Sur diverses extensions des chaînes de Markov cachées avec application au traitement des signaux radar

Jérôme Lapuyade-Lahorgue

Ecole doctorale SMPC

Thèse présentée pour l'obtention du diplôme de
DOCTEUR DE l'INSTITUT NATIONAL DES TELECOMMUNICATIONS

Doctorat délivré conjointement par
Telecom Sud-Paris et l'Université Pierre et Marie Curie-Paris 6

Spécialité : MATHÉMATIQUES APPLIQUÉES

Présentée par

Jérôme LAPUYADE-LAHORGUE

Sujet de la thèse

SUR DIVERSES EXTENSIONS DES CHAÎNES DE MARKOV CACHÉES AVEC APPLICATION AU TRAITEMENT DES SIGNAUX RADAR.

Soutenue le 10 décembre 2008.
devant le jury composé de

Président	Michel BRONIATOWSKI	Professeur à l'Université Paris VI.
Rapporteur	Alain HILLION	Professeur à Telecom Bretagne.
Rapporteur	Nikolaos LIMNIOS	Professeur à l'Université de Technologie de Compiègne.
Examinateur	Christophe AMBROISE	Professeur à l'Université d'Evry.
Examinateur	Michel BRONIATOWSKI	Professeur à l'Université Paris VI.
Examinateur	Stéphane DERRODE	Maître de Conférence (HDR) à l'Ecole Centrale de Marseille.
Examinateur	François ROUEFF	Professeur à Telecom ParisTech.
Directeur de thèse	Wojciech PIECZYNSKI	Professeur à Telecom SudParis.
Invité du jury	Frédéric BARBARESCO	Co-encadrant, Expert Radar à Thales Air Systems.
Invité du jury	Philippe GUGUEN	Ingénieur à la DGA de Rennes.

Thèse n° 2008INT0024

Remerciements

Je tiens en premier lieu à remercier Wojciech Pieczynski, mon directeur de thèse, d'avoir accepté de diriger mes travaux durant mes trois années de thèse et de m'avoir communiqué diverses connaissances mathématiques utiles pour la suite de mon cursus professionnel. Je le remercie particulièrement pour m'avoir soutenu tout au long de ma thèse.

Je remercie également Frédéric Barbaresco, avec qui j'ai entretenu de nombreuses discussions sur des sujets mathématiques très divers.

Je remercie avec reconnaissance Nikolaos Limnios et Alain Hillion qui ont accepté la tâche de rapporteurs, ainsi que les autres membres du jury Christophe Ambroise, Michel Broniatowski, Stéphane Derrode et François Roueff, pour l'intérêt qu'ils ont porté à mes travaux.

Je remercie la DGA de m'avoir donné les moyens financiers indispensables à la réalisation de cette thèse.

Je remercie mes collègues de travail, doctorants, chercheurs et membres du personnel que j'ai pû cotoyer durant ces trois ans. Je remercie en particulier Marc Oudin et Boujemaa Ait-El-Fquih qui ont apporté la bonne ambiance également indispensable dans tout lieu de travail.

Plus personnellement, je remercie ma famille dont mon père pour m'avoir soutenu dans les moments difficiles de cette thèse. Je remercie également mes amis et plus particulièrement Stéphane et Joachim qui m'ont supporté (mais dans la bonne humeur) durant ces trois ans.

Table des matières

Notations

Symboles

$\mathbb{P}$:	Mesure de probabilité.
$\mathbb{E}$:	Espérance associée.
X, Y :	Variables aléatoires en majuscule.
x, y :	Leurs réalisations respectives en minuscule.
$p(x;\theta)$, $p(y;\theta)$:	Modèles statistiques paramétriques respectifs.
Θ :	Espace des paramètres d'un modèle statistique.
μ_Θ :	Mesure a priori sur Θ.
$\mathcal{Y}$:	Espace vectoriel de dimension finie, espace des observations.
ν :	Mesure de référence sur $\mathcal{Y}$.
L :	Fonction de perte.
$\mathcal{B}_\Theta$:	Tribu borélienne sur Θ.
$\mathcal{B}_\mathcal{Y}$:	Tribu borélienne sur $\mathcal{Y}$.
$R(\theta,\varphi)$:	Risque moyen associé à l'estimateur φ.
$\rho(\mu,\varphi)$:	Risque bayésien associé à la mesure a priori μ et à l'estimateur φ.
$r_\mu(y,\varphi)$:	Risque a posteriori associé à la mesure a priori μ et à l'estimateur φ.
$\lambda_{\mathbb{R}^k}$:	Mesure de Lebesgue sur $\mathbb{R}^k$.
$\bar{K}(p,q)$:	Information de Kullback entre p et q.
$\mathcal{X}$:	Espace des états cachés fini.
$\nu_\mathcal{X}$:	Mesure de décompte sur $\mathcal{X}$.
$p(x)$:	Pour x discret, probabilité de $X = x$.
$p(y)$:	Pour $y \in \mathcal{Y}$, densité de Y par rapport à ν.
$p(x,y)$:	Pour x discret et $y \in \mathcal{Y}$ continu, densité par rapport à $\nu_\mathcal{X} \otimes \nu$.
U :	Processus auxiliaire d'un modèle à données latentes.
$U^1, \ldots, U^l$:	l processus auxiliaires d'un modèle à données latentes.
$q(.\|.)$:	Transition de la chaîne immergée.
$d(.,.)$:	Loi du temps de séjour.
Cov :	Covariance d'un processus.
$\bar{Z}$:	Conjugué de Z.
C :	Fonction de répartition d'une copule.
c :	Densité d'une copule.
$(\gamma(k))_{k\in\mathbb{Z}}$:	Famille de covariance d'un processus.
$\mathbb{P}_{\text{fa}}$:	Probabilité de fausses alarmes.
$\mathbb{P}_{\text{d}}$:	Probabilité de détection.
dl :	Elément de longueur.
$d\tau$:	Elément de volume.

Acronymes

MAP :	Maximum A Posteriori.
MPM :	Maximum Marginal a Posteriori.
ML :	Maximum de Vraisemblance.
CMC :	Chaîne de Markov cachée.
CMCouple :	Chaîne de Markov couple.
CMT :	Chaîne de Markov triplet.
CMC-MS :	Chaîne de Markov cachée M-stationnaire.
CSMC :	Chaîne semi-markovienne cachée.
CMC-ML :	Chaîne de Markov cachée à mémoire longue.
EM :	Expectation Maximization.
ICE :	Iterative Conditional Estimation.
BI :	Bruit Indépendant.
ML :	Mémoire Longue.
CCPM :	Chaîne Couple Partiellement de Markov.
CTPM :	Chaîne Triplet Partiellement de Markov.
TFAC :	Taux de Fausses Alarmes Constant.

Introduction générale

La segmentation d'un signal ou d'une image est la technique permettant de diviser ce signal ou cette image en un nombre fini de zones. Le résultat de la segmentation est une "cartographie" du signal ou de l'image d'origine permettant ainsi de faciliter son analyse. Il existe deux grands types de segmentation [34] : les segmentations par contour et les segmentations par région. Les segmentations par contour s'appuient sur les discontinuités de l'image afin de la découper en zones. Tandis que dans la segmentation par région, nous regroupons les pixels selon les caractéristiques de chaque zone. C'est ce type de segmentation que nous utiliserons dans cette thèse. Plus exactement, les propriétés sur lesquelles s'appuiera notre segmentation seront de nature statistique et les techniques de segmentation utilisées seront les méthodes d'inférence bayésienne [16]. La technique de segmentation sera alors qualifiée de "segmentation bayésienne". En segmentation bayésienne, le signal ou l'image est considéré comme la réalisation y d'un champ aléatoire $Y = (Y_u)_{u \in \mathcal{S}}$ et on estime la réalisation cachée x d'un champ aléatoire $X = (X_u)_{u \in \mathcal{S}}$, cette dernière correspondant à la segmentation.

Au plan mathématique, la problématique traitée dans cette thèse est donc celle de l'estimation d'une grande quantité de variables aléatoires inobservables $X = (X_u)_{u \in \mathcal{S}}$ à partir des variables observées $Y = (Y_u)_{u \in \mathcal{S}}$. La taille de l'ensemble fini d'indices $\mathcal{S}$ est supposée être trop grande pour que l'on puisse définir et manipuler la loi $p(x, y)$ du couple (X, Y) dans toute sa généralité ; on est alors obligé de considérer les lois $p(x, y)$ de forme particulière. Les chaînes de Markov cachés (CMC) sont parmi les modèles les plus simples et les plus utilisés pour modéliser $p(x, y)$; ils permettent la recherche de $X = x$ par l'application de différentes méthodes bayésiennes. Dans de tels modèles, la loi $p(x)$ de X est markovienne et la loi $p(y|x)$ modélise le "bruit". Ces modèles sont appliqués dans de très nombreux problèmes et le nombre de publications scientifiques sur le sujet est en croissance exponentielle. Citons quelques publications récentes concernant les biosciences [64, 86, 89], la climatologie [10], les communications [27], l'écologie [71], l'économétrie et les finances [51, 115], le traitement de l'écriture [31], des signaux musicaux [106], ou encore des images [48, 77]. Cependant, ces modèles ont leur limite pour la raison suivante. Les traitements bayésiens utilisants les CMC sont rendus possibles par la markovianité de la loi $p(x|y)$. Or, lorsque l'on a supposé la markovianité de $p(x)$, la markovianité de $p(x|y)$ nécessaire aux traitements n'est obtenue qu'au prix de la simplification, qui peut être excessive dans certaines applications, de la loi $p(y|x)$. Cet inconvénient a été contourné par la généralisation des CMC aux chaînes de Markov "Couples" (CMCouples, [96]), dans lesquelles on suppose directement la markovianité de $p(x, y)$. La markovianité étant conservée par le conditionnement, les deux lois $p(x|y)$ et $p(y|x)$ sont alors markoviennes. La markovianité de $p(x|y)$ autorise alors les mêmes traitements que dans les CMC classiques, et la markovianité de $p(y|x)$ autorise des modélisations du bruit bien plus complètes que dans les CMC classiques. On obtient ainsi un modèle plus général dont les applications peuvent notablement améliorer les résultats obtenus avec les CMC classiques

[39]. Par la suite, les CMCouples ont été étendues aux chaînes de Markov triplets (CMT [95]), dans lesquelles on considère une troisième chaîne aléatoire $U = (U_u)_{u \in \mathcal{S}}$ et l'on suppose la markovianité de la loi $p(x, u, y)$ du triplet (X, U, Y). On obtient ainsi un modèle générique très riche, donnant lieu à des multiples possibilités particulières [100, 102]. Enfin, les CMT ont été généralisées aux chaînes triplet partiellement de Markov (CTPM [97]) qui sont des triplets (X, U, Y) markoviens par rapport à (X, U), mais non nécessairement markoviens par rapport à Y.

Notre travail se situe dans le contexte général des CMT et CTPM. Nous apportons un certain nombre de contributions théoriques ou expérimentales, en proposant des CMT et CTPM originaux et en validant l'intérêt des nouveaux modèles par des expérimentations. Nous proposons diverses applications au traitement de l'image et du signal, notamment du signal radar ; cependant, les nouveaux modèles ont une portée très générale et peuvent s'appliquer partout où s'appliquent les CMC classiques.

Le contenu de la thèse est le suivant.

Les rappels sur l'inférence bayésienne classique sont présentés dans le **chapitre 1**. En particulier, nous y introduisons les mesures de Jeffreys, qui sont étroitement liées à la notion de métrique dans les ensembles fonctionnels de densités de probabilité. Cette métrique sera définie au chapitre 6 et utilisée dans les applications au radar.

Le **chapitre 2** est consacré à l'inférence bayésienne classique dans les modèles de Markov cachés. Nous rappelons les diverses markovianités sur graphes et détaillons le calcul des deux estimateurs bayésiens classiques, qui sont le "Maximum a Posteriori" (MAP) et le "Maximum des Marginales a Posteriori" (MPM), dans les cas des chaînes et des arbres de Markov cachés. Nous précisons également deux méthodes d'estimation des paramètres qui sont "Maximisation-Espérance" (abréviation anglaise EM) et "Estimation conditionnelle itérative" (abréviation anglaise ICE).

Les CMCouples et les CMT sont rappelés dans le **chapitre 3** et nous y présentons deux modèles originaux. Le premier concerne l'introduction des chaînes semi-markoviennes cachées originales, qui sont des CMT particuliers, et qui permettent des calculs plus rapides que les chaînes semi-markoviennes cachées classiques. Le deuxième consiste en l'introduction des CMT modélisant simultanément la semi-markovianité et la non stationnarité de la chaîne cachée. Nous présentons également des simulations et des segmentations des images réelles validant l'intérêt des nouveaux modèles et des traitements non supervisés associés.

Dans le **chapitre 4**, nous nous intéressons aux bruits $p(y|x)$ non gaussiens. Un cadre très général, que nous reprenons, a été récemment exploré par les travaux de N. Brunel, qui ont permis l'introduction des copules dans le contexte des chaînes de Markov cachées [21, 22, 23]. Nous reprenons ces idées en les étendant à des CMT permettant de traiter, en non supervisé, des chaînes non stationnaires cachées par du bruit non gaussien.

Le **chapitre 5** est consacré aux CTPM permettant de prendre en compte des bruits à mémoire longue. Nous y proposons plusieurs modèles originaux dans lesquels la chaîne inobservable est markovienne ou semi-markovienne et où le bruit, dont les marginales sont gaussiennes ou pas, est à dépendance longue. Nous proposons également une méthode d'estimation des paramètres de type ICE originale, dont l'extension aux modèles à mémoire longue est non triviale, et validons l'intérêt des démarches non supervisées réunissant ICE et MPM par des expériences informatiques.

Dans le **chapitre 6**, nous appliquons l'inférence bayésienne au traitement du signal radar. Dans ce contexte, un élément de $\mathcal{S}$ est un couple $s = (d, \theta)$, où d représente la distance au

radar et θ l'angle de visée, appelé également azimut. La réalisation y_u de Y_u est le signal reçu de la distance d et de l'azimut θ. La réalisation estimée x de $X = (X_u)_{u \in \mathcal{S}}$ est une segmentation du signal. Nous verrons comment la segmentation bayésienne peut être utilisée, comme "résumé" du signal, par d'autres traitements postérieurs. Le traitement que nous proposerons est celui de la détection. Nous proposerons dans ce chapitre un détecteur original utilisant une distance entre lois de probabilité. Ce détecteur généralise le détecteur classique du "Taux de Fausses Alarmes Constant" (TFAC), nous définirons ainsi la moyenne de lois de probabilité et la distance entre deux lois de probabilité. Les améliorations apportées par ce détecteur sont validées par des expérimentations sur données réelles.

Chapitre 1

Généralités sur l'inférence bayésienne

Nous présentons dans ce chapitre les fondements de l'inférence bayésienne. L'inférence bayésienne est un domaine des statistiques très riche et ouvrant la voie à diverses applications. Nous parlons d'inférence "bayésienne" lorsque l'on se donne une distribution a priori sur ce qu'on cherche à inférer. Ce qu'on cherche à inférer peut être le paramètre θ d'un modèle paramétrique $p(y;\theta)$, mais on peut également utiliser l'inférence bayésienne pour estimer la réalisation cachée d'un modèle à données latentes. Dans un modèle à données latentes, le signal observé est considéré comme la réalisation y d'un processus Y et ce que l'on cherche est la réalisation x d'un processus X, les deux processus étant liés par une loi de probabilité $p(x,y)$.
Ce chapitre se divise en deux sections. La première amène le formalisme bayésien dans sa généralité. Dans la deuxième section, nous nous intéressons au choix de l'a priori et au choix de la loi $p(x,y)$ dans les modèles à données latentes. Concernant l'inférence bayésienne dans un modèle paramétrique $p(y;\theta)$, nous y abordons deux types de lois a priori. Les deux lois a priori abordées sont les lois conjuguées à une famille paramétrique et les mesures de Jeffreys. Les premières présentent un intérêt algorithmique car la loi a posteriori est dans la même famille paramétrique que la loi a priori. Les lois conjuguées sont souvent utilisées en estimation des paramètres par échantillonnage de Gibbs car la règle de Bayes est simple à implémenter [36]. Quant aux mesures de Jeffreys, elles font partie de la catégorie des mesures a priori dites "non informatives". On choisit d'utiliser un a priori non informatif lorsque l'on ne dispose d'aucune connaissance sur le paramètre. Concernant les modèles à données latentes, les lois a priori seront choisies de façon à ce que les modèles $p(x,y)$ permettent d'utiliser les algorithmes d'inférence bayésienne tels que les algorithmes de Baum-Welsh et de Viterbi. Ces modèles devront être suffisamment simples pour pouvoir utiliser ce type d'algorithmes, et suffisamment riches pour pouvoir modéliser certaines propriétés comme la markovianité, la semi-markovianité ou la dépendance longue dans les observations.

1.1 Principe de l'inférence bayésienne

On considère Y une variable aléatoire à valeurs dans un $\mathbb{R}$-espace vectoriel de dimension finie $\mathcal{Y}$ muni de sa tribu borélienne $\mathcal{B}_{\mathcal{Y}}$. Un modèle statistique paramétrique pour la loi de Y est une famille de densités de probabilité $\{y \in \mathcal{Y} \to p(y;\theta) : \theta \in \Theta\}$ par rapport à une mesure ν sur $\mathcal{Y}$. La fonction de $\mathcal{Y} \times \Theta$ dans $\mathbb{R}^+$ qui à (y,θ) associe $p(y;\theta)$ est appelée vraisemblance. L'ensemble Θ est l'ensemble des paramètres du modèle ; on considèrera dans la suite que

$\Theta \subset \mathbb{R}^k$. Muni de sa tribu borélienne $\mathcal{B}_\Theta$, $(\Theta, \mathcal{B}_\Theta)$ est un espace mesurable, il sera également muni d'une mesure de référence. Lorsque Θ est un sous-ensemble discret, la mesure de référence est la mesure de décompte et lorsque Θ est un ouvert non vide de $\mathbb{R}^k$, ce sera la mesure induite par la mesure de Lebesgue $\lambda_{\mathbb{R}^k}$.
Une stratégie de décision est une fonctionnelle φ de $\mathcal{Y}$ dans Θ. On l'appelle aussi estimateur de θ.

1.1.1 Fonction de coût et risque

Une fonction $L : \Theta \times \Theta \to \mathbb{R}^+$ est dite "fonction de coût" si elle vérifie $L(\theta, \hat{\theta}) = 0$ lorsque $\hat{\theta} = \theta$. Soit $\varphi : \mathcal{Y} \to \Theta$ une stratégie de décision. On appelle risque la quantité

$$R(\theta, \varphi) = \mathbb{E}_\theta \left[L(\theta, \varphi(Y)) \right], \tag{1.1}$$

où $\mathbb{E}_\theta$ est l'espérance.
Pour tout $\theta \in \Theta$, le risque est le coût moyen induit par φ.

1.1.2 Des stratégies admissibles aux stratégies bayésiennes

Définition 1.1.1 (Relation de préférence et stratégies admissibles)**.** *Notons Φ l'ensemble des stratégies de décisions. Une relation de préférence est une relation d'ordre sur Φ. La relation "φ_1 préférée à φ_2" (resp. strictement préférée) est notée $\varphi_1 \gtrsim \varphi_2$ (resp. $\varphi_1 > \varphi_2$) et on dit que φ_1 et φ_2 sont équivalentes si $\varphi_1 \gtrsim \varphi_2$ et $\varphi_2 \gtrsim \varphi_1$. On dit que φ est une stratégie admissible s'il n'existe pas de stratégie qui lui soit strictement préférée.*

La relation "$\varphi \gtrsim \varphi' \Leftrightarrow \forall \theta,\ R(\theta, \varphi) \leq R(\theta, \varphi')$" n'est pas une relation d'ordre total ; en effet, pour certaines valeurs de θ, on peut avoir $R(\theta, \varphi) \leq R(\theta, \varphi')$ tandis que pour d'autres valeurs de θ, on a $R(\theta, \varphi) \geq R(\theta, \varphi')$. On peut alors considérer le risque bayésien qui ne dépend pas de θ mais d'une mesure μ sur l'espace des paramètres $(\Theta, \mathcal{B}_\Theta)$ appelée "mesure a priori". On notera f la densité de la mesure a priori par rapport à la mesure de référence de Θ. La mesure a priori n'est pas obligatoirement une mesure de probabilité. De plus, elle peut vérifier $\mu(\Theta) = +\infty$, on dit alors qu'elle est impropre. On exigera par contre que la quantité $p_\mu(y) =^{\text{def}} \displaystyle\int_\Theta p(y;\theta) d\mu(\theta)$ soit finie. Dans ce cas, la densité $\theta \to \dfrac{p(y;\theta) f(\theta)}{p_\mu(y)}$ définit une mesure de probabilité sur $(\Theta, \mathcal{B}_\Theta)$ appelée mesure a posteriori que l'on notera $\mu(.|y)$. Sa densité sera notée $f(.|y)$. La formule :

$$f(\theta|y) = \frac{p(y;\theta) f(\theta)}{p_\mu(y)}$$

est parfois appelée "la règle de Bayes".
Le risque bayésien est ensuite défini par :

$$\rho(\mu, \varphi) = \mathbb{E}_\mu \left[R(\theta, \varphi) \right], \tag{1.2}$$

où $\mathbb{E}_\mu$ est l'intégration sous la mesure μ.
La mesure a priori quantifie la connaissance que l'on a avant toute expérience sur le paramètre

θ. Nous détaillerons à la section 1.2 le choix de cet a priori.
Dans le cadre bayésien, on dit que φ est préférée à φ' si $\rho(\mu,\varphi) \leq \rho(\mu,\varphi')$. Dans ce cas la stratégie admissible φ_μ, si elle existe, est appelée "stratégie bayésienne".
Le risque bayésien s'écrit également :

$$\rho(\mu,\varphi) = \int_{\mathcal{Y}} \mathbb{E}_\mu \left[L(\theta,\varphi(y))|y\right] p_\mu(y) d\nu(y),$$

où $\mathbb{E}_\mu[.|y]$ est l'intégration sous la mesure a posteriori $\mu(.|y)$.
La quantité $r_\mu(y,\varphi) = \mathbb{E}_\mu \left[L(\theta,\varphi(y))|y\right]$ est appelée risque a posteriori et on a le résultat classique suivant :

Proposition 1.1.1. *Si φ est une stratégie de décision telle que :*

$$\forall y \in \mathcal{Y},\ \forall \varphi' \in \Phi,\ r_\mu(y,\varphi) \leq r_\mu(y,\varphi'), \tag{1.3}$$

alors φ est une stratégie bayésienne.

Preuve.
Voir [44]. □

Dans le cas où la stratégie bayésienne est unique, il suffit alors de chercher la décision satisfaisant (1.3).
Donnons quelques exemples de stratégies bayésiennes couramment rencontrées :

- si Θ est continu et si $L(\theta,\hat{\theta}) = (\theta-\hat{\theta})^2$, alors $r_\mu(y,\varphi) = \mathbb{E}_\mu\left[(\theta-\varphi(y))^2|y\right]$. La stratégie bayésienne est l'espérance a posteriori $\hat{\theta}(y) = \mathbb{E}_\mu(\theta|y)$;
- si Θ est discret et si $L(\theta,\hat{\theta})$ vaut 0 si $\theta = \hat{\theta}$ et 1 sinon, alors la stratégie bayésienne est donnée par $\hat{\theta}_{MAP}(y) = \arg\max_\theta \mu(\theta|y)$. On appelle cet estimateur "estimateur du Maximum A Posteriori" (MAP) ;
- soit $\epsilon > 0$. Si Θ est continu et si $L(\theta,\hat{\theta})$ vaut 0 lorsque $\left|\theta-\hat{\theta}\right| < \epsilon$ et 1 sinon, alors la stratégie bayésienne est donnée par $\hat{\theta}_\epsilon(y) = \arg\max_\theta \mu([\theta-\epsilon,\theta+\epsilon]|y)$.

L'estimateur du MAP est généralisé au cas où Θ est continu par $\hat{\theta}_{MAP}(y) = \arg\max_\theta f(\theta|y)$; cependant, il n'est pas défini à partir de la même fonction de perte que dans le cas discret. Le lien avec l'estimateur $\hat{\theta}_\epsilon$ et l'étude de l'estimateur du MAP dans le cas continu figurent dans [15] pages 258-259.
Notons l'exemple classique suivant :

Exemple 1.1.1. Soit $\{p(y;\theta) : \theta \in \Theta\}$ un modèle paramétrique. Considérons les deux estimateurs de θ suivants :

- maximum de vraisemblance : $\hat{\theta}_{ML} = \arg\max p(y;\theta)$;
- maximum a posteriori : $\hat{\theta}_{MAP} = \arg\max f(\theta|y)$.

Si la densité de la loi a priori est $f(\theta) \propto 1$ pour tout $\theta \in \Theta$, alors la densité de la loi a posteriori est $f(\theta|y) \propto \dfrac{p(y;\theta)}{p_\mu(y)}$ et donc $\hat{\theta}_{ML} = \hat{\theta}_{MAP}$.

Remarque : La loi uniforme a longtemps été considérée comme la loi non informative. Le caractère non informatif de cette loi a été énoncé pour la première fois par T. Bayes dans un contexte particulier. En effet, celui-ci considère qu'en absence de connaissance sur le paramètre θ, nous n'avons aucune raison de privilégier un événement $\theta \in A$ plutôt qu'un autre. Mais comme l'a souligné R. A. Fisher, lorsqu'on effectue un changement de paramétrage $\eta = g(\theta)$, ne pas connaître θ est équivalent à ne pas connaître η. Cependant si θ suit une loi uniforme, η ne suit pas en général une loi uniforme. Le contexte particulier dans lequel travaillait T. Bayes fut celui où Θ est discret. Dans ce cas, l'image d'une loi uniforme par une fonctionnelle est encore une loi uniforme. Lorsque l'espace Θ est continu, nous devons alors choisir un autre type de loi non informative. Plus exactement, nous disons qu'une loi est non informative lorsqu'elle maximise la quantité d'information manquante. Nous devons pour cela définir ce qu'est la quantité d'information. Lorsque θ prend ses valeurs dans $\{\theta_1, \ldots, \theta_m\}$ avec les probabilités $\mu(\theta_j) = \mu_j$, celle-ci est définie comme l'entropie de Shannon :

$$H(\mu) = -\sum_{j=1}^{m} \log(\mu_j)\mu_j. \tag{1.4}$$

Nous voyons que cette quantité est bien maximale lorsque θ suit la loi uniforme.
Lorsque Θ est continu et θ suit une loi de densité f par rapport à la mesure de Lebesgue $\lambda_{\mathbb{R}^k}$, l'entropie de Shannon est généralisée par :

$$H(\mu) = -\int_{\Theta} \log f(\theta) f(\theta) d\lambda_{\mathbb{R}^k}(\theta).$$

Cependant, ce n'est pas la bonne mesure d'information que nous devons maximiser. La maximisation de cette quantité fournit des lois uniformes et nous avons souligné que la loi uniforme n'est pas la bonne mesure non informative dans le cas continu. De la même façon, une quantité d'information ne doit pas dépendre du paramétrage, ainsi la quantité d'information sur θ doit être égale à la quantité d'information sur $\eta = g(\theta)$. Cependant, l'entropie de Shannon est invariante par changement de paramétrage uniquement dans le cas discret. Nous verrons dans la sous-section 1.2.1 quelle quantité d'information nous devrons maximiser dans le cas continu et quelle mesure non informative devra être utilisée.

1.2 Choix de l'a priori

1.2.1 Mesures de Jeffreys

Comme nous l'avons discuté dans l'exemple ci-dessus, choisir une loi a priori uniforme comme loi non informative n'est pas judicieux lorsque l'espace des paramètres est continu. Nous devons alors choisir une autre loi non informative. Pour cela, nous allons commencer par définir la quantité d'information "manquante" que l'on doit maximiser. Ensuite, nous montrerons que cette quantité d'information est indépendante du paramétrage.
Soient $\Theta \subset \mathbb{R}^k$ et $\Xi \subset \mathbb{R}^k$ deux ensembles de paramètres, ouverts de $\mathbb{R}^k$. Soit Y une variable aléatoire prennant ses valeurs dans un espace vectoriel de dimension finie $\mathcal{Y}$ muni d'une mesure de référence ν. On considère qu'il existe un $\mathcal{C}^1$-difféomorphisme g de Θ dans Ξ. Rappelons qu'un $\mathcal{C}^1$-difféomorphisme est une application de classe $\mathcal{C}^1$ bijective et dont

l'application réciproque est également de classe $\mathcal{C}^1$. Ainsi, si $\{p(y;\theta) : \theta \in \Theta\}$ est un modèle paramétrique de Y dans le paramétrage Θ, le modèle paramétrique dans le paramétrage Ξ est $\{q(y;\eta) : \eta \in \Xi\}$ où :

$$q(y;\eta) = p(y;\theta), \text{ avec } \eta = g(\theta).$$

Soit μ_Θ une mesure a priori sur Θ de densité f_Θ par rapport à la mesure de Lebesgue. La quantité d'information manquante sur le paramètre θ lorsque la loi a priori est μ_Θ est définie dans [15] pages 157-158, comme l'information de Kullback :

$$\bar{K}(\mu_\Theta(.|y), \mu_\Theta) = \int_\Theta \log\left(\frac{f_\Theta(\theta|y)}{f_\Theta(\theta)}\right) d\mu_\Theta(\theta|y). \tag{1.5}$$

Contrairement à l'entropie de Shannon, cette quantité d'information dépend de l'observation y. Elle s'interprète comme l'information manquante dans la loi a priori et disponible dans l'observation.
Ecrivons cette quantité d'information dans le paramétrage Ξ, μ_Ξ sera la mesure a priori correspondante à μ_Θ dans le paramétrage Ξ et f_Ξ sa densité. On a :

$$f_\Theta(\theta) = |\mathrm{Jac}_\theta(g)|\, f_\Xi(g(\theta)),$$

où $|\mathrm{Jac}_\theta(g)|$ est le déterminant jacobien de g.
On note $p(y) = \displaystyle\int_\Theta p(y;\theta) d\mu_\Theta(\theta)$ (resp. $q(y) = \displaystyle\int_\Theta q(y;\eta) d\mu_\Xi(\eta)$).
On a :

$$\begin{aligned}\bar{K}(\mu_\Xi(.|y), \mu_\Xi) &= \int_\Xi \log\left(\frac{f_\Xi(\eta|y)}{f_\Xi(\eta)}\right) d\mu_\Xi(\eta|y) \\ &= \int_\Xi \log\left(q(y;\eta)\right) \frac{q(y;\eta)}{q(y)} d\mu_\Xi(\eta) - \log\left(q(y)\right).\end{aligned}$$

Effectuant le changement de variable $\eta = g(\theta)$, on a :

$$q(y) = \int_\Theta q(y;g(\theta)) \underbrace{|\mathrm{Jac}_\theta(g)|\, f_\Xi(g(\theta)) d\lambda_{\mathbb{R}^k}(\theta)}_{d\mu_\Theta(\theta)} = p(y),$$

On a également :

$$\begin{aligned}&\int_\Xi \log\left(q(y;\eta)\right) \frac{q(y;\eta)}{q(y)} d\mu_\Xi(\eta) \\ =\ &\int_\Theta \log\left(q(y;g(\theta))\right) \frac{q(y;g(\theta))}{q(y)} |\mathrm{Jac}_\theta(g)|\, f_\Xi(g(\theta)) d\lambda_{\mathbb{R}^k}(\theta) \\ =\ &\int_\Theta \log\left(p(y;\theta)\right) \frac{p(y;\theta)}{p(y)} d\mu_\Theta(\theta).\end{aligned}$$

Ainsi :

$$\bar{K}(\mu_\Xi(.|y), \mu_\Xi) = \bar{K}(\mu_\Theta(.|y), \mu_\Theta). \tag{1.6}$$

Cette quantité ne dépend donc pas du paramètrage choisi.
Soit $y_{1:N} = (y_1, \ldots, y_N)$ un échantillon de réalisations indépendantes de $p(y;\theta)$ et $p(y_{1:N};\theta) = \prod_{n=1}^{N} p(y_n;\theta)$ sa vraisemblance. On définit la quantité d'information moyenne par :

$$\mathcal{J}_N(\mu_\Theta) = \int_{\Theta\times\mathcal{Y}^N} \bar{K}(\mu_\Theta(.|y_{1:N}), \mu_\Theta) p(y_{1:N};\theta) d\mu_\Theta(\theta) d\nu(y_{1:N}).$$

On va choisir comme mesure a priori une mesure maximisant cette quantité d'information. L'expression d'une telle mesure, appelée mesure de Jeffreys, est donnée par la proposition 1.2.1. Considérons pour cela la matrice d'information de Fisher $I_Y(\theta)$ de $p(y;\theta)$. Sous les conditions d'interversion "dérivation" et "intégrale", le coefficient (i,j) de $I_Y(\theta)$ vérifie :

$$\begin{aligned}(I_Y(\theta))_{i,j} &= \mathbb{E}_\theta\left(\frac{\partial}{\partial\theta_i}\log p(Y;\theta) \times \frac{\partial}{\partial\theta_j}\log p(Y;\theta)\right)\\ &= -\mathbb{E}_\theta\left(\frac{\partial^2}{\partial\theta_i\theta_j}\log p(Y;\theta)\right).\end{aligned}$$

Proposition 1.2.1 (Forme des a priori de Jeffreys)**.** *Soient Θ un ouvert non vide de $\mathbb{R}^k$, $\{p(y;\theta) : \theta \in \Theta\}$ une famille de densités de probabilité sur $\mathcal{Y}$ par rapport à une mesure ν, et $y_{1:N} = (y_1, \ldots, y_N)$ un échantillon de $p(y;\theta)$.*
Si les conditions suivantes sont vérifiées :

- *la loi a posteriori $\mu_\Theta(.|y_{1:N})$ converge en loi, lorsque N tend vers l'infini, vers la loi normale de $\mathbb{R}^k$, notée $\hat{\mu}_\Theta(.|y_{1:N})$, dont la moyenne est l'estimateur du maximum de vraisemblance $\hat{\theta} = \hat{\theta}(y_{1:N})$ et dont la matrice de covariance est $\frac{1}{N}\left[I_Y(\hat{\theta})\right]^{-1}$;*
- *pour tout compact K de Θ, on a :*

$$\lim_{N\to+\infty} \int_{K\times\mathcal{Y}^N} \bar{K}(\mu_\Theta(.|y_{1:N}), \hat{\mu}_\Theta(.|y_{1:N})) p(y_{1:N};\theta) d\mu_\Theta(\theta) d\nu(y_{1:N}) = 0.$$

Alors, il existe un entier $N \geq 1$ tel que pour tout $n \geq N$, la quantité

$$\mathcal{J}_n(\mu_\Theta) = \int_{\Theta\times\mathcal{Y}^n} \bar{K}(\mu_\Theta(.|y_{1:n}), \mu_\Theta) p(y_{1:n};\theta) d\mu_\Theta(\theta) d\nu(y_{1:n}),$$

est maximale pour la mesure μ_Θ de densité $\theta \to \sqrt{\det I_Y(\theta)}$ par rapport à la mesure de Lebesgue. Cette mesure est appelée mesure de Jeffreys.

