernoulli distributions are in fact *generically identifiable* shedding some new light on these models (see Section A.4).

Here, ‘generic’ is used in the sense of algebraic geometry. It means that the set of points for which identifiability does not hold has zero-measure (more details are given below). In this sense, any observed data set has probability one of being drawn from an identifiable model. This is exactly the *practical identifiability* statisticians are looking for.

An important remark is that many discrete space models (like Markov or hidden Markov models) involve parameterization maps which are in fact polynomials in the scalar parameters. Thus, identifiability issues have been recently studied by algebraic geometers. They present the problem in different terminology, describing the image of the parameterization map as a higher secant variety of a Segre variety. When the dimension of this variety is less than expected, the variety is termed defective. Recent works such as [1, 22, 23] have made much progress on determining when defects occur.

However, as pointed out in [41], focusing on dimensions is not sufficient for a complete understanding of the identifiability question. Indeed, even if the dimensions of the parameter space and the image match, the parameterization might be a generically k -to-one map where k cannot be characterized with dimensional approaches. In the following, we will focus on latent-class models assuming the number r of

classes is known. In this context, we might have a k -to-one map with $k > r!$.

This possibility was already raised in the context of psychological studies by Kruskal [81] whose work in [82] provides a strong result ensuring generic $r!$ -to-oneness of the parameterization map for latent r -class models, under certain conditions. Kruskal's work, however is focused on models with only 3 observed (manifest) variables (in other terms, on secant varieties of Segre products with 3 factors, or on 3-way arrays).

In [41], Elmore, Hall, and Neeman address the question of $r!$ -to-oneness for latent-class models with many binary observed variables (i.e., for secant varieties of Segre products with many factors, or on $2 \times 2 \times \dots \times 2$ tables). They show that with sufficiently many observed variables, the image of the parameterization map is birationally equivalent to a symmetrization of the parameter space under the symmetric group Σ_r . Thus, for sufficiently many observed variables, the parameterization map is generically $r!$ -to-one. From a statistical perspective, a consequence of this work (pointed out as the main result by these authors) is that for nonparametric multivariate mixtures, generic identifiability is ensured as soon as there are sufficiently enough variates. Moreover, their proof is constructive enough to give a numerical understanding of how many observed variables are sufficient, though this number's growth in r is much larger than seems necessary (see Corollary A.3 for more details). Note that the authors do not emphasize the *generic* aspect of their theorem, and state that the statistical result holds "without making any assumption on the distributions". Their result also has a consequence (which was not pointed out by these authors, and we explain why in Section A.4) on the identifiability of finite mixtures of multivariate Bernoulli distributions, that we shall discuss here.

In this work, we give a rather different approach to establishing that with sufficiently many observed variables the parameterization map of r -latent class models is generically $r!$ -to-one. Our approach is based on a preliminary result by Kruskal [81] and applies not only to binary variables, but as easily to ones with more states as well. In the case of binary variables (multivariate Bernoulli mixtures), we obtain a much lower figure than what can be obtained using [41], for a sufficient number of variables to ensure generic identifiability and in fact one that has the correct growth order of $\log_2 r$. (The constant factor we obtain is however still unlikely to be optimal.)

Thus, an immediate consequence of our results is the identifiability of finite mixtures of discrete multivariate distributions, as soon as the dimension p of the observed variable is large enough. But the method has further consequences on more sophisticated models with a latent structure. Our approach is very simple : If there are many observed variables, group them into 3 collections, and view the composite states of a collection as the states of a single clumped variable. Then Kruskal's result on 3-way tables can be applied, after a little work to show that the clumping process results in a sufficiently generic model of 3 observed variables.

Note that application of Kruskal's result is limited to finite mixtures of discrete multivariate (with dimension larger than 3) distributions or to latent variables models where the observations are not univariate independent and identically distributed (i.i.d.) random variables. Indeed, the method needs at

least three observed random variables for one latent one, which is never the case for classical univariate mixtures models where the observations are i.i.d. and thus the problem reduces to one observed and one latent random variables. Note also that we always assume the number of latent classes to be known, which is crucial in using Kruskal's approach. Identification of the number of classes is an important issue that we do not consider here.

Algebraic terminology

Polynomials will play an important role throughout our arguments, and so we introduce some basic terminology and facts from algebraic geometry that we will use.

An *algebraic variety* defined by a finite collection of multivariate polynomials $\{f_i\}_{i=1}^n \subset \mathbb{C}[x_1, x_2, \dots, x_k]$ is their simultaneous zero-set,

$$V(f_1, \dots, f_n) = \{\mathbf{a} \in \mathbb{C}^k \mid f_i(\mathbf{a}) = 0, 1 \leq i \leq n\}.$$

A variety is all of \mathbb{C}^k only when all f_i are 0; otherwise, a variety is called a *proper subvariety* and must be of dimension less than k , and hence of Lebesgue measure 0 in \mathbb{C}^k . Analogous statements hold if we replace \mathbb{C}^k by \mathbb{R}^k , or even by any subset $S \subseteq \mathbb{R}^k$ containing an open k -dimensional ball. This last possibility is of course most relevant for our statistical model of interest, since the parameter space is naturally identified with a full-dimensional subset of $[0, 1]^L$ for some L (see Section A.2 for more details). We will often use that intersections (resp. unions) of algebraic varieties are algebraic varieties as they correspond to simultaneous zero-set of sums (resp. products) of the original polynomials.

Given such a set $S \subseteq \mathbb{R}^k$, we will often need to say some property holds for all points in S except possibly for those on some proper subvariety $S \cap V(f_1, \dots, f_n)$. We express this by saying the property holds *generically* on S . We emphasize that the set of exceptional points of S , where the property need not hold, are thus of Lebesgue measure zero.

Roadmap

We first present the r -latent class model in Section A.2. Then, Kruskal's result and consequences are presented in Section A.3. Applications for identifiability of finite mixtures of discrete multivariate distributions appear in Section A.4. Application of these results to dependent variables models, like hidden Markov models or the mixture model for random graphs is an ongoing work. The proofs are postponed to Section A.5.

In the following, for any integer n , the notation $[n]$ is used for the set $\{1, 2, \dots, n\}$.

A.2. The discrete latent class model

Consider a set of observed random variables $\{X_j\}_{1 \leq j \leq p}$ where X_j has finite state space with cardinality κ_j . Note that these variables are not assumed to have the same state space nor to be i.i.d.. To model

the distribution of these variables, we will use a latent (non-observed) random variable Z with values in $[r]$ where r is assumed to be known. We interpret Z as denoting an (unobservable) class, and assume that conditional on Z , the X_j 's are independent random variables. The probability distribution of Z is given by the vector $\boldsymbol{\pi} = (\pi_i) \in \mathbb{R}_{\geq 0}^r$. Moreover, the probability distribution of X_j conditional on $Z = i$ is specified by a vector $\mathbf{p}_{i,j} \in \mathbb{R}_{\geq 0}^{\kappa_j}$ with $\mathbf{p}_{i,j} \mathbf{1}^t = 1$, where $\mathbf{1} = (1 \ 1 \ \dots \ 1)$. We will use the notation $\mathbf{p}_{i,j}(l)$ for the l -th coordinate of this vector ($1 \leq l \leq \kappa_j$).