Preuve.
Voir [47] pages 127-128. □

La mesure de Jeffreys maximise l'information manquante moyenne lorsque la taille de l'échantillon est suffisamment grande, on la considérera donc comme non informative.
Montrons maintenant qu'une mesure de Jeffreys est transformée en une autre mesure de Jeffreys par changement de paramétrage. Sachant que l'information de Fisher $I_Y(\theta)$ dépend du paramétrage de la loi de Y, on notera cette matrice $I_{Y,\Theta}$ (resp. $I_{Y,\Xi}$) lorsque la loi est paramétrée par Θ (resp. Ξ). Si η suit la loi de Jeffreys de densité $\eta \to \sqrt{\det I_{Y,\Xi}(\eta)}$ par rapport

à la mesure de Lebesgue, alors en utilisant le théorème de changement de variable, $\theta = g^{-1}(\eta)$ suit la loi de densité $\theta \to \sqrt{\det I_{Y,\Xi}(g(\theta))}\,|\mathrm{Jac}_\theta(g)|$. De plus :

$$\begin{pmatrix} \dfrac{\partial \log p(y;\theta)}{\partial \theta_1} \\ \vdots \\ \dfrac{\partial \log p(y;\theta)}{\partial \theta_k} \end{pmatrix} = (\mathrm{Jac}_\theta(g))^T \begin{pmatrix} \dfrac{\partial \log q(y;\eta := g(\theta))}{\partial \eta_1} \\ \vdots \\ \dfrac{\partial \log q(y;\eta := g(\theta))}{\partial \eta_k} \end{pmatrix},$$

où $\mathrm{Jac}_\theta(g)$ est la matrice jacobienne de g. Ainsi :

$$I_{Y,\Theta}(\theta) = (\mathrm{Jac}_\theta(g))^T\, I_{Y,\Xi}(g(\theta))\mathrm{Jac}_\theta(g).$$

On en déduit que la loi de θ est la loi de Jeffreys de densité $\theta \to \sqrt{\det I_{Y,\Theta}(\theta)}$ par rapport à la mesure de Lebesgue.
Nous verrons au chapitre 6 une autre façon d'introduire les mesures de Jeffreys ainsi que leur relation avec les lois uniformes.

1.2.2 Lois a priori conjuguées

Définition 1.2.1 (Lois conjuguées)**.** *Soit $\{p(y;\theta) : \theta \in \Theta\}$ un modèle paramétrique. Une loi a priori μ appartenant à un modèle paramétrique est "conjuguée" à ce modèle si la loi a posteriori $\mu(\theta|y) \propto \mu(\theta)p(y;\theta)$ appartient au même modèle paramétrique que la loi a priori.*

Le paramètre de la loi a priori est couramment appelé hyperparamètre. Lorsqu'on utilise les lois conjuguées, la règle de Bayes revient à remettre à jour l'hyperparamètre.
Le tableau 1.1 donne quelques exemples de lois conjuguées pour des familles de lois usuelles, IG désigne l'inverse d'une loi gamma. Nous donnons dans ce tableau les lois a priori conjugées et les lois a posteriori $\mu(\theta|y_{1:N}) \propto \mu(\theta)\prod_{n=1}^{N} p(y_n;\theta)$ correspondantes, où $y_{1:N} = (y_1, \ldots, y_N)$ est un échantillon de réalisations indépendantes de $p(y;\theta)$.

Famille paramétrique	Loi a priori conjuguée	Loi a posteriori
Loi normale $\mathcal{N}_{\mathbb{R}}(m,s)$, paramètre m	$\mu(m) \sim \mathcal{N}_{\mathbb{R}}(\mu,\gamma)$	$\mu(m\|y_{1:N}) \sim \mathcal{N}_{\mathbb{R}}\left(\dfrac{\mu s+\gamma\sum_{n=1}^{N} y_n}{s+N\gamma}, \dfrac{\gamma s}{s+N\gamma}\right)$
Loi normale $\mathcal{N}_{\mathbb{R}}(m,s)$, paramètre s	$\mu(s) \sim IG(a,b)$	$\mu(s\|y_{1:N}) \sim IG\left(a+\dfrac{N}{2}, \dfrac{2b}{2+b\sum_{n=1}^{N}(y_n-m)^2}\right)$
Loi exponentielle $\mathcal{E}(\lambda)$	$\mu(\lambda) \sim \Gamma(a,b)$	$\mu(\lambda\|y_{1:N}) \sim \Gamma\left(a+N, \dfrac{b}{1+b\sum_{n=1}^{N} y_n}\right)$
Loi binômiale de paramètre $q \in [0,1]$	$\mu(q) \sim \beta(a,b)$	$\mu(q\|y_{1:N}) \sim \beta\left(a+\sum_{n=1}^{N} y_n, b+NK-\sum_{n=1}^{N} y_n\right)$

TAB. 1.1 – Familles conjuguées de familles paramétriques et paramètre de la loi a posteriori fonction de celui de la loi a priori.

1.2.3 Modèles à données latentes

Considérons un couple de processus $Z = (X_s, Y_s)_{s\in\mathcal{S}}$ où Y est observable et X ne l'est pas, chaque X_s prend ses valeurs dans un ensemble fini $\mathcal{X} = \{\omega_1, \ldots, \omega_K\}$ et chaque Y_s prend ses valeurs dans un $\mathbb{R}$-espace vectoriel de dimension finie. La réalisation y de Y représente donc le signal observé et nous souhaitons estimer la réalisation x de X. Les valeurs prises par X_s sont appelées "classes" ou "étiquettes" et la réalisation x est parfois appelée segmentation du signal y. $\mathcal{S}$ est appelé l'ensemble des sites, il peut être un sous-ensemble de $\mathbb{N}$ comme dans le cas des chaînes, un ensemble muni d'une structure d'arbre dans le cas des arbres ou un sous-ensemble de $\mathbb{Z}^p$ dans le cas des champs, il représente ainsi la structure "topologique" du signal. Les deux processus sont liés par une densité de probabilité du type $p(x,y;\theta)$, où θ est le paramètre du modèle. La loi a posteriori $p(x|y;\theta)$ représente la connaissance que l'on a sur x à partir de l'observation y. Les modèles statistiques $p(x,y;\theta)$ seront choisis de façon à ce que la loi a posteriori $p(x|y;\theta)$ soit calculable dans un temps raisonnable pour des processus de "grande taille". Par exemple, pour une image de taille 256×256 que l'on souhaite segmenter en 2 classes, il existe $2^{256\times256} \approx 2\times10^{19728}$ valeurs pour $p(x|y;\theta)$. Dans le cas général, le calcul de $p(x|y;\theta)$ par la formule de Bayes est de complexité algorithmique trop élevée. Dans le cas où Z est choisi comme un vecteur à composantes indépendantes, $p(x|y;\theta) = \prod_{s\in\mathcal{S}} p(x_s|y_s;\theta)$ et chaque $p(x_s|y_s;\theta)$ ne prend que K valeurs, le calcul de $p(x|y;\theta)$ se fait très simplement. Cependant ce modèle ne permet pas de prendre en compte les éventuelles dépendances au sein des états cachés et des observations. Ainsi le modèle devra être choisi suffisamment simple pour permettre le calcul rapide de la loi a posteriori, et suffisamment riche pour modéliser les situations de dépendance les plus réalistes possible.

Remarque : Dans la démarche bayésienne classique, il est courant de se donner la loi a priori $p(x;\theta)$ qui représente les dépendances au sein du processus caché et la loi d'attache aux données $p(y|x;\theta)$, ainsi $p(x,y;\theta) = p(x;\theta)p(y|x;\theta)$. Cependant, comme nous le verrons, se donner directement la loi jointe $p(x,y;\theta)$ permet de modéliser des situations de dépendance

plus complexes.
Donnons quelques exemples de modèles à données latentes, on omettra le paramètre θ. Dans les modèles présentés, on prendra $\mathcal{S} = \{1, \ldots, N\}$ et on notera $z_{1:N}$ la réalisation de Z.

Mélange indépendant

Soit $\mathcal{S} = \{1, \ldots, N\}$ et considérons les variables $Z_n = (X_n, Y_n)$ indépendantes. Nous avons :

$$p(z_{1:N}) = \prod_{n=1}^{N} p(x_n, y_n) = \prod_{n=1}^{N} p(x_n)p(y_n|x_n).$$

Chaînes de Markov cachées à bruit indépendant

La loi jointe $p(z_{1:N})$ est donnée par :

$$p(z_{1:N}) = p(x_1)p(y_1|x_1) \prod_{n=1}^{N-1} p(x_{n+1}|x_n)p(y_{n+1}|x_{n+1}),$$

ainsi :

- $p(x_{1:N}) = p(x_1) \prod_{n=1}^{N-1} p(x_{n+1}|x_n)$, soit X est une chaîne de Markov ;
- $p(y_{1:N}|x_{1:N}) = \prod_{n=1}^{N} p(y_n|x_n)$.

Chaînes de Markov couples

La loi jointe $p(z_{1:N})$ est donnée par :

$$p(z_{1:N}) = p(x_1, y_1) \prod_{n=1}^{N} p(x_{n+1}, y_{n+1}|x_n, y_n).$$

Dans ce modèle, X n'est plus obligatoirement une chaîne de Markov, les variables aléatoires Y_n ne sont plus obligatoirement indépendantes conditionnellement à X et l'égalité $p(y_n|x_{1:N}) = p(y_n|x_n)$ n'a plus obligatoirement lieu. Nous reviendrons sur ce modèle au chapitre 3.

Chaînes de Markov triplets

Dans les chaînes de Markov triplets, nous introduisons un troisième processus U, dit processus auxiliaire, tel que le triplet (X, U, Y) soit une chaîne de Markov. Ce modèle généralise celui des chaînes de Markov couples ; une chaîne de Markov couple est une chaîne de Markov triplet telle que $U = X$. De plus, si on note $V = (X, U)$, le processus (V, Y) est une chaîne de Markov couple ; ainsi les algorithmes d'inférence bayésienne étudiés au chapitre 2 dans le cas des chaînes couples restent utilisables dans les chaînes de Markov triplets.
Comme nous le verrons dans les chapitres suivants, U peut avoir différentes interprétations. Il peut modéliser la non stationnarité de X [65, 68], ou la semi-markovianité, auquel cas U_n est le temps de séjour restant pour X dans la valeur x_n [68]. D'autres interprétations de U sont présentées dans [1, 2, 22, 24, 65]. Ce modèle est particulièrement riche car aucune des chaînes X, U, Y, (X, U), (X, Y) ou (U, Y) n'est nécessairement une chaîne de Markov.

Conclusion

Nous avons présenté dans ce chapitre les notions classiques de l'inférence bayésienne. L'approche classique de l'inférence bayésienne consiste à se donner une loi a priori sur un paramètre ou une réalisation cachée que l'on veut estimer. Cette loi a priori représente la connaissance avant toute expérience sur le paramètre. L'estimateur de ce paramètre inconnu est ensuite choisi selon un critère d'optimalité en se donnant une fonction de perte. Le choix du critère d'optimalité dépend de la nature du problème considéré, ce qui confère aux méthodes bayésiennes une grande souplesse. Nous avons introduit, dans une deuxième section, le choix de la loi a priori. La loi a priori peut être choisie conformément à la connaissance que l'on a sur le paramètre ou la réalisation cachée. Ainsi, si on ne dispose d'aucune connaissance, on est amené à considérer des mesures a priori non informatives telles que les mesures de Jeffreys. La forme de la loi a priori peut être également choisie de façon subjective, de façon à pouvoir calculer facilement la loi a posteriori. C'est le cas notamment des lois a priori conjuguées, mais c'est aussi le cas de certaines lois $p(x)$ dans les modèles à données latentes. Cependant, comme nous l'avons vu aux travers des exemples des chaînes de Markov couples et triplets, se donner la loi jointe $p(x, y)$ au lieu de se donner la loi a priori $p(x)$ et la loi conditionnelle $p(y|x)$ permet de considérer des situations de dépendance plus complexes. Cette loi jointe devra être choisie suffisamment simple de façon à pouvoir calculer facilement la loi a posteriori $p(x|y)$ grâce aux algorithmes que nous verrons au chapitre suivant. Elle devra également être choisie suffisamment riche de façon à modéliser une gamme variée de comportement. Nous verrons au cours de cette thèse différents modèles à données latentes, dont certains originaux, pour lesquels la loi a posteriori est calculable, ce qui rend possible l'utilisation des méthodes bayésiennes de recherche du processus caché x.

Chapitre 2

Inférence bayésienne dans les modèles de Markov cachés

Dans ce chapitre, nous détaillons l'estimation des paramètres et des états cachés dans certains modèles à données latentes classiques. Dans toute la suite, on considère un modèle paramétrique $\{z \to p(z;\theta) : \theta \in \Theta\}$ pour le couple de processus $Z = (X, Y) = (X_u, Y_u)_{u\in\mathcal{S}}$, où $\mathcal{S}$ est un ensemble fini de sites. Dans un modèle à données latentes, nous observons y une réalisation de Y, et nous devons estimer x la réalisation cachée du processus X. Le modèle paramétrique représente la relation probabiliste entre la réalisation cachée et l'observation. La réalisation x de X sera estimée à partir de la loi a posteriori $p(x|y;\theta)$ selon un critère d'optimalité déterminé par une fonction de perte L. Dans la suite, chaque X_u prendra ses valeurs dans un ensemble fini $\mathcal{X} = \{\omega_1, \ldots, \omega_K\}$ et chaque Y_u dans un $\mathbb{R}$-espace vectoriel $\mathcal{Y}$ de dimension finie. Nous nous limiterons aux cas où $z \to p(z;\theta)$ est une densité par rapport à la mesure produit $(\nu \otimes \lambda_{\mathcal{Y}})^{\otimes|\mathcal{S}|}$, où ν est la mesure de décompte sur $\mathcal{X}$ et $\lambda_{\mathcal{Y}}$ est la mesure de Lebesgue sur $\mathcal{Y}$. Les fonctions de perte que nous utiliserons sont :

1. $L(x, \hat{x}) = \sum_{u\in\mathcal{S}} I(x_u \neq \hat{x}_u)$;
2. $L(x, \hat{x}) = I(x \neq \hat{x})$,

où $I(A) = 1$ si A est vraie et 0 sinon.
Pour la première fonction de perte, l'estimateur obtenu est celui du "Maximum des Marginales a Posteriori" (MPM). Il est défini par $(\hat{x}_{MPM})_u = \arg\max_{x_u} p(x_u|y;\theta)$ pour tout $u \in \mathcal{S}$. Pour la seconde fonction de perte, l'estimateur bayésien est celui du "Maximum A Posteriori" (MAP), défini par $\hat{x}_{MAP} = \arg\max_x p(x|y)$. Dans le cas général, le calcul des estimateurs du MPM et du MAP nécessitent de connaître toutes les probabilités a posteriori $p(x|y;\theta)$, soit $K^{|\mathcal{S}|}$ valeurs. Comme nous le verrons, lorsque le modèle est suffisamment simple, il existe des algorithmes permettant le calcul du MPM et du MAP même pour des ensembles d'indices très riches, pouvant dépasser un million d'éléments. Ces algorithmes, appelés algorithmes d'inférence bayésienne, s'appuient sur la factorisation de la loi $p(z;\theta)$. Nous introduirons la relation entre factorisation et dépendances au travers des modèles graphiques. Dans un second temps, nous aborderons les algorithmes d'inférence bayésienne dans les cas plus spécifiques des chaînes et des arbres de Markov couples [39, 67, 68, 81, 96, 98]. Pour finir, nous présenterons dans cette section deux algorithmes d'estimation du paramètre θ : l'algorithme "Expectation Maximisation" (EM) et l'algorithme "Iterative Conditional Estimation" (ICE). Le premier

est parmi les méthodes les plus utilisées dans les modèles de Markov cachés. Le deuxième, que nous retiendrons dans la suite de notre thèse, se prête mieux aux divers modèles généralisant les chaînes de Markov cachées proposés dans notre manuscrit.

2.1 Algorithmes d'inférence bayésienne et modèles graphiques de dépendance

Les modèles graphiques de dépendance abordés dans cette section sont détaillés dans [59, 60, 70].

2.1.1 Graphes de dépendance non orientés et markovianité

Nous appelons graphe un couple $G = (\mathcal{S}, \mathcal{E})$, où $\mathcal{S}$ est un ensemble fini et $\mathcal{E}$ est un sous-ensemble de $\mathcal{S} \times \mathcal{S}$. L'ensemble $\mathcal{S}$ sera appelé ensemble des sommets ou sites du graphe et $\mathcal{E}$ l'ensemble des arêtes. Le graphe sera qualifié de non orienté lorsque pour $(u, v) \in \mathcal{E}$, le couple (v, u) appartient aussi à $\mathcal{E}$, il sera qualifié d'orienté dans le cas contraire. Les graphes considérés par la suite seront non orientés. Dans un graphe non orienté, il existe une arête entre u et v si $(u, v) \in \mathcal{E}$, on dira alors que v (resp. u) est voisin de u (resp. v). L'ensemble des voisins de u sera noté $\mathcal{V}_u$ et appelé voisinage de u. On dira qu'il existe un chemin entre u et v s'il existe des sites $u_1, \ldots, u_n$ et des arêtes entre u et u_1, u_1 et u_2, ..., u_n et v. On dira alors que le chemin de u à v passe par u_k pour $k \in \{1, \ldots, n\}$. Deux ensembles de sites a et b sont séparés par un ensemble de sites c si tout chemin d'un site de a vers un site de b passe par au moins un site de c. Un sous-graphe c de G est une clique si et seulement si :

- ou bien il existe un sommet u tel que $c = (\{u\}, \{(u, u)\})$, c est alors appelé "singleton" et sera noté par abus $c = \{u\}$;
- ou bien deux sommets de c sont deux sommets mutuellement voisins dans G.

Markovianité et graphes de dépendance

Soit $G = (\mathcal{S}, \mathcal{E})$ un graphe non orienté et soit $Z = (Z_u)_{u \in \mathcal{S}}$ un ensemble indexé par $\mathcal{S}$ de variables aléatoires, qui sera appelé "champ aléatoire", à valeurs dans $\mathcal{Z} = \prod_{u \in \mathcal{S}} \mathcal{Z}_u$ muni d'une mesure de référence $\nu_{\mathcal{Z}} = \bigotimes_{u \in \mathcal{S}} \nu_{\mathcal{Z}_u}$. On distingue trois types de dépendance pouvant se déduire de la lecture d'un graphe :

- la markovianité par paires ;
- la markovianité globale ;
- la markovianité locale.

Définition 2.1.1. *Soit $G = (\mathcal{S}, \mathcal{E})$ un graphe et $Z = (Z_u)_{u \in \mathcal{S}}$ un champ aléatoire. Z satisfait vis-à-vis du graphe G :*

- *la propriété de markovianité par paires si pour tous sites u et v, s'il n'existe pas d'arête entre u et v, alors Z_u et Z_v sont indépendantes conditionnellement à l'ensemble de variables aléatoires $\{Z_t : t \notin \{u, v\}\}$;*
- *la propriété de markovianité globale si pour trois sous-ensembles a, b et c non vides et disjoints de $\mathcal{S}$, si c sépare a et b, alors les ensembles de variables aléatoires $Z_a = (Z_t)_{t \in a}$*

et $Z_b = (Z_t)_{t\in b}$ sont indépendants conditionnellement à l'ensemble de variables aléatoires $Z_c = (Z_t)_{t\in c}$;
- *la propriété de markovianité locale si pour tout u de voisinage $\mathcal{V}_u$, les ensembles de variables aléatoires $\{Z_u\}$ et $\{Z_v : v \neq u, v \notin \mathcal{V}_u\}$ sont indépendants conditionnellement à l'ensemble de variables aléatoires $\{Z_t : t \in \mathcal{V}_u\}$.*

Lorsque le champ aléatoire Z satisfait la propriété de markovianité par paires vis-à-vis du graphe G, G est appelé graphe de dépendance de Z.

Remarque : Il est important de noter que, sous l'hypothèse de markovianité globale, Z_a et Z_b peuvent être indépendants conditionnellement à Z_c sans que c ne sépare les ensembles a et b.

Proposition 2.1.1 (Equivalence des markovianités)**.** *Soit $Z = (Z_u)_{u\in\mathcal{S}}$ un champ aléatoire indexé sur l'ensemble fini de sites $\mathcal{S}$ d'un graphe G et à valeurs dans un ensemble mesurable $\mathcal{Z} = \prod_{u\in\mathcal{S}} \mathcal{Z}_u$ muni d'une mesure de référence $\nu_{\mathcal{Z}} = \bigotimes_{u\in\mathcal{S}} \nu_{\mathcal{Z}_u}$.*
- *Si Z satisfait la propriété de markovianité globale vis-à-vis de G, alors il satisfait la propriété de markovianité locale vis-à-vis de G ;*
- *si Z satisfait la propriété de markovianité locale vis-à-vis de G, alors il satisfait la propriété de markovianité par paires vis-à-vis de G.*

Soit p la densité de Z par rapport à $\nu_{\mathcal{Z}}$. Si $p(z)$ est strictement positif pour tout $z \in \mathcal{Z}$, alors les trois markovianités sont équivalentes.

Preuve.
Voir [70] pages 32-33. □

Nous donnons ci-après les deux exemples les plus couramment utilisés de modèles markoviens. Lorsque l'ensemble $\mathcal{S}$ est un sous-ensemble de $\mathbb{N}$, les champs aléatoires seront qualifiés de "chaîne" ou "processus".

Exemple 2.1.1 (Chaîne de Markov)**.** Un processus $Z = (Z_n)_{1\leq n\leq N}$ est une chaîne de Markov si pour tout $n > 1$, les ensembles $\{Z_k : k < n\}$ et $\{Z_k : k > n\}$ sont indépendants conditionnellement à Z_n.

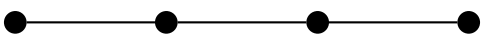


FIG. 2.1 – Graphe de dépendance d'une chaîne de Markov.

Considérons le processus marginal $Z_{1:N} = (Z_n)_{1\leq n\leq N}$. Comme $(Z_1, \ldots, Z_{N-2})$ et Z_N sont indépendants conditionnellement à Z_{N-1}, alors :

$$p(z_{1:N}) = p(z_{1:N-1}) \times p(z_N|z_1, \ldots, z_{N-1}) = p(z_{1:N-1}) \times p(z_N|z_{N-1}).$$

En réitérant le raisonnement pour le processus marginal $Z_{1:N-1}$ puis pour $Z_{1:N-2}$ et ainsi de suite, on en déduit :

$$p(z_{1:N}) = p(z_1) \prod_{n=1}^{N} p(z_{n+1}|z_n) \text{ pour tout } N \geq 1.$$

Exemple 2.1.2 (Champ de Markov sur $\mathbb{Z}^2$)**.** Soit $G = (\mathcal{S}, \mathcal{E})$ un graphe tel que $\mathcal{S} = \{1, \ldots, M\} \times \{1, \ldots, N\}$. Un champ aléatoire Z est un champ de Markov s'il satisfait la propriété de markovianité locale vis-à-vis de G.
Si les voisins d'un site (m, n) sont les sites $(m+1, n)$, $(m-1, n)$, $(m, n+1)$ et $(m, n-1)$, on obtient le graphe représenté à la figure 2.2.

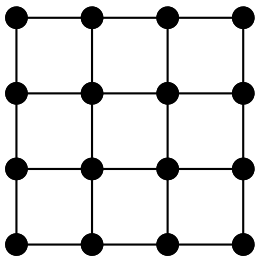


FIG. 2.2 – Graphe de dépendance d'un champ de Markov par rapport aux quatres plus proches voisins.

2.1.2 Factorisation d'une loi selon un graphe

Nous précisons dans cette sous-section les liens entre la markovianité globale d'un champ aléatoire et la factorisation de sa loi.

Définition 2.1.2 (Factorisation d'une distribution)**.** *Soit $Z = (Z_u)_{u \in \mathcal{S}}$ un champ aléatoire indexé sur l'ensemble fini des sites $\mathcal{S}$ d'un graphe G et à valeurs dans $\mathcal{Z} = \prod_{u \in \mathcal{S}} \mathcal{Z}_u$ muni de la mesure de référence $\nu_{\mathcal{Z}} = \bigotimes_{u \in \mathcal{S}} \nu_{\mathcal{Z}_u}$. La densité p de la loi de $Z = (Z_u)_{u \in \mathcal{S}}$ par rapport à la mesure $\nu_{\mathcal{Z}}$ se factorise selon le graphe $G = (\mathcal{S}, \mathcal{E})$ si elle s'écrit :*

$$p(z) = \prod_{c \in \mathcal{C}} f_c(z_c),$$

où $\mathcal{C}$ est l'ensemble des cliques du graphe et pour toute clique c, f_c est une fonction de $\prod_{u \text{ site de } c} \mathcal{Z}_u$ dans $\mathbb{R}^+$.

Proposition 2.1.2. *Si la distribution de Z se factorise selon le graphe G, alors Z satisfait la propriété de markovianité globale vis-à-vis de ce graphe.*

Preuve.
Voir [70]. □

La réciproque de cette proposition est donnée par le théorème d'Hammersley-Clifford. Rappelons tout d'abord la définition d'un champ de Gibbs.

Définition 2.1.3 (Champ de Gibbs)**.** *Soit $G = (\mathcal{S}, \mathcal{E})$ un graphe et soit $\mathcal{C}$ l'ensemble de ses cliques.*
Un champ aléatoire $Z = (Z_u)_{u\in\mathcal{S}}$ est un champ de Gibbs vis-à-vis du graphe $G = (\mathcal{S}, \mathcal{E})$ s'il existe une famille d'applications à valeurs réelles appelée "potentiel de Gibbs" $\{\phi_c : c \in \mathcal{C}\}$ telle que la densité de sa loi par rapport à une mesure de référence s'écrit :

$$p(z) \propto \exp\left(-\sum_{c\in\mathcal{C}} \phi_c(z_c)\right).$$

Si Z est un champ de Gibbs, alors il se factorise selon le graphe considéré.

Théorème 2.1.1 (Théorème d'Hammersley-Clifford)**.** *Soit $Z = (Z_u)_{u\in\mathcal{S}}$ un champ aléatoire indexé sur l'ensemble de sites $\mathcal{S}$ d'un graphe G et à valeurs dans un ensemble mesurable $\mathcal{Z} = \prod_{u\in\mathcal{S}} \mathcal{Z}_u$ muni d'une mesure de référence $\nu_{\mathcal{Z}} = \bigotimes_{u\in\mathcal{S}} \nu_{\mathcal{Z}_u}$. Soit p la densité de Z par rapport à $\nu_{\mathcal{Z}}$.*
Si Z est un champ de Gibbs vis-à-vis de G, alors il vérifie la propriété de markovianité globale vis-à-vis du graphe G.
Si $p(z)$ est strictement positif pour tout $z \in \mathcal{Z}$ et si Z satisfait la propriété de markovianité locale, alors Z est un champ de Gibbs pour le graphe G.

Preuve. Pour la preuve, voir [54]. □

Ainsi, sous la condition de positivité de la loi de Z, nous avons l'équivalence entre les quatres propriétés :
- la loi de Z se factorise selon le graphe G ;
- la loi de Z satisfait la propriété de markovianité globale ;
- la loi de Z satisfait la propriété de markovianité locale ;
- la loi de Z satisfait la propriété de markovianité par paires.

2.2 Algorithmes d'inférence bayésienne dans les modèles de Markov couples

Nous présentons les algorithmes d'inférence bayésienne dans le cas des chaînes et des arbres de Markov couples. Un champ aléatoire $Z = (X_u, Y_u)_{u\in\mathcal{S}}$ est une chaîne de Markov couple si $\mathcal{S} = \{1, \ldots, N\}$ et si le processus Z est une chaîne de Markov. On notera alors $x_{1:N}$ et $y_{1:N}$ les réalisations respectives de X et de Y. Les algorithmes d'inférence étudiés dans cette section utilisent la factorisation de la loi selon son graphe de dépendance. Il permettent ainsi, comme

nous allons le voir, le calcul rapide des estimateurs du MPM et du MAP. Le modèle de Markov couple généralise le modèle classique de Markov cachées à bruit indépendant [81, 96, 99, 104]. Il permet de modéliser certaines situations ne pouvant pas être prises en compte par ce dernier. Par ailleurs, l'égalité $p(y_n|x_{1:N}) = p(y_n|x_n)$ n'a pas toujours lieu dans les chaînes couples. Les chaînes de Markov cachées à bruit indépendant étant des chaînes couples particulières, les algorithmes d'inférence bayésienne peuvent être encore utilisés dans les chaînes de Markov cachés à bruit indépendant.

2.2.1 Algorithme de Baum-Welsh

Soit $Z = (X_n, Y_n)_{n\in\{1,\dots,N\}}$ une chaîne de Markov telle que chaque X_n prend ses valeurs dans $\mathcal{X} = \{\omega_1, \dots, \omega_K\}$ et chaque marginale Y_n prend ses valeurs dans un espace vectoriel de dimension finie $\mathcal{Y}$. L'algorithme de Baum-Welsh permet de calculer les lois a posteriori $p(x_n|y_{1:N})$ et $p(x_n, x_{n+1}|y_{1:N})$ à partir d'un échantillon observé $y_{1:N}$ de Y. Nous allons voir dans cette sous-section deux algorithmes de Baum-Welsh. La première version est la plus classique ; elle est toutefois sujette à des problèmes numériques. La deuxième version, dite "conditionnelle", permet d'éviter ces problèmes.

Algorithme de Baum-Welsh classique

L'algorithme de Baum-Welsh est constitué de deux étapes. La première, appelée "étape directe", consiste à calculer les sommations successives $\sum_{x_k} p(x_{k:N}, y_{1:N})$ pour k de 1 à n et dans la seconde, appelée "étape rétrograde", on calcule les sommations successives $\sum_{x_k} p(x_{n:k}, y_{1:N})$ pour k de N à $n+1$. On obtient ainsi les quantités $p(x_n, x_{n+1}, y_{1:N})$ et $p(x_n, y_{1:N})$. Comme Z est une chaîne de Markov, les étapes directes et rétrogrades s'écrivent facilement.

1. Etape directe :
   - initialisation : $\alpha_1(x_1) = p(x_1, y_1)$;
   - itération pour n de 1 à $N-1$:

$$\alpha_{n+1}(x_{n+1}) = \sum_{x_n} \alpha_n(x_n) p(x_{n+1}, y_{n+1}|x_n, y_n). \tag{2.1}$$

2. Etape rétrograde :
   - initialisation : $\beta_N(x_N) = 1$;
   - itération pour n de N à 2 :

$$\beta_{n-1}(x_{n-1}) = \sum_{x_n} \beta_n(x_n) p(x_n, y_n|x_{n-1}, y_{n-1}). \tag{2.2}$$

Finalement, on a :
- $p(x_n|y_{1:N}) \propto p(x_n, y_{1:N}) = \alpha_n(x_n)\beta_n(x_n)$;
- $p(x_n, x_{n+1}|y_{1:N}) \propto p(x_n, x_{n+1}, y_{1:N}) = \alpha_n(x_n)\beta_{n+1}(x_{n+1})p(x_{n+1}, y_{n+1}|x_n, y_n)$;
- $p(x_{n+1}|x_n, y_{1:N}) = \dfrac{\beta_{n+1}(x_{n+1})}{\beta_n(x_n)} p(x_{n+1}, y_{n+1}|x_n, y_n)$.

Algorithme de Baum-Welsh conditionnel

L'algorithme de Baum-Welsh conditionnel a été proposé par P. Devijver dans le cas des chaînes de Markov cachées classiques dans [40] et généralisé aux chaînes couples par S. Derrode dans [39]. Comme $\alpha_n(x_n) = p(x_n, y_{1:n})$ et $\beta_n(x_n) = p(y_{n+1:N}|x_n, y_n)$; ainsi, si le processus Y est à valeurs réelles, et si n est suffisamment grand, $\alpha_n(x_n)$ devient très petit. De même si n est suffisamment petit par rapport à N, $\beta_n(x_n)$ devient très petit. Ainsi, lorsque l'on programme l'algorithme de Baum-Welsh de cette façon, on rencontre des problèmes numériques car l'ordinateur considère comme nulle les valeurs trop petites. Afin de remédier à ce problème, nous modifions l'algorithme de Baum-Welsh en divisant les quantités $\alpha_n(x_n)$ et $\beta_n(x_n)$ par des quantités du même ordre de grandeur.
On pose :

$$\tilde{\alpha}_n(x_n) = \frac{\alpha_n(x_n)}{p(y_{1:n})} = p(x_n|y_{1:n}),$$

X_n étant à valeurs finies, cette quantité ne pose aucun problème numérique. On a alors $p(x_n, y_{1:N}) = p(y_{1:n})\tilde{\alpha}_n(x_n)\beta_n(x_n)$, ainsi $p(x_n|y_{1:N}) = \dfrac{\tilde{\alpha}_n(x_n)\beta_n(x_n)}{p(y_{n+1:N}|y_{1:n})}$. Comme $p(x_n|y_{1:N})$ ne pose aucun problème numérique, on pose :

$$\tilde{\beta}_n(x_n) = \frac{\beta_n(x_n)}{p(y_{n+1:N}|y_{1:n})} = \frac{p(y_{n+1:N}|x_n, y_n)}{p(y_{n+1:N}|y_{1:n})}.$$

L'algorithme de Baum-Welsh modifié s'écrit :

1. Etape directe :
   - initialisation : $\tilde{\alpha}_1(x_1) = p(x_1|y_1)$;
   - itération :

$$\tilde{\alpha}_{n+1}(x_{n+1}) = \frac{1}{p(y_{n+1}|y_{1:n})} \sum_{x_n} \tilde{\alpha}_n(x_n) p(x_{n+1}, y_{n+1}|x_n, y_n), \tag{2.3}$$

   et

$$p(y_{n+1}|y_{1:n}) = \sum_{x_{n+1}} \sum_{x_n} \tilde{\alpha}_n(x_n) p(x_{n+1}, y_{n+1}|x_n, y_n) \ ;$$

2. Etape rétrograde :
   - initialisation : $\tilde{\beta}_N(x_N) = 1$;
   - itération :

$$\tilde{\beta}_n(x_n) = \frac{\displaystyle\sum_{x_{n+1}} \tilde{\beta}_{n+1}(x_{n+1}) p(x_{n+1}, y_{n+1}|x_n, y_n)}{\displaystyle\sum_{x_{n+1}} \sum_{x_n} \tilde{\alpha}_n(x_n) p(x_{n+1}, y_{n+1}|x_n, y_n)} \ ; \tag{2.4}$$

On a ainsi :
- $p(x_n|y_{1:N}) = \tilde{\alpha}_n(x_n)\tilde{\beta}_n(x_n)$;
- $p(x_n, x_{n+1}|y_{1:N}) \propto \tilde{\alpha}_n(x_n)\tilde{\beta}_{n+1}(x_{n+1}) p(x_{n+1}, y_{n+1}|x_n, y_n)$;
- $p(x_{n+1}|x_n, y_{1:N}) \propto \dfrac{\tilde{\beta}_{n+1}(x_{n+1})}{\tilde{\beta}_n(x_n)} p(x_{n+1}, y_{n+1}|x_n, y_n)$.

2.2.2 Algorithme de Viterbi

L'algorithme de Viterbi permet de calculer l'estimateur du MAP de manière itérative et rapide. Soit $Z = (X_n, Y_n)_{n\in\{1,\dots,N\}}$ une chaîne de Markov couple.

– Initialisation :

$$\hat{x}_N(x_{N-1}) = \arg\max_{x_N} p(x_N, y_N|x_{N-1}, y_{N-1}) \ ;$$

et

$$\psi_{N-1}(x_{N-1}) = \max_{x_N} p(x_N, y_N|x_{N-1}, y_{N-1}) \ ;$$

– Itération (étape rétrograde) pour n de $N-1$ à 1 :

$$\hat{x}_n(x_{n-1}) = \arg\max_{x_n} [p(x_n, y_n|x_{n-1}, y_{n-1})\psi_n(x_n)] \ ;$$

et

$$\psi_{n-1}(x_{n-1}) = \max_{x_n} [p(x_n, y_n|x_{n-1}, y_{n-1})\psi_n(x_n)] \ ;$$

– Etape directe :

$$(\hat{x}_{MAP})_1 = \arg\max_{x_1} p(x_1, y_1)\psi_1(x_1) \text{ et } (\hat{x}_{MAP})_{n+1} = \hat{x}_{n+1}((\hat{x}_{MAP})_n).$$

On obtient ainsi, après les étapes "rétrograde" et "directe", $\hat{x}_{MAP}$ maximisant $p(x_{1:N}|y_{1:N})$. Dans [39], il est également proposé un algorithme de Viterbi conditionnel afin d'éviter les problèmes numériques.

2.2.3 Algorithme de Baum-Welsh adapté aux arbres de Markov

Soit $\mathcal{S}$ un ensemble fini de sites. Soit $\mathcal{S}_1, \dots, \mathcal{S}_P$ une partition de $\mathcal{S}$ telle que $\mathcal{S}_1 = \{1\}$, $|\mathcal{S}_1| \leq |\mathcal{S}_2| \leq \dots |\mathcal{S}_P|$. A chaque $u \in \mathcal{S}_k$ pour $k \neq 1$, il existe un unique élément noté $\rho(u)$ dans $\mathcal{S}_{k-1}$. Cet élément est appelé parent de u. Les éléments v tels que u soit parent de v sont appelés fils de u, on note $c(u)$ leur ensemble. Les voisins d'un site u sont le site parent $\rho(u)$ et l'ensemble des fils $c(u)$. L'ensemble $\mathcal{E}$ des arêtes est l'ensemble des couples (u, v) tels que $u \in c(v)$ ou $v \in c(u)$. Un champ aléatoire $Z = (Z_u)_{u\in\mathcal{S}}$ est un arbre de Markov si son graphe de dépendance est $(\mathcal{S}, \mathcal{E})$; sa distribution s'écrit alors :

$$p(z) = p(z_1) \prod_{u\neq 1} p(z_u|z_{\rho(u)}).$$

Considérons $Z = (X_u, Y_u)_{u\in\mathcal{S}}$ un arbre de Markov couple. Comme pour les chaînes couples, les arbres de Markov couples généralisent les arbres de Markov cachés dans lesquels le champ caché X est un arbre de Markov. Dans [94], on donne des conditions pour qu'un arbre de Markov couple soit un arbre de Markov caché.
L'algorithme de Baum-Welsh dans le cas des arbres se déroule en deux temps appelés "récursion ascendante" et "récursion descendante" et fonctionne de la manière suivante.

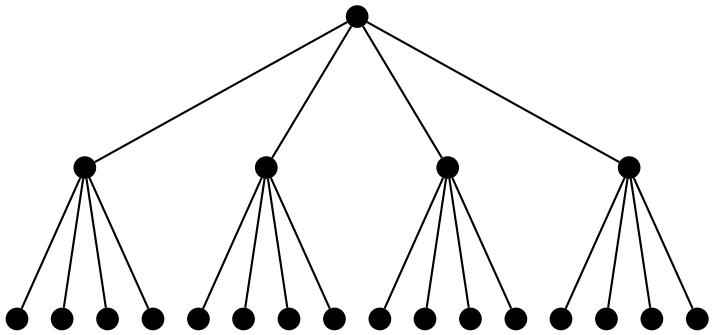


FIG. 2.3 – Graphe de dépendance d'un arbre de Markov.

1. Etape ascendante :
   - initialisation :
     Pour $u \in \mathcal{S}_P$, $\beta_u(x_u) = 1$;
   - itération pour k de $P-1$ à 1 :
     Pour $u \in \mathcal{S}_k$ et $t \in c(u)$, $\beta_{t,u}(x_u) = \sum_{x_t} \beta_t(x_t) p(x_t, y_t | x_u, y_u)$ et $\beta_u(x_u) = \prod_{t \in c(u)} \beta_{t,u}(x_u)$.
2. Etape descendante :
   - initialisation :
     $\alpha_1(x_1) = p(x_1, y_1)$;
   - itération pour k de 2 à P :
     Pour $u \in \mathcal{S}_k$, $\alpha_u(x_u) = \sum_{x_{\rho(u)}} \alpha_{\rho(u)}(x_{\rho(u)}) \frac{\beta_{\rho(u)}(x_{\rho(u)})}{\beta_{u,\rho(u)}(x_{\rho(u)})} p(x_u, y_u | x_{\rho(u)}, y_{\rho(u)})$.

La distribution $p(x|y)$ est également une distribution markovienne et on a :
- $p(x_u|y) \propto p(x_u, y) = \alpha_u(x_u)\beta_u(x_u)$;
- $p(x_u|x_{\rho(u)}, y) = \frac{\beta_u(x_u)}{\beta_{u,\rho(u)}(x_{\rho(u)})} p(x_u, y_u | x_{\rho(u)}, y_{\rho(u)})$.

La version conditionnelle de l'algorithme de Baum-Welsh dans le cas des arbres figure dans [49].

Nous n'étudierons pas les modèles d'arbre de Markov dans cette thèse. En revanche, tous les modèles de chaînes abordés dans cette thèse peuvent se généraliser au cas des arbres.
Dans la section suivante, nous abordons l'estimation des paramètres d'un modèle à données latentes $p(z;\theta)$.

2.3 Estimation des paramètres

2.3.1 Algorithme EM

L'algorithme "Expectation Maximisation" (EM) est parmi les plus utilisés pour estimer les paramètres d'une chaîne de Markov cachée ; ainsi que ceux de différents autres modèles à données latentes. Cet algorithme est issu des travaux de A. P. Dempster, N. M. Laird et D. B. Rubin [38] sur l'estimation par maximum de vraisemblance à partir de données incomplètes. Nous allons exposer dans cette sous-section l'algorithme EM dans sa généralité puis nous détaillerons l'algorithme EM dans le cas des chaînes de Markov cachées à bruit indépendant

de type exponentiel.
On considère un couple de processus $(X,Y)=(X_u,Y_u)_{u\in\mathcal{S}}$ tel que chaque X_u soit à valeurs dans $\mathcal{X}=\{\omega_1,\ldots,\omega_K\}$ et chaque Y_u à valeurs dans un espace vectoriel de dimension finie $\mathcal{Y}$. On notera $p(x,y;\theta)$ la distribution de (X,Y) de paramètre θ. Nous observons la réalisation y de Y et nous devons estimer le paramètre θ.
Soit

$$Q(\theta|\theta_q)=\mathbb{E}_{\theta_q}\left(\log\left(p(X,y;\theta)\right)|y\right)=\sum_{x\in\mathcal{X}^{|\mathcal{S}|}}\log\left(p(x,y;\theta)\right)p(x|y;\theta_q), \tag{2.5}$$

l'espérance sachant $Y=y$ de la log-vraisemblance en données complètes lorsque le paramètre vaut θ_q. Le but de l'algorithme EM est de construire une suite $(\theta_q)_{q\in\mathbb{N}}$ telle que :

$$\log\left(p(y;\theta_{q+1})\right)\geq\log\left(p(y;\theta_q)\right).$$

On a :

$$\log\left(p(y;\theta)\right)=Q(\theta|\theta_q)-\sum_{x\in\mathcal{X}^{|\mathcal{S}|}}\log\left(p(x|y;\theta)\right)p(x|y;\theta_q).$$

La fonction $\theta\rightarrow\sum_{x\in\mathcal{X}^{|\mathcal{S}|}}\log\left(p(x|y;\theta)\right)p(x|y;\theta_q)$ est maximale pour $\theta=\theta_q$, soit :

$$\sum_{x\in\mathcal{X}^{|\mathcal{S}|}}\log\left(p(x|y;\theta)\right)p(x|y;\theta_q)\leq\sum_{x\in\mathcal{X}^{|\mathcal{S}|}}\log\left(p(x|y;\theta_q)\right)p(x|y;\theta_q).$$

Ainsi

$$\log\left(p(y;\theta)\right)\geq Q(\theta|\theta_q)-\sum_{x\in\mathcal{X}^{|\mathcal{S}|}}\log\left(p(x|y;\theta_q)\right)p(x|y;\theta_q),$$

et donc si $\theta_{q+1}=\arg\max Q(\theta|\theta_q)$, on en déduit :

$$\begin{aligned}\log\left(p(y;\theta_{q+1})\right) &\geq Q(\theta_{q+1}|\theta_p)-\sum_{x\in\mathcal{X}^{|\mathcal{S}|}}\log\left(p(x|y;\theta_q)\right)p(x|y;\theta_q),\\ &\geq Q(\theta_q|\theta_q)-\sum_{x\in\mathcal{X}^{|\mathcal{S}|}}\log\left(p(x|y;\theta_q)\right)p(x|y;\theta_q)=\log\left(p(y;\theta_q)\right).\end{aligned}$$

L'algorithme EM fonctionne de la manière suivante :
- étape E (Expectation) :
  calcul de $Q(\theta|\theta_q)=\mathbb{E}_{\theta_q}\left(\log\left(p(X,y;\theta)\right)|y\right)$;
- étape M (Maximisation) :
  maximisation de $\theta\rightarrow Q(\theta|\theta_q)$.

Il suffit en fait de choisir θ_{q+1} tel que $Q(\theta_{q+1}|\theta_q)\geq Q(\theta_q|\theta_q)$ pour avoir $\log\left(p(y;\theta_{q+1})\right)\geq\log\left(p(y;\theta_q)\right)$. L'algorithme est alors appelé "Generalized Expectation Maximisation"(GEM). Dans [79], on peut trouver d'autres versions de l'algorithme EM. Parmi celles-ci, on peut citer l'algorithme "Stochastic Expectation Maximisation" (SEM) [20, 29] qui est une version stochastique de l'algorithme EM. L'objectif de l'algorithme SEM est d'éviter la convergence de la suite $(\theta_q)_{q\in\mathbb{N}}$ vers un maximum local de $\theta\rightarrow\log\left(p(y;\theta)\right)$ en introduisant une perturbation stochastique.

Donnons maintenant un exemple illustrant le fonctionnement de l'algorithme EM. On considère une chaîne Markov cachée $(X, Y) = (X_n, Y_n)_{1 \leq n \leq N}$ classique avec la distribution :

$$p(x_{1:N}, y_{1:N}) = p(x_1)p(y_1|x_1) \prod_{n=1}^{N-1} p(x_{n+1}|x_n)p(y_{n+1}|x_{n+1}).$$

Prenons le cas particulier où $\mathcal{Y} = \mathbb{R}$ et $p(y_n|x_n = \omega_j)$ est une loi de type exponentiel de paramètre fonctionnel a et de paramètre scalaire λ_j, les détails sur les modèles exponentiels sont donnés au chapitre 4. Nous avons :

$$p(y_n|x_n = \omega_j) = \frac{1}{Z(\lambda_j)} \exp\left(\lambda_j a(y_n)\right), \text{ avec } \int_{\mathcal{Y}} \exp\left(\lambda_j a(y_n)\right) dy_n = Z(\lambda_j).$$

Le paramètre θ que l'on cherche à estimer a pour composantes les paramètres de $p(x_{1:N})$, qui sont la loi initiale $p(x_1)$ et la transition $p(x_{n+1}|x_n)$ que l'on supposera indépendante de n, et le paramètre λ_j de $p(y_n|x_n = \omega_j)$, également supposé indépendant de n.
Considérons l'étape E (on omettra le paramètre θ par mesure de simplicité).
On a :

$$\begin{aligned} \log\left(p(X, y_{1:N})\right) &= \log\left(p(X_1)\right) + \sum_{n=1}^{N-1} \log\left(p(X_{n+1}|X_n)\right) \\ &+ \sum_{n=1}^{N} \log\left(p(y_n|X_n)\right). \end{aligned}$$

Ainsi, en intégrant sous la loi a posteriori $p(x_{1:N}|y_{1:N}; \theta_q)$ et sachant que les quantités $p(x_{n+1} = \omega_j|x_n = \omega_i)$ ne dépendent pas de n, on a :

$$\begin{aligned} Q(\theta|\theta_q) &= \sum_{j=1}^{K} \log\left(p(x_1 = \omega_j)\right) p(x_1 = \omega_j|y_{1:N}; \theta_q) \\ &+ \sum_{i=1}^{K} \sum_{j=1}^{K} \sum_{n=1}^{N-1} \log\left(p(x_{n+1} = \omega_j|x_n = \omega_i)\right) p(x_n = \omega_i, x_{n+1} = \omega_j|y_{1:N}; \theta_q) \\ &+ \sum_{j=1}^{K} \sum_{n=1}^{N} \log\left(p(y_n|x_n = \omega_j)\right) p(x_n = \omega_j|y_{1:N}; \theta_q). \end{aligned}$$

L'étape M consiste alors à maximiser cette quantité. On notera $p(x_1; \theta_{q+1})$, $p(x_{n+1}|x_n; \theta_{q+1})$ la loi initiale et la transition de X sous le paramètre θ_{q+1} et $\lambda_{q+1,j}$ les paramètres de $p(y_n|x_n = \omega_j; \theta_{q+1})$. Les composantes de θ_{q+1} sont alors les K valeurs $(p(x_1 = \omega_j; \theta_{q+1}))_{j \in \{1,\ldots,K\}}$, les K^2 valeurs $(p(x_{n+1} = \omega_j|x_n = \omega_i; \theta_{q+1}))_{(i,j) \in \{1,\ldots,K\}^2}$ et les K valeurs $(\lambda_{q+1,j})_{j \in \{1,\ldots,K\}}$. Nous devons maximiser chacun des trois termes de la somme. La quantité

$$\sum_{j=1}^{K} \log\left(p(x_1 = \omega_j)\right) p(x_1 = \omega_j|y_{1:N}; \theta_q)$$

est maximale pour :

$$p(x_1 = \omega_j; \theta_{q+1}) = p(x_1 = \omega_j|y_{1:N}; \theta_q).$$

Soit $i \in \{1, \ldots, K\}$, la quantité

$$\sum_{j=1}^{K} \sum_{n=1}^{N-1} \log\left(p(x_{n+1} = \omega_j | x_n = \omega_i)\right) p(x_n = \omega_i, x_{n+1} = \omega_j | y_{1:N}; \theta_q) \quad =$$
$$\sum_{j=1}^{K} \log\left(p(x_{n+1} = \omega_j | x_n = \omega_i)\right) \sum_{n=1}^{N-1} p(x_n = \omega_i, x_{n+1} = \omega_j | y_{1:N}; \theta_q)$$

est maximale pour :

$$p(x_{n+1} = \omega_j | x_n = \omega_i; \theta_{q+1}) = \frac{\sum_{n=1}^{N-1} p(x_n = \omega_i, x_{n+1} = \omega_j | y_{1:N}; \theta_q)}{\sum_{n=1}^{N-1} p(x_n = \omega_i | y_{1:N}; \theta_q)}.$$

L'algorithme EM nécessite de connaître les probabilités a posteriori $p(x_n = \omega_i | y_{1:N}; \theta_q)$ et $p(x_n = \omega_i, x_{n+1} = \omega_j | y_{1:N}; \theta_q)$ qui sont calculées par l'algorithme de Baum-Welsh.
Il reste la maximisation du dernier terme. Dans le cas des modèles exponentiels, l'étape M et la maximisation de la vraisemblance sont analogues.
En effet, soit $y_{1:N}$ un échantillon d'une loi de type exponentiel de densité $p(y; \lambda) = \frac{1}{Z(\lambda)} \exp(\lambda a(y))$, alors l'estimateur du maximum de vraisemblance $\hat{\lambda}_{MV}(y_{1:N})$ est solution de l'équation de vraisemblance :

$$\frac{\partial}{\partial \lambda} \log\left(Z(\lambda)\right) = \frac{1}{N} \sum_{n=1}^{N} a(y_n).$$

La maximisation de chacune des quantités

$$\sum_{n=1}^{N} \log\left(p(y_n | x_n = \omega_j)\right) p(x_n = \omega_j | y_{1:N}; \theta_q) =$$
$$-\log\left(Z(\lambda_j)\right) \sum_{n=1}^{N} p(x_n = \omega_j | y_{1:N}; \theta_q) + \lambda_j \sum_{n=1}^{N} a(y_n) p(x_n = \omega_j | y_{1:N}; \theta_q).$$

se fait en résolvant les équations :

$$\frac{\partial}{\partial \lambda_j} \log\left(Z(\lambda_j)\right) = \frac{1}{\sum_{n=1}^{N} p(x_n = \omega_j | y_{1:N}; \theta_q)} \times \sum_{n=1}^{N} a(y_n) p(x_n = \omega_j | y_{1:N}; \theta_q).$$

Considérons les exemples suivants de modèles exponentiels, avec les équations de vraisemblance correspondantes.

- Loi exponentielle standard de densité sur $\mathbb{R}^+$ donnée par $p(y) = \lambda \exp(-\lambda y)$.
  On a $Z(\lambda) = \frac{1}{\lambda}$ et $a(y) = -y$, ainsi l'estimateur du maximum de vraisemblance $\hat{\lambda}_{MV}$ et

l'estimée $\lambda_{q+1,j}$ sont donnés par :

$$\text{MV} : \quad \hat{\lambda}_{MV} = \frac{N}{\sum_{n=1}^{N} y_n}.$$

$$\text{EM} : \quad \lambda_{q+1,j} = \frac{\sum_{n=1}^{N} p(x_n = \omega_j | y_{1:N}; \theta_q)}{\sum_{n=1}^{N} y_n p(x_n = \omega_j | y_{1:N}; \theta_q)}.$$

– Loi normale $\mathcal{N}_{\mathbb{R}}(m, s)$, les estimations de la moyenne et de la variances sont données par :

$$\text{MV} : \begin{cases} \hat{m}_{MV} = \dfrac{1}{N}\displaystyle\sum_{n=1}^{N} y_n, \\ \hat{s}_{MV} = \dfrac{1}{N}\displaystyle\sum_{n=1}^{N} (y_n - \hat{m}_{MV})^2, \end{cases}$$

$$\text{EM} : \begin{cases} m_{q+1,j} = \dfrac{\displaystyle\sum_{n=1}^{N} y_n p(x_n = \omega_j | y_{1:N}; \theta_q)}{\displaystyle\sum_{n=1}^{N} p(x_n = \omega_j | y_{1:N}; \theta_q)}, \\ s_{q+1,j} = \dfrac{\displaystyle\sum_{n=1}^{N} (y_n - m_{q+1,j})^2 p(x_n = \omega_j | y_{1:N}; \theta_q)}{\displaystyle\sum_{n=1}^{N} p(x_n = \omega_j | y_{1:N}; \theta_q)}. \end{cases}$$

– Loi $\Gamma(a, b)$ de densité sur $\mathbb{R}^+$ donnée par $p(y; a, b) = \dfrac{1}{\Gamma(a) b^a} y^{a-1} \exp\left(-\dfrac{y}{b}\right)$, le paramétrage canonique est donné par :

$$\begin{aligned} a(y) &= \begin{pmatrix} \log(y) \\ -y \end{pmatrix}, \\ \lambda(a, b) &= \begin{pmatrix} a - 1 \\ \frac{1}{b} \end{pmatrix} = \begin{pmatrix} \lambda^{(1)} \\ \lambda^{(2)} \end{pmatrix}, \\ \text{et } \log\left(Z(\lambda)\right) &= \log\left(\Gamma(\lambda^{(1)} + 1)\right) - (\lambda^{(1)} + 1) \log(\lambda^{(2)}). \end{aligned}$$

Ainsi :

$$\text{MV} : \begin{cases} \psi(\hat{\lambda}^{(1)}_{MV}+1) - \log(\hat{\lambda}^{(2)}_{MV}) = \dfrac{1}{N}\displaystyle\sum_{n=1}^{N}\log(y_n), \\ \dfrac{\hat{\lambda}^{(1)}_{MV}+1}{\hat{\lambda}^{(2)}_{MV}} = \dfrac{1}{N}\displaystyle\sum_{n=1}^{N} y_n, \end{cases}$$

$$\text{EM} : \begin{cases} \psi(\lambda^{(1)}_{q+1,j}+1) - \log(\lambda^{(2)}_{q+1,j}) = \dfrac{\displaystyle\sum_{n=1}^{N}\log(y_n)p(x_n=\omega_j|y_{1:N};\theta_q)}{\displaystyle\sum_{n=1}^{N}p(x_n=\omega_j|y_{1:N};\theta_q)}, \\ \dfrac{\lambda^{(1)}_{q+1,j}+1}{\lambda^{(2)}_{q+1,j}} = \dfrac{\displaystyle\sum_{n=1}^{N} y_n p(x_n=\omega_j|y_{1:N};\theta_q)}{\displaystyle\sum_{n=1}^{N}p(x_n=\omega_j|y_{1:N};\theta_q)}, \end{cases}$$

où ψ est la fonction digamma, dérivée logarithmique de la fonction Γ (voir Annexe A). Si on exprime ces équations avec le paramétrage (a, b), on trouve :

$$\text{MV} : \begin{cases} \psi(\hat{a}_{MV}) + \log(\hat{b}_{MV}) = \dfrac{1}{N}\displaystyle\sum_{n=1}^{N}\log(y_n), \\ \hat{a}_{MV}\hat{b}_{MV} = \dfrac{1}{N}\displaystyle\sum_{n=1}^{N} y_n, \end{cases}$$

$$\text{EM} : \begin{cases} \psi(a_{q+1,j}) + \log(b_{q+1,j}) = \dfrac{\displaystyle\sum_{n=1}^{N}\log(y_n)p(x_n=\omega_j|y_{1:N};\theta_q)}{\displaystyle\sum_{n=1}^{N}p(x_n=\omega_j|y_{1:N};\theta_q)}, \\ a_{q+1,j}b_{q+1,j} = \dfrac{\displaystyle\sum_{n=1}^{N} y_n p(x_n=\omega_j|y_{1:N};\theta_q)}{\displaystyle\sum_{n=1}^{N}p(x_n=\omega_j|y_{1:N};\theta_q)}. \end{cases}$$

Malgré l'intérêt indéniable de l'algorithme EM et son très bon comportement dans nombreuses situations réelles, notons qu'il présente également des faiblesses qui vont en partie motiver l'utilisation de l'algorithme ICE. Tout d'abord, la suite $(\theta_q)_{q\in\mathbb{N}}$ construite par EM ne converge pas obligatoirement vers le maximum global de la vraisemblance de Y et il n'existe pas de théorèmes généraux donnant des conditions d'une telle convergence. De plus, l'étape de maximisation peut être délicate, notamment dans certains modèles comprenant des lois Γ ou K.

2.3.2 Algorithme ICE

Soit $Z = (X_u, Y_u)_{u\in\mathcal{S}}$ un modèle à données latentes tel que chaque X_u prend ses valeurs dans l'ensemble fini $\mathcal{X} = \{\omega_1, \ldots, \omega_K\}$ et chaque Y_u dans un espace vectoriel de dimension finie $\mathcal{Y}$. Soit $\theta = (\theta_1, \ldots, \theta_k) \in \mathbb{R}^k$ le paramètre de $p(z;\theta)$. Pour utiliser ICE, on a besoin de deux conditions :

- l'existence d'un estimateur $T = (T_1, \ldots, T_k)$ explicite à partir des données complètes ;
- la possibilité de simulation, pour tout θ, de X selon $p(x|y;\theta)$.

L'algorithme ICE est itératif et fonctionne de la manière suivante :

1. donnée d'un paramètre initial θ_0 ;
2. à partir de θ_q, on calcule :

$$\theta_{q+1,j} = \mathbb{E}_{\theta_q}\left(T_j(X,y)|y\right),$$

pour les composantes T_j pour lesquelles ce calcul est possible. Pour les autres composantes T_i, on pose :

$$\theta_{q+1,i} = \frac{\sum_{l=1}^{L} T_i(x^l, y)}{L},$$

où $x^1, \ldots, x^L$ sont simulés selon $p(x|y;\theta_q)$.

Nous voyons que ICE fonctionne sous des hypothèses très faibles et il n'existe pas de problème de maximisation. Notons que l'estimateur T peut être l'estimateur de vraisemblance ou pas. Soit $Z = (X_n, Y_n)_{1\leq n\leq N}$ une chaîne de Markov cachée à bruit indépendant de distribution :

$$p(x_{1:N}, y_{1:N}) = p(x_1)p(y_1|x_1)\prod_{n=1}^{N-1} p(x_{n+1}|x_n)p(y_{n+1}|x_{n+1}).$$

On supposera dans cet exemple que $p(y_n|x_n = \omega_j)$ est la densité d'une loi $\Gamma(a_j, b_j)$:

$$p(y_n|x_n = \omega_j) = \frac{1}{\Gamma(a_j)b_j^{a_j}} y_n^{a_j-1}\exp\left(-\frac{y_n}{b_j}\right).$$

Nous supposerons que la chaîne X est stationnaire et réversible, ainsi $p(x_n = \omega_i, x_{n+1} = \omega_j) = p(x_n = \omega_j, x_{n+1} = \omega_i)$ et est indépendant de n. Les paramètres du modèle sont :

- les $K^2 - 1$ paramètres $p(x_n = \omega_i, x_{n+1} = \omega_j)$;
- les $2K$ paramètres (a_j, b_j) de la loi $p(y_n|x_n = \omega_j)$.

L'estimateur à partir des données complètes est

$$T = \left((\hat{R}_{i,j}(x_{1:N}, y_{1:N}))_{1\leq i\leq K, 1\leq j\leq K}, (\hat{a}_j(x_{1:N}, y_{1:N}))_{1\leq j\leq K}, (\hat{b}_j(x_{1:N}, y_{1:N}))_{1\leq j\leq K}\right),$$

où $\hat{R}_{i,j}(x_{1:N}, y_{1:N})$ est l'estimateur de $p(x_n = \omega_i, x_{n+1} = \omega_j)$, $\hat{a}_j(x_{1:N}, y_{1:N})$ et $\hat{b}_j(x_{1:N}, y_{1:N})$ sont les estimateurs de a_j et b_j. Les estimateurs $\hat{a}_j(x_{1:N}, y_{1:N})$ et $\hat{b}_j(x_{1:N}, y_{1:N})$ sont donnés par :

$$\hat{b}_j(x_{1:N}, y_{1:N}) = \frac{\hat{s}_j}{\hat{m}_j} \text{ et } \hat{a}_j(x_{1:N}, y_{1:N}) = \frac{\hat{m}_j}{\hat{b}_j},$$

où

$$\begin{aligned}\hat{m}_j &= \frac{\sum_{n=1}^{N} y_n I(x_n = \omega_j)}{\sum_{n=1}^{N} I(x_n = \omega_j)}, \\ \hat{s}_j &= \frac{\sum_{n=1}^{N} (y_n - \hat{m}_j)^2 I(x_n = \omega_j)}{\sum_{n=1}^{N} I(x_n = \omega_j)}.\end{aligned}$$

Enfin, $\hat{R}_{i,j}(x_{1:N}, y_{1:N})$ est classiquement donné par :

$$\hat{R}_{i,j}(x_{1:N}, y_{1:N}) = \frac{1}{2(N-1)} \sum_{n=1}^{N-1} \left[I(x_n = \omega_i, x_{n+1} = \omega_j) + I(x_n = \omega_j, x_{n+1} = \omega_i) \right],$$

où $I(A) = 1$ si A est vraie et 0 sinon.
L'espérance $\theta_{q+1} = \mathbb{E}_{\theta_q}\left(T(X, y_{1:N})|y_{1:N}\right)$ est calculable pour les composantes $\hat{R}_{i,j}$, ce qui donne :

$$\begin{aligned}p(x_n = \omega_i, x_{n+1} = \omega_j; \theta_{q+1}) &= \frac{1}{2(N-1)} \sum_{n=1}^{N-1} \left[p(x_n = \omega_i, x_{n+1} = \omega_j | y_{1:N}; \theta_q) \right. \\ &+ \left. p(x_{n+1} = \omega_i, x_n = \omega_j | y_{1:N}; \theta_q) \right].\end{aligned}$$

L'espérance $\theta_{q+1} = \mathbb{E}\left(T(X, y_{1:N})|y_{1:N}; \theta_q\right)$ n'est pas calculable pour les composantes a_j et b_j. Ainsi, on simule L réalisations $x^{(1)}, \ldots, x^{(L)}$ de X selon la loi a posteriori $p(x_{1:N}|y_{1:N}; \theta_q)$. Ensuite, pour chaque l de 1 à L, on calcule les estimées $\hat{a}_j^{(l)} = \hat{a}_j(x_{1:N}^{(l)}, y_{1:N})$ et $\hat{b}_j^{(l)} = \hat{b}_j(x_{1:N}^{(l)}, y_{1:N})$. Pour finir, on pose :

$$a_{q+1,j} = \frac{1}{L} \sum_{l=1}^{L} \hat{a}_j^{(l)} \text{ et } b_{q+1,j} = \frac{1}{L} \sum_{l=1}^{L} \hat{b}_j^{(l)}.$$

Les algorithmes ICE et EM se valent dans les cas classiques de chaînes de Markov cachées à bruit indépendant et gaussien [14].