For each class i , the joint distribution of the variables X_1, \dots, X_p conditional on $Z = i$ is then given by a p -dimensional $\kappa_1 \times \dots \times \kappa_p$ table

$$\mathbb{P}_i = \bigotimes_{j=1}^p \mathbf{p}_{i,j},$$

whose (l_1, l_2, \dots, l_p) -entry is $\prod_{j=1}^p \mathbf{p}_{i,j}(l_j)$. Let

$$(A.1) \quad \mathbb{P} = \sum_{i=1}^r \pi_i \mathbb{P}_i.$$

Then \mathbb{P} is the distribution of a discrete *latent structure model*. The π_i are interpreted as probabilities that a draw from the population is in the i th of r classes. Conditioned on the class, the p discrete variables are independent. However, since the class is not discernable, the p feature variables X_j described by one-dimensional marginalizations of \mathbb{P} are generally not independent.

We refer to the model described above as the r -class, p -feature model, with state space $[\kappa_1] \times \dots \times [\kappa_p]$ as $\mathcal{M}(r; \kappa_1, \kappa_2, \dots, \kappa_p)$. Identifying the parameter space of this model with a subset S of $\mathbb{R}_{\geq 0}^L$ where $L = (r-1) + r \sum_{i=1}^p (\kappa_i - 1)$ and letting $K = \prod_{i=1}^p \kappa_i$, we denote the parameterization map for this model by

$$\Psi_{r,p,(\kappa_i)} : S \rightarrow [0, 1]^K.$$

We will not be explicit in describing the parameter space, but instead work with vectors such as $\boldsymbol{\pi}$ and $\mathbf{p}_{i,j}$, always implicitly assuming their entries add to 1.

As previously noted, this model is not identifiable if $r > 1$, since the sum in equation (A.1) can always be reordered without changing \mathbb{P} . Even modulo label swapping, there are certainly special instances when identifiability will not hold. For instance, if $\mathbb{P}_i = \mathbb{P}_j$, then the parameters π_i and π_j can be varied, as long as their sum $\pi_i + \pi_j$ is held fixed, without effect on the distribution \mathbb{P} . Slightly more elaborate ‘special’ instances of non-identifiability can be constructed, but in full generality this issue remains poorly understood. Ideally, one would know for which choices of $r, p, (\kappa_i)$ the model is identifiable up to permutation of the terms in (A.1) for *generic* parameters, along with a characterization of the exceptional set on which identifiability fails.

A.3. Kruskal’s theorem and its consequences

The basic identifiability result on which we build our later arguments is a result of J. Kruskal in the context of factor analyses for $p = 3$ features. Kruskal’s result deals with a three-way contingency table

(or array) which cross-classifies a sample of n individuals with respect to 3 polytomous variables (each one taking values in $\{1, \dots, \kappa_i\}$). If there is some latent variable Z with values in $\{1, \dots, r\}$ so that each of the n individuals belongs to one of the r latent classes and within the l th latent class, the 3 observed variables are mutually independent, then this r -class latent structure would serve as a simple explanation of the observed relationships among the variables in the 3-way contingency table. This latent structure analysis corresponds exactly to using a mixture with r classes to model the distribution of 3 random variables.

Let us mention that four-way contingency tables are studied in [63] but no result about unicity of the decomposition is given in this setup.

For $j = 1, 2, 3$, let M_j be a matrix of size $r \times \kappa_j$, with $\mathbf{m}_i^j = (m_i^j(1), \dots, m_i^j(\kappa_j))$ the i th row of M_j . Let $[M_1, M_2, M_3]$ denote the $\kappa_1 \times \kappa_2 \times \kappa_3$ tensor defined by

$$[M_1, M_2, M_3] = \sum_{i=1}^r \mathbf{m}_i^1 \otimes \mathbf{m}_i^2 \otimes \mathbf{m}_i^3.$$

In other words, $[M_1, M_2, M_3]$ is a three-dimensional array whose (u, v, w) element is

$$[M_1, M_2, M_3]_{u,v,w} = \sum_{i=1}^r m_i^1(u) \times m_i^2(v) \times m_i^3(w),$$

for any $1 \leq u \leq \kappa_1, 1 \leq v \leq \kappa_2, 1 \leq w \leq \kappa_3$. Note that simultaneously permuting the rows of all the M_j and/or rescaling the rows so the the scaling factors used for the $\mathbf{m}_i^j, j = 1, 2, 3$ multiply to 1 leaves $[M_1, M_2, M_3]$ unchanged.

The keypoint in our work is that the probability distribution in a latent-class model with three observed variables is exactly described by such a tensor. Indeed, if we let M_j be the matrix whose i th row is $\mathbf{p}_{ij} = \mathbb{P}(X_j = \cdot | Z = i)$ and if moreover we let \tilde{M}_1 be the matrix whose i th row is $\pi_i \mathbf{p}_{i1}$, we obtain that the (u, v, w) element of the tensor $[\tilde{M}_1, M_2, M_3]$ equals $\mathbb{P}(X_1 = u, X_2 = v, X_3 = w)$. Thus, the knowledge of the distribution of (X_1, X_2, X_3) is equivalent to the knowledge of the tensor $[\tilde{M}_1, M_2, M_3]$. Note that the M_i s are stochastic matrices and thus the scaling factor in this tensor is the vector of π_i s.

For a matrix M , the *Kruskal rank* of M will mean the largest number I such that every set of I rows of M are independent. Note that this concept would change if we replaced ‘row’ by ‘column,’ but we will only use the row version in this paper. With the Kruskal rank of M denoted by $\text{rang}_K M$, note that

$$\text{rang}_K M \leq \text{rang} M.$$

Moreover, in the particular case where a matrix M of size $p \times q$ has rank p , its Kruskal rank is also equal to p .

Theorem A.1. (*Kruskal [81]*) *Let $I_j = \text{rank}_K M_j$. If*

$$I_1 + I_2 + I_3 \geq 2r + 2,$$

then $[M_1, M_2, M_3]$ uniquely determines the M_j , up to simultaneously permutation and rescaling of the rows.

A slight reformulation, using the above mentioned equivalence between the knowledge of the distribution of three latent-class variables and tensors and using the fact that rows of stochastic matrices sum to 1, gives

Corollary A.1. *Consider the model $\mathcal{M}(r; \kappa_1, \kappa_2, \kappa_3)$. Using the parameterization above, suppose all entries of $\boldsymbol{\pi}$ are positive. For each $j = 1, 2, 3$, let M_j denote the matrix whose rows are $\mathbf{p}_{i,j}$, $i = 1, \dots, r$, and I_j its Kruskal rank. Then if*

$$I_1 + I_2 + I_3 \geq 2r + 2,$$

the parameters of the model are uniquely identifiable, up to label swapping.