Conclusion

Dans ce chapitre, nous avons présenté les principaux algorithmes d'inférence bayésienne permettant d'estimer les états inobservés dans le contexte de chaînes et d'arbres de Markov cachés. Ces algorithmes permettent le calcul rapide des lois marginales a posteriori, ce qui rend possible la mise en place de l'estimateur du MPM. Nous avons également présenté deux algorithmes généraux d'estimation des paramètres, qui sont EM et ICE. Dans la suite, nous utiliserons sauf mention contraire l'algorithme ICE. En effet, dans certains modèles, la loi des observations conditionnellement aux états cachés est complexe et l'étape M de

l'algorithme EM peut être délicate. Par ailleurs, le principe de ICE nous permettra d'en proposer une extension dans le cas des bruits à mémoire longue étudiés dans le chapitre 5. De plus, nous montrerons au travers de simulations présentées au cours de cette thèse le bon comportement de l'algorithme ICE, qui s'avère capable d'estimer correctement les paramètres en présence de bruits importants. Joint aux estimateurs du MAP ou du MPM, la méthode ICE permet ainsi, une fois le modèle fixé, de proposer des algorithmes de recherche des états cachés de manière automatique, ce qui revêt une importance primordiale dans de nombreuses applications. Concernant la convergence de l'algorithme ICE, des premiers résultats ont été démontrés dans [101]. Une autre approche abordée dans la thèse de N. Brunel [21] est celle des fonctions "estimantes". La théorie des fonctions estimantes développée dans [57] permet de réunir dans le même formalisme de nombreuses méthodes d'estimation.

Chapitre 3

Inférence dans les chaînes semi-markoviennes cachées M-stationnaires

Dans ce chapitre, nous étudions les méthodes d'estimation et de segmentation vues au chapitre précédent dans le cadre de deux modèles généralisant les chaînes de Markov cachées classiques. Ces deux modèles font partie de la famille générale de chaînes de Markov "triplets". Nous commençons par rappeler la famille des modèles de Markov "couples", qui généralisent les modèles de Markov cachés. Nous introduisons ensuite les chaînes "triplets" [95, 102] et nous rappelons un de leurs intérêts, important pour les applications pratiques et développé récemment dans la thèse de P. Lanchantin [65], qui réside dans leur aptitude à traiter des données cachées non stationnaires. Dans un second temps, nous précisons les deux modèles particuliers introduits dans cette thèse. Le premier modèle est celui des chaînes semi-markoviennes, qui peuvent être vues comme des CMT particuliers. Nous présentons des résultats théoriques montrant que la semi-markovianité classique, où le temps aléatoire de séjour dans un état est à valeurs dans $\mathbb{N}$, peut également se modéliser par un temps de séjour "minimal". Nous définissons ainsi une sous-famille des modèles semi-markoviens cachés classiques et montrons leur intérêt, au niveau du temps de calcul, par rapport aux démarches fondées sur les chaînes semi-markoviennes classiques. Ensuite, nous étendons ce modèle au cas non stationnaire. L'intérêt des nouveaux modèles est validé par des expérimentations.

3.1 Chaînes de Markov couples et chaînes de Markov cachées

Soit $Z = (X_n, Y_n)_{1 \leq n \leq N}$ un processus où chaque X_n prend ses valeurs dans un ensemble fini $\mathcal{X} = \{\omega_1, \ldots, \omega_K\}$ et chaque Y_n prend ses valeurs dans un $\mathbb{R}$-espace vectoriel de dimension finie $\mathcal{Y}$. Notons $p(z_{1:N})$ la densité de Z par rapport à la mesure produit $(\nu \otimes \lambda_{\mathcal{Y}})^N$, où ν est la mesure de décompte sur $\mathcal{X}$ et $\lambda_{\mathcal{Y}}$ est la mesure de Lebesgue sur $\mathcal{Y}$. Supposons que Z est une chaîne de Markov couple. Nous avons :

$$p(z_{1:N}) = p(z_1) \prod_{n=1}^{N-1} p(z_{n+1}|z_n) = \frac{p(z_1, z_2)p(z_2, z_3) \ldots p(z_{N-1}, z_N)}{p(z_2)p(z_3) \ldots p(z_{N-1})}. \tag{3.1}$$

D'après la factorisation de la loi de Z et l'équivalence entre factorisation et markovianité vue au chapitre 2, on en déduit que $p(x_{1:N}|y_{1:N})$ et $p(y_{1:N}|x_{1:N})$ sont des distributions markoviennes. La chaîne sera dite stationnaire si les densités $p(z_n, z_{n+1})$ ne dépendent pas de n. La proposition suivante donne des conditions pour que la chaîne cachée X soit une chaîne de Markov :

Proposition 3.1.1. *Soit $Z = (X_n, Y_n)_{1\leq n\leq N}$ une chaîne de Markov couple vérifiant :*

1. *$p(z_n, z_{n+1})$ ne dépend pas de $n \in \{1, \ldots, N-1\}$;*
2. *$p(z_n = a, z_{n+1} = b) = p(z_n = b, z_{n+1} = a)$ pour tout $n \in \{1, \ldots, N-1\}$ et pour tout a et b.*

Alors les trois conditions :

- *X est une chaîne de Markov ;*
- *pour tout $n \in \{2, \ldots, N\}$, $p(y_n|x_n, x_{n-1}) = p(y_n|x_n)$;*
- *pour tout $n \in \{1, \ldots, N\}$, $p(y_n|x_{1:N}) = p(y_n|x_n)$*

sont équivalentes.

Preuve. Voir [99]. □

Nous constatons ainsi que le modèle des chaînes de Markov couples est plus général que celui des chaînes de Markov cachées dans lequel le processus caché est une chaîne de Markov. Il permet ainsi de modéliser des bruits complexes et de considérer des processus cachés non markoviens.

3.2 Chaînes de Markov cachées M-stationnaires

3.2.1 Le modèle

Dans cette section, chaque X_n prend ses valeurs dans l'ensemble fini $\mathcal{X} = \{\omega_1, \ldots, \omega_K\}$, chaque U_n prend ses valeurs dans l'ensemble fini $\Lambda = \{\lambda_1, \ldots, \lambda_M\}$ et chaque Y_n dans un espace vectoriel de dimension finie $\mathcal{Y}$. Le processus (X, U, Y) est une chaîne M-stationnaire cachée si sa distribution est de la forme :

$$p(x_{1:N}, u_{1:N}, y_{1:N}) = p(u_1)p(x_1|u_1)p(y_1|x_1)\prod_{n=1}^{N-1} p(u_{n+1}|u_n)p(x_{n+1}|x_n, u_{n+1})p(y_{n+1}|x_{n+1}). \quad (3.2)$$

Dans ce modèle X n'est pas une chaîne de Markov et les processus U et Y sont indépendants conditionnellement à X. Lorsque $u_1 = \ldots = u_N = \lambda_k$, $p(x_{1:N}|u_{1:N})$ est la distribution d'une chaîne de Markov de loi initiale $p(x_1|u_{1:N}) = p(x_1|u_1)$ et de transition $p(x_{n+1}|u_{1:N}, x_n) = p(x_{n+1}|u_{n+1}, x_n)$.

3.2.2 Inférence dans le modèle de chaînes de Markov cachée M-stationnaires

Nous détaillerons uniquement l'estimation des paramètres de la loi $p(x_{1:N}, u_{1:N})$, l'estimation des paramètres de $p(y_n|x_n)$ étant déjà étudiée au chapitre 2.

Soit $y_{1:N} = (y_1, \ldots, y_N)$ la réalisation observée de Y. Les paramètres à estimer sont $p(u_{n+1}|u_n)$ et $p(x_{n+1}|x_n, u_{n+1})$. Les estimations de $p(u_{n+1}|u_n)$ et $p(x_{n+1}|x_n, u_{n+1})$ sont obtenues à partir de celle de $p(x_n, u_n, x_{n+1}, u_{n+1})$. Nous supposerons que la chaîne (X, U) est stationnaire et réversible, soit $p(x_n = \omega_i, u_n = \lambda_k, x_{n+1} = \omega_j, u_{n+1} = \lambda_l) = p(x_n = \omega_j, u_n = \lambda_l, x_{n+1} = \omega_i, u_{n+1} = \lambda_k)$ et indépendant de n. Notons θ_q le vecteur paramètre obtenu à l'itération q de ICE. Afin d'estimer $p(x_n, u_n, x_{n+1}, u_{n+1})$, nous devons nous donner un estimateur à partir des données complètes $(x_{1:N}, u_{1:N}, y_{1:N})$. On choisit comme estimateur :

$$\begin{aligned}\hat{p}(x_n = \omega_i, u_n = \lambda_k, x_{n+1} = \omega_j, u_{n+1} = \lambda_l)(x_{1:N}, u_{1:N}, y_{1:N}) = \\ \frac{1}{2(N-1)} \sum_{n=1}^{N-1} [I(x_n = \omega_i, u_n = \lambda_k, x_{n+1} = \omega_j, u_{n+1} = \lambda_l) + \\ I(x_n = \omega_j, u_n = \lambda_l, x_{n+1} = \omega_i, u_{n+1} = \lambda_k)].\end{aligned}$$

L'espérance de cette quantité sachant $Y = y$ et sous le paramètre θ_q peut se calculer exactement et donne :

$$\begin{aligned}p(x_n = \omega_i, u_n = \lambda_k, x_{n+1} = \omega_j, u_{n+1} = \lambda_l; \theta_{q+1}) = \\ \frac{1}{2(N-1)} \sum_{n=1}^{N-1} [p(x_n = \omega_i, u_n = \lambda_k, x_{n+1} = \omega_j, u_{n+1} = \lambda_l | y_{1:N}; \theta_q) + \\ p(x_n = \omega_j, u_n = \lambda_l, x_{n+1} = \omega_i, u_{n+1} = \lambda_k | y_{1:N}; \theta_q)],\end{aligned}$$

et on a :

$$\begin{aligned}p(u_{n+1} = \lambda_l | u_n = \lambda_k; \theta_{q+1}) &\propto p(x_n = \omega_i, u_n = \lambda_k, x_{n+1} = \omega_j, u_{n+1} = \lambda_l; \theta_{q+1}), \\ p(x_{n+1} = \omega_j | x_n = \omega_i, u_{n+1} = \lambda_l; \theta_{q+1}) &\propto p(x_n = \omega_i, u_n = \lambda_k, x_{n+1} = \omega_j, u_{n+1} = \lambda_l; \theta_{q+1}).\end{aligned}$$

3.3 Chaînes semi-markoviennes cachées

Dans cette section, nous présentons les chaînes semi-markoviennes cachées et nous proposons d'écrire celles-ci comme un cas particulier de chaîne de Markov triplet. Les chaînes semi-markoviennes généralisent les chaînes de Markov pour des temps de séjour non nécessairement distribués suivant une loi géométrique. Parmi les applications du modèle semi-markovien, on peut citer la génétique [18, 25, 33], la reconnaissance vocale [42, 73, 83, 108], la reconnaissance de caractères [91, 109, 121], la segmentation d'images [41] ou l'analyse de modèles graphiques [52].

Nous commençons par donner la définition générale d'une chaîne semi-markovienne cachée ainsi que certaines de ses propriétés. Une étude plus approfondie des propriétés des chaînes semi-markoviennes est disponible dans [8, 32, 75]. Dans un second temps, nous proposons un modèle semi-markovien particulier original. Nous verrons que ce nouveau modèle permet l'utilisation de l'algorithme de Baum-Welsh avec une complexité algorithmique réduite. Nous conclurons le chapitre par des expérimentations utilisant le modèle original de chaînes semi-markoviennes cachées M-stationnaires.

3.3.1 Définition et propriétés d'une chaîne semi-markovienne

Nous choisissons de définir une chaîne semi-markovienne à valeurs dans un espace fini comme la marginale d'une chaîne de Markov. Une définition plus générale peut se trouver dans [74].

Définition 3.3.1 (Chaîne semi-markovienne)**.** *Soit $\mathcal{X}$ un ensemble fini. On définit les quantités suivantes :*

- *la probabilité π sur $\mathcal{X}$;*
- *la transition $q(.|.)$ sur $\mathcal{X}^2$ vérifiant $q(x|x) = 0$;*
- *pour tout $x \in \mathcal{X}$, les densités de probabilité $d(x, .)$ sur $\mathbb{N}^*$.*

Un processus $X = (X_n)_{n\geq 1}$ tel que chaque X_n prend ses valeurs dans $\mathcal{X}$ est une chaîne semi-markovienne de loi initiale π, de transition q et de loi de durée d s'il existe un processus $U = (U_n)_{n\geq 1}$, où chaque U_n prend ses valeurs dans $\mathbb{N}^*$, tel que (X, U) soit une chaîne de Markov homogène donnée par :

- *la loi initiale :*

$$p(x_1, u_1) = \pi(x_1)d(x_1, u_1) \ , \tag{3.3}$$

- *et les transitions :*

$$p(x_{n+1}, u_{n+1}|x_n, u_n) = p(x_{n+1}|x_n, u_n) \times p(u_{n+1}|x_{n+1}, u_n), \tag{3.4}$$

avec :

$$p(x_{n+1}|x_n, u_n) = \begin{cases} \delta_{x_n}(x_{n+1}) \ si \ u_n > 1, \\ q(x_{n+1}|x_n) \ si \ u_n = 1, \end{cases} \tag{3.5}$$

et

$$p(u_{n+1}|x_{n+1}, u_n) = \begin{cases} \delta_{u_n-1}(u_{n+1}) \ si \ u_n > 1, \\ d(x_{n+1}, u_{n+1}) \ si \ u_n = 1, \end{cases} \tag{3.6}$$

où $\delta_x(x) = 1$ et $\delta_x(x') = 0$ pour $x' \neq x$.

Dans cette définition, on remarque que la transition q et la loi de temps de séjour d ne dépendent pas de n, la chaîne semi-markovienne est alors qualifiée d'"homogène".
Notons que u_n représente le temps de séjour restant de la chaîne dans $X_n = x_n$, la variable aléatoire U_n est appelée "temps de récurrence avant".
Considérons la définition suivante :

Définition 3.3.2. *Soit $X = (X_n)_{n\geq 1}$ un processus à valeurs dans un ensemble fini $\mathcal{X}$. On définit :*

- *la suite des instants de saut $V = (V_n)_{n\geq 1}$ par :*

$$V_1 = 1 \ et \ V_n = \inf\left\{k > V_{n-1} : X_k \neq X_{V_{n-1}}\right\} \ ;$$

- *la chaîne immergée $\tilde{X} = (\tilde{X}_n)_{n\geq 1}$ par :*

$$\forall n \geq 1, \ \tilde{X}_n = X_{V_n} \ ;$$

– *pour tout $n \geq 1$, le temps de séjour dans $\tilde{X}_n$ par :*

$$T_n = V_{n+1} - V_n.$$

La réalisation de la chaîne immergée correspond à la suite des valeurs visitées par le processus X. Lorsque X est une chaîne semi-markovienne de transition q, on montre que sa chaîne immergée est une chaîne de Markov de transition q. Plus exactement :

Proposition 3.3.1. *Soit $X = (X_n)_{n\geq 1}$ un processus à valeurs dans un ensemble fini $\mathcal{X} = \{\omega_1, \ldots, \omega_K\}$ de chaîne immergée $\tilde{X}$, soit $V = (V_n)_{n\geq 1}$ la suite des instants de saut de X et soit $T = (T_n)_{n\geq 1}$ le processus des temps de séjour.*
X est une chaîne semi-markovienne homogène de transition q et de loi de durée d si et seulement si les deux conditions suivantes sont satisfaites :

1. *pour tout n, $\tilde{X}_{n+1}$ est indépendante des variables aléatoires $\tilde{X}_1, \ldots, \tilde{X}_n, T_1, \ldots, T_n$ conditionnellement à $\tilde{X}_n$ et la loi de $\tilde{X}_{n+1}$ sachant $\tilde{X}_n$ ne dépend pas de n ;*
2. *pour tout n, T_{n+1} est indépendante des variables aléatoires $\tilde{X}_1, \ldots, \tilde{X}_{n+1}, T_1, \ldots, T_n$ conditionnellement à $\tilde{X}_{n+1}$ et la loi de T_{n+1} sachant $\tilde{X}_{n+1}$ ne dépend pas de n.*

Sous ces conditions, la transition q et la loi de durée d sont données par :
– $q(\omega_j|\omega_i) = p(\tilde{x}_{n+1} = \omega_j|\tilde{x}_n = \omega_i)$;
– $d(\omega_j, u) = p(t_{n+1} = u|\tilde{x}_{n+1} = \omega_j)$.

Preuve.
La première implication est évidente. Montrons la deuxième implication. Notons $D(\omega_j, u) = \sum_{t\geq u} d(\omega_j, t)$, le processus X est une chaîne semi-markovienne de transition q et de loi de durée d si et seulement si :

$$p(\underbrace{x_1 = \tilde{x}_1, \ldots, x_{t_1} = \tilde{x}_1}_{t_1 \text{ fois}}, \underbrace{x_{t_1+1} = \tilde{x}_2, \ldots, x_{t_1+t_2} = \tilde{x}_2}_{t_2 \text{ fois}},$$
$$\ldots, \underbrace{x_{t_1+\ldots+t_{L-2}+1} = \tilde{x}_{L-1}, \ldots, x_{t_1+\ldots+t_{L-1}} = \tilde{x}_{L-1}}_{t_{L-1} \text{ fois}}, \underbrace{x_{t_1+\ldots+t_{L-1}+1} = \tilde{x}_L, \ldots, x_N = \tilde{x}_L}_{N-(t_1+\ldots+t_{L-1}) \text{ fois}}) =$$
$$\pi(\tilde{x}_1)d(\tilde{x}_1, t_1) \times \prod_{l=1}^{L-2} q(\tilde{x}_{l+1}|\tilde{x}_l)d(\tilde{x}_{l+1}, t_{l+1}) \times q(\tilde{x}_L|\tilde{x}_{L-1})D(\tilde{x}_L, N - (t_1 + \ldots + t_{L-1})) \quad (3.7)$$

Supposons que le processus X satisfasse les conditions 1. et 2. et montrons (3.7). La loi de tout processus X s'exprime en fonction de celles de T et de $\tilde{X}$ par :

$$p(\underbrace{x_1 = \tilde{x}_1, \ldots, x_{t_1} = \tilde{x}_1}_{t_1 \text{ fois}}, \underbrace{x_{t_1+1} = \tilde{x}_2, \ldots, x_{t_1+t_2} = \tilde{x}_2}_{t_2 \text{ fois}},$$
$$\ldots, \underbrace{x_{t_1+\ldots+t_{L-2}+1} = \tilde{x}_{L-1}, \ldots, x_{t_1+\ldots+t_{L-1}} = \tilde{x}_{L-1}}_{t_{L-1} \text{ fois}}, \underbrace{x_{t_1+\ldots+t_{L-1}+1} = \tilde{x}_L, \ldots, x_N = \tilde{x}_L}_{N-(t_1+\ldots+t_{L-1}) \text{ fois}}) =$$
$$\pi(\tilde{x}_1)p(t_1|\tilde{x}_1) \times \prod_{l=1}^{L-2} p(\tilde{x}_{l+1}|\tilde{x}_1, \ldots, \tilde{x}_l, t_1, \ldots, t_l)p(t_{l+1}|\tilde{x}_1, \ldots, \tilde{x}_{l+1}, t_1, \ldots, t_l)$$
$$\times p(\tilde{x}_L|\tilde{x}_1, \ldots, \tilde{x}_{L-1}, t_1, \ldots, t_{L-1})p(t_L \geq N - (t_1 + \ldots + t_{L-1})|\tilde{x}_1, \ldots, \tilde{x}_L, t_1, \ldots, t_{L-1}).$$

En utilisant les conditions 1. et 2., on en déduit le résultat. □

Chaînes semi-markoviennes comme généralisation des chaînes de Markov

La proposition suivante donne les conditions nécessaire et suffisante pour qu'une chaîne semi-markovienne soit une chaîne de Markov.

Proposition 3.3.2. *On reprend les notations de la définition 3.3.1. Une chaîne semi-markovienne $X = (X_n)_{n\geq 1}$ à valeurs dans un espace d'état fini $\mathcal{X}$ est une chaîne de Markov de matrice de transition $Q(x, x') = p\left(x_{n+1} = x' | x_n = x\right)$ si et seulement si :*

$$d(x, u) = Q(x, x)^{u-1}\left(1 - Q(x, x)\right) \text{ pour tout } x \text{ et pour tout } u \geq 1.$$

De plus, si cette condition est vraie, on a :

$$q(x'|x) = \frac{Q(x, x')}{1 - Q(x, x)} \text{ pour tout } x,\ x' \text{ tels que } x \neq x'.$$

Conditions de stationnarité d'une chaîne semi-markovienne

Nous donnons des conditions suffisantes pour qu'une chaîne semi-markovienne à valeurs dans un espace fini soit stationnaire. Dans [74] il y est étudié la stationnarité de chaînes semi-markoviennes dans un cadre plus général. Lorsqu'un processus $Z = (Z_n)_{n\geq 1}$ est stationnaire, nous appelerons "mesure stationnaire" la mesure de probabilité commune de la loi des Z_n. Si Z est une chaîne de Markov à espace d'états fini de transition Q, sa loi initiale π définit une mesure stationnaire si et seulement si $\pi Q = \pi$, auquel cas, π est appelée "mesure invariante". Soit (X, U) la chaîne de Markov introduite dans la définition 3.3.1 avec $\mathcal{X} = \{\omega_1, \ldots, \omega_K\}$. Supposons que (X, U) admette une mesure invariante. La mesure invariante μ de la chaîne (X, U) vérifie alors :

$$\begin{aligned}\mu(x_{n+1}, u_{n+1}) &= \sum_{x_n, u_n} \mu(x_n, u_n) p(x_{n+1}|x_n, u_n) p(u_{n+1}|x_{n+1}, u_n) \\ &= \sum_{x_n \neq x_{n+1}} \mu(x_n, 1) q(x_{n+1}|x_n) d(x_{n+1}, u_{n+1}) + \mu(x_{n+1}, u_{n+1} + 1). \qquad (3.8)\end{aligned}$$

La mesure μ^{semi} obtenue par marginalisation $\mu^{\text{semi}}(x) = \sum_u \mu(x, u)$ est une mesure stationnaire pour X. Celle-ci satisfait :

$$\mu^{\text{semi}}(x_{n+1}) = \sum_{x_n \neq x_{n+1}} \mu(x_n, 1) q(x_{n+1}|x_n) + \left(\mu^{\text{semi}}(x_{n+1}) - \mu(x_{n+1}, 1)\right).$$

Les $\mu_{x_n} = \mu(x_n, 1)$ sont alors solutions du système :

$$\mu(x_{n+1}, 1) = \sum_{x_n \neq x_{n+1}} \mu(x_n, 1) q(x_{n+1}|x_n). \qquad (3.9)$$

Ainsi la mesure $(\mu_{x_n})_{x_n \in \mathcal{X}}$ est une mesure invariante de la chaîne immergée. On peut également exprimer à partir de la formule (3.8), la mesure stationnaire μ^{semi} en fonction de la mesure

$(\mu_{x_n})_{x_n \in \mathcal{X}}$.
D'après (3.8), on a :

$$\mu(\omega_j, 2) = \mu_{\omega_j} - \sum_{i \neq j} \mu_{\omega_i} q(\omega_j|\omega_i) d(\omega_j, 1),$$

et comme les μ_{x_n} sont solutions de (3.9), on en déduit :

$$\mu(\omega_j, 2) = \mu_{\omega_j} \left(1 - d(\omega_j, 1)\right).$$

Procédant ainsi de manière récursive en exprimant $\mu(\omega_j, k)$ en fonction de $\mu(\omega_j, k-1)$ grâce à la formule (3.8), on en déduit que pour tout $k \geq 2$,

$$\mu(\omega_j, k) = \mu_{\omega_j} \left(1 - \sum_{l=1}^{k-1} d(\omega_j, l)\right).$$

La quantité $1 - \sum_{l=1}^{k-1} d(\omega_j, l)$ est la probabilité que le temps de séjour dans l'état ω_j soit supérieur ou égal à k. Par ailleurs, soit V_j la variable aléatoire de densité $d(\omega_j, .)$, on a alors :

$$\mu^{\text{semi}}(\omega_j) = \mu_{\omega_j} \sum_{k=1}^{+\infty} \mathbb{P}\left(V_j \geq k\right),$$

Remarquant que $\sum_{k=1}^{+\infty} \mathbb{P}\left(V_j \geq k\right) = \mathbb{E}(V_j)$, on a finalement la proposition suivante :

Proposition 3.3.3. *Soit la chaîne de Markov (X, U) dont la loi est définie par (3.3), (3.4), (3.5) et (3.6). Si (X, U) admet une mesure invariante μ, alors la chaîne immergée de la chaîne semi-markovienne X admet une mesure invariante $\tilde{\mu}$ donnée par :*

$$\tilde{\mu}(x) = \mu(x, 1) \text{ pour tout } x \in \mathcal{X}.$$

Définissons la mesure de probabilité π^{semi} par :

$$\pi^{semi}(x) \propto \tilde{\mu}(x)\mathbb{E}(V_x) \text{ pour tout } x \in \mathcal{X},$$

où V_x est une variable aléatoire de densité $u \in \mathbb{N}^* \to d(x, u)$. Si la loi de X_1 a pour distribution π^{semi}, alors la chaîne semi-markovienne est stationnaire.

Exemple 3.3.1. Supposons que la chaîne semi-markovienne X prenne deux valeurs ω_1 et ω_2 et supposons que (X, U) admette une mesure invariante. La chaîne immergée est alors de matrice de transition $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ et donc toute mesure invariante $\tilde{\mu}$ de la chaîne immergée $\tilde{X}$ satisfait $\tilde{\mu}(\omega_1) = \tilde{\mu}(\omega_2)$. Notons λ_j le temps de séjour moyen dans l'état ω_j, la probabilité π^{semi} définie par :

$$\begin{aligned} \pi^{\text{semi}}(\omega_1) &= \frac{\lambda_1}{\lambda_1 + \lambda_2} \\ \pi^{\text{semi}}(\omega_2) &= \frac{\lambda_2}{\lambda_1 + \lambda_2}, \end{aligned}$$

est une probabilité invariante pour la chaîne semi-markovienne X.

3.3.2 Etudes des chaînes semi-markoviennes à temps fini

Dans cette section, nous nous intéressons à des processus de la forme $X_{1:N} = (X_n)_{1\leq n\leq N}$. La proposition suivante permet de considérer un nombre fini de valeurs pour chaque U_n tout en conservant la loi voulue pour le processus $X_{1:N}$.

Proposition 3.3.4. *Soit (X, U) la chaîne de Markov définie par (3.4), (3.5) et (3.6). Notons $D(x,n) = \sum_{k\geq n} d(x,k)$ la probabilité que le temps de séjour dans l'état x soit supérieure ou égale à n.*
Le processus à temps fini $X_{1:N} = (X_1, \ldots, X_N)$ est marginal de la chaîne de Markov $(X_{1:N}, W_{1:N})$ à valeurs dans $\mathcal{X} \times \{1, \ldots, N\}$ dont la loi est donnée par :

- *Loi initiale :*

$$p(x_1, w_1) = \pi(x_1)p(w_1|x_1), \tag{3.10}$$

avec :

$$p(w_1|x_1) = \begin{cases} d(x_1, w_1) \text{ si } 1 \leq w_1 \leq N-1, \\ D(x_1, N) \text{ si } w_1 = N. \end{cases} \tag{3.11}$$

- *Transition :*

$$p(x_{n+1}, w_{n+1}|x_n, w_n) = p(x_{n+1}|x_n, w_n) \times p(w_{n+1}|x_{n+1}, w_n), \tag{3.12}$$

avec :

$$p(x_{n+1}|x_n, w_n) = \begin{cases} \delta_{x_n}(x_{n+1}) \text{ si } w_n > 1, \\ q(x_{n+1}|x_n) \text{ si } w_n = 1, \end{cases} \tag{3.13}$$

et

$$p(w_{n+1}|x_{n+1}, w_n) = \begin{cases} \delta_{w_n-1}(w_{n+1}) \text{ si } w_n > 1, \\ d(x_{n+1}, w_{n+1}) \text{ si } w_n = 1 \text{ et } 1 \leq w_{n+1} \leq N-1, \\ D(x_{n+1}, N) \text{ si } w_n = 1 \text{ et } w_{n+1} = N, \end{cases} \tag{3.14}$$

où $\delta_x(x) = 1$ et $\delta_x(x') = 0$ pour $x' \neq x$.

Preuve de la proposition 3.3.4. Notons $(X^W_{1:N}, W_{1:N})$ le processus dont la loi est définie par (3.10), (3.11), (3.12), (3.13) et (3.14).
Tout d'abord, intéressons nous à la loi de $X_{1:N}$. On notera L le nombre de valeurs visitées par la chaîne $X_{1:N}$ et $\tilde{x}_1, \ldots, \tilde{x}_L$ la suite des valeurs visitées. On a alors :

$$\begin{aligned} p(x_{1:N}) &= p(\underbrace{\tilde{x}_1, \ldots, \tilde{x}_1}_{u_1 \text{ fois}}, \underbrace{\tilde{x}_2, \ldots, \tilde{x}_2}_{u_2 \text{ fois}}, \ldots, \underbrace{\tilde{x}_L, \ldots, \tilde{x}_L}_{N-(u_1+\ldots+u_{L-1}) \text{ fois}}) \\ &= \pi(\tilde{x}_1) \prod_{j=1}^{L-1} q(\tilde{x}_{j+1}|\tilde{x}_j) d(\tilde{x}_j, u_j) \times D(\tilde{x}_L, N - (u_1 + \ldots + u_{L-1})). \end{aligned}$$

Concernant la loi de $X^W_{1:N}$, l'événement :

$$x^W_{1:N} = (\underbrace{\tilde{x}_1, \ldots, \tilde{x}_1}_{u_1 \text{ fois}}, \underbrace{\tilde{x}_2, \ldots, \tilde{x}_2}_{u_2 \text{ fois}}, \ldots, \underbrace{\tilde{x}_L, \ldots, \tilde{x}_L}_{N-(u_1+\ldots+u_{L-1}) \text{ fois}})$$

est la réunion des événements A_j avec $j \in \{0, \ldots, u_1 + \ldots + u_{L-1}\}$ suivants :

$$(x^W_1 = \tilde{x}_1, w_1 = u_1, x^W_{u_1+1} = \tilde{x}_2, w_{u_1+1} = u_2, \ldots, x^W_{u_1+\ldots+u_{L-2}+1} = \tilde{x}_{L-1}, w_{u_1+\ldots+u_{L-2}+1} = u_{L-1},$$
$$\ldots, x^W_{u_1+\ldots+u_{L-1}+1} = \tilde{x}_L, w_{u_1+\ldots+u_{L-1}+1} = N - (u_1 + \ldots + u_{L-1}) + j).$$

Pour tout $j < u_1 + \ldots + u_{L-1}$, la probabilité de l'événement A_j est égale à :

$$\pi(\tilde{x}_1) \prod_{l=1}^{L-1} q(\tilde{x}_l, \tilde{x}_{l+1}) d(\tilde{x}_l, u_l) \times d(\tilde{x}_L, N - (u_1 + \ldots + u_{L-1}) + j),$$

et pour $j = u_1 + \ldots + u_{L-1}$, est égale à :

$$\pi(\tilde{x}_1) \prod_{l=1}^{L-1} q(\tilde{x}_l, \tilde{x}_{l+1}) d(\tilde{x}_l, u_l) \times D(\tilde{x}_L, N).$$

La réunion des événements A_j est donc de probabilité :

$$\pi(\tilde{x}_1) \prod_{j=1}^{L-1} q(\tilde{x}_j, \tilde{x}_{j+1}) d(\tilde{x}_j, u_j) \times D(\tilde{x}_L, N - (u_1 + \ldots + u_{L-1})).$$

On conclut que $X_{1:N}$ et $X^W_{1:N}$ ont les mêmes lois. □

De cette proposition, on conclut que toute chaîne semi-markovienne peut être considérée comme la marginale d'une chaîne de Markov à espace d'état fini, pourvu que l'on ne considère qu'une partie finie $X_{1:N} = (X_n)_{1 \leq n \leq N}$ du processus. Ainsi un modèle latent $(X_n, Y_n)_{1 \leq n \leq N}$ tel que $(X_n)_{1 \leq n \leq N}$ soit une chaîne semi-markovienne peut être considérée comme une chaîne de Markov triplet telle que chaque (X_n, U_n) prend un nombre fini de valeurs, nous appelerons un tel modèle "chaîne semi-markovienne cachée". Cependant, même avec cette considération, l'algorithme de Baum-Welsh possède une forte complexité algorithmique, étant donné que si le processus X prend K états et le processus U prend N états, l'espace d'états a un cardinal égal à $N \times K$. Nous proposons dans la sous-section suivante un modèle semi-markovien particulier qui permet de palier ces difficultés algorithmiques.

3.3.3 Un modèle semi-markovien particulier

Dans cette sous-section, nous proposons un modèle semi-markovien particulier original. Dans ce nouveau modèle, on peut restreindre le nombre de valeurs du processus auxiliaire U tout en s'autorisant des temps de séjour arbitrairement longs. Soit $(X, U) = (X_n, U_n)_{1 \leq n \leq N}$ tel que chaque X_n prend ses valeurs dans $\mathcal{X} = \{\omega_1, \ldots, \omega_K\}$ et chaque U_n prend ses valeurs dans un ensemble fini Λ. On considère :

– π une distribution de probabilité sur $\mathcal{X}$;

– des distributions de probabilité $\bar{d}(x,.)$ sur Λ pour tout $x \in \mathcal{X}$;
– des transitions $r(.|.)$ sur $\mathcal{X}^2$.

La distribution $p(x_{1:N}, u_{1:N})$ de (X,U) est définie par :

$$p(x_{1:N}, u_{1:N}) = \pi(x_1)\bar{d}(x_1,u_1)\prod_{n=1}^{N-1} p(x_{n+1}|x_n,u_n)p(u_{n+1}|x_{n+1},u_n), \tag{3.15}$$

où :

$$p(x_{n+1}|x_n,u_n) = \begin{cases} \delta_{x_n}(x_{n+1}) \text{ si } u_n > 1, \\ r(x_{n+1}|x_n) \text{ si } u_n = 1, \end{cases} \tag{3.16}$$

et

$$p(u_{n+1}|x_{n+1},u_n) = \begin{cases} \delta_{u_n-1}(u_{n+1}) \text{ si } u_n > 1, \\ \bar{d}(x_{n+1},u_{n+1}) \text{ si } u_n = 1. \end{cases} \tag{3.17}$$

A la différence avec le modèle semi-markovien classique, ici la transition $r(x|x)$ peut très bien être non nulle.
Nous allons montrer que ce modèle est un cas particulier de modèle semi-markovien et nous allons exprimer la loi du temps de séjour qui nous sera utile par la suite.

Semi-markovianité de X et loi du temps de séjour

Introduisons la chaîne $\check{X}$ construite de la manière suivante :

$$\begin{aligned} &\check{X}_k = X_{n_k}, \quad \text{où} \\ &\quad n_1 = 1 \quad \text{et} \quad \forall k \geq 2,\ n_k = \min\{n \geq n_{k-1} : U_n = 1\} + 1. \end{aligned}$$

La chaîne $\check{X}$ est une chaîne de Markov de transition r et possède la même chaîne immergée que le processus X ; notons $\tilde{X}$ cette chaîne immergée. Soient $T = (T_n)_{n\geq 1}$ et $\check{T} = (\check{T}_n)_{n\geq 1}$ les processus respectifs de temps de séjour de X et $\check{X}$. Considérons également la chaîne $\check{U}$ définie par :

$$\begin{aligned} &\check{U}_k = U_{n_k}, \quad \text{où} \\ &\quad n_1 = 1 \quad \text{et} \quad \forall k \geq 2,\ n_k = \min\{n \geq n_{k-1} : U_n = 1\} + 1. \end{aligned}$$

L'événement $(T_n = s)$ est égal à la réunion disjointe :

$$(T_n = s) = \coprod_{k=1}^{s} \left[(\check{T}_n = k) \cap (\check{U}_{\check{T}_1+\ldots+\check{T}_{n-1}+1} + \check{U}_{\check{T}_1+\ldots+\check{T}_{n-1}+2} + \ldots + \check{U}_{\check{T}_1+\ldots+\check{T}_{n-1}+k} = s)\right].$$

D'après la définition du modèle, $p(\tilde{x}_{n+1}|\tilde{x}_1,\ldots,\tilde{x}_n,t_1,\ldots,t_n) = \dfrac{r(\tilde{x}_{n+1}|\tilde{x}_n)}{1-r(\tilde{x}_n|\tilde{x}_n)}$ et la propriété 1. de la proposition 3.3.1 est satisfaite. Les variables aléatoires $\check{U}_{\check{T}_1+\ldots+\check{T}_{n-1}+l}$ correspondent aux valeurs de U_k lorsque $U_{k-1} = 1$ et $k \in \{T_1+\ldots+T_{n-1}+1, T_1+\ldots+T_n\}$, elles sont donc indépendantes conditionnellement à $\tilde{X}_n$ et indépendantes de $\tilde{X}_1,\ldots,\tilde{X}_{n-1},T_1,\ldots,T_{n-1},\check{T}_n$ conditionnellement à $\tilde{X}_n$. De plus, la loi de $\check{U}_{\check{T}_1+\ldots+\check{T}_{n-1}+l}$ sachant $\tilde{X}_n = \tilde{x}_n$ est donnée par $\bar{d}(\tilde{x}_n,.)$. De même, $\check{T}_n$ est le temps de séjour de la chaîne de Markov $\check{X}$ dans l'état $\tilde{X}_n$, ainsi

$\check{T}_n$ est indépendante de $\tilde{X}_1, \ldots, \tilde{X}_{n-1}, T_1, \ldots, T_{n-1}$ conditionnellement à $\tilde{X}_n$. De plus, comme $\check{X}$ est une chaîne de Markov de transition r, la loi de $\check{T}_n$ conditionnellement à $\tilde{X}_n$ est donnée par $p(\check{t}_n|\tilde{x}_n) = (1 - r(\tilde{x}_n|\tilde{x}_n))r(\tilde{x}_n|\tilde{x}_n)^{\check{t}_n - 1}$. On en déduit que la propriété 2. de la proposition 3.3.1 est satisfaite. Ainsi, X est une chaîne semi-markovienne, sa transition et sa loi de durée sont données par :

$$\forall \omega_i \neq \omega_j,\ q(\omega_j|\omega_i) = \frac{r(\omega_j|\omega_i)}{1 - r(\omega_i|\omega_i)}\ ; \tag{3.18}$$

$$d(\omega_j, s) = (1 - r(\omega_j|\omega_j)) \sum_{k=1}^{s} r(\omega_j|\omega_j)^{k-1} \sum_{v_1 + \ldots + v_k = s} \bar{d}(\omega_j, v_1) \ldots \bar{d}(\omega_j, v_k) \text{ pour tout } s \geq 1. \tag{3.19}$$

L'expression ci-dessus ne permet pas de retrouver n'importe quelle loi de temps de séjour $d(\omega_j, .)$, le modèle ainsi défini est donc un modèle semi-markovien particulier. Par contre, il est possible de trouver des paramètres $(\bar{d}(\omega_j, .), r(\omega_j|\omega_j))$ tels que la loi du temps de séjour soit géométrique. Notre modèle reste plus général que le modèle de chaîne de Markov.

Identifiabilité du modèle

Dans ce paragraphe, nous abordons l'identifiabilité du modèle dans le cas où chaque U_n prend ses valeurs dans $\Lambda = \{1, 2\}$. Ainsi, les paramètres de notre modèle sont les $K(K-1)$ transitions $r(\omega_j|\omega_i)$ où $i \in \{1, \ldots, K\}$ et $j \in \{1, \ldots, K-1\}$ et les K paramètres $\bar{d}(\omega_j, 1)$. Le modèle étant inclus dans celui des chaînes semi-markoviennes, il est déterminé par la loi du temps de séjour donnée par $(d(\omega_j, .))_{1 \leq j \leq K}$ et par les transitions $(q(\omega_j|\omega_i))_{1 \leq i \leq K, 1 \leq j \leq K}$. Le modèle sera identifiable si pour une loi de temps de séjour et des transitions $q(\omega_j|\omega_i)$ données, il existe un unique jeu de paramètres $(\bar{d}(\omega_j, 1))_{1 \leq j \leq K}$, $(r(\omega_j|\omega_i))_{1 \leq i \leq K, 1 \leq j \leq K}$. De plus, lorsque $\omega_i \neq \omega_j$, $q(\omega_j|\omega_i) = \dfrac{r(\omega_j|\omega_i)}{1 - r(\omega_i|\omega_i)}$, car X et $\check{X}$ ont même chaîne immergée et $\check{X}$ est une chaîne de Markov de transition $(r(\omega_j|\omega_i))_{1 \leq i \leq K, 1 \leq j \leq K}$. Ainsi, la donnée de $q(\omega_j|\omega_i)$ et de $r(\omega_i|\omega_i)$ permet de déduire la valeur de $r(\omega_j|\omega_i)$. Le modèle sera donc identifiable si pour tout j, les probabilités $r(\omega_j|\omega_j)$ et $\bar{d}(\omega_j, 1)$ sont uniques pour une loi de temps de séjour et une transition $(q(\omega_j|\omega_i))_{1 \leq i \leq K, 1 \leq j \leq K}$ données.
Soient $(q(\omega_j|\omega_i))_{1 \leq i \leq K, 1 \leq j \leq K}$ et $d(\omega_j, .)_{1 \leq j \leq K}$ les paramètres définissant la loi d'une chaîne semi-markovienne. D'après (3.19), le modèle défini par les formules (3.15), (3.16) et (3.17) existera pour les paramètres $(q(\omega_j|\omega_i))_{1 \leq i \leq K, 1 \leq j \leq K}$ et $d(\omega_j, .)_{1 \leq j \leq K}$ si pour tout $j \in \{1, \ldots, K\}$ le système suivant :

$$S_j : \begin{cases} d(\omega_j, 1) = & (1 - r(\omega_j|\omega_j))\bar{d}(\omega_j, 1) \\ d(\omega_j, 2) = & (1 - r(\omega_j|\omega_j)) \left[1 - \bar{d}(\omega_j, 1) + r(\omega_j|\omega_j)\bar{d}(\omega_j, 1)^2\right] \\ \vdots & \\ d(\omega_j, 2m) = & (1 - r(\omega_j|\omega_j)) \left[\sum_{l=0}^{m} C_{m+l}^{2l} r(\omega_j|\omega_j)^{m+l-1} \bar{d}(\omega_j, 1)^{2l} (1 - \bar{d}(\omega_j, 1))^{m-l}\right] \\ d(\omega_j, 2m+1) = & (1 - r(\omega_j|\omega_j)) \left[\sum_{l=1}^{m+1} C_{m+l}^{2l-1} r(\omega_j|\omega_j)^{m+l-1} \bar{d}(\omega_j, 1)^{2l-1} (1 - \bar{d}(\omega_j, 1))^{m-l+1}\right] \\ \vdots & \end{cases}$$

a une solution. Il sera identifiable si pour tout $j \in \{1, \ldots, K\}$, la solution est unique.
On supposera dans la suite que pour tout j, $r(\omega_j|\omega_j) \neq 1$; dans le cas contraire, le temps de séjour dans l'état ω_j serait infini. Ainsi, en remplaçant dans la deuxième équation $\bar{d}(\omega_j, 1)$ par $\dfrac{d(\omega_j, 1)}{1 - r(\omega_j|\omega_j)}$, on prouve que toute solution du système S_j est solution du système suivant :

$$S'_j : \begin{cases} d(\omega_j, 1) = (1 - r(\omega_j|\omega_j))\bar{d}(\omega_j, 1) \\ r(\omega_j|\omega_j)^2 - (2 - d(\omega_j, 1) - d(\omega_j, 2) - d(\omega_j, 1)^2)\, r(\omega_j|\omega_j) + 1 - d(\omega_j, 1) - d(\omega_j, 2) = 0 \end{cases}$$

Soit f la fonction définie sur $[0, 1]$ par :

$$f(r) = r^2 - \left(2 - d(\omega_j, 1) - d(\omega_j, 2) - d(\omega_j, 1)^2\right) r + 1 - d(\omega_j, 1) - d(\omega_j, 2),$$

celle-ci est décroissante pour

$$r \leq 1 - \frac{d(\omega_j, 1) + d(\omega_j, 2) + d(\omega_j, 1)^2}{2},$$

et croissante pour

$$r \geq 1 - \frac{d(\omega_j, 1) + d(\omega_j, 2) + d(\omega_j, 1)^2}{2}.$$

Notons

$$r_{\min} = 1 - \frac{d(\omega_j, 1) + d(\omega_j, 2) + d(\omega_j, 1)^2}{2},$$

on a $r_{\min} \in [0, 1]$, $f(0) = 1 - d(\omega_j, 1) - d(\omega_j, 2) \geq 0$, $f(1) = d(\omega_j, 1)^2 \geq 0$ et

$$f(r_{\min}) = \left(\frac{3d(\omega_j, 1) + d(\omega_j, 2) + d(\omega_j, 1)^2}{2}\right)\left(\frac{d(\omega_j, 1) - d(\omega_j, 2) - d(\omega_j, 1)^2}{2}\right).$$

Nous avons trois possibilités :
- soit $d(\omega_j, 1)(1 - d(\omega_j, 1)) > d(\omega_j, 2)$, alors $f(r_{\min}) > 0$ et donc le système S'_j n'a pas de solution ; ce qui implique que le système S_j n'a pas de solution non plus ;
- soit $d(\omega_j, 1)(1 - d(\omega_j, 1)) = d(\omega_j, 2)$, alors $f(r_{\min}) = 0$. Si $f(0) > 0$ et $f(1) > 0$, le système S'_j a une unique solution. Cette solution est donnée par $r(\omega_j|\omega_j) = r_{\min} = 1 - d(\omega_j, 1)$ et $\bar{d}(\omega_j, 1) = 1$. Pour que cette solution soit également solution de S_j, nécessairement $d(\omega_j, .)$ doit être une loi géométrique. Si $f(1) = 0$, alors $d(\omega_j, 1) = 0$ et sous la condition $d(\omega_j, 1)(1 - d(\omega_j, 1)) = d(\omega_j, 2)$, on a $d(\omega_j, 2) = 0$. Une solution $(r(\omega_j|\omega_j), \bar{d}(\omega_1, 1))$ de S_j, si elle existe, doit nécessairement satisfaire $r(\omega_j|\omega_j) = 1$, ce qui est exclu. Si $f(0) = 0$, alors sous la condition $d(\omega_j, 1)(1 - d(\omega_j, 1)) = d(\omega_j, 2)$, on a $d(\omega_j, 1)d(\omega_j, 2) = d(\omega_j, 2)$. Ainsi, $d(\omega_j, 1) = 1$ ou $d(\omega_j, 2) = 0$ et donc nécessairement $d(\omega_j, 2) = 0$. Alors, si S_j a une solution $(r(\omega_j|\omega_j), \bar{d}(\omega_1, 1))$ celle-ci doit satisfaire $1 - \bar{d}(\omega_j, 1) + r(\omega_j|\omega_j)\bar{d}(\omega_j, 1)^2 = 0$. Cette dernière équation implique $\bar{d}(\omega_j, 1) = 1$ et $r(\omega_j|\omega_j) = 0$ et donc $d(\omega_j, 1) = 1$. On en déduit que $r_{\min} = 0$. Ainsi le système S'_j a une unique solution donnée par $r(\omega_j|\omega_j) = r_{\min} = 0$ et $\bar{d}(\omega_j, 1) = 1$. Sous la condition $d(\omega_j, 1) = 1$, la solution de S_j existe et est l'unique solution de S'_j ;
- soit $d(\omega_j, 1)(1 - d(\omega_j, 1)) < d(\omega_j, 2)$, alors $f(r_{\min}) < 0$ et le système S'_j a deux solutions

$(r_1(\omega_j|\omega_j), \bar{d}_1(\omega_j,1))$ et $(r_2(\omega_j|\omega_j), \bar{d}_2(\omega_j,1))$. On prouve que toute solution du système S_j est solution du système suivant :

$$S_j'' : \begin{cases} d(\omega_j,1) = (1-r(\omega_j|\omega_j))\bar{d}(\omega_j,1) \\ r(\omega_j|\omega_j)^2 - (2-d(\omega_j,1)-d(\omega_j,2)-d(\omega_j,1)^2)\, r(\omega_j|\omega_j) + 1 - d(\omega_j,1) - d(\omega_j,2) = 0 \\ d(\omega_j,1) r(\omega_j|\omega_j)^2 + [d(\omega_j,1)\,(d(\omega_j,1)-d(\omega_j,2)-1) - d(\omega_j,3)]\, r(\omega_j|\omega_j) + d(\omega_j,3) = 0 \end{cases}$$

et toute solution de S_j'' est solution de S_j'. Si aucune solution de S_j' n'est solution de S_j'', alors le système S_j n'a pas de solution. Si une seule solution de S_j' est solution de S_j'', alors S_j'' possède une unique solution. Notons $(r^0(\omega_j|\omega_j), \bar{d}^0(\omega_j,1))$ la solution de S_j''. Le système S_j possède une solution si et seulement si le temps de séjour $d(\omega_j,.)$ satisfait :

$$\begin{cases} d(\omega_j,1) = & (1-r^0(\omega_j|\omega_j))\bar{d}^0(\omega_j,1) \\ d(\omega_j,2) = & (1-r^0(\omega_j|\omega_j))\left[1-\bar{d}^0(\omega_j,1)+r^0(\omega_j|\omega_j)\bar{d}^0(\omega_j,1)^2\right] \\ \vdots & \\ d(\omega_j,2m) = & (1-r^0(\omega_j|\omega_j))\left[\sum_{l=0}^{m} C_{m+l}^{2l} r^0(\omega_j|\omega_j)^{m+l-1}\bar{d}^0(\omega_j,1)^{2l}(1-\bar{d}^0(\omega_j,1))^{m-l}\right] \\ d(\omega_j,2m+1) = & (1-r^0(\omega_j|\omega_j))\left[\sum_{l=1}^{m+1} C_{m+l}^{2l-1} r^0(\omega_j|\omega_j)^{m+l-1}\bar{d}^0(\omega_j,1)^{2l-1}(1-\bar{d}^0(\omega_j,1))^{m-l+1}\right] \\ \vdots & \end{cases}$$

dans ce cas la solution de S_j est l'unique solution $(r^0(\omega_j|\omega_j), \bar{d}^0(\omega_j,1))$ de S_j''. Si les deux solutions $(r_1(\omega_j|\omega_j), \bar{d}_1(\omega_j,1))$ et $(r_2(\omega_j|\omega_j), \bar{d}_2(\omega_j,1))$ de S_j' sont solutions de S_j'', alors les deux dernières équations de S_j'' d'inconnue $r(\omega_j|\omega_j)$ sont proportionnelles. Ainsi :

$$\begin{cases} d(\omega_j,3) = d(\omega_j,1)\,(1-d(\omega_j,1)-d(\omega_j,2)) \\ d(\omega_j,3) - d(\omega_j,1)\,(d(\omega_j,1)-d(\omega_j,2)-1) = d(\omega_j,1)\left(2-d(\omega_j,1)-d(\omega_j,2)-d(\omega_j,1)^2\right) \end{cases}$$

Ce dernier système implique que $d(\omega_j,1)\,(d(\omega_j,1)^2 - d(\omega_j,1) - d(\omega_j,2)) = 0$ et donc $d(\omega_j,1) = 0$. Ainsi, $\bar{d}(\omega_j,1) = 0$ ou $r(\omega_j|\omega_j) = 1$. La solution de S_j'' telle que $r(\omega_j|\omega_j) = 1$ est exclue. Ainsi $\bar{d}(\omega_j,1) = 0$, ce qui implique que le temps de séjour dans ω_j doit nécessairement être pair et le temps de séjour de la chaîne $(X_{2n+1})_{n\geq 0}$ dans ω_j doit nécessairement être une loi géométrique pour que cette solution soit également solution de S_j.
On en déduit que sous les conditions d'existence des paramètres $(\bar{d}(\omega_j,1))_{1\leq j\leq K}$ et $(r(\omega_j|\omega_i))_{1\leq i\leq K, 1\leq j\leq K}$, le modèle est identifiable.

Condition de markovianité

Nous donnons les conditions sur $(r(\omega_j|\omega_i))_{i\in\{1,\ldots,K\}, j\in\{1,\ldots,K\}}$ et $\bar{d}(\omega_j,.)_{j\in\{1,\ldots,K\}}$ pour que X soit une chaîne de Markov.

Proposition 3.3.5. *Soit (X,U) la chaîne de Markov définie par les formules (3.15), (3.16) et (3.17). La chaîne semi-markovienne X est une chaîne de Markov de matrice de transition Q si et seulement si :*

$$\begin{cases} \bar{d}(\omega_j,1) = 1 \\ r(\omega_j|\omega_i) = Q(\omega_i,\omega_j) \end{cases} \tag{3.20}$$

Preuve. Nous allons tout d'abord démontrer que X est une chaîne de Markov si et seulement si :

$$\forall j,\ \forall s \geq 1,\ \bar{d}(\omega_j, s) = \bar{d}(\omega_j, 1)(1 - \bar{d}(\omega_j, 1))^{s-1}. \tag{3.21}$$

Supposons d'abord que X soit une chaîne de Markov de transition Q.
Nous allons démontrer (3.21) par récurrence sur s. Il est clair que (3.21) est vraie pour $s = 1$. Supposons le vrai pour tout $k \leq s - 1$. On a alors :

$$d(\omega_j, s) = (1 - r(\omega_j|\omega_j)) \left[\bar{d}(\omega_j, s) + \sum_{k=2}^{s} r(\omega_j|\omega_j)^{k-1} \bar{d}(\omega_j, 1)^k \sum_{v_1+\ldots+v_k=s} (1 - \bar{d}(\omega_j, 1))^{s-k} \right]. \tag{3.22}$$

Par un calcul de dénombrement classique, on montre que le nombre d'éléments $(v_1, \ldots, v_k)$ tels que $v_j \geq 1$ pour tout j et $v_1 + \ldots + v_k = s$ vaut C_{s-1}^{k-1}. Ainsi la formule (3.22) devient :

$$d(\omega_j, s) = (1 - r(\omega_j|\omega_j)) \left[\bar{d}(\omega_j, s) + \sum_{k=2}^{s} C_{s-1}^{k-1} r(\omega_j|\omega_j)^{k-1} \bar{d}(\omega_j, 1)^k (1 - \bar{d}(\omega_j, 1))^{s-k} \right]. \tag{3.23}$$

Sous la condition de markovianité on montre que $r(\omega_j|\omega_j)\bar{d}(\omega_j, 1) = Q(\omega_j, \omega_j) - 1 + \bar{d}(\omega_j, 1)$, d'où :

$$d(\omega_j, s) = \frac{1 - Q(\omega_j, \omega_j)}{\bar{d}(\omega_j, 1)} \left[\bar{d}(\omega_j, s) + \bar{d}(\omega_j, 1) \sum_{k=1}^{s-1} C_{s-1}^{k} (Q(\omega_j, \omega_j) - 1 + \bar{d}(\omega_j, 1))^k (1 - \bar{d}(\omega_j, 1))^{s-1-k} \right],$$

et donc comme de plus $d(\omega_j, s) = Q(\omega_j, \omega_j)^{s-1}(1 - Q(\omega_j, \omega_j))$, alors :

$$\bar{d}(\omega_j, s) + \bar{d}(\omega_j, 1) \sum_{k=1}^{s-1} C_{s-1}^{k} (Q(\omega_j, \omega_j) - 1 + \bar{d}(\omega_j, 1))^k (1 - \bar{d}(\omega_j, 1))^{s-1-k} = Q(\omega_j, \omega_j)^{s-1} \bar{d}(\omega_j, 1),$$

ainsi :

$$\bar{d}(\omega_j, s) + \bar{d}(\omega_j, 1) \left[Q(\omega_j, \omega_j)^{s-1} - (1 - \bar{d}(\omega_j, 1))^{s-1} \right] = Q(\omega_j, \omega_j)^{s-1} \bar{d}(\omega_j, 1),$$

on en déduit le résultat.
Etudions maintenant la réciproque, supposons (3.21) vraie pour tout s et montrons alors que X est une chaîne de Markov. Pour cela, on introduit $Q(\omega_j, \omega_j)$ tel que :

$$d(\omega_j, 1) = (1 - r(\omega_j|\omega_j))\bar{d}(\omega_j, 1) = 1 - Q(\omega_j, \omega_j).$$

En remplaçant alors dans l'expression de $d(\omega_j, s)$, $\bar{d}(\omega_j, s)$ par $\bar{d}(\omega_j, 1)(1 - \bar{d}(\omega_j, 1))^{s-1}$, $r(\omega_j|\omega_j)\bar{d}(\omega_j, 1)$ par $Q(\omega_j, \omega_j) - 1 + \bar{d}(\omega_j, 1)$ et $1 - r(\omega_j|\omega_j)$ par $\dfrac{1 - Q(\omega_j, \omega_j)}{\bar{d}(\omega_j, 1)}$, on en déduit :

$$\begin{aligned} d(\omega_j, s) &= \frac{1 - Q(\omega_j, \omega_j)}{\bar{d}(\omega_j, 1)} \left[\bar{d}(\omega_j, 1) \sum_{k=0}^{s-1} C_{s-1}^{k} (Q(\omega_j, \omega_j) - 1 + \bar{d}(\omega_j, 1))^k (1 - \bar{d}(\omega_j, 1))^{s-1-k} \right] \\ &= (1 - Q(\omega_j, \omega_j)) Q(\omega_j, \omega_j)^{s-1}, \end{aligned}$$

d'où le résultat.

Si $\Lambda = \{1, \dots, L\}$, on a $\sum_{s=1}^{L} \bar{d}(\omega_j, s) = 1$. Si X est de Markov de transition Q, comme $\bar{d}(\omega_j, s) = \bar{d}(\omega_j, 1)(1 - \bar{d}(\omega_j, 1))^{s-1}$, ainsi $\left(1 - \bar{d}(\omega_j, 1)\right)^L = 0$ et donc $\bar{d}(\omega_j, 1) = 1$ et par conséquent $r(\omega_j|\omega_i) = Q(\omega_i, \omega_j)$. Réciproquement, si $\bar{d}(\omega_j, 1) = 1$, X est une chaîne de Markov de transition $(r(\omega_j|\omega_i))_{1 \leq i,j \leq K}$. □

3.3.4 Inférence dans le modèle semi-markovien

Nous présentons maintenant l'estimation des états cachés et des paramètres dans le cas des chaînes semi-markoviennes cachées. De manière analogue aux chaînes de Markov cachées, les chaînes semi-markoviennes cachées sont représentées par un couple de processus $(X, Y) = (X_n, Y_n)_{1 \leq n \leq N}$ où la réalisation $y_{1:N}$ de Y est observée et celle de X, soit $x_{1:N}$, doit être estimée. Dans le modèle de chaînes semi-markoviennes cachées, X est une chaîne semi-markovienne et donc la marginale d'une chaîne de Markov (X, U). En raison de la complexité algorithmique nous limiterons le nombre de valeurs prises par U à 10 et nous utiliserons le modèle semi-markovien particulier défini par les formules (3.15), (3.16) et (3.17). Le processus (X, U) étant alors une chaîne de Markov à espace d'états fini, on peut alors utiliser l'algorithme de Baum-Welsh pour estimer les états cachés. Nous allons commencer par étudier les chaînes semi-markoviennes cachées à bruit indépendant, puis nous les étendrons aux chaînes semi-markoviennes M-stationnaires à bruit indépendant, ce qui suppose l'introduction d'un deuxième processus auxiliaire.

Algorithme de Baum-Welsh dans les chaînes semi-markoviennes

Soit $(X, Y) = (X_n, Y_n)_{1 \leq n \leq N}$ la chaîne de Markov définie par (3.15), (3.16) et (3.17), où chaque X_n prend ses valeurs dans $\mathcal{X} = \{\omega_1, \dots, \omega_K\}$ et chaque U_n prend ses valeurs dans $\Lambda = \{1, \dots, L\}$. La distribution d'une chaîne semi-markovienne cachée à bruit indépendant $(X, U, Y) = (X_n, U_n, Y_n)_{n \in \{1,\dots,N\}}$ s'écrit :

$$p(x_{1:N}, u_{1:N}, y_{1:N}) = \pi(x_1)\bar{d}(x_1, u_1)p(y_1|x_1) \prod_{n=1}^{N-1} p(x_{n+1}|x_n, u_n)p(u_{n+1}|x_{n+1}, u_n)p(y_{n+1}|x_{n+1}),$$

où $p(x_{n+1}|x_n, u_n)$ et $p(u_{n+1}|x_{n+1}, u_n)$ sont donnés par les formules (3.16) et (3.17). Nous détaillerons uniquement la version classique de l'algorithme de Baum-Welsh, la version conditionnelle s'en déduisant sans difficulté particulière.

Remarquons que si $u_n > 1$ la transition $p(x_{n+1}, u_{n+1}|x_n, u_n)$ est non nulle seulement lorsque $x_{n+1} = x_n$ et $u_{n+1} = u_n - 1$; l'algorithme de Baum-Welsh est ainsi de complexité algorithmique plus faible que dans le cas d'une chaîne de Markov triplet générale pour laquelle X_n prend K valeurs et U_n prend L valeurs. Plus exactement, on a :

– Etape directe :

Pour $u_{n+1} \leq L - 1$:

$$\alpha_{n+1}(x_{n+1}, u_{n+1}) =$$

$$\sum_{x_n} \alpha_n(x_n, 1)r(x_{n+1}|x_n)\bar{d}(x_{n+1}, u_{n+1})p(y_{n+1}|x_{n+1}) + \alpha_n(x_{n+1}, u_{n+1} + 1)p(y_{n+1}|x_{n+1}),$$

soit $K \times (L-1)$ quantités à calculer nécessitant $K+1$ sommations.
Pour $u_{n+1} = L$:

$$\alpha_{n+1}(x_{n+1}, L) = \sum_{x_n} \alpha_n(x_n, 1) r(x_{n+1}|x_n) \bar{d}(x_{n+1}, L) p(y_{n+1}|x_{n+1}),$$

soit K quantités à calculer nécessitant K sommations.

– Etape rétrograde :
Pour $u_n > 1$:

$$\beta_n(x_n, u_n) = \beta_{n+1}(x_n, u_n - 1) p(y_{n+1}|X_{n+1} = x_n),$$

soit $K \times (L-1)$ quantités à calculer nécessitant 1 sommation.
Pour $u_n = 1$:

$$\beta_n(x_n, 1) = \sum_{x_{n+1}, u_{n+1}} \beta_{n+1}(x_{n+1}, u_{n+1}) r(x_{n+1}|x_n) \bar{d}(x_{n+1}, u_{n+1}) p(y_{n+1}|x_{n+1}),$$

soit K quantités à calculer nécessitant $K \times L$ sommations.

La complexité algorithmique de l'algorithme de Baum-Welsh est alors égale à $\mathcal{O}(N \times K^2 \times L)$ au lieu de $\mathcal{O}(N \times K^2 \times L^2)$ dans le cas des chaînes de Markov triplets générales et au lieu de $\mathcal{O}(N^2 \times K^2)$ si on avait utilisé le modèle semi-markovien caché général.
Nous avons classiquement :

$$p(x_n, u_n|y_{1:N}) \propto \alpha_n(x_n, u_n) \beta_n(x_n, u_n),$$

et

$$p(x_{n+1}, u_{n+1}|x_n, u_n, y_{1:N}) = \frac{\beta_{n+1}(x_{n+1}, u_{n+1})}{\beta_n(x_n, u_n)} p(x_{n+1}|x_n, u_n) p(u_{n+1}|x_{n+1}, u_n) p(y_{n+1}|x_{n+1}).$$

Les transitions vérifient :

$$p(x_{n+1}, u_{n+1}|x_n, u_n, y_{1:N}) \propto \left\{ \begin{array}{c} \delta_{x_n}(x_{n+1}) \delta_{u_n - 1}(u_{n+1}) \text{ si } u_n > 1 \ ; \\ \beta_{n+1}(x_{n+1}, u_{n+1}) r(x_{n+1}|x_n) \bar{d}(x_{n+1}, u_{n+1}) p(y_{n+1}|x_{n+1}) \text{ si } u_n = 1. \end{array} \right.$$

On remarque que la distribution $p(x_{1:N}, u_{1:N}|y_{1:N})$ peut être donnée par les formules (3.15), (3.16) et (3.17), avec $r(x_{n+1}|x_n)$ et $\bar{d}(x_n, .)$ à remplacer par :

$$r(x_{n+1}|x_n, y_{1:N}) \propto r(x_{n+1}|x_n) p(y_{n+1}|x_{n+1}) \sum_{u_{n+1}} \beta_{n+1}(x_{n+1}, u_{n+1}) \bar{d}(x_{n+1}, u_{n+1}),$$

et

$$\bar{d}(x_{n+1}, u_{n+1}|y_{1:N}) \propto \beta_{n+1}(x_{n+1}, u_{n+1}) \bar{d}(x_{n+1}, u_{n+1}).$$

On en déduit que la distribution $p(x_{1:N}|y_{1:N})$ est également celle d'une chaîne semi-markovienne.