Corollary A.2. *The model $\mathcal{M}(r; \kappa_1, \kappa_2, \kappa_3)$ is generically identifiable, up to label swapping.*

A.4. Finite mixtures of discrete multivariate distributions

Finite mixtures of discrete multivariate distributions are widely used to model data, for instance in biologic taxonomy, medical diagnosis or classification of text documents [60, 100]. The identifiability issue in these models has been first addressed forty years ago in a paper by Teicher [118]. Teicher's result states the equivalence between identifiability of product measures distributions and the corresponding one-dimensional models. As a consequence, finite mixtures of multivariate Bernoulli distributions are not identifiable in a strict sense [67]. This result is valid for finite mixtures with an unknown number of components but it can easily be seen that non identifiability occurs in this case even with a known number of components [20, Section 1]. The equivalence condition stated by Teicher has prevented the statisticians to look further at this issue for years. We believe in particular that this is the reason why the result on generic identifiability of multivariate Bernoulli distributions that can be obtained with Elmore *et al.* results [41] were not emphasized by these authors.

Here, we prove that finite mixtures of multivariate Bernoulli distributions (with a known number of components) are in fact generically identifiable, explaining why statisticians find these models interesting despite their non-(strict)-identifiability [20].

Theorem A.2. *Consider the model $\mathcal{M}(r; s_1, s_2, \dots, s_p)$ where $p \geq 3$. Suppose there exists a tripartition of the set $S = [p] = \{1, 2, 3, \dots, p\}$ into three disjoint non-empty subsets S_1, S_2, S_3 , such that if $\sigma_i = \prod_{j \in S_i} s_j$ then*

$$(A.2) \quad \min(r, \sigma_1) + \min(r, \sigma_2) + \min(r, \sigma_3) \geq 2r + 2.$$

Then the model is generically identifiable, up to label swapping.

Considering finite mixtures of p -variate Bernoulli distributions (which correspond exactly to r -class, p -binary feature model $\mathcal{M}(r; 2, 2, \dots, 2)$), we therefore are interested in choosing a tripartition that maximizes the left hand side of inequality (A.2).

Corollary A.3. *The finite mixture of r different p -variate Bernoulli distributions is generically identifiable, up to label swapping, provided*

$$p \geq 2 \lceil \log_2 r \rceil + 1.$$

Note that this identifiability result was already a consequence of [41], with a larger value of the lower bound on p from which the result is valid. However, this had not been noted by these authors. Moreover, our result outperforms the one obtained by [41]. Indeed, letting $C(r)$ be the minimal integer such that if $p > C(r)$ then the r -class, p binary feature model is generically identifiable, then [41] established that

$$\log_2 r \leq C(r) \leq c_2 r \log_2 r$$

for some effectively computable constant c_2 . The lower bound here is easy to obtain from the necessity that the dimension of the parameter space $rp + (r - 1)$ be no larger than that of the distribution space $2^p - 1$, but the upper bound required substantial work. Here, we establish the stronger result that $C(r) \leq 2 \lceil \log_2 r \rceil + 1$.

A.5. Proofs

Proof of Corollary A.1. Let \tilde{M}_1 be the matrix of size $r \times \kappa_1$ such that its i th row is $\pi_i \mathbf{p}_{i,1} = \pi_i \mathbb{P}(X_1 = \cdot | Z = i)$. Its Kruskal rank is denoted \tilde{I}_1 . We already saw that the tensor $[\tilde{M}_1, M_2, M_3]$ describes the probability distribution of the observations (X_1, X_2, X_3) . Kruskal's result states that, as soon as Kruskal ranks satisfy the condition $\tilde{I}_1 + I_2 + I_3 \geq 2r + 2$, this probability distribution uniquely determines the matrices \tilde{M}_1, M_2 and M_3 up to rescaling of the rows and label swapping. Note that as π has positive entries, Kruskal rank \tilde{I}_1 is equal to I_1 . Moreover, using that the matrices M_1, M_2 and M_3 are stochastic and that the entries of π are positive, we get the result. \square

Proof of Corollary A.2. We need to see that for any fixed choice of a positive integer I_j , those matrices M_j whose Kruskal rank is strictly less than I_j form an algebraic variety. This is because the matrices for which a specific set of I_j rows are dependent is the zero set of all $I_j \times I_j$ minors obtained from those rows. Then, by taking appropriate unions of these sets of minors for different numbers of rows we may obtain polynomials whose zero set is precisely those matrices of Kruskal rank less than I_j . Moreover, the matrices M_j have size $r \times \kappa_j$. Thus, the set of matrices whose Kruskal rank is strictly less than r for instance, forms a proper sub-variety and the matrices not in this set satisfy $I_1 + I_2 + I_3 \geq 3r \geq 2r + 2$. To conclude the proof, note also that the set of vectors π admitting zero entries is also a proper sub-variety of the parameter set. \square

Proof of Theorem A.2. Our goal is to apply Kruskal's result to models with more than 3 observed variables by means of a 'grouping' argument.

First, given an $n \times a_1$ matrix A_1 and an $n \times a_2$ matrix A_2 , define the $n \times a_1 a_2$ matrix $A = A_1 \otimes^{row} A_2$, as the row-wise tensor product, so that

$$A(i, a_2(j - 1) + k) = A_1(i, j)A_2(i, k).$$

For each $j \in \{1, \dots, p\}$, we denote by M_j the $r \times s_j$ matrix whose i th row is $\mathbb{P}(X_j = \cdot | Z = i)$. Now, we introduce the three matrices

$$N_i = \bigotimes_{j \in S_i}^{\text{row}} M_j, \quad i = 1, 2, 3,$$

and the tensor $N = [\tilde{N}_1, N_2, N_3]$, where the i th row of \tilde{N}_1 is π_i times the i th row of N_1 . It is easily seen that N contains the probabilities of the triple clumped variables $(\{X_j\}_{j \in S_1}, \{X_j\}_{j \in S_2}, \{X_j\}_{j \in S_3})$. More precisely N_j is the $r \times \sigma_j$ size matrix, whose i -th row is $\mathbb{P}(\{X_k\}_{k \in S_j} = \cdot | Z = i)$ and N is the three-dimensional array of size $\sigma_1 \times \sigma_2 \times \sigma_3$ with entries $N(u, v, w)$ equal to $\mathbb{P}(\{X_j\}_{j \in S_1} = u, \{X_j\}_{j \in S_2} = v, \{X_j\}_{j \in S_3} = w)$. Thus, the knowledge of the distribution of the observations is equivalent to the knowledge of N . Moreover, for parameters π having positive entries (which is a generic condition), Kruskal ranks of \tilde{N}_1 and N_1 are equal. In the next Lemma, we characterize Kruskal rank of the row-tensor product obtained from generic matrices A_i .