Estimation des paramètres du modèle

Dans ce paragraphe, nous abordons l'estimation des paramètres de $p(x_{1:N}, u_{1:N})$, l'estimation de la loi conditionnelle $p(y_{1:N}|x_{1:N})$ ayant été décrite dans la section 2.3 du chapitre 2. Dans les expérimentations, nous utiliserons l'algorithme ICE. La chaîne de Markov (X, U) sera stationnaire. La mesure invariante π de (X, U) sera obtenue en résolvant $\pi Q_{X,U} = \pi$, $Q_{X,U}$ désignant la transition de (X, U). Les estimations de $\bar{d}(\omega_j, .)$ et de $r(\omega_j|\omega_i)$ seront obtenues à partir de celle de $p(x_n, u_n = 1, x_{n+1}, u_{n+1})$. L'estimateur de $p(x_n, u_n = 1, x_{n+1}, u_{n+1})$ à partir des données complètes $(x_{1:N}, u_{1:N}, y_{1:N}) = (x_n, u_n, y_n)_{1 \leq n \leq N}$ est donné par :

$$\hat{p}(x_n = \omega_i, u_n = 1, x_{n+1} = \omega_j, u_{n+1} = u)(x_{1:N}, u_{1:N}, y_{1:N}) = \frac{1}{N-1} \sum_{n=1}^{N-1} I(x_n = \omega_i, u_n = 1, x_{n+1} = \omega_j, u_{n+1} = u).$$

Soit θ_q le vecteur paramètre obtenu à l'étape q de ICE, l'espérance sachant $y_{1:N}$ sous le paramètre θ_q est alors calculable et donne :

$$p(x_n = \omega_i, u_n = 1, x_{n+1} = \omega_j, u_{n+1} = u; \theta_{q+1}) = \frac{1}{N-1} \sum_{n=1}^{N-1} p(x_n = \omega_i, u_n = 1, x_{n+1} = \omega_j, u_{n+1} = u|y_{1:N}; \theta_q).$$

Les ré-estimées $\bar{d}^{q+1}(\omega_j, u)$ et $r^{q+1}(\omega_j|\omega_i)$ de $\bar{d}(\omega_j, u)$ et $r(\omega_j|\omega_i)$ sont données par :

$$\bar{d}^{q+1}(\omega_j, u) = \frac{\sum_{\omega_i} p(x_n = \omega_i, u_n = 1, x_{n+1} = \omega_j, u_{n+1} = u; \theta_{q+1})}{\sum_{\omega_i, u} p(x_n = \omega_i, u_n = 1, x_{n+1} = \omega_j, u_{n+1} = u; \theta_{q+1})},$$

$$r^{q+1}(\omega_j|\omega_i) = \frac{\sum_{u} p(x_n = \omega_i, u_n = 1, x_{n+1} = \omega_j, u_{n+1} = u; \theta_{q+1})}{\sum_{\omega_j, u} p(x_n = \omega_i, u_n = 1, x_{n+1} = \omega_j, u_{n+1} = u; \theta_{q+1})}.$$

3.3.5 Expérimentations

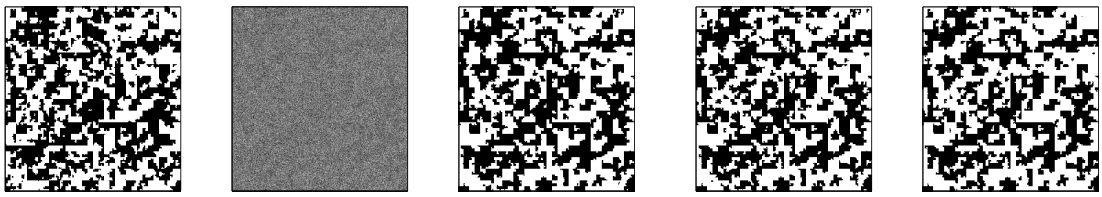
Nous présentons dans cette sous-section quatres expériences. Dans la première expérience, les données sont issues du modèle de chaînes de Markov cachées et segmentées suivant les modèles de chaînes de Markov cachées et de chaînes semi-markoviennes cachées. Dans la seconde expérience, les données sont issues du modèle de chaînes semi-markoviennes cachées et segmentées, comme ci-dessus, par les mêmes deux méthodes. Le but de cette deuxième expérience est de savoir si négliger la semi-markovianité a des conséquences sur la qualité de la segmentation. Dans la troisième expérience, les données suivent le modèle semi-markovien M-stationnaire et nous les segmentons suivant le modèle de chaînes de Markov cachées M-stationnaires et le modèle de chaînes semi-markoviennes cachées M-stationnaires. Dans la dernière expérience, nous appliquons ces deux derniers modèles sur une image réelle. Dans toutes les expériences, la taille des processus est égale à $N = 256 \times 256$ et les processus monodimensionnels sont transformés en images bi-dimensionnelles grâce au parcours d'Hilbert-Peano figurant en annexe B. Le nombre d'itérations de l'algorithme ICE est égal à 100 et les

paramètres sont initialisés par la méthode du K-means. Les modèles présentés dans toute la sous-section seront notés :

- chaînes de Markov cachées à bruit indépendant : CMC ;
- chaînes de Markov cachées M-stationnaires : CMC-M-S ;
- chaînes semi-markoviennes cachées à bruit indépendant : CSMC ;
- chaînes semi-markoviennes cachées M-stationnaires : CSMC-M-S.

Segmentation de données issues d'un modèle de chaîne de Markov cachées à bruit indépendant

Dans ce paragraphe, les données sont issues d'un modèle de chaîne de Markov cachées à bruit indépendant. Les états cachés sont ensuite estimés en utilisant les modèles CMC et CSMC. La matrice de transition de la chaîne de Markov X est $Q = \begin{pmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{pmatrix}$ et $p(y_n|x_n)$ suit une loi normale $\mathcal{N}_{\mathbb{R}}(0, 10)$ lorsque $x_n = \omega_1$ et $\mathcal{N}_{\mathbb{R}}(1, 10)$ lorsque $x_n = \omega_2$.



(a) Réalisation de X. (b) Réalisation de Y. (c) Restauration avec les vrais paramètres : 17.22% d'erreur. (d) Par CMC : 17.66% d'erreur. (e) Par CSMC : 17.91% d'erreur.

FIG. 3.1 – Simulation d'une chaîne de Markov cachée et sa restauration.

Modèle	Moyennes		Variances		Taux d'erreur
	ω_1	ω_2	ω_1	ω_2	
CMC	-0.04	1.00	9.99	9.98	17.66%
CSMC	-0.05	0.96	9.82	10.16	17.91%
Vraies valeurs	0	1	10	10	17.22%

TAB. 3.1 – Estimation des paramètres de la loi d'observation.

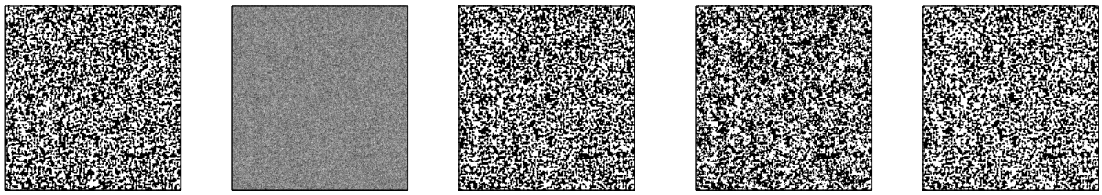
Concernant l'estimation de Q, le modèle CMC donne $\hat{Q}_{\text{CMC}} = \begin{pmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{pmatrix}$. L'estimation de la transition $r(\omega_j|\omega_i)$ par CSMC avec $L = 10$ est donnée par la matrice $\begin{pmatrix} 0.98 & 0.02 \\ 0.02 & 0.98 \end{pmatrix}$ et celle de $\bar{d}$ est donnée par :

$$\begin{aligned} \hat{\bar{d}}(\omega_1, .) &= (0.82, 0.01, 0.0, 0.04, 0.01, 0.04, 0.03, 0.01, 0.04, 0), \\ \hat{\bar{d}}(\omega_2, .) &= (0.85, 0, 0.01, 0.01, 0.01, 0.02, 0.05, 0.02, 0.02, 0.01). \end{aligned}$$

Des résultats de l'estimation, nous voyons que le modèle semi-markovien caché est capable de retrouver la markovianité des données. En effet la probabilité $\hat{\bar{d}}(\omega_j, u)$ est élevée pour $u = 1$ et faible pour les autres états. Par conséquent, les estimations des paramètres de la loi d'observation par le modèle CMC et par le modèle CSMC donnent des résultats comparables. Ainsi l'utilisation des CSMC à la place des CMC ne dégrade pas, dans le cadre de notre étude, la qualité des résultats en non supervisé. Notons également le degré élevé du bruitage et le très bon comportement de ICE.

Segmentation de données issues d'un modèle de chaînes semi-markoviennes cachées

Considérons une CSMC telle que chaque X_n prend ses valeurs dans $\mathcal{X} = \{\omega_1, \omega_2\}$, chaque U_n prend ses valeurs dans $\Lambda = \{1, \ldots, 10\}$, la transition $r(x_{n+1}|x_n)$ est donnée par la matrice $R = \begin{pmatrix} 0.3 & 0.7 \\ 0.7 & 0.3 \end{pmatrix}$ et pour tout x_{n+1}, $\bar{d}(x_{n+1}, 10) = \dfrac{3}{4}$ et $\bar{d}(x_{n+1}, u_{n+1}) = \dfrac{1}{36}$ si $u_{n+1} \neq 10$. Les lois d'observation $p(y_n|x_n)$ sont les lois normales $\mathcal{N}_{\mathbb{R}}(0, 2)$ lorsque $x_n = \omega_1$ et $\mathcal{N}_{\mathbb{R}}(1, 2)$ lorsque $x_n = \omega_2$.



(a) Réalisation de X. (b) Réalisation de Y. (c) Restauration avec les vrais paramètres : 21.39% d'erreur. (d) Par CMC : 24.68% d'erreur. (e) Par CSMC : 21.76% d'erreur.

FIG. 3.2 – Simulation d'une chaîne semi-markovienne cachée et sa restauration.

Des résultats de la segmentation présentés dans la figure 3.2, nous constatons que, dans le cadre de notre étude, les CMC sont relativement robustes par rapport à la semi-markovianité des données cachées ; cependant, la différence entre les résultats n'est pas tout à fait négligeable.

Modèle	Moyennes		Variances		Taux d'erreur
	ω_1	ω_2	ω_1	ω_2	
CMC	-0.04	1.06	1.97	1.95	24.68%
CSMC	-0.03	0.96	2.02	2.00	21.76%
Vraies valeurs	0	1	2	2	21.39%

TAB. 3.2 – Estimation des paramètres de la loi d'observation.

L'estimation de R par le modèle semi-markovien caché est donnée par :

$$\hat{R}_{\text{CSMC-BI}} = \begin{pmatrix} 0.29 & 0.71 \\ 0.71 & 0.29 \end{pmatrix}, \tag{3.24}$$

et celle de $\bar{d}(x_{n+1}, u_{n+1})$ est donnée par :

$$\begin{aligned}\hat{\bar{d}}(\omega_1, .) &= (0.00, 0.04, 0.00, 0.05, 0.00, 0.04, 0.08, 0.00, 0.00, 0.79)\,,\\ \hat{\bar{d}}(\omega_2, .) &= (0.01, 0.04, 0.00, 0.08, 0.00, 0.01, 0.03, 0.00, 0.02, 0.81)\,.\end{aligned}$$

Nous constatons un bon comportement de l'ICE, en particulier dans l'estimation des paramètres du bruit.

Segmentation de données issues du modèle semi-markovien caché M-stationnaire

Dans le modèle étudié lors de cette expérience, nous introduisons deux processus auxiliaires U^1 et U^2. Le processus U^1 modélise la semi-markovianité et le second processus U^2 modélise la M-stationnarité. Nous supposons que chaque X_n prend ses valeurs dans $\mathcal{X} = \{\omega_1, \ldots, \omega_K\}$, chaque U_n^1 prend ses valeurs dans $\Lambda_1 = \{1, \ldots, L\}$, et chaque U_n^2 prend ses valeurs dans $\Lambda_2 = \{\lambda_1, \ldots, \lambda_M\}$. Le modèle semi-markovien M-stationnaire que nous étudions est donné par :

$$\begin{aligned}p(x_{1:N}, u^1_{1:N}, u^2_{1:N}, y_{1:N}) &= p(u_1^2)p(x_1|u_1^2)p(u_1^1|x_1, u_1^2)p(y_1|x_1)\\ &\times \prod_{n=1}^{N-1} p(u^2_{n+1}|u^1_n, u^2_n)p(x_{n+1}|x_n, u^1_n, u^2_{n+1})p(u^1_{n+1}|x_{n+1}, u^1_n, u^2_{n+1})\\ &\times \prod_{n=1}^{N-1} p(y_{n+1}|x_{n+1}),\end{aligned} \tag{3.25}$$

où

$$p(u^2_{n+1}|u^1_n, u^2_n) = \left\{ \begin{array}{c} \delta_{u^2_n}(u^2_{n+1}) \text{ si } u^1_n > 1,\\ p(u^2_{n+1}|u^2_n) \text{ si } u^1_n = 1,\end{array}\right. \tag{3.26}$$

$$p(x_{n+1}|x_n, u^1_n, u^2_{n+1}) = \left\{ \begin{array}{c} \delta_{x_n}(x_{n+1}) \text{ si } u^1_n > 1,\\ p(x_{n+1}|x_n, u^2_{n+1}) \text{ si } u^1_n = 1,\end{array}\right. \tag{3.27}$$

et

$$p(u^1_{n+1}|x_{n+1}, u^1_n, u^2_{n+1}) = \left\{ \begin{array}{c} \delta_{u^1_n - 1}(u^1_{n+1}) \text{ si } u^1_n > 1,\\ p(u^1_{n+1}|x_{n+1}, u^2_{n+1}) \text{ si } u^1_n = 1.\end{array}\right. \tag{3.28}$$

Le but de cette expérience est de montrer l'importance simultanée de la M-stationnarité et de la semi-markovianité lorsque les données simulées sont issues d'un modèle semi-markovien caché M-stationnaire. Les données sont simulées selon le modèle de chaînes semi-markoviennes cachées M-stationnaires avec $K = 2$, $M = 2$ et $L = 10$ et :

– $p(u^2_{n+1}|u^2_n)$ a pour matrice de transition :

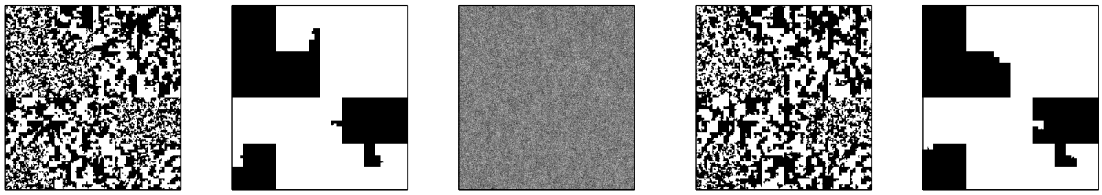
$$S = \left(\begin{array}{cc} 0.9999 & 0.0001\\ 0.0001 & 0.9999\end{array}\right)\,;$$

– pour tout u^2_{n+1}, $p(x_{n+1}|x_n, u^1_n = 1, u^2_{n+1})$ a pour matrice de transition :

$$Q = \left(\begin{array}{cc} 0.9 & 0.1\\ 0.1 & 0.9\end{array}\right)\,;$$

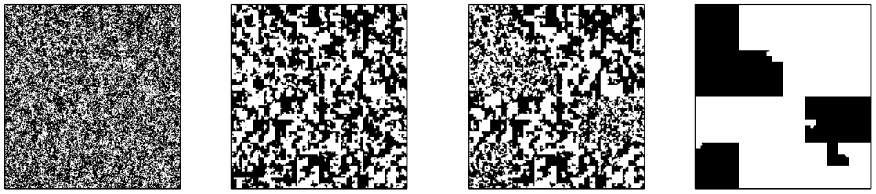
– si $u^2_{n+1} = \lambda_1$, alors pour tout x_{n+1}, $p(u^1_{n+1}|x_{n+1}, u^2_{n+1}, u^1_n = 1) = \delta_1(u^1_{n+1})$;
– si $u^2_{n+1} = \lambda_2$, alors pour tout x_{n+1}, $p(u^1_{n+1}|x_{n+1}, u^2_{n+1}, u^1_n = 1)$ est la distribution d'une loi uniforme sur Λ_1 ;
– $p(y_n|x_n)$ est la densité d'une loi normale $\mathcal{N}_{\mathbb{R}}(0, 5)$ si $x_n = \omega_1$ et $\mathcal{N}_{\mathbb{R}}(1, 5)$ si $x_n = \omega_2$.

Nous choisissons ensuite d'estimer les paramètres par les méthodes ICE correspondantes et les états cachés par le MPM en utilisant les modèles CMC, CMC-M-S, CSMC et CSMC-M-S.

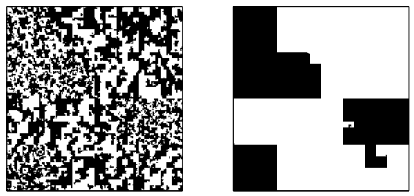


(a) Réalisation de X. (b) Réalisation de U^2. (c) Réalisation de Y. (d) Estimation de $x_{1:N}$: 23.32% d'erreur. (e) Estimation de $u^2_{1:N}$: 2.23% d'erreur.

FIG. 3.3 – Simulation d'une chaîne semi-markovienne cachée M-stationnaire et sa restauration utilisant les vrais paramètres.



(a) Estimation de $x_{1:N}$ par CMC : 36.3% d'erreur. (b) Estimation de $x_{1:N}$ par CSMC : 24.2% d'erreur. (c) Estimation de $x_{1:N}$ par CMC-M-S : 23.43% d'erreur. (d) Estimation de $u^2_{1:N}$ par CMC-M-S : 2.28% d'erreur.



(e) Estimation de $x_{1:N}$ par CSMC-M-S : 23.39% d'erreur. (f) Estimation de $u^2_{1:N}$ par CSMC-M-S : 2.31% d'erreur.

FIG. 3.4 – Restauration en utilisant les modèles CMC, CMC-M-S, CSMC et CSMC-M-S.

Modèle	Moyenne		Variance		Taux d'erreur pour X	Taux d'erreur pour U_2
	ω_1	ω_2	ω_1	ω_2		
CMC	-0.30	1.46	4.49	4.12	36.3%	-
CMC-M-S	0.00	1.00	4.92	5.03	23.43%	2.28%
CSMC	0.02	0.99	4.87	5.12	24.2%	-
CSMC-M-S	0.02	1.01	4.96	5.01	23.39%	2.31%
Vraies valeurs	0	1	5	5	-	-

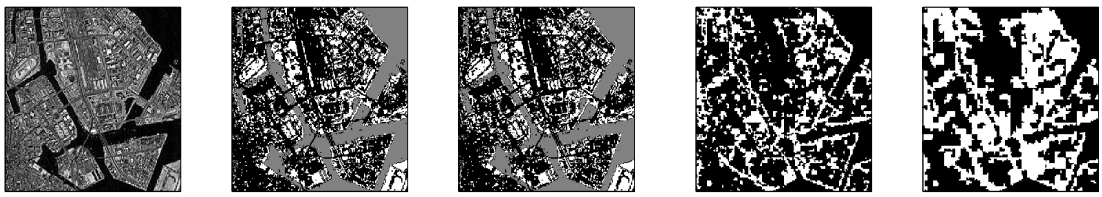
TAB. 3.3 – Estimation des paramètres de la loi d'observation.

De cette expérience, nous constatons que négliger la semi-markovianité du processus caché est peu pénalisant lorsque l'on a tenu compte de la M-stationnarité. En effet les résultats obtenus en segmentant par le modèle CMC-M-S et par le modèle CSMC-M-S sont comparables. Cependant négliger l'hypothèse de M-stationnarité est plus pénalisant mais nous pouvons constater des figures 3.4,(a) et (b) que le modèle semi-markovien estime mieux les états cachés que celui des chaînes de Markov cachées.

Notons que si les taux d'erreur sont comparables dans les segmentations (b) et (c) à la figure 3.4, la (c), qui utilise le vrai modèle, est plus proche de la vraie image sur la figure 3.3. Quant à l'estimation des paramètres par ICE présentée dans le tableau 3.3, nous constatons des résultats similaires pour les 4 modèles utilisés.

Segmentation d'une image réelle

Dans cette dernière expérience, nous proposons de segmenter une image réelle SAR (Synthetic Aperture Radar). Nous comparons les modèles CMC-M-S et CSMC-M-S sur la photographie aérienne de Tokyo présentée à la Figure 3.5,(a). Dans cette expérience, on prendra $K = 3$, $M = 2$ et $L = 5$.



(a) Image observée. (b) Estimation de $x_{1:N}$ par CMC-M-S. (c) Estimation de $x_{1:N}$ par CSMC-M-S. (d) Estimation de $u^2_{1:N}$ par CMC-M-S. (e) Estimation de $u^2_{1:N}$ par CSMC-M-S.

FIG. 3.5 – Segmentation d'une image réelle SAR.

De cette expérience, on voit des résultats similaires pour l'estimation de x. Cependant, l'estimation de u^2 semble être meilleure en utilisant le modèle semi-markovien caché M-stationnaire. Le modèle semi-markovien est capable de considérer des lois de temps de séjour plus générales que le modèle markovien, ce qui peut expliquer la qualité de l'estimation de

u^2. Dans la thèse de P. Lanchantin [65], cette image a été segmentée par les modèles CMC et CMC-M-S et il y est constaté des améliorations notables apportées par le modèle CMC-M-S, par rapport au modèle CMC, au niveau des estimations des paramètres.

Conclusion

Nous avons utilisé des modèles de Markov triplets particuliers pour proposer dans ce chapitre un modèle semi-markovien caché original, permettant des gains notables de temps de calcul. Nous avons ensuite étendu le modèle des chaînes de Markov cachées M-stationnaires introduit par P. Lanchantin [65] à celui des chaînes semi-markoviennes cachées M-stationnaires. Pour cela, nous avons considéré deux processus auxiliaires ; le premier modélisant la semi-markovianité du processus caché, et le second sa non stationnarité. Le modèle proposé fait ainsi partie des chaînes triplets "multi-variées", où la chaîne auxiliaire est composée de plus d'un processus, chacune des composantes modélisant une propriété particulière. Nous avons proposé une méthode ICE d'estimation des paramètres adéquate et les divers algorithmes de segmentation non supervisée, correspondants aux divers modèles, ont été testés via les simulations informatiques dans le contexte de bruits importants. Les résultats obtenus montrent que les deux chaînes auxiliaires sont nécessaires et l'absence d'aucune d'entre elles ne peut être compensée par l'autre. Ils ont également mis en évidence le très bon comportement de l'algorithme ICE, qui peut alors être utilisé dans des situations très bruitées.

Chapitre 4

Modèles de Markov triplets à observations non gaussiennes

Dans le chapitre précédent, nous nous sommes concentrés sur la loi du processus caché et nous avons étendu le modèle de chaînes de Markov cachées à celui des chaînes semi-markoviennes cachées M-stationnaires. Dans le présent chapitre, nous étudions la loi d'observation. Pour cela, nous rappelons diverses lois non gaussiennes telles que les lois de type exponentiel, elliptiques ou les "Vecteurs aléatoires sphériquement invariants" (abréviation anglaise SIRV) fréquemment rencontrés en traitement du signal radar. Nous nous intéressons ensuite aux copules, qui sont tout d'abord présentées brièvement dans leur contexte historique. Ensuite nous décrivons leur introduction récente dans le contexte des chaînes de Markov cachées et couples [21, 23, 24, 98], qui permettent la conception de très nombreux modèles particuliers du bruit. Pour finir, nous présentons dans la dernière sous-section un modèle triplet de chaînes de Markov non stationnaires cachées avec du bruit corrélé non gaussien original. Une méthode originale d'estimation des paramètres de type ICE est proposée et la méthode non supervisée correspondante de segmentation est validée par des expériences informatiques. En particulier, nous montrons l'importance du choix de la bonne copule ; toute chose égale par ailleurs (en particulier, les mêmes marginales des observations conditionnellement aux classes), l'utilisation d'une copule différente de celle correspondant aux données peut dégrader de manière significative les résultats des segmentations.

4.1 Lois elliptiques, modèles exponentiels et lois de Von Mises-Fisher

Dans cette section, nous donnons deux familles de lois généralisant la loi normale. La première est celle des lois de type exponentiel et la seconde est celle des lois elliptiques. Nous commençons par donner la définition générale d'un modèle exponentiel. Dans un second temps, nous donnons l'exemple de la loi de Von Mises-Fisher qui sera utilisée dans les applications au radar au chapitre 6.

4.1.1 Modèles exponentiels

Soit $(\mathcal{Y}, \mathcal{B}, \nu)$ un ensemble mesuré, où $\mathcal{B}$ est une tribu sur $\mathcal{Y}$ et ν une mesure définie sur $\mathcal{B}$.

Définition 4.1.1 (Modèles exponentiels)**.** *Soit a une application de $\mathcal{Y}$ dans $\mathbb{R}^k$, Θ un sous-ensemble de $\mathbb{R}^d$, $\theta \in \Theta$ et α une application de Θ dans $\mathbb{R}^k$. Une variable aléatoire à valeurs dans $(\mathcal{Y}, \mathcal{B})$ suit une loi de type exponentiel de paramètre numérique θ et de paramètres fonctionnels a et α si sa densité par rapport à ν s'écrit :*

$$\forall y \in \mathcal{Y},\ p(y) \propto \exp\left(\langle \alpha(\theta), a(y) \rangle\right),$$

où $\langle ., . \rangle$ est le produit scalaire euclidien de $\mathbb{R}^k$.

Lorsque le modèle exponentiel est paramétré par $\lambda = \alpha(\theta)$, on dit que le paramétrage est canonique.

4.1.2 Loi de Von Mises-Fisher

Rappelons d'abord comment est définie la mesure de Lebesgue sur la sphère $\mathcal{S}^{d-1} = \left\{x \in \mathbb{R}^d : \|x\| = 1\right\}$, que l'on notera $\sigma_{\mathcal{S}^{d-1}}$. La sphère sera munie de sa tribu borélienne $\mathcal{B}_{\mathcal{S}^{d-1}}$. Notons également $\mathcal{B}_{[0,+\infty[}$ la tribu borélienne de $[0, +\infty[$. La tribu borélienne de $\mathbb{R}^d$ est la tribu produit $\mathcal{B}_{[0,+\infty[} \otimes \mathcal{B}_{\mathcal{S}^{d-1}}$; soit celle engendrée par les boréliens, dit élémentaires, $B_1 \times B_2$ où $B_1 \in \mathcal{B}_{[0,+\infty[}$ et $B_2 \in \mathcal{B}_{\mathcal{S}^{d-1}}$. La mesure de Lebesgue sur la sphère est définie par :

$$\sigma_{\mathcal{S}^{d-1}}(B_2) = \frac{\lambda_{\mathbb{R}^d}(B_1 \times B_2)}{\displaystyle\int_{B_1} r^{d-1} dr},$$

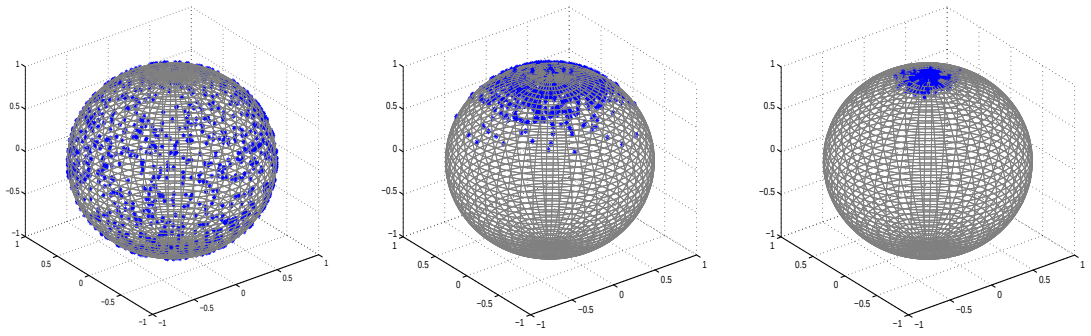
qui est aussi l'intégrale de la fonction indicatrice de B_2 par rapport à la mesure $\sigma_{\mathcal{S}^{d-1}}$. On définit ensuite l'intégrale des fonctions étagées (combinaison de fonctions indicatrices), l'intégrale d'une fonction mesurable de $\mathcal{B}_{\mathcal{S}^{d-1}}$ dans $\mathbb{R}$ est finalement définie en utilisant le théorème classique de Beppo-Levy. On a la formule de changement de variable suivante :

$$\int_B f(y) d\lambda_{\mathbb{R}^d}(y) = \int_{\psi^{-1}(B)} f(ru) r^{d-1} dr \otimes d\sigma_{\mathcal{S}^{d-1}}(u),$$

où ψ est l'application de $[0, +\infty[\times \mathcal{S}^{d-1}$ dans $\mathbb{R}^d$ qui à (r, u) associe $y = ru$ et B est un borélien de $\mathbb{R}^d$.
Soit $\mu \in \mathcal{S}^{d-1}$ et $\kappa \in \mathbb{R}^+$. La loi de Von Mises-Fisher de paramètre de direction μ et de paramètre de concentration κ est une loi sur la sphère de $\mathcal{S}^{d-1}$ de densité par rapport à la mesure de Lebesgue $\sigma_{\mathcal{S}^{d-1}}$ donnée par :

$$u \in \mathcal{S}^{d-1} \to \frac{\kappa^{\frac{d}{2}-1}}{(2\pi)^{\frac{d}{2}} I_{\frac{d}{2}-1}(\kappa)} \exp\left(\kappa \langle \mu, u \rangle\right), \tag{4.1}$$

où I_ν est la fonction de Bessel modifiée de première espèce de paramètre ν (voir Annexe A) et $\langle ., . \rangle$ est le produit scalaire de $\mathbb{R}^d$.



(a) Loi uniforme sur la sphère $\kappa = 0$. (b) Concentration $\kappa = 10$. (c) Concentration $\kappa = 100$.

FIG. 4.1 – Exemples de trois lois de Von Mises-Fisher.

Estimation des paramètres

Nous rappelons ci-après la méthode d'estimation par maximum de vraisemblance proposée dans [5], des paramètres d'une loi de Von Mises-Fisher.
Soit $u_{1:N} = (u^{(1)}, \dots, u^{(N)})$ un échantillon de la loi de Von Mises-Fisher. La vraisemblance s'écrit :

$$p(u_{1:N}) = C \frac{\kappa^{N\left(\frac{d}{2}-1\right)}}{\left(I_{\frac{d}{2}-1}(\kappa)\right)^N} \exp\left(\kappa \left\langle \mu, \sum_{n=1}^{N} u^{(n)} \right\rangle\right),$$

où C est une constante indépendante de κ et de μ. Ainsi la log-vraisemblance est :

$$L(u_{1:N}) = \text{Cste} + N\left(\frac{d}{2}-1\right)\log(\kappa) - N\log\left(I_{\frac{d}{2}-1}(\kappa)\right) + \kappa \left\langle \mu, \sum_{n=1}^{N} u^{(n)} \right\rangle.$$

La maximisation en μ ne dépend pas du paramètre κ et est atteinte en :

$$\hat{\mu}_{MV} = \frac{\sum_{n=1}^{N} u^{(n)}}{\| \sum_{n=1}^{N} u^{(n)} \|}.$$

Pour l'estimation du paramètre κ, on doit résoudre :

$$\frac{\frac{\partial}{\partial\kappa} I_{\frac{d}{2}-1}(\kappa)}{I_{\frac{d}{2}-1}(\kappa)} - \left(\frac{d}{2}-1\right) \times \frac{1}{\kappa} = \frac{1}{N} \left\langle \hat{\mu}_{MV}, \sum_{n=1}^{N} u^{(n)} \right\rangle,$$

soit

$$\frac{\frac{\partial}{\partial\kappa} I_{\frac{d}{2}-1}(\kappa)}{I_{\frac{d}{2}-1}(\kappa)} - \left(\frac{d}{2}-1\right) \times \frac{1}{\kappa} = \frac{1}{N} \| \sum_{n=1}^{N} u^{(n)} \|.$$

En utilisant le développement de Laurent de la fonction de Bessel figurant en annexe A, on montre que la dérivée de I_ν satisfait :

$$I'_\nu(x) = \frac{\nu I_\nu(x)}{x} + I_{\nu+1}(x),$$

ainsi le maximum de vraisemblance $\hat{\kappa}_{MV}$ satisfait :

$$\frac{I_{\frac{d}{2}}(\hat{\kappa}_{MV})}{I_{\frac{d}{2}-1}(\hat{\kappa}_{MV})} = \frac{1}{N} \parallel \sum_{n=1}^{N} u^{(n)} \parallel,$$

il est unique car la fonction $\kappa \rightarrow \dfrac{I_{\frac{d}{2}}(\kappa)}{I_{\frac{d}{2}-1}(\kappa)}$ est croissante.

La loi de Von Mises-Fisher étant de type exponentiel, l'étape de maximisation de l'algorithme EM aboutit à des équations similaires. Détaillons l'algorithme EM dans le cas des chaînes de Markov cachées à bruit indépendant.