Lemma A.1. *Let $A_i, i = 1, \dots, p$ denote $r \times a_i$ matrices, $a = \prod a_i$ and*

$$A = \bigotimes_{i=1, \dots, p}^{\text{row}} A_i$$

the $r \times a$ matrix obtained by taking tensor products of the corresponding rows of the A_i . Then for generic A_i s,

$$\text{rank}_K A = \text{rang} A = \min(r, a),$$

More specifically, there exists a finite set of polynomials in the entries of the A_i s, the non-vanishing of which ensures A has full rank and full Kruskal rank. Moreover, the zero set of these polynomials in $\mathbb{C}^{\sum_i r a_i}$ is of dimension strictly less than $\sum_i r a_i$, and hence is of Lebesgue measure zero.

Proof of the lemma. The condition that a matrix A not have full rank (resp. full Kruskal rank) is equivalent to the simultaneous vanishing of its maximal minors (resp. *idem* when $r \leq a$ and equivalent to the existence of one vanishing maximal minor when $r > a$). Composing the map sending $\{A_i\} \rightarrow A$ with these minors gives polynomials in the entries of the A_i . To see that the polynomials in the entries of the A_i are non-zero, it is enough to exhibit a single choice of the A_i for which A has full rank (resp. full Kruskal rank).

Let $x_{ij}, i = 1, \dots, p, j = 1, \dots, a_i$ be distinct prime numbers. Consider A_i defined by

$$A_i = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_{i1} & x_{i2} & \dots & x_{ia_i} \\ x_{i1}^2 & x_{i2}^2 & \dots & x_{ia_i}^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_{i1}^{r-1} & x_{i2}^{r-1} & \dots & x_{ia_i}^{r-1} \end{pmatrix}.$$

For any vector $\mathbf{y} \in \mathbb{C}^t$, let $V(\mathbf{y}) = V(y_1, y_2, \dots, y_t)$ denote the $t \times t$ Vandermonde matrix, with entries y_j^{i-1} . Suppose first that $a \geq r$. Then the rows of A are the first r rows of the Vandermonde matrix $V(\tilde{\mathbf{y}})$,

where $\tilde{\mathbf{y}}$ is a vector whose entries are $\prod_i x_{ij_i}$ for choices of $1 \leq j_i \leq a_i$. As the products $\prod_i x_{ij_i}$ are distinct by choice of the x_{ij} , $V(\tilde{\mathbf{y}})$ is non-singular, so A has rank and Kruskal rank equal to r .

If instead $r > a$, then the first a rows of A form an invertible Vandermonde matrix. Thus, A is of rank a .

Now, we need a more elaborate argument to conclude on full kruskal rank. If $r > a$, compose the map $\{A_i\} \rightarrow A$ with the $a \times a$ minor from the first a rows of A . This gives us a polynomial in the entries of the A_i , the non-vanishing of which ensures the first a rows of A are independent. We can show that this polynomial is not identically zero as follows : First consider a specific choice of A_i s such that $\text{rang} A = a$, as we have already shown exists. Then by permuting rows of all A_i simultaneously and hence permuting the rows of A in the same way, we may assume the first a rows on A are independent, and so our polynomial is non-zero on these A_i s, and hence is not identically zero.

Now similarly construct non-zero polynomials whose non-vanishing ensures all other sets of a rows of A are independent. The proper subvariety defined by the product of all these polynomials then is precisely those choices of $\{A_i\}$ for which A is not of full Kruskal rank. This concludes the proof of the lemma. \square

Note that to apply this result to generic matrices M_j , we have to deal with the fact that each row of each M_j sums to 1. However, as both rank and Kruskal rank are unaffected by multiplying rows by non-zero scalars, and rows sums not being zero is a generic condition (defined by the non-vanishing of linear polynomials) the above theorem immediately implies the same conclusion holds when all the M_j are assumed to have row sums of 1. We thus obtain that for the above defined matrices N_i , their Kruskal rank I_i is equal to $\min(r, \sigma_i)$. Thus, by assumption, the matrices N_i satisfy the condition of Kruskal's Theorem with $I_i = \min(r, \sigma_i)$. This implies that the tensor $N = [\tilde{N}_1, N_2, N_3]$ uniquely determines the matrices N_i and the vector $\boldsymbol{\pi}$ (up to permutation of the rows). We now need the following lemma.

Lemma A.2. *Suppose $A = \bigotimes_{i=1, \dots, p}^{\text{row}} A_i$ where the A_i are stochastic matrices. Then the A_i are uniquely determined by A .*

Proof of the lemma. Since each row of each A_i sums to 1, one easily sees that the entries in A_i can be recovered as sums of certain entries in the same row of A . \square

Using this lemma, we obtain that each N_i uniquely determines the matrices M_j and the desired result follows. \square

Proof of Corollary A.3. It is enough to consider the case where $p = 2 \lceil \log_2 r \rceil + 1$. With $k = \lceil \log_2 r \rceil$, we have that $2^{k-1} < r \leq 2^k$. Choosing

$$\sigma_1 = \sigma_2 = 2^k, \quad \sigma_3 = 2,$$

inequality (A.2) in Theorem A.2 holds. \square

Références

- [1] ABO, H., OTTAVIANI, G., AND PETERSON, C. (2008). Induction for secant varieties of Segre varieties. *Trans. Amer. Math. Soc.*. To appear. [arXiv.org :math/0607191](http://arxiv.org/math/0607191).
- [2] AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control AC-19*, 716–723. System identification and time-series analysis. MR0423716
- [3] ANDERSON, B. D. O. (1999). The realization problem for hidden Markov models. *Math. Control Signals Systems* **12**, 1, 80–120. MR1685090
- [4] ARNDT, P. F., BURGE, C. B., AND HWA, T. (2003). DNA sequence evolution with neighbor-dependent mutation. *J. Comput. Biol.* **10**, 3/4, 313–322.
- [5] ARRIBAS-GIL, A. (2007). Estimation dans des modèles à variables cachées : alignement de séquences biologiques et modèles d'évolution. Ph.D. thesis, Université Paris-Sud, France. http://www.math.u-psud.fr/~arribas/these_arribas_gil.pdf.
- [6] BANERJEE, O., EL GHAOUI, L., AND D'ASPREMONT, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* **9**, 485–516.
- [7] BARBU, V. AND LIMNIOS, N. (2006). Maximum likelihood estimation for hidden semi-Markov models. *C. R. Math. Acad. Sci. Paris* **342**, 3, 201–205. MR2198194
- [8] BESAG, J. (1975). Statistical analysis of non-lattice data. *The Statistician* **24**, 3, 179–195.
- [9] BIRGÉ, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete* **65**, 2, 181–237. MR722129
- [10] BIRGÉ, L. (2001). A new look at an old result : Fano's lemma. Tech. rep., Prépublication 632, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7.
- [11] BISHOP, M. AND THOMPSON, E. (1986). Maximum likelihood alignment of DNA sequences. *J. Mol. Biol.* **190**, 159–165.
- [12] BLISCHKE, W. R. (1964). Estimating the parameters of mixtures of binomial distributions. *J. Amer. Statist. Assoc.* **59**, 510–528. MR0162310
- [13] BOYS, R. J. AND HENDERSON, D. A. (2001). A comparison of reversible jump MCMC algorithms for DNA sequence segmentation using hidden Markov models. *Comp. Sci. and Statist.* **33**, 35–49.
- [14] BUTUCEA, C. (2007). Goodness-of-fit testing and quadratic functional estimation from indirect observations. *Ann. Statist.* **35**, 5, 1907–1930. MR2363957
- [15] BUTUCEA, C. AND TSYBAKOV, A. B. (2007a). Sharp optimality in density deconvolution with dominating bias. I. *Theory Probab. Appl* **52**, 1, 111–128. MR2354572
- [16] BUTUCEA, C. AND TSYBAKOV, A. B. (2007b). Sharp optimality in density deconvolution with dominating bias. II. *Theory Probab. Appl* **52**, 2, 336–349.
- [17] BÉRARD, J., GOUÉRÉ, J.-B., AND PIAU, D. (2008). Solvable models of neighbor-dependent nucleotide substitution processes. *Mathematical Biosciences* **211**, 56–88.
- [18] BÉRARD, J. AND PIAU, D. (2008). Coupling times with ambiguities for particle systems and applications to context-dependent DNA substitution models. Tech. rep., arXiv :0712.0072.
- [19] CAPPÉ, O., MOULINES, E., AND RYDÉN, T. (2005). *Inference in hidden Markov models*. Springer

- Series in Statistics. Springer, New York. MR2159833
- [20] CARREIRA-PERPIÑÁN, M. Á. AND RENALS, S. (2000). Practical identifiability of finite mixtures of multivariate bernoulli distributions. *Neural Comp.* **12**, 1, 141–152.
- [21] CASTILLO, I. (2007). Semi-parametric second-order efficient estimation of the period of a signal. *Bernoulli* **13**, 4, 910–932. MR2364219
- [22] CATALISANO, M. V., GERAMITA, A. V., AND GIMIGLIANO, A. (2002). Ranks of tensors, secant varieties of Segre varieties and fat points. *Linear Algebra Appl.* **355**, 263–285. MR1930149
- [23] CATALISANO, M. V., GERAMITA, A. V., AND GIMIGLIANO, A. (2005). Higher secant varieties of the Segre varieties $\mathbb{P}^1 \times \cdots \times \mathbb{P}^1$. *J. Pure Appl. Algebra* **201**, 1-3, 367–380. MR2158764
- [24] CAVALIER, L. AND TSYBAKOV, A. B. (2001). Penalized blockwise Stein’s method, monotone oracles and sharp adaptive estimation. *Math. Methods Statist.* **10**, 3, 247–282. Meeting on Mathematical Statistics (Marseille, 2000). MR1867161
- [25] CELISSE, A., GUEDJ, M., NUEL, G., AND ROBIN, S. (2008). kerfdr : A semi-parametric kernel-based approach to local FDR estimations. Tech. rep., INRA.
- [26] CHEN, S. S., DONOHO, D. L., AND SAUNDERS, M. A. (2001). Atomic decomposition by basis pursuit. *SIAM Rev.* **43**, 1, 129–159 (electronic). MR1854649
- [27] CHRISTENSEN, O. (2006). Pseudo-likelihood for non-reversible nucleotide substitution models with neighbour dependent rates. *Stat. Appl. Genet. Mol. Biol.* **5**, 1, Article 18.
- [28] CHRISTENSEN, O., HOBOLTH, A., AND JENSEN, J. (2005). Pseudo-likelihood analysis of codon substitution models with neighbor-dependent rates. *J. Comput. Biol.* **12**, 9, 1166–82.
- [29] ÇINLAR, E. (1969). Markov renewal theory. *Advances in Appl. Probability* **1**, 123–187. MR0268975
- [30] CIUPERCA, G. (2002). Likelihood ratio statistic for exponential mixtures. *Ann. Inst. Statist. Math.* **54**, 3, 585–594. MR1932403
- [31] COOK, A. AND RUSSELL, M. (1986). Improved duration modelling in hidden markov models using series-parallel configurations of states. In *Autumn Conference on Speech and Hearing*. Vol. **8**. Proc. Institute of Acoustics, Windermere, 299–306.
- [32] CSISZÁR, I. AND SHIELDS, P. C. (2000). The consistency of the BIC Markov order estimator. *Ann. Statist.* **28**, 6, 1601–1619. MR1835033
- [33] CSISZÁR, I. AND TALATA, Z. (2006). Context tree estimation for not necessarily finite memory processes, via BIC and MDL. *IEEE Trans. Inform. Theory* **52**, 3, 1007–1016. MR2238067
- [34] DAUDIN, J.-J., PICARD, F., AND ROBIN, S. (2008). A mixture model for random graphs. *Statist. Comput.* **18**, 2, 173–183.
- [35] DEMPSTER, A. P. (1972). Covariance selection. *Biometrics, Special Multivariate Issue* **28**, 157–175.
- [36] DEMPSTER, A. P., LAIRD, N. M., AND RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39**, 1, 1–38. MR0501537
- [37] DURBIN, R., EDDY, S. R., KROGH, A., AND MITCHISON, G. (1998). *Biological sequence analysis. Probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge.
- [38] EDDY, S. R. (1998). Profile hidden Markov models. *Bioinformatics Review* **14**, 9, 755–763.
- [39] EFROMOVICH, S. (1997). Density estimation for the case of supersmooth measurement error. *J.*