Soit $u_{1:N} = (u^{(1)}, \ldots, u^{(N)})$ la réalisation observée et $x_{1:N} = (x_1, \ldots, x_N)$ la réalisation cachée, μ_j, κ_j les paramètres de la loi de Von Mises-Fisher $p(u^{(n)}|x_n = \omega_j)$ et θ_q le vecteur paramètre obtenu à l'étape q de EM. Alors les paramètres $\mu_{q+1,j}$ et $\kappa_{q+1,j}$ obtenus à l'étape $q+1$ sont :

$$\mu_{q+1,j} = \frac{\displaystyle\sum_{n=1}^{N} u^{(n)} p\left(x_n = \omega_j | u_{1:N}; \theta_q\right)}{\parallel \displaystyle\sum_{n=1}^{N} u^{(n)} p\left(x_n = \omega_j | u_{1:N}; \theta_q\right) \parallel};$$

$$\frac{I_{\frac{d}{2}}(\kappa_{q+1,j})}{I_{\frac{d}{2}-1}(\kappa_{q+1,j})} = \frac{\parallel \displaystyle\sum_{n=1}^{N} u^{(n)} p\left(x_n = \omega_j | u_{1:N}; \theta_q\right) \parallel}{\displaystyle\sum_{n=1}^{N} p\left(x_n = \omega_j | u_{1:N}; \theta_q\right)}.$$

Divers résultats concernant la segmentation et l'estimation des paramètres dans les modèles où les observations suivent des lois de Von Mises-Fisher sont présentés au chapitre 6 sur des données réelles issues de radar bande HF.

4.1.3 Lois elliptiques

Une variable aléatoire prenant ses valeurs dans $\mathbb{R}^d$ suit une loi elliptique si les isodensités sont des ellipsoïdes de $\mathbb{R}^d$. En d'autres termes, la densité d'une loi elliptique est définie par :

Définition 4.1.2. *Soit $m \in \mathbb{R}^d$, Σ une matrice réelle symétrique définie positive de dimension $d \times d$ et h une fonction de $\mathbb{R}^+$ dans $\mathbb{R}^+$. Une variable aléatoire Y à valeurs dans $\mathbb{R}^d$ suit une loi elliptique de paramètres euclidiens $m \in \mathbb{R}^d$ et Σ et de paramètre fonctionnel h si sa densité s'écrit :*

$$\forall y \in \mathbb{R}^d,\ p(y) = \frac{1}{\sqrt{\det \Sigma}} \times h\left(\parallel \Sigma^{-\frac{1}{2}}(y-m) \parallel^2\right).$$

La proposition suivante donne le lien entre lois elliptiques et lois uniformes sur la sphère :

Proposition 4.1.1. *Une variable aléatoire Y suit une loi elliptique de paramètres euclidiens m et Σ si et seulement si Y s'écrit :*

$$Y = R\Sigma^{\frac{1}{2}}U + m,$$

où U et R sont indépendantes, U suit une loi uniforme sur la sphère $\mathcal{S}^{d-1}$ et R est une variable aléatoire à valeurs dans $\mathbb{R}^+$.
Dans ce cas, le paramètre fonctionnel de la loi de Y est donné par :

$$\forall r \in \mathbb{R}^+,\ h(r) = \frac{\Gamma\left(\frac{d}{2}+1\right)}{d\pi^{\frac{d}{2}} r^{\frac{d-1}{2}}} f(\sqrt{r}),$$

où f est la densité de la variable aléatoire R et Γ est la fonction eulérienne rappelée en annexe A.

Par ailleurs, il est possible de calculer les moments d'une loi elliptique. En effet, nous avons :

Proposition 4.1.2. *Soit Y une variable aléatoire elliptique de paramètres euclidiens m et Σ et de paramètre fonctionnel h. Sous la condition :*

$$\int_0^{+\infty} h(r^2) r^{d+1} dr < +\infty,$$

Y est de carré intégrable, sa moyenne et sa matrice de covariance sont alors données par :

$$\begin{aligned} \mathbb{E}(Y) &= m, \\ Cov(Y) &= \frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}+1\right)} \left(\int_0^{+\infty} h(r^2) r^{d+1} dr \right) \Sigma. \end{aligned}$$

Dans la section suivante, nous présentons des lois elliptiques particulières : les vecteurs aléatoires sphériquement invariants (SIRV). Ces vecteurs aléatoires sont le produit d'une variable aléatoire positive et d'un vecteur gaussien. Un vecteur gaussien est elliptique ; il est le produit de la loi uniforme sur la sphère et de la racine carrée d'une loi du χ^2 ; il s'ensuit que la loi d'un SIRV est elliptique.

4.2 Vecteurs aléatoires sphériquement invariants

Dans cette section, nous étudions un cas particulier de distributions elliptiques, celui des vecteurs aléatoires sphériquement invariants (SIRV). Un SIRV de dimension d est le produit de deux variables aléatoires : la "texture" qui est une variable aléatoire à valeurs positives, et le "speckle" qui est un vecteur aléatoire gaussien de $\mathbb{R}^d$ ou $\mathbb{C}^d$, selon que le SIRV soit réel ou complexe. La terminologie est empruntée à celle des spécialistes du signal radar [55, 105, 118, 119]. En traitement du signal radar, à une distance et un angle de visée donnés, le signal réfléchi est un vecteur complexe appelé "données In Phase-Quadrature (IQ)". Ce vecteur complexe est souvent considéré comme un SIRV, le "speckle" pouvant s'interpréter physiquement comme le chatoiement optique dû à l'excitation des électrons et la texture comme les fluctuations spatiales macroscopiques du signal réfléchi. Nous détaillerons l'acquisition des données IQ au chapitre 6.

4.2.1 Lois SIRV à valeurs dans $\mathbb{R}^d$: définition et exemples

Définition 4.2.1 (Vecteurs aléatoires sphériquement invariants réels)**.** *Une variable aléatoire Y à valeurs dans $\mathbb{R}^d$ est un vecteur sphériquement invariant s'il s'écrit :*

$$Y = R \times Z + m,$$

où R est une variable aléatoire réelle positive appelée texture, Z est un vecteur aléatoire gaussien centré à valeurs dans $\mathbb{R}^d$ appelé "speckle", et m est un vecteur réel appelé paramètre de position.

A partir de maintenant, on notera Σ la matrice de covariance du "speckle". Lorsque le paramètre de position $m = 0$ et Σ est la matrice identité, on dira que le SIRV est centré réduit. Les deux principaux exemples de SIRV réels sont :

Lois de Student sur $\mathbb{R}^d$

La texture est l'inverse de la racine carré d'une variable aléatoire de loi Γ de paramètre de forme ν, sa densité est définie sur $\mathbb{R}^+$ par :

$$p(r) = \frac{2}{\Gamma(\nu)} \frac{1}{r^{2\nu+1}} \exp\left(-\frac{1}{r^2}\right).$$

La loi de Student de dimension d de paramètres ν, m et Σ est notée $\mathcal{S}_{\mathbb{R}^d}(\nu, m, \Sigma)$ et sa densité est :

$$y \in \mathbb{R}^d \rightarrow \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{\det\Sigma}} \times \frac{\Gamma\left(\nu + \frac{d}{2}\right)}{\Gamma(\nu)} \times \frac{1}{\left[1 + \frac{1}{2}(y-m)^T\Sigma^{-1}(y-m)\right]^{\nu+\frac{d}{2}}}$$

Lois K sur $\mathbb{R}^d$

La texture est la racine carrée d'une loi Γ de paramètre de forme ν, sa densité est définie sur $\mathbb{R}^+$ par :

$$p(r) = \frac{2}{\Gamma(\nu)} r^{2\nu-1} \exp\left(-r^2\right).$$

La densité d'une loi K de paramètres ν, m et Σ, notée $K_{\mathbb{R}^d}(\nu, m, \Sigma)$, est :

$$y \in \mathbb{R}^d \rightarrow \frac{2}{(2\pi)^{\frac{d}{2}}\sqrt{\det(\Sigma)}\Gamma(\nu)} \left(\frac{(y-m)^T\Sigma^{-1}(y-m)}{2}\right)^{\frac{\nu}{2}-\frac{d}{4}} K_{\nu-\frac{d}{2}}\left(2\sqrt{\frac{(y-m)^T\Sigma^{-1}(y-m)}{2}}\right),$$

où K_ν désigne la fonction de Bessel de deuxième espèce modifiée à ν degrés de liberté (voir Annexe A).

4.2.2 Lois gaussiennes complexes circulaires et lois SIRV complexes

Nous introduisons ci-après, les lois gaussiennes complexes circulaires, importantes pour la suite, notamment dans les algorithmes de détection sur les données radar. A partir de ces lois gaussiennes complexes, nous définirons les lois SIRV à valeurs dans un $\mathbb{C}$-espace vectoriel.

Lois gaussiennes complexes circulaires

Définition 4.2.2 (Variable aléatoire gaussienne complexe circulaire)**.** *Une variable aléatoire Z à valeurs dans $\mathbb{C}$ est gaussienne si le vecteur $(Re(Z), Im(Z))$ est un vecteur gaussien de $\mathbb{R}^2$. Elle est dite :*

- *centrée si $\mathbb{E}(Z) = 0$;*
- *circulaire si $Re(Z)$ et $Im(Z)$ sont indépendantes et de même variance.*

On note $\sigma_{\mathbb{C}}^2$ la variance $\mathbb{E}\left(|Z - \mathbb{E}(Z)|^2\right)$ et $m = \mathbb{E}(Z) = \mathbb{E}(Re(Z)) + i\mathbb{E}(Im(Z))$ la moyenne de Z.

La densité d'une variable aléatoire gaussienne complexe circulaire Z de moyenne m et de variance $\sigma_{\mathbb{C}}^2$ est :

$$z \in \mathbb{C} \to \frac{1}{\pi\sigma_{\mathbb{C}}^2} \exp\left(-\frac{|z-m|^2}{\sigma_{\mathbb{C}}^2}\right).$$

On définit également :

Définition 4.2.3 (Vecteurs aléatoires gaussiens complexes circulaires)**.** *Un vecteur aléatoire complexe $Z = (Z_1, \ldots, Z_d)$ est gaussien si pour tout $(\lambda_1, \ldots, \lambda_d) \in \mathbb{C}^d$, la variable aléatoire $\lambda_1 Z_1 + \ldots + \lambda_d Z_d$ est une variable aléatoire gaussienne complexe.*
Le vecteur aléatoire Z est circulaire si $\mathbb{E}\left((Z - \mathbb{E}(Z))(Z - \mathbb{E}(Z))^T\right) = 0$. On définit sa matrice de covariance complexe par :

$$\Sigma_{\mathbb{C}} = \mathbb{E}\left((Z - \mathbb{E}(Z))(\bar{Z} - \mathbb{E}(\bar{Z}))^T\right).$$

La densité d'un vecteur aléatoire gaussien complexe circulaire de moyenne m et de covariance complexe $\Sigma_{\mathbb{C}}$ est donnée par :

$$z \in \mathbb{C}^d \to \frac{1}{\pi^d \det \Sigma_{\mathbb{C}}} \exp\left(-\overline{(z-m)}^T \Sigma_{\mathbb{C}}^{-1}(z-m)\right).$$

Les SIRV réels sont étendus aux SIRV complexes pour lesquels le "speckle" est une variable aléatoire gaussienne complexe circulaire. Nous donnons ci-dessous les densités des lois de Student et K complexes.

Lois de Student sur $\mathbb{C}^d$

La densité d'une loi de Student complexe est donnée par :

$$y \in \mathbb{C}^d \to \frac{\Gamma(\nu+d)}{\pi^d \Gamma(\nu) \det(\Sigma_{\mathbb{C}})} \times \frac{1}{\left[1 + \overline{(y-m)}^T \Sigma_{\mathbb{C}}^{-1}(y-m)\right]^{\nu+d}}.$$

Lois K sur $\mathbb{C}^d$

La densité d'une loi K complexe est donnée par :

$$y \in \mathbb{C}^d \to \frac{2}{\pi^d \Gamma(\nu) \det(\Sigma_{\mathbb{C}})} \left(\overline{(y-m)^T \Sigma_{\mathbb{C}}^{-1}(y-m)}\right)^{\frac{\nu-d}{2}} K_{\nu-d}\left(2\sqrt{\overline{(y-m)^T \Sigma_{\mathbb{C}}^{-1}(y-m)}}\right).$$

4.3 Copules et lois multivariées non gaussiennes

Nous allons présenter dans cette section des lois multivariées qui peuvent se définir explicitement à partir des lois marginales et d'un terme d'agrégation appelé "copule".
L'introduction des copules est historiquement issue des travaux de M. Fréchet et de ceux de A. Sklar. Les travaux de M. Fréchet [45] concernaient les familles de vecteurs aléatoires ayant mêmes lois marginales. Il a ainsi donné le nom de "classes de Fréchet" aux classes de la relation d'équivalence $[Y = (Y_1, \ldots, Y_d) \sim \tilde{Y} = (\tilde{Y}_1, \ldots, \tilde{Y}_d)] \Leftrightarrow [Y \text{ et } \tilde{Y} \text{ ont mêmes lois marginales}]$. Cependant, M. Fréchet n'a pas établi la forme générale des fonctions de répartition jointes. Il a fallu attendre le théorème établi par A. Sklar [110, 112], qui porte actuellement son nom, pour pouvoir exprimer la fonction de répartition jointe en fonction des fonctions de répartition marginales. Nous présenterons dans la sous-section 4.3.1 ci-après, la problématique qui a conduit A. Sklar à son théorème.

4.3.1 Copules et théorème de Sklar

Les travaux de A. Sklar ne concernaient pas directement l'écriture de la loi jointe fonction des lois marginales. Il travaillait avec B. Schweizer sur la généralisation de la conjonction "et" en logique floue [43, 50, 63, 88] et sur les espaces métriques probabilisés [80, 110]. En logique floue, la valeur de vérité d'une formule est étendue en une fonction à valeurs dans $[0, 1]$ mesurant la croyance que l'on a d'une proposition. Cette fonction de croyance peut être une probabilité, auquel cas on parle de logique floue probabilisée. La valeur de vérité de la conjonction a été étendue par A. Sklar et B. Schweizer à l'aide des normes triangulaires :

Définition 4.3.1 (Normes triangulaires)**.** *Une application $T : [0, 1]^2 \to [0, 1]$ est une norme triangulaire bivariée si elle satisfait :*
- *$T(u, v) = T(v, u)$ (commutativité) ;*
- *$T(T(u, v), w) = T(u, T(v, w))$ (associativité) ;*
- *$T(u, 1) = u$;*
- *Si $u_1 \leq u_2$ et $v_1 \leq v_2$ alors $T(u_1, v_1) \leq T(u_2, v_2)$.*

Une norme triangulaire de dimension d est une application T_d de $[0, 1]^d$ dans $[0, 1]$ définie récursivement à partir d'une norme triangulaire bivariée T par :

1. *$T_3(u, v, w) = T(T(u, v), w)$;*
2. *$T_{n+1}(u_1, \ldots, u_{n+1}) = T(T_n(u_1, \ldots, u_n), u_{n+1})$.*

La croyance de l'expression "A et B", notée $A \wedge B$, est alors donnée par $\text{Bel}(A \wedge B) = T(\text{Bel}(A), \text{Bel}(B))$, où $\text{Bel}(A)$ (resp. $\text{Bel}(B)$) désigne le degré de croyance de A (resp. B). Les espaces métriques probabilisés sur lesquels travaillèrent A. Sklar et B. Schweizer sont des ensembles de variables aléatoires $(W_{x,y})_{(x,y)\in\mathcal{X}^2}$ à valeurs dans $\mathbb{R}^+$ et indexées sur un espace métrique $\mathcal{X}$, tels que $W_{x,y}$ a les propriétés d'une distance :
- $\forall (x, y, t)$, $\mathbb{P}(W_{x,y} \leq t) = \mathbb{P}(W_{y,x} \leq t)$ (symétrie) ;
- $\mathbb{P}(W_{x,y} = 0) = 1$ si et seulement si $x = y$;
- $\forall (x, y, z, t, s)$, $\mathbb{P}\left(W_{x,z} \leq t; W_{z,y} \leq s\right) \leq \mathbb{P}\left(W_{x,y} \leq t + s\right)$ (inégalité triangulaire).

La distance entre deux points x et y de $\mathcal{X}$ est ainsi généralisée en utilisant la fonction de répartition $t \to F_{x,y}(t) = \mathbb{P}(W_{x,y} \leq t)$. La problématique qui a conduit A. Sklar à son

théorème est la suivante. Dans un espace métrique classique, si on connaît les distances $d(x,z)$ et $d(z,y)$, on sait majorer la distance $d(x,y)$. Dans un espace métrique probabilisé, on voudrait également minorer la fonction de répartition $F_{x,y}$ lorsque l'on connaît les fonctions de répartitions $F_{x,z}$ et $F_{z,y}$. L'idée est alors d'exprimer $\mathbb{P}(W_{x,z} \leq t; W_{z,y} \leq s)$ en fonction de $F_{x,z}$ et $F_{z,y}$. Le théorème formulé par A. Sklar affirme qu'il est possible d'écrire $\mathbb{P}(W_{x,z} \leq t; W_{z,y} \leq s)$ en fonction de $F_{x,z}$ et $F_{z,y}$ grâce à un terme d'agrégation appelée copule.
Nous donnons ci-dessous la définition d'une copule telle qu'elle est énoncée par R. B. Nelsen [85]. On définit pour cela l'opérateur $\Delta_{u_k}^{v_k}$ qui à une fonction C' de $[0,1]^{d'}$ dans $[0,1]$, pour $d' \leq d$, associe la fonction de $[0,1]^{d'-1}$ dans $[0,1]$ définie par :

$$\forall x \in [0,1]^{d'-1},\ \Delta_{u_k}^{v_k} C'(x) = C'(x_1, \ldots, v_k, \ldots, x_{d'}) - C'(x_1, \ldots, u_k, \ldots, x_{d'}).$$

Définition 4.3.2 (Copule de dimension d)**.** *Une fonction $C : [0,1]^d \to [0,1]$ est une copule si elle vérifie :*

1. *pour tout i et pour tout $(u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_d) \in [0,1]^{d-1}$, $C(u_1, \ldots, u_{i-1}, 0, u_{i+1}, \ldots, u_d) = 0$;*
2. *pour tout i et tout $u_i \in [0,1]$, $C(1, \ldots, 1, u_i, 1, \ldots, 1) = u_i$;*
3. *pour tout $u, v \in [0,1]^d$ tel que pour tout i, $u_i \leq v_i$, on a $C = \Delta_{u_d}^{v_d} \Delta_{u_{d-1}}^{v_{d-1}} \ldots \Delta_{u_1}^{v_1} C \geq 0$.*

Remarque : Une copule de dimension d peut s'interpréter comme la fonction de répartition d'une loi à valeurs dans $[0,1]^d$ dont les lois marginales sont uniformes sur $[0,1]$.

Enonçons maintenant le théorème de Sklar :

Théorème 4.3.1 (Théorème de Sklar)**.** *Soit F une fonction de répartition sur $\mathbb{R}^d$ de fonctions de répartitions marginales $F_1, \ldots, F_d$. Alors il existe une copule C sur $[0,1]^d$ telle que :*

$$\forall y \in \mathbb{R}^d, F(y_1, \ldots, y_d) = C(F_1(y_1), \ldots, F_d(y_d)). \tag{4.2}$$

De plus, si les marges $F_1, \ldots, F_d$ sont continues, alors la copule est unique.

La réciproque de ce théorème est donnée par le théorème de Deheuvels :

Théorème 4.3.2 (Théorème de Deheuvels)**.** *Si F_1, ..., F_d sont d fonctions de répartition à support réel et si C est une copule de dimension d, alors il existe un vecteur aléatoire Y de dimension d de marginales F_j et de copule C tels que la fonction de répartition F de Y soit donnée par :*

$$\forall y \in \mathbb{R}^d,\ F(y_1, \ldots, y_d) = C(F_1(y_1), \ldots, F_d(y_d)).$$

Preuve. Voir [37]. □

Remarque : En plus de répondre aux attentes de A. Sklar, les copules répondent aux attentes de M. Fréchet. Elles permettent d'exprimer la loi jointe de vecteurs aléatoires à partir des lois marginales. Sous l'hypothèse de continuité des marginales, l'unique copule est

la fonction de répartition de $U = (F_1(Y_1), \ldots, F_d(Y_d))$. Si de plus, la fonction de répartition F est dérivable, nous avons :

$$\forall y \in \mathbb{R}^d, f(y_1, \ldots, y_d) = f_1(y_1) \ldots f_d(y_d) c(F_1(y_1), \ldots, F_d(y_d)) \tag{4.3}$$

où f est la densité jointe, $f_1, \ldots, f_d$ sont les densités marginales et c est la densité de la copule, soit la densité du vecteur $U = (F_1(Y_1), \ldots, F_d(Y_d))$.

Après avoir formulé son théorème, A. Sklar voulait trouver des conditions pour que $\mathbb{P}(W_{x,z} \leq t, W_{z,y} \leq s)$ représente la croyance sur $(W_{x,z} \leq t, W_{z,y} \leq s)$. Il voulait ainsi trouver des conditions pour que la copule C soit une norme triangulaire. Dans ce cas, l'espace métrique probabilisé est appelé espace métrique flou ou espace de Menger [80]. Les conditions pour qu'une copule soit une norme triangulaire sont données par la proposition 4.3.1. Rappelons d'abord la définition d'application lipschitzienne.

Définition 4.3.3 (Application lipschitzienne)**.** *Une application $T : I \subset \mathbb{R}^d \to \mathbb{R}$ est lipschitzienne si :*

$$\textit{Pour tout } u \in I \textit{ et tout } v \in I, \; |T(u) - T(v)| \leq \sum_{k=1}^{d} |u_k - v_k|.$$

Nous avons :

Proposition 4.3.1.
- *Une copule bivariée est une norme triangulaire si et seulement si elle satisfait $C(C(u, v), w) = C(u, C(v, w))$;*
- *Une norme triangulaire est une copule si et seulement si elle est lipschitzienne.*

Nous donnons dans la sous-section suivante les différents exemples de copules que nous utiliserons par la suite.

4.3.2 Exemples de copules

Copule produit

La copule produit est la copule d'un vecteur aléatoire à composantes indépendantes. Cette copule est donnée par :

$$C(u_1, \ldots, u_d) = \prod_{j=1}^{d} u_j.$$

Bornes de Fréchet

L'expression de la borne supérieure de Fréchet est :

$$C(u_1, u_2, \ldots, u_d) = \min(u_1, u_2, \ldots, u_d).$$

La borne inférieure de Fréchet est une copule uniquement dans le cas bivarié et est donnée par :

$$C(u_1, u_2, \ldots, u_d) = \max(u_1 + \ldots + u_d - d + 1, 0).$$

La terminologie "supérieure" et "inférieure" provient de l'inégalité suivante :

$$\max(u_1 + \ldots + u_d - d + 1, 0) \leq C(u_1, \ldots, u_d) \leq \min(u_1, \ldots, u_d),$$

où C est une copule de dimension d.

Copule gaussienne

La copule gaussienne est l'unique copule telle que si chaque marginale Y_k d'un vecteur aléatoire $(Y_1, \ldots, Y_d)$ suit une loi normale $\mathcal{N}_{\mathbb{R}}(m_k, \sigma_k^2)$, alors le vecteur joint suit une loi normale $\mathcal{N}_{\mathbb{R}^d}((m_1, \ldots, m_d), \Sigma)$.
Elle est donnée par :

$$c(u_1, \ldots, u_d) = \frac{1}{\sqrt{\det R}} \exp\left(-\frac{1}{2}(\phi^{-1}(u_1), \ldots, \phi^{-1}(u_d))^T \left(R^{-1} - \mathrm{Id}\right) (\phi^{-1}(u_1), \ldots, \phi^{-1}(u_d))\right),$$

où ϕ est la fonction de répartition d'une loi normale réelle centrée et réduite et R est la matrice de corrélation de Pearson donnée par :

$$R = \begin{pmatrix} \frac{1}{\sigma_1} & 0 & \ldots & \ldots & 0 \\ 0 & \frac{1}{\sigma_2} & 0 & \ldots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \ldots & \ldots & 0 & \frac{1}{\sigma_d} \end{pmatrix} \Sigma \begin{pmatrix} \frac{1}{\sigma_1} & 0 & \ldots & \ldots & 0 \\ 0 & \frac{1}{\sigma_2} & 0 & \ldots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \ldots & \ldots & 0 & \frac{1}{\sigma_d} \end{pmatrix}.$$

Copule de Student

La copule de Student est l'unique copule telle que si pour tout $k \in \{1, \ldots, d\}$, Y_k suit une loi de Student de paramètre de forme ν alors le vecteur aléatoire $(Y_1, \ldots, Y_d)$ suit une loi de Student de même paramètre de forme.
La densité de la copule de Student de paramètre de forme ν et de matrice de corrélation R est donnée par :

$$\begin{aligned} c(u_1, \ldots, u_d) &= \frac{1}{\sqrt{\det R}} \times \frac{\Gamma\left(\nu + \frac{d}{2}\right) \Gamma(\nu)^{d-1}}{\Gamma\left(\nu + \frac{1}{2}\right)^d} \times \\ &\times \frac{\left[\left(1 + \frac{\left(\phi^{-1}(u_1)\right)^2}{2}\right) \ldots \left(1 + \frac{\left(\phi^{-1}(u_d)\right)^2}{2}\right)\right]^{\nu + \frac{1}{2}}}{\left[1 + \frac{1}{2}(\phi^{-1}(u_1), \ldots, \phi^{-1}(u_d))^T R^{-1} (\phi^{-1}(u_1), \ldots, \phi^{-1}(u_d))\right]^{\nu + \frac{d}{2}}}. \quad (4.4) \end{aligned}$$

où ϕ est la fonction de répartition de la loi de Student monovariée de paramètre de forme ν, centrée et réduite.

Copule K

La copule K est l'unique copule telle que si les lois marginales sont des lois K de paramètre de forme ν, alors le vecteur joint suit encore une loi K de même paramètre de forme.
La densité de la copule K s'écrit :

$$c(u_1,\ldots,u_d) = \frac{\Gamma(\nu)^{d-1}}{2^{d-1}} \times \frac{2^{\frac{\nu}{2}(d-1)}}{\sqrt{\det R}} \times \frac{\left[\xi^T R^{-1}\xi\right]^{\frac{\nu}{2}-\frac{d}{4}}}{\left[\xi_1\ldots\xi_d\right]^{\nu-\frac{1}{2}}} \times \frac{K_{\nu-\frac{d}{2}}(\sqrt{2\xi^T R^{-1}\xi})}{\prod_{j=1}^{d} K_{\nu-\frac{1}{2}}(\sqrt{2}\xi_j)}, \tag{4.5}$$

où $\xi = (\phi_\nu^{-1}(u_1),\ldots,\phi_\nu^{-1}(u_d))$ avec ϕ_ν fonction de répartition d'une loi K monovariée de paramètre de forme ν, centrée et réduite et K_ν est la fonction de Bessel modifiée de seconde espèce de paramètre de forme ν.

Parmi les autres exemples de copules, on peut citer également les copules archimédiennes [46, 76, 117], qui présentent surtout un intérêt dans le cas bivarié.

4.3.3 Mesures de dépendance

Dans cette sous-section, nous abordons les différentes mesures de dépendance, appelées aussi mesures d'association entre variables aléatoires. Une mesure de dépendance est un moyen de quantifier la dépendance entre variables aléatoires, ou encore la manière dont les deux variables aléatoires sont liées. La mesure de dépendance la plus classique entre deux variables aléatoires réelles Y_1 et Y_2 est le coefficient de corrélation de Pearson donné par :

$$\rho = \frac{\sigma_{Y_1,Y_2}}{\sqrt{\sigma^2_{Y_1}\sigma^2_{Y_2}}},$$

où $\sigma^2_{Y_i}$ est la variance de la variable aléatoire Y_i et σ_{Y_1,Y_2} est la covariance entre Y_1 et Y_2.
Cependant, le coefficient de corrélation de Pearson a ses limites. Premièrement, il n'existe que si les variables aléatoires sont de carré intégrable. Deuxièmement, il s'avère inefficace pour modéliser certaines situations de dépendance. Par exemple, on aimerait que si $Y_2 = f(Y_1)$, alors la corrélation vaut 1 si f est croissante ou -1 si f est décroissante, mais ce n'est pas le cas. En effet, à titre d'exemple, si Y_1 suit une loi uniforme sur $[1,2]$ et si $Y_2 = \dfrac{1}{Y_1}$, alors $\rho = \sqrt{12} \times \dfrac{2-3\log(2)}{\sqrt{2-4\left(\log(2)\right)^2}} \simeq -0.98$. La corrélation de Pearson s'avère être une bonne mesure de corrélation si f est linéaire.

Corrélation ρ_S de Spearman

Définition 4.3.4 (Corrélation de Spearman)**.** *Soit (Y_1, Y_2) un vecteur aléatoire à valeurs dans $\mathbb{R}^2$, de copule C et de fonctions de répartition marginales continues F_1 et F_2. La corrélation de Spearman entre Y_1 et Y_2 est la corrélation de Pearson entre les variables aléatoires $U_1 = F_1(Y_1)$ et $U_2 = F_2(Y_2)$ définie par :*

$$\rho_S = 12\sigma_{U_1,U_2},$$

où σ_{U_1,U_2} est la covariance entre U_1 et U_2.

Remarque : Lorsque la copule C d'un vecteur aléatoire (Y_1, Y_2) admet une densité c, la corrélation de Spearman est également donnée par :

$$\rho_S = \rho(U, V) = 12 \int_{[0,1]^2} uvc(u, v)dudv - 3.$$

La proposition suivante nous montre qu'une dépendance du type $Y_2 = f(Y_1)$, où f est un $\mathcal{C}^1$-difféomorphisme de $\mathbb{R}$ dans $\mathbb{R}$, est équivalente à $\rho_S = 1$ ou -1.

Proposition 4.3.2. *Soit ρ_S la corrélation de Spearman entre deux variables aléatoires réelles Y_1 et Y_2 de fonctions de répartition respectives F_1 et F_2, qui sont des $\mathcal{C}^1$-difféomorphismes de $\mathbb{R}$ dans $[0, 1]$. Nous avons :*

- *$\rho_S = 1$ si et seulement si $Y_2 = f(Y_1)$ où f est un $\mathcal{C}^1$-difféomorphisme de $\mathbb{R}$ dans $\mathbb{R}$ strictement croissant ;*
- *$\rho_S = -1$ si et seulement si $Y_2 = g(Y_1)$ où g est un $\mathcal{C}^1$-difféomorphisme de $\mathbb{R}$ dans $\mathbb{R}$ strictement décroissant.*

Preuve. Notons $U_1 = F_1(Y_1)$ et $U_2 = F_2(Y_2)$. Si $\rho_S = 1$, ρ_S étant la corrélation de Pearson du couple (U_1, U_2), alors $U_2 = \lambda U_1 + b$, avec $\lambda > 0$. Comme U_1 et U_2 suivent toutes les deux la même loi, alors $\lambda = 1$ et $b = 0$. On en déduit que $Y_2 = (F_2^{-1} \circ F_1)(Y_1)$. Si $\rho_S = -1$, alors par le même raisonnement, $U_2 = 1 - U_1$, d'où $Y_2 = F_2^{-1}(1 - F_1(Y_1))$.
Pour la réciproque, on se sert du lemme :

Lemme 4.3.1. *Soient U et V deux variables aléatoires suivant une loi uniforme sur $[0, 1]$. Si $V = g(U)$ où g est un $\mathcal{C}^1$-difféomorphisme de $[0, 1]$ dans $[0, 1]$, alors nécessairement $g(u) = u$ pour tout $u \in [0, 1]$ ou $g(u) = 1 - u$ pour tout $u \in [0, 1]$.*

Preuve du lemme. Il suffit d'écrire la densité de U fonction de celle de V. Si V suit la loi uniforme, pour que U suive la loi uniforme, il est nécessaire que $|g'(u)| = 1$ pour tout u et que $g([0, 1]) = [0, 1]$. On en déduit le résultat. □

Ainsi si $Y_2 = f(Y_1)$, où f est un $\mathcal{C}^1$-difféomorphisme, alors $U_2 = (F_2 \circ f \circ F_1^{-1})(U_1)$ d'où, en utilisant le lemme, $U_2 = U_1$ si f est croissante et $U_2 = 1 - U_1$ si f est décroissante. Ainsi $\rho_S = 1$ si f est croissante et -1 si f est décroissante. □

Remarque : Par construction, la corrélation de Spearman ne dépend que de la copule. Ainsi, la corrélation de Spearman de $(f(Y_1), g(Y_2))$ où f et g sont deux $\mathcal{C}^1$-difféomorphismes croissants est égale à la corrélation de Spearman de (Y_1, Y_2), les deux vecteurs aléatoires ayant même copule.

Mesure de concordance-discordance de Kendall et corrélation τ de Kendall

La corrélation de Kendall dérive de la notion de concordance-discordance définie ci-après et est très utilisée en finance et économétrie.