- Amer. Statist. Assoc.* **92**, 438, 526–535. MR1467846
- [40] EFRON, B., HASTIE, T., JOHNSTONE, I., AND TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32**, 2, 407–499. MR2060166
- [41] ELMORE, R., HALL, P., AND NEEMAN, A. (2005). An application of classical invariant theory to identifiability in nonparametric mixtures. *Ann. Inst. Fourier (Grenoble)* **55**, 1, 1–28. MR2141286
- [42] EPHRAIM, Y. AND MERHAV, N. (2002). Hidden Markov processes. *IEEE Trans. Inform. Theory* **48**, 6, 1518–1569. Special issue on Shannon theory : perspective, trends, and applications. MR1909472
- [43] ERDŐS, P. AND RÉNYI, A. (1959). On random graphs. I. *Publ. Math. Debrecen* **6**, 290–297. MR0120167
- [44] ERDŐS, P. AND RÉNYI, A. (1961a). On the evolution of random graphs. *Bull. Inst. Internat. Statist.* **38**, 343–347. MR0148055
- [45] ERDŐS, P. AND RÉNYI, A. (1961b). On the strength of connectedness of a random graph. *Acta Math. Acad. Sci. Hungar.* **12**, 261–267. MR0130187
- [46] ERDŐS, P. AND RÉNYI, A. (1966). On the existence of a factor of degree one of a connected random graph. *Acta Math. Acad. Sci. Hungar.* **17**, 359–368. MR0200186
- [47] FANO, R. M. (1952). *Class notes for Transmission of Information*,. Course 6.574. MIT, Cambridge.
- [48] FELSENSTEIN, J. (1981). Evolutionary trees from DNA sequences : A maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376.
- [49] FELSENSTEIN, J. AND CHURCHILL, G. (1996). A Hidden Markov Model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.* **13**, 1, 93–104.
- [50] FERGUSON, J. D. (1980). Variable duration models for speech. In *Proc. symp. on the applications of hidden Markov models to text and speech*, J. D. Ferguson, Ed. Princeton, New Jersey, 143–179.
- [51] FINESSO, L. (1991). Consistent estimation of the order for Markov and hidden Markov chains. Ph.D. thesis, University of Maryland, ISR, USA. <http://www.isr.umd.edu/~baras/publications/dissertations/1975-1995/90-PhD-Finesso.pdf>.
- [52] FRANK, O. AND HARARY, F. (1982). Cluster inference by using transitivity indices in empirical graphs. *J. Amer. Statist. Assoc.* **77**, 380, 835–840. MR686407
- [53] FRIEDMAN, J., HASTIE, T., HÖFLING, H., AND TIBSHIRANI, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics* **1**, 2, 302–332.
- [54] FRIEDMAN, J., HASTIE, T., AND TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 3, 432–441.
- [55] FUH, C.-D. (2006). Efficient likelihood estimation in state space models. *Ann. Statist.* **34**, 4, 2026–2068. MR2283726
- [56] GAMBIN, A., TIURYN, J., AND TYSZKIEWICZ, J. (2006). Alignment with context dependent scoring function. *J. Comput. Biol.* **13**, 1, 81–101 (electronic). MR2253542
- [57] GASSIAT, E. AND BOUCHERON, S. (2003). Optimal error exponents in hidden Markov models order estimation. *IEEE Trans. Inform. Theory* **49**, 4, 964–980. MR1984482
- [58] GASSIAT, E. AND LÉVY-LEDUC, C. (2006). Efficient semiparametric estimation of the periods

- in a superposition of periodic functions with unknown shape. *J. Time Ser. Anal.* **27**, 6, 877–910. MR2328546
- [59] GILBERT, E. J. (1959). On the identifiability problem for functions of finite Markov chains. *Ann. Math. Statist.* **30**, 688–697. MR0107304
- [60] GLICK, N. (1973). Sample-based multinomial classification. *Biometrics* **29**, 241–256. MR0347009
- [61] GOLDMAN, N. AND YANG, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**, 5, 725–736.
- [62] GOLUBEV, G. (1988). Estimating the period of a signal of unknown shape corrupted by white noise. *Probl. Inf. Transm.* **24**, 4, 288–299.
- [63] GOODMAN, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61**, 215–231. MR0370936
- [64] GUÉDON, Y. (2003). Estimating hidden semi-Markov chains from discrete sequences. *J. Comput. Graph. Statist.* **12**, 3, 604–639. MR2002638
- [65] GUÉDON, Y. (2005). Hidden hybrid Markov/semi-Markov chains. *Comput. Statist. Data Anal.* **49**, 3, 663–688. MR2141411
- [66] GUÉDON, Y. (2007). Exploring the state sequence space for hidden Markov and semi-Markov chains. *Comput. Statist. Data Anal.* **51**, 5, 2379–2409. MR2338978
- [67] GYLLENBERG, M., KOSKI, T., REILINK, E., AND VERLAAN, M. (1994). Nonuniqueness in probabilistic numerical identification of bacteria. *J. Appl. Probab.* **31**, 2, 542–548. MR1274807
- [68] HAMILTON, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* **57**, 2, 357–384. MR996941
- [69] HEIN, J., WIUF, C., KNUDSEN, B., MOLLER, M., AND WIBLING, G. (2000). Statistical alignment : computational properties, homology testing and goodness-of-fit. *J. Mol. Biol.* **302**, 265–279.
- [70] HELLER, A. (1965). On stochastic processes derived from Markov chains. *Ann. Math. Statist.* **36**, 1286–1291. MR0176520
- [71] HOLLAND, P. W. (1968). Some properties of an algebraic representation of stochastic processes. *Ann. Math. Statist.* **39**, 164–170. MR0221574
- [72] HOLZMANN, H., BISSANTZ, N., AND MUNK, A. (2007). Density testing in a contaminated sample. *J. Multivariate Analysis* **98**, 1, 57–75. MR2292917
- [73] HUANG, X. (1994). A context dependent method for comparing sequences. In *Combinatorial pattern matching (Asilomar, CA, 1994)*. Lecture Notes in Comput. Sci., Vol. **807**. Springer, Berlin, 54–63. MR1289202
- [74] JENSEN, J. L. AND PEDERSEN, A.-M. K. (2000). Probabilistic models of DNA sequence evolution with context dependent rates of substitution. *Adv. in Appl. Probab.* **32**, 2, 499–517. MR1778577
- [75] KNUDSEN, B. AND MIYAMOTO, M. (2003). Sequence alignments and pair hidden Markov models using evolutionary history. *J. Mol. Biol.* **333**, 453–460.
- [76] KOLTCHINSKII, V. I. (2000). Empirical geometry of multivariate data : a deconvolution approach. *Ann. Statist.* **28**, 2, 591–629. MR1790011
- [77] KOOPMANS, T. C., Ed. (1950). *Statistical Inference in Dynamic Economic Models*. Cowles Com-

- mission Monograph No. 10. John Wiley & Sons Inc., New York, N.Y. MR0038640
- [78] KOOPMANS, T. C. AND REIERSØL, O. (1950). The identification of structural characteristics. *Ann. Math. Statistics* **21**, 165–181. MR0039967
- [79] KRICHEVSKY, R. E. AND TROFIMOV, V. K. (1981). The performance of universal encoding. *IEEE Trans. Inform. Theory* **27**, 2, 199–207. MR633417
- [80] KROGH, A., BROWN, M., MIAN, I., SJOLANDER, K., AND HAUSSLER, D. (1994). Hidden Markov models in computational biology : Applications to protein modelling. *J. Mol. Biol.* **235**, 1501–1531.
- [81] KRUSKAL, J. B. (1976). More factors than subjects, tests and treatments : an indeterminacy theorem for canonical decomposition and individual differences scaling. *Psychometrika* **41**, 3, 281–293. MR0488592
- [82] KRUSKAL, J. B. (1977). Three-way arrays : rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and Appl.* **18**, 2, 95–138. MR0444690
- [83] LEVINSON, S. E. (1986). Continuously variable duration hidden Markov models for automatic speech recognition. *Computer speech and language* **1**, 29–45.
- [84] LEVY, P. (1956). Processus semi-markoviens. In *Proceedings of the International Congress of Mathematicians, 1954, Amsterdam, vol. III*. Erven P. Noordhoff N.V., Groningen, 416–426. MR0088105
- [85] LI, W.-H., WU, C.-I., AND LUO, C.-C. (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* **2**, 2, 150–174.
- [86] LINDSAY, B. G. (1995). *Mixture models : theory, geometry and applications*. NSF-CBMS Regional Conference Series in Probability and Statistics Vol. 5, Institute of Mathematical Statistics, Hayward.
- [87] LUNTER, G. AND HEIN, J. (2004). A nucleotide substitution model with nearest-neighbour interactions. *Bioinformatics* **20**, 1, 216–223.
- [88] MALLOWS, C. L. (1995). More comments on C_p . *Technometrics* **37**, 4, 362–372. MR1365719
- [89] MCLACHLAN, G. AND PEEL, D. (2000). *Finite mixture models*. Wiley Series in Probability and Statistics : Applied Probability and Statistics. Wiley-Interscience, New York. MR1789474
- [90] MEINSHAUSEN, N. AND BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34**, 3, 1436–1462. MR2278363
- [91] MEISTER, A. (2004). On the effect of misspecifying the error density in a deconvolution problem. *Can. J. Stat.* **32**, 4, 439–449. MR2125855
- [92] MEISTER, A. (2005). Non-estimability in spite of identifiability in density deconvolution. *Math. Methods Statist.* **14**, 4, 479–487 (2006). MR2210543
- [93] MEISTER, A. (2006). Density estimation with normal measurement error with unknown variance. *Statist. Sinica* **16**, 1, 195–211. MR2256087
- [94] MEISTER, A. (2007). Deconvolving compactly supported densities. *Math. Methods Statist.* **16**, 1, 63–76. MR2319471
- [95] METZLER, D. (2003). Statistical alignment based on fragment insertion and deletion models. *Bioinformatics* **19**, 4, 490–499.
- [96] MIKLOS, I., LUNTER, G. A., AND HOLMES, I. (2004). A "Long Indel" Model For Evolutionary

- Sequence Alignment. *Mol. Biol. Evol.* **21**, 3, 529–540.
- [97] MITROPHANOV, A. Y. AND BORODOVSKY, M. (2006). Statistical significance in biological sequence analysis. *Briefings in Bioinformatics* **7**, 1, 2–24.
- [98] MUSE, S. AND GAUT, B. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* **11**, 5, 715–724.
- [99] NEUMANN, M. H. (1997). On the effect of estimating the error density in nonparametric deconvolution. *J. Nonparametr. Statist.* **7**, 4, 307–330. MR1460203
- [100] NIGAM, K., MCCALLUM, A. K., THRUN, S., AND MITCHELL, T. M. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning* **39**, 2/3, 103–134.
- [101] NOWICKI, K. AND SNIJDERS, T. A. B. (2001). Estimation and prediction for stochastic block-structures. *J. Amer. Statist. Assoc.* **96**, 455, 1077–1087. MR1947255
- [102] OSBORNE, M. R., PRESNELL, B., AND TURLACH, B. A. (2000a). A new approach to variable selection in least squares problems. *IMA J. Numer. Anal.* **20**, 3, 389–403. MR1773265
- [103] OSBORNE, M. R., PRESNELL, B., AND TURLACH, B. A. (2000b). On the LASSO and its dual. *J. Comput. Graph. Statist.* **9**, 2, 319–337. MR1822089
- [104] PACTER, L. E. AND STURMFELS, B. E. (2005). *Algebraic statistics for computational biology*. Cambridge University Press, New York, NY, USA.
- [105] PEDERSEN, A.-M. K. AND JENSEN, J. L. (2001). A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. *Mol. Biol. Evol.* **18**, 5, 763–776.
- [106] PINSKER, M. (1980). Optimal filtration of square-integrable signals in Gaussian noise. *Probl. Inf. Transm.* **16**, 120–133.
- [107] PYKE, R. (1961a). Markov renewal processes : definitions and preliminary properties. *Ann. Math. Statist.* **32**, 1231–1242. MR0133888
- [108] PYKE, R. (1961b). Markov renewal processes with finitely many states. *Ann. Math. Statist.* **32**, 1243–1259. MR0154324
- [109] RISSANEN, J. (1978). Modelling by shortest data description. *Automatica* **14**, 465–471.
- [110] ROTHENBERG, T. J. (1971). Identification in parametric models. *Econometrica* **39**, 577–591. MR0436944
- [111] RUSSEL, M. J. AND MOORE, R. K. (1985). Explicit modeling of state occupancy in hidden Markov models for automatic speech recognition. In *Proc. of the int. conference on acoustics, speech and signal processing*. IEEE, New York, 5–8.
- [112] SCHÖNIGER, M. AND VON HAESELER, A. (1994). A stochastic model for the evolution of autocorrelated DNA sequences. *Molecular Phylogenetics and Evolution* **3**, 3, 240–247.
- [113] SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 2, 461–464. MR0468014
- [114] SIEPEL, A. AND HAUSSLER, D. (2004). Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.* **21**, 3, 468–488.
- [115] SMITH, W. L. (1954). Asymptotic renewal theorems. *Proc. Roy. Soc. Edinburgh. Sect. A.* **64**, 9–48.

- MR0060755
- [116] SNIJDERS, T. A. B. AND NOWICKI, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *J. Classification* **14**, 1, 75–100. MR1449742
- [117] TALLBERG, C. (2005). A Bayesian approach to modeling stochastic blockstructures with covariates. *Journal of Mathematical Sociology* **29**, 1, 1–23.
- [118] TEICHER, H. (1967). Identifiability of mixtures of product measures. *Ann. Math. Statist* **38**, 1300–1302. MR0216635
- [119] THORNE, J., KISHINO, H., AND FELSENSTEIN, J. (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.* **33**, 114–124.
- [120] THORNE, J., KISHINO, H., AND FELSENSTEIN, J. (1992). Inching toward reality : an improved likelihood model of sequence evolution. *J. Mol. Evol.* **34**, 3–16.
- [121] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 1, 267–288. MR1379242
- [122] TSYBAKOV, A. B. (2004). *Introduction à l'estimation non-paramétrique. (Introduction to nonparametric estimation)*. Mathématiques & Applications. 41. Springer, Paris.
- [123] VAN DER VAART, A. W. AND WELLNER, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York. MR1385671
- [124] VIDYASAGAR, M. (2005). Realization theory for hidden Markov models : The complete realization problem. In *IEEE Conference on Decision and Control*. IEEE, Sevilla, Spain.
- [125] VON HAESELER, A. AND SCHONIGER, M. (1998). Evolution of DNA or amino acid sequences with dependent sites. *J. Comput. Biol.* **5**, 1, 149–164.
- [126] WHELAN, S. AND GOLDMAN, N. (2004). Estimating the frequency of events that cause multiple-nucleotide changes. *Genetics* **167**, 4, 2027–2043.
- [127] WILBUR, W. J. AND LIPMAN, D. J. (1984). The context dependent comparison of biological sequences. *SIAM J. Appl. Math.* **44**, 3, 557–567. MR745112
- [128] YANG, Z. (1995). A space-time process model for the evolution of DNA sequences. *Genetics* **139**, 2, 993–1005.
- [129] YU, Y.-K., BUNDSCHUH, R., AND HWA, T. (2002). Statistical significance and extremal ensemble of gapped local hybrid alignment. In *Biological Evolution and Statistical Physics*. Lecture Notes in Physics, Vol. **585**. Springer, Berlin/Heidelberg, 3–21.
- [130] YU, Y.-K. AND HWA, T. (2001). Statistical significance of probabilistic sequence alignment and related local hidden Markov models. *J. Comput. Biol.* **8**, 3, 249–282.
- [131] ZOLOTAREV, V. M. (1986). *One-dimensional stable distributions*. Translations of Mathematical Monographs, Vol. **65**. American Mathematical Society, Providence, RI. MR854867
- [132] ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 476, 1418–1429. MR2279469

Liste des travaux

Revue avec comité de lecture

- [M1] DOUC, RANDAL ET MATIAS, CATHERINE (2001). Asymptotics of the maximum likelihood estimator for general hidden Markov models. *Bernoulli*, **7**, 3, 381–420. MR1836737
- [M2] MATIAS, CATHERINE (2002). Semiparametric deconvolution with unknown noise variance. *ESAIM Probab. & Stat.*, **6**, 271–292. MR1943151
- [M3] MATIAS, CATHERINE ET TAUPIN, MARIE-LUCE (2004). Minimax estimation of linear functionals in the convolution model. *Mathematical Methods of Statistics*, **13**, 3, 282–328. MR2109461
- [M4] BUTUCEA, CRISTINA ET MATIAS, CATHERINE (2005). Minimax estimation of the noise level and of the deconvolution density in a semiparametric convolution model. *Bernoulli* **11**, 2, 309–340. MR2132729
- [M5] MATIAS, C. ET SCHBATH, S. ET BIRMELE, E. ET DAUDIN, J-J. ET ROBIN, S. (2006). Networks motifs : mean and variance for the count. *Revstat* , **4**, 1, 31–51. MR2259363
- [M6] CASTILLO, ISMAËL ET LÉVY-LEDUC, CÉLINE ET MATIAS, CATHERINE (2006). Exact adaptive estimation of the shape of a periodic function with unknown period corrupted by white noise. *Mathematical Methods of Statistics*, **15**, 2, 146–175. MR2256473
- [M7] ARRIBAS-GIL, ANA ET GASSIAT, ELISABETH ET MATIAS, CATHERINE (2006). Parameter estimation in pair hidden Markov models. *Scandinavian Journal of Statistics* **33**, 4, 651–671. MR2300909
- [M8] CHAMBAZ, ANTOINE ET MATIAS, CATHERINE (2008). Number of hidden states and memory : a joint order estimation problem for Markov chains with Markov regime. À paraître dans *ESAIM Probab. & Stat.*.
- [M9] BUTUCEA, CRISTINA ET MATIAS, CATHERINE ET POUET, CHRISTOPHE (2008). Adaptive goodness-of-fit testing from indirect observations. À paraître dans les *Annales de l'IHP, Probabilités et Statistiques*.
- [M10] BUTUCEA, CRISTINA ET MATIAS, CATHERINE ET POUET, CHRISTOPHE (2008). Adaptivity in convolution models with partially known noise distribution. *Electronic Journal of Statistics* **2**, 897-915.

Prépublications et manuscrits en préparation

- [M11] ALLMAN, ELIZABETH ET MATIAS, CATHERINE ET RHODES, JOHN. Identifiability of latent class models with many observed variables. ArXiv :0809.5032, 2008.
- [M12] AMBROISE, CHRISTOPHE ET CHIQUET, JULIEN ET MATIAS, CATHERINE. Penalized Maximum Likelihood Inference for Sparse Gaussian Graphical Models with Hidden Structure. Manuscrit en préparation.

Communications avec actes et comité de lecture

- GERENCSÉR, LÁSZLÓ ET MATIAS, CATHERINE. Self-exciting point processes with applications in finance and medicine. 18th International Symposium on Mathematical Theory of Networks and Systems, 2008.
- BUTUCEA, CRISTINA ET MATIAS, CATHERINE (2004). Estimation minimax dans un modèle de mélange semi paramétrique. *Actes des XXXVIèmes journées de statistique*.
- MATIAS, C. ET SCHBATH, S. ET BIRMELÉ, E. ET DAUDIN, J-J. ET ROBIN, S. (2005). Networks motifs : mean and variance for the count. *Proceedings CompBioNets2005*.

Publications dans des revues sans comité, notes

- P.Y. BOURGUIGNON, M. GUEDJ, F. KÉPÈS, C. MATIAS, G. NUEL, N. OMONT, B. PRUM (2004). Recherche de gènes impliqués dans une maladie. *Matapli*, **74**.
- DOUC, RANDAL ET MATIAS, CATHERINE (2000). Propriétés asymptotiques de l'estimateur de maximum de vraisemblance pour des modèles de Markov cachés généraux. *C.R.A.S.*, **330**, Série I, 135–138.

Liste des co-auteurs

Elizabeth Allman, University of Fairbanks, Alaska (US).
Christophe Ambroise, Université d'Évry Val d'Essonne, Évry.
Ana Arribas-Gil, Universidad Carlos III, Madrid (España).
Etienne Birmelé, Université d'Évry Val d'Essonne, Évry.
Cristina Butucea, Université des Sciences et Technologies de Lille 1.
Ismaël Castillo, Vrije Universiteit, Amsterdam (Pays Bas).
Antoine Chambaz, Université René Descartes, Paris.
Julien Chiquet, Université d'Évry Val d'Essonne, Évry.
Jean-Jacques Daudin, Agro Paris Tech, Paris.
Randal Douc, Telecom Sud Paris, Évry.
Elisabeth Gassiat, Université Paris Sud, Orsay.
László Gerencsér, Sztaki Institute, Budapest (Hongrie).
Céline Lévy-Leduc, Telecom Paris Tech, Paris.
Christophe Pouet, Université de Provence, Marseille.
John Rhodes, University of Fairbanks, Alaska (US).
Stéphane Robin, Agro Paris Tech, Paris.
Sophie Schbath, INRA, Jouy-en-Josas.
Marie-Luce Taupin, Université René Descartes, Paris.