Définition 4.3.5 (Corrélation de Kendall)**.** *Soient deux vecteurs aléatoires* (Y_1, Y_2) *et* $(\tilde{Y}_1, \tilde{Y}_2)$ *de* $\mathbb{R}^2$ *indépendants suivant la même loi bivariée. On appelle "coefficient de corrélation de Kendall" de* (Y_1, Y_2)*, la quantité :*

$$\tau = \mathbb{P}((Y_1 - \tilde{Y}_1)(Y_2 - \tilde{Y}_2) > 0) - \mathbb{P}((Y_1 - \tilde{Y}_1)(Y_2 - \tilde{Y}_2) < 0).$$

Remarque : Si deux variables aléatoires Y_1 et Y_2 sont liées par une relation du type $Y_2 = f(Y_1)$ où f est une fonction monotone de $\mathbb{R}$ dans $\mathbb{R}$, alors $\tau = 1$ si f est croissante et $\tau = -1$ si f est décroissante.
Si la copule d'un vecteur aléatoire bivarié (Y_1, Y_2) admet une densité, alors sa corrélation de Kendall est donnée par :

$$\tau = 4 \int_{[0,1]^2} C(u, v)c(u, v)dudv - 1.$$

La corrélation de Spearman et celle de Kendall sont des cas particuliers d'une mesure de dépendance appelée mesure de concordance-discordance et définie par :

Définition 4.3.6. *Soient* (Y_1, Y_2) *et* $(\tilde{Y}_1, \tilde{Y}_2)$ *deux vecteurs aléatoires bivariés indépendants de mêmes lois marginales et de copules respectives* C *et* $\tilde{C}$*. La mesure de concordance-discordance* $Q(C, \tilde{C})$ *est définie par*

$$Q(C, \tilde{C}) = \mathbb{P}((Y_1 - \tilde{Y}_1)(Y_2 - \tilde{Y}_2) > 0) - \mathbb{P}((Y_1 - \tilde{Y}_1)(Y_2 - \tilde{Y}_2) < 0).$$

Remarque : Si les copules respectives C et $\tilde{C}$ de deux vecteurs aléatoires bivariés (Y_1, Y_2) et $(\tilde{Y}_1, \tilde{Y}_2)$ de mêmes lois marginales admettent des densités, alors la mesure de concordance-discordance est donnée par :

$$Q(C, \tilde{C}) = 4 \int_{[0,1]^2} C(u, v)\tilde{c}(u, v)dudv - 1.$$

On en déduit les propriétés suivantes :
- $\rho_S = 3Q(C, \Pi)$, où Π désigne la copule produit ;
- $\tau = Q(C, C)$.

Coefficients de mélangeance

Les coefficients de mélangeance sont des mesures de dépendance asymptotique. L'étude de la mélangeance de processus stationnaires permet de déduire leurs propriétés asymptotiques [58, 116]. Soient $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé, $\mathcal{F}$ et $\mathcal{G}$ deux sous-tribus de $\mathcal{A}$. Notons $\mathbb{P}(A|\mathcal{G})$ l'espérance sachant $\mathcal{G}$ de la variable aléatoire qui à $\omega \in \Omega$ associe 1 si $\omega \in A$ et 0 sinon. Les différents coefficients de mélangeance entre $\mathcal{F}$ et $\mathcal{G}$ sont définis par :
- α-mélangeance, ou mélangeance forte : $\alpha(\mathcal{F}, \mathcal{G}) = \sup_{A \in \mathcal{F}, B \in \mathcal{G}} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|$;
- β-mélangeance : $\beta(\mathcal{F}, \mathcal{G}) = \mathbb{E}\left(\sup_{A \in \mathcal{F}} |\mathbb{P}(A) - \mathbb{P}(A|\mathcal{G})|\right)$;
- ρ-mélangeance : $\rho(\mathcal{F}, \mathcal{G}) = \sup\{|\rho(X, Y)| : X\ \mathcal{F}\text{-mesurable}, Y\ \mathcal{G}\text{-mesurable}\}$, où ρ est la corrélation de Pearson ;

– ϕ-mélangeance : $\phi(\mathcal{F},\mathcal{G}) = \sup_{A\in\mathcal{F},B\in\mathcal{G},\mathbb{P}(B)>0} |\mathbb{P}(A) - \mathbb{P}(A|B)|$.

Nous avons les inégalités $2\alpha(\mathcal{F},\mathcal{G}) \leq \beta(\mathcal{F},\mathcal{G}) \leq \phi(\mathcal{F},\mathcal{G})$ et $4\alpha(\mathcal{F},\mathcal{G}) \leq \rho(\mathcal{F},\mathcal{G}) \leq 2\sqrt{\phi(\mathcal{F},\mathcal{G})}$.
Soit c un des coefficients de mélangeance. On dit qu'un processus $(Y_n)_{n\geq 0}$ est c-mélangeant, si $\lim_{n\to+\infty} c(\sigma(Y_0), \sigma((Y_k)_{k\geq n})) = 0$, où $\sigma(Z)$ est la tribu engendrée par la variable aléatoire Z.
Soient f et g deux $\mathcal{C}^1$-difféomorphismes croissants. Comme Y et $f(Y)$ engendrent la même tribu et (Y_1, Y_2) et $(f(Y_1), g(Y_2))$ ont la même copule, ainsi les coefficients de mélangeance entre deux tribus $\sigma(Y_1)$ et $\sigma(Y_2)$ ne dépendent que de la copule.
Par conséquent, la corrélation d'un processus $Y = (Y_n)_{n\geq 1}$ aura le même comportement asymptotique que celle de tout processus $Z = (f_n(Y_n))_{n\geq 1}$, où chaque f_n est une fonction croissante. Nous reviendrons sur cette remarque au chapitre 5 lorsque l'on définira la dépendance longue dans les processus non gaussiens.

4.4 Chaînes de Markov cachées M-stationnaires à bruit corrélé non gaussien

Nous proposons dans cette section un modèle original et montrons son intérêt en segmentation non supervisée des données non stationnaires, corrélées et non gaussiennes.

4.4.1 Copules dans les chaînes de Markov cachées M-stationnaires

Considérons un processus triplet $(X,U,Y) = (X_n,U_n,Y_n)_{1\leq n\leq N}$ tel que chaque X_n prend ses valeurs dans l'ensemble fini $\mathcal{X} = \{\omega_1,\ldots,\omega_K\}$, chaque U_n prend ses valeurs dans l'ensemble fini $\Lambda = \{\lambda_1,\ldots,\lambda_M\}$, et chaque Y_n prend ses valeurs dans $\mathbb{R}$. Nous proposons de généraliser dans cette sous-section le modèle de chaînes de Markov cachées M-stationnaires vu au chapitre 3 au cas d'observations corrélées et non gaussiennes. La loi du triplet (X,U,Y) sera de la forme :

$$\begin{aligned} p(x_{1:N},u_{1:N},y_{1:N}) &= p(u_1)p(x_1|u_1)p(y_1|x_1) \\ &\times \prod_{n=1}^{N-1} p(u_{n+1}|u_n)p(x_{n+1}|x_n,u_{n+1})p(y_{n+1}|x_n,x_{n+1},y_n). \end{aligned} \tag{4.6}$$

On remarque que les processus U et (X,U) sont des chaînes de Markov. On supposera que la chaîne de Markov (X,U) est stationnaire réversible, soit $p(x_n=\omega_i,u_n=\lambda_k,x_{n+1}=\omega_j,u_{n+1}=\lambda_l) = p(x_n=\omega_j,u_n=\lambda_l,x_{n+1}=\omega_i,u_{n+1}=\lambda_k)$ et ne dépend pas de n. Les lois initiales $p(u_1)$ et $p(x_1|u_1)$ et les transitions $p(u_{n+1}|u_n)$ et $p(x_{n+1}|x_n,u_{n+1})$ sont définies à partir de $p(x_n,u_n,x_{n+1},u_{n+1})$. On supposera également que $p(y_n,y_{n+1}|x_n=\omega_i,x_{n+1}=\omega_j) = p(y_{n+1},y_n|x_n=\omega_j,x_{n+1}=\omega_i)$ et ne dépend pas de n, ainsi la chaîne de Markov (X,U,Y) est également stationnaire et réversible. Plus exactement, la loi jointe $p(y_n,y_{n+1}|x_n,x_{n+1})$ sera donnée par :

$$p(y_n,y_{n+1}|x_n,x_{n+1}) = p(y_n|x_n)p(y_{n+1}|x_{n+1})c_{x_n,x_{n+1}}\left(F_{x_n}(y_n),F_{x_{n+1}}(y_{n+1})\right), \tag{4.7}$$

où $c_{x_n,x_{n+1}}$ est la copule de la loi jointe $p(y_n,y_{n+1}|x_n,x_{n+1})$ et F_{x_n} (resp. $F_{x_{n+1}}$) est la fonction de répartition de la loi marginale $p(y_n|x_n)$ (resp. $p(y_{n+1}|x_{n+1})$).
Afin d'avoir l'égalité $p(y_n,y_{n+1}|x_n=\omega_i,x_{n+1}=\omega_j) = p(y_{n+1},y_n|x_n=\omega_j,x_{n+1}=\omega_i)$,

on supposera $c_{\omega_i,\omega_j}\left(F_{\omega_i}(y_n), F_{\omega_j}(y_{n+1})\right) = c_{\omega_j,\omega_i}\left(F_{\omega_j}(y_{n+1}), F_{\omega_i}(y_n)\right)$. La loi conditionnelle $p(y_{n+1}|x_n, x_{n+1}, y_n)$ est exprimée par :

$$p(y_{n+1}|x_n, x_{n+1}, y_n) = p(y_{n+1}|x_{n+1}) c_{x_n,x_{n+1}}\left(F_{x_n}(y_n), F_{x_{n+1}}(y_{n+1})\right). \tag{4.8}$$

4.4.2 Estimation des paramètres

Nous proposons dans cette sous-section de détailler l'algorithme ICE dans le cas du modèle de chaînes de Markov cachées M-stationnaires à bruit corrélé non gaussien. Les paramètres à estimer sont les lois $p(x_n, u_n, x_{n+1}, u_{n+1})$, les lois marginales $p(y_n|x_n)$ et la copule de chaque loi jointe $p(y_n, y_{n+1}|x_n, x_{n+1})$. L'estimation de $p(x_n, u_n, x_{n+1}, u_{n+1})$ a déjà été étudiée dans la sous-section 3.2.2 du chapitre 3, ainsi nous nous consacrons uniquement à l'estimation de la loi jointe $p(y_n, y_{n+1}|x_n, x_{n+1})$. Celle-ci se déroule en deux temps. Notons θ_q le paramètre estimé à l'étape q de l'algorithme ICE. Nous commençons par ré-estimer les lois marginales, puis avec les nouvelles valeurs courantes des lois marginales, nous ré-estimons la copule.
Considérons le cas où les marginales $p(y_n|x_n = \omega_i)$ sont des lois normales $\mathcal{N}_{\mathbb{R}}(m_i, \sigma_i^2)$. Soit $y_{1:N} = (y_1, \ldots, y_N)$ la réalisation observée. Les estimateurs de m_i et σ_i^2 à partir des données complètes $(x_{1:N}, u_{1:N}, y_{1:N})$ sont :

$$\begin{aligned}
\hat{m}_i(x_{1:N}, y_{1:N}) &= \frac{\sum_{n=1}^{N} y_n I(x_n = \omega_i)}{\sum_{n=1}^{N} I(x_n = \omega_i)}, \\
\hat{\sigma}_i^2(x_{1:N}, y_{1:N}) &= \frac{\sum_{n=1}^{N} (y_n - \hat{m}_i(x_{1:N}, y_{1:N}))^2 I(x_n = \omega_i)}{\sum_{n=1}^{N} I(x_n = \omega_i)}.
\end{aligned}$$

L'espérance sachant $y_{1:N}$ sous le paramètre θ_q n'est pas calculable explicitement. Ainsi, on simule L échantillons indépendants $(x_{1:N}^{(m)}, u_{1:N}^{(m)})$ pour $1 \leq m \leq L$ selon la loi a posteriori $p(x_{1:N}, u_{1:N}|y_{1:N}; \theta_q)$. Les ré-estimation de m_i et σ_i^2 sont alors données par :

$$\begin{aligned}
m_i^{q+1} &= \frac{1}{L}\sum_{m=1}^{L} \hat{m}_i(x_{1:N}^{(m)}, u_{1:N}^{(m)}, y_{1:N}^{(m)}), \\
(\sigma_i^{q+1})^2 &= \frac{1}{L}\sum_{m=1}^{L} \hat{\sigma}_i^2(x_{1:N}^{(m)}, u_{1:N}^{(m)}, y_{1:N}^{(m)}).
\end{aligned}$$

Dans les expérimentations présentées dans la sous-section suivante, nous prendrons $L = 1$.
Notons maintenant $F_{\omega_i}^{q+1}$ la fonction de répartition de la loi normale $\mathcal{N}_{\mathbb{R}}\left(m_i^{q+1}, (\sigma_i^{q+1})^2\right)$. La deuxième étape de ICE consiste à estimer la copule. Dans le cas des copules gaussiennes (resp. de Student), notons ϕ la fonction de répartition de la loi normale centrée réduite (resp. loi de Student centrée réduite). Le paramètre de corrélation $\rho_{x_n,x_{n+1}}$ de la copule $c_{x_n,x_{n+1}}$ de $p(y_n, y_{n+1}|x_n, x_{n+1})$ est estimé à partir de l'échantillon $z_{1:N} = (z_1, \ldots, z_N)$ où $z_n = \phi^{-1} \circ F_{\omega_i}^{q+1}(y_n)$. En effet, $p(z_n, z_{n+1}|x_n, x_{n+1})$ suit une loi normale (resp. de Student) de corrélation

$\rho_{x_n,x_{n+1}}$. Un estimateur de ρ_{ω_i,ω_j} à partir des données complètes $(x_{1:N}, u_{1:N}, z_{1:N})$ est donné par :

$$\hat{\rho}_{\omega_i,\omega_j}(x_{1:N}, z_{1:N}) = \frac{\hat{\sigma}^{(1,2)}_{\omega_i,\omega_j}}{\hat{\sigma}^{(1)}_{\omega_i,\omega_j}\hat{\sigma}^{(2)}_{\omega_i,\omega_j}},$$

où

$$\begin{aligned}
(\hat{\sigma}^{(1)}_{\omega_i,\omega_j})^2 &= \frac{\sum_{n=1}^{N-1}(z_n - \hat{m}^{(1)}_{\omega_i,\omega_j})^2 I(x_n = \omega_i, x_{n+1} = \omega_j)}{\sum_{n=1}^{N-1} I(x_n = \omega_i, x_{n+1} = \omega_j)}, \\
(\hat{\sigma}^{(2)}_{\omega_i,\omega_j})^2 &= \frac{\sum_{n=1}^{N-1}(z_{n+1} - \hat{m}^{(2)}_{\omega_i,\omega_j})^2 I(x_n = \omega_i, x_{n+1} = \omega_j)}{\sum_{n=1}^{N-1} I(x_n = \omega_i, x_{n+1} = \omega_j)}, \\
\hat{\sigma}^{(1,2)}_{\omega_i,\omega_j} &= \frac{\sum_{n=1}^{N-1}(z_n - \hat{m}^{(1)}_{\omega_i,\omega_j})(z_{n+1} - \hat{m}^{(2)}_{\omega_i,\omega_j}) I(x_n = \omega_i, x_{n+1} = \omega_j)}{\sum_{n=1}^{N-1} I(x_n = \omega_i, x_{n+1} = \omega_j)},
\end{aligned}$$

avec

$$\begin{aligned}
\hat{m}^{(1)}_{\omega_i,\omega_j} &= \frac{\sum_{n=1}^{N-1} z_n I(x_n = \omega_i, x_{n+1} = \omega_j)}{\sum_{n=1}^{N-1} I(x_n = \omega_i, x_{n+1} = \omega_j)}, \\
\hat{m}^{(2)}_{\omega_i,\omega_j} &= \frac{\sum_{n=1}^{N-1} z_{n+1} I(x_n = \omega_i, x_{n+1} = \omega_j)}{\sum_{n=1}^{N-1} I(x_n = \omega_i, x_{n+1} = \omega_j)}.
\end{aligned}$$

Comme pour l'estimation des lois marginales, l'espérance sachant $y_{1:N}$ (resp. $z_{1:N}$) de $\hat{\rho}_{\omega_i,\omega_j}$ n'est pas calculable explicitement. Ainsi on simule L échantillons indépendants $(x^{(m)}_{1:N}, u^{(m)}_{1:N})$ pour $1 \leq m \leq L$ selon la loi a posteriori $p(x_{1:N}, u_{1:N}|y_{1:N}; \theta_q)$ et la ré-estimation de ρ_{ω_i,ω_j} est :

$$\rho^{q+1}_{\omega_i,\omega_j} = \frac{1}{L}\sum_{m=1}^{L} \hat{\rho}_{\omega_i,\omega_j}(x^{(m)}_{1:N}, u^{(m)}_{1:N}, z^{(m)}_{1:N}).$$

4.4.3 Expérimentations

Nous présentons dans cette sous-section deux expériences, dans lesquelles nous considérons une chaîne de Markov M-stationnaire (X, U, Y) dont la distribution est donnée par (4.6) et (4.7). La réalisation de processus mono-dimensionnels sera transformée en image bi-dimensionnelle grâce au parcours d'Hilbert-Peano figurant en annexe B. En reprenant les notations de la sous-section 4.4.1, on prendra pour ces deux expériences, $K = 2$ et $M = 2$. Les paramètres de (X, U) seront donnés par :

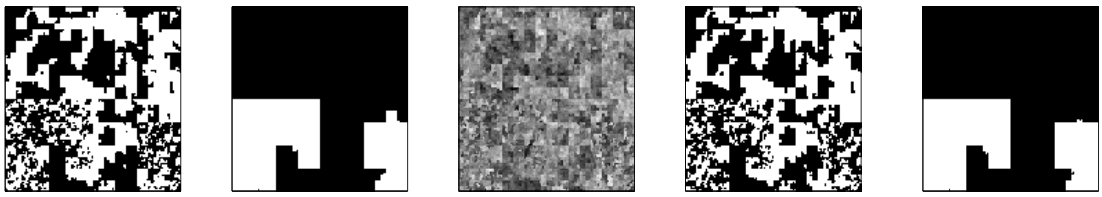
$$\begin{aligned} p(u_n = \lambda_1, u_{n+1} = \lambda_1) &= p(u_n = \lambda_2, u_{n+1} = \lambda_2) &= 0.49995, \\ p(u_n = \lambda_1, u_{n+1} = \lambda_2) &= p(u_n = \lambda_2, u_{n+1} = \lambda_1) &= 0.00005, \end{aligned}$$

$$\begin{aligned} p(x_n = \omega_1, x_{n+1} = \omega_1 | u_{n+1} = \lambda_1) &= p(x_n = \omega_2, x_{n+1} = \omega_2 | u_{n+1} = \lambda_1) &= 0.495, \\ p(x_n = \omega_2, x_{n+1} = \omega_1 | u_{n+1} = \lambda_1) &= p(x_n = \omega_2, x_{n+1} = \omega_1 | u_{n+1} = \lambda_1) &= 0.005, \\ p(x_n = \omega_1, x_{n+1} = \omega_1 | u_{n+1} = \lambda_2) &= p(x_n = \omega_2, x_{n+1} = \omega_2 | u_{n+1} = \lambda_2) &= 0.45, \\ p(x_n = \omega_2, x_{n+1} = \omega_1 | u_{n+1} = \lambda_2) &= p(x_n = \omega_2, x_{n+1} = \omega_1 | u_{n+1} = \lambda_2) &= 0.05. \end{aligned}$$

Dans les deux expériences, les lois marginales $p(y_n|x_n)$ seront respectivement la loi normale $\mathcal{N}_{\mathbb{R}}(0, 1)$ lorsque $x_n = \omega_1$ et la loi normale $\mathcal{N}_{\mathbb{R}}(2, 1)$ lorsque $x_n = \omega_2$. Dans la première expérience, nous simulons les données avec une copule gaussienne et dans la seconde, les nous simulons avec une copule de Student. Les données seront ensuite segmentées, dans chacune de ces expériences, en utilisant le modèle avec copule gaussienne et le modèle avec copule de Student. Le but de ces expériences est de savoir si le choix de la copule a de l'importance lorsque l'on ne connaît pas la manière dont les données sont corrélées, et si le choix d'une copule différente de celle du vrai modèle dégrade considérablement les résultats.

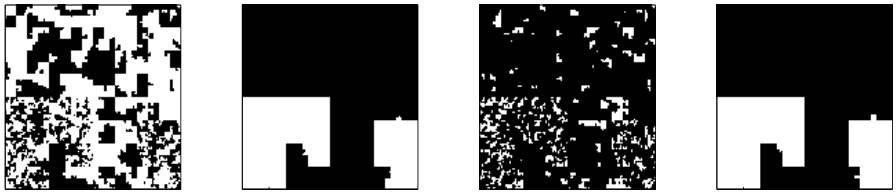
Segmentation de données simulées avec une copule gaussienne

Dans cette première expérience, les données sont simulées à l'aide d'une copule gaussienne de paramètres $\rho_{\omega_1,\omega_1} = \rho_{\omega_2,\omega_2} = 0.9$ et $\rho_{\omega_1,\omega_2} = \rho_{\omega_2,\omega_1} = 0$. Ainsi, $p(y_n, y_{n+1}|x_n, x_{n+1})$ est une distribution normale de corrélation $\rho_{x_n,x_{n+1}}$. Les données sont ensuite segmentées par trois méthodes. La première méthode utilise les vrais paramètres du modèle et les états cachés sont estimés par MPM. La seconde méthode utilise le vrai modèle, celui dont la copule est gaussienne, les paramètres sont estimés par ICE et les états cachés estimés par MPM. Quant à la dernière méthode, elle utilise le modèle dont la copule est de Student de paramètre de forme $\nu = 10$, les autres paramètres étant estimés par ICE, et les états cachés par MPM.



(a) Réalisation de X. (b) Réalisation de U. (c) Réalisation de Y. (d) Estimation de $x_{1:N}$: 5.49% d'erreur. (e) Estimation de $u_{1:N}$: 0.45% d'erreur.

FIG. 4.2 – Segmentation avec les vrais paramètres de données simulées avec la copule gaussienne.



(a) Estimation de $x_{1:N}$ (CG) : 9.43% d'erreur. (b) Estimation de $u_{1:N}$ (CG) : 0.51% d'erreur. (c) Estimation de $x_{1:N}$ (CS) : 38.44% d'erreur. (d) Estimation de $u_{1:N}$ (CS) : 0.66% d'erreur.

FIG. 4.3 – Segmentation non supervisée de données simulées avec la copule gaussienne en utilisant le vrai modèle avec copule gaussienne (CG) et le modèle avec copule de Student (CS).

		CG		CS		Vraies valeurs	
		ω_1	ω_2	ω_1	ω_2	ω_1	ω_2
m_{ω_i}		-0.05	2.04	0.41	1.75	0	2
$\sigma^2_{\omega_i}$		0.91	0.96	1.63	1.04	1	1
ρ	ω_1	0.89	-0.05	0.93	0.18	0.9	0
	ω_2	-0.05	0.92	0.18	0.78	0	0.9
Taux d'erreur pour X		9.43%		38.44%		5.49%	
Taux d'erreur pour U		0.51%		0.66%		0.45%	

TAB. 4.1 – Estimation de la loi d'observation en utilisant le vrai modèle avec copule gaussienne (CG) et le modèle avec copule de Student (CS).

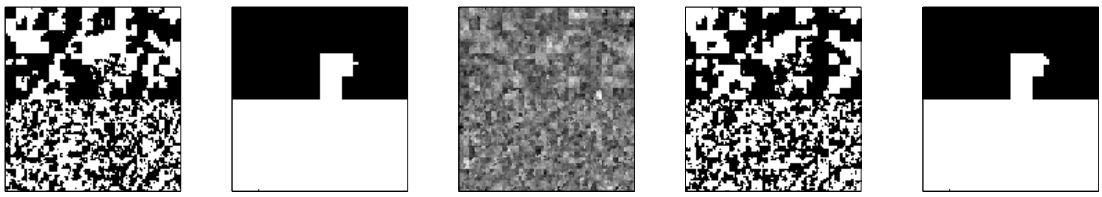
		CG				CS			
		λ_1		λ_2		λ_1		λ_2	
$p(u_n = \lambda_k, u_{n+1} = \lambda_l)$	λ_1	0.4999		0.0001		0.4999		0.0001	
	λ_2	0.0001		0.4999		0.0001		0.4999	
		ω_1	ω_2	ω_1	ω_2	ω_1	ω_2	ω_1	ω_2
$p(x_n = \omega_i, x_{n+1} = \omega_j \mid u_{n+1} = \lambda_l)$	ω_1	0.49	0.01	0.45	0.05	0.49	0.01	0.48	0.02
	ω_2	0.01	0.49	0.05	0.45	0.01	0.49	0.02	0.48

TAB. 4.2 – Estimation de la loi de (X, U) en utilisant le vrai modèle avec copule gaussienne (CG) et le modèle avec copule de Student (CS).

De ces résultats, nous constatons que le choix d'une copule différente de celle qui a permis de simuler les données peut dégrader fortement les résultats. La moyenne et la variance sont mal estimées par le modèle utilisant une copule de Student, ce qui a pour conséquence une mauvaise estimation des états cachés. Cependant, on peut remarquer que la loi de U ainsi que sa réalisation demeurent bien estimées.

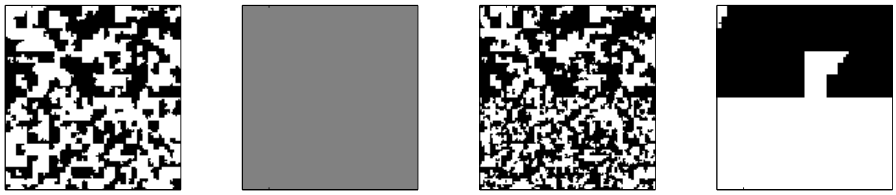
Segmentation de données simulées avec une copule de Student

Dans cette expérience, les données sont simulées à l'aide d'une copule de Student de paramètre de forme $\nu = 10$ et de paramètres de corrélation $\rho_{\omega_1,\omega_1} = \rho_{\omega_2,\omega_2} = 0.9$ et $\rho_{\omega_1,\omega_2} = \rho_{\omega_2,\omega_1} = 0$. Les données sont ensuite segmentées par trois méthodes : la première utilise les vrais paramètres, la seconde utilise le modèle avec copule gaussienne et les paramètres sont estimés avec ICE et la dernière méthode utilise le vrai modèle avec copule de Student et les paramètres également estimés par ICE. Dans tous les cas, la segmentation est faite par l'estimateur du MPM.



(a) Réalisation de X. (b) Réalisation de U. (c) Réalisation de Y. (d) Estimation de $x_{1:N}$: 15.39% d'erreur. (e) Estimation de $u_{1:N}$: 0.73% d'erreur.

FIG. 4.4 – Segmentation avec les vrais paramètres de données simulées avec la copule de Student.



(a) Estimation de $x_{1:N}$ (CG) : 26.04% d'erreur. (b) Estimation de $u_{1:N}$ (CG) : 54.10% d'erreur. (c) Estimation de $x_{1:N}$ (CS) : 21.91% d'erreur. (d) Estimation de $u_{1:N}$ (CS) : 0.79% d'erreur.

FIG. 4.5 – Segmentation de données simulées avec la copule de Student en utilisant le modèle avec copule gaussienne (CG) et le vrai modèle avec copule de Student (CS).

		CG		CS		Vraies valeurs	
		ω_1	ω_2	ω_1	ω_2	ω_1	ω_2
m_{ω_i}		0.36	1.61	−0.13	1.94	0	2
$\sigma^2_{\omega_i}$		2.66	2.88	0.91	1.18	1	1
ρ	ω_1	0.92	−0.01	0.89	0.01	0.9	0
	ω_2	−0.01	0.94	0.01	0.91	0	0.9
Taux d'erreur pour X		26.4%		21.91%		15.39%	
Taux d'erreur pour U		54.10%		0.79%		0.73%	

TAB. 4.3 – Estimation de la loi d'observation en utilisant le modèle avec copule gaussienne (CG) et le vrai modèle avec copule de Student (CS).

		CG				CS			
		λ_1		λ_2		λ_1		λ_2	
$p(u_n = \lambda_k, u_{n+1} = \lambda_l)$	λ_1	~ 0.5		~ 0		0.4999		0.0001	
	λ_2	~ 0		~ 0.5		0.0001		0.4999	
		ω_1	ω_2	ω_1	ω_2	ω_1	ω_2	ω_1	ω_2
$p(x_n = \omega_i, x_{n+1} = \omega_j \mid u_{n+1} = \lambda_l)$	ω_1	0.46	0.04	−	−	0.49	0.01	0.45	0.05
	ω_2	0.04	0.46	−	−	0.01	0.49	0.05	0.45

TAB. 4.4 – Estimation de la loi de (X, U) en utilisant le modèle avec copule gaussienne (CG) et le vrai modèle avec copule de Student (CS).

On constate que remplacer la copule de Student par la copule gaussienne peut, en mode non supervisé, dégrader notablement les résultats. La nature dont les marginales sont corrélées doit être considérée avec précaution. Dans les deux expériences que nous avons présentées, les marginales étaient des lois normales, seule la copule était différente.

Conclusion

Dans ce chapitre, nous nous sommes intéressés, afin de mieux modéliser les différentes variétés des bruits, aux diverses extensions des lois normales. Nous avons tout d'abord rappelé différentes généralisations de la loi normale telles que les modèles exponentiels, les lois elliptiques ou sphériquement invariantes (SIRV). En guise d'exemple de loi de type exponentiel, nous avons cité la loi de Von Mises-Fisher qui sera utilisée dans les applications radar au chapitre 6. Un autre moyen de considérer des lois multivariées non gaussiennes est d'utiliser des copules, qui permettent de modéliser un grand nombre de lois multivariées à lois marginales données. Nous avons ensuite utilisé les copules pour modéliser des observations dépendantes dans le cadre du modèle de Markov caché non stationnaire. Nous avons présenté des simulations, au sein desquelles nous avons utilisé une méthode d'estimation des paramètres de type ICE originale, montrant qu'il est important, dans le cadre de ce modèle et en non supervisé, d'utiliser la vraie copule. Nous verrons au chapitre suivant, comment utiliser les copules pour générer des observations à dépendance longue.

Chapitre 5

Chaînes de Markov triplets avec bruit à dépendance longue

Dans ce chapitre, nous introduisons la notion de dépendance longue dans les modèles de Markov triplets. La dépendance longue, appelée également mémoire longue, se traduit par une corrélation persistante entre les marginales d'un processus. Cette persistance peut être due à des phénomènes fractals, comme dans les bruits gaussiens fractionnaires [78]. Elle peut être également due à des phénomènes saisonniers, comme dans les processus de Gegenbauer [30, 53]. Les phénomènes fractals et saisonniers se retrouvent notamment dans les données financières [56, 114], ou dans les images naturelles [66, 69, 92]. Parmi les autres applications de la dépendance longue, on peut citer le traitement des protocoles TCP/IP [87, 61, 120], ou encore la modélisation des précipitations et intempéries [82].
Nous commençons par définir la notion de processus à dépendance longue et donnons les principaux exemples. Ensuite, nous verrons comment modéliser des observations à dépendance longue à l'aide du modèle des "chaînes couples (resp. triplets) partiellement de Markov" introduit dans [98]. En particulier, nous proposons un nouveau modèle de chaîne semi-markovienne cachée par du bruit à mémoire longue. Nous proposons un algorithme ICE original, demandant des aménagements importants de son principe général, permettant d'estimer les paramètres de notre modèle à observations à dépendance longue. Pour finir, nous proposerons la modélisation de la dépendance longue dans des processus non gaussiens à l'aide des copules. Diverses simulations informatiques valident l'intérêt des nouveaux modèles et traitements non supervisés associés.

5.1 Processus à dépendance longue

5.1.1 Définition

Soit $Y = (Y_n)_{n\in\mathbb{Z}}$ un processus réel. On dira qu'il est du second ordre si pour tout $n \in \mathbb{Z}$, Y_n est de carré intégrable, et qu'il est stationnaire du second ordre si sa covariance $\mathbb{E}(Y_nY_{n-k}) - \mathbb{E}(Y_n)\mathbb{E}(Y_{n-k})$ ne dépend pas de n. On la note alors $\gamma(k)$ et on a $\gamma(k) = \gamma(-k)$. La famille $(\gamma(k))_{k\in\mathbb{Z}}$ sera appelée famille de covariances.

Définition 5.1.1 (Dépendance longue)**.** *Soit $Y = (Y_n)_{n\in\mathbb{Z}}$ un processus réel stationnaire du second ordre et de famille de covariances $(\gamma(k))_{k\in\mathbb{Z}}$.*

Le processus Y est à dépendance longue s'il existe $\alpha \in]0,1]$ et $C \in \mathbb{R}$ tels que :

$$\lim_{k \to +\infty} k^{\alpha} \gamma(k) = C.$$

On notera alors $\gamma(k) \sim_{+\infty} Ck^{-\alpha}$.

Remarque : La covariance d'un processus à mémoire longue satisfait $\sum_{k \in \mathbb{N}} \gamma(k) = +\infty$. Lorsque $\gamma(k) \sim_{+\infty} Ck^{-\alpha}$ pour $\alpha > 1$, on dit que Y est à dépendance intermédiaire.

Dans les deux sous-sections suivantes, nous donnons les principaux exemples de processus à dépendance longue et leurs propriétés.

5.1.2 Processus auto-similaires et bruits gaussiens fractionnaires

Un bruit gaussien fractionnaire est défini comme le processus d'accroissements d'un mouvement brownien fractionnaire, ce dernier étant défini comme un processus à temps continu possédant des propriétés d'auto-similarité. Nous préciserons ces notions ci-après.

Mouvements browniens fractionnaires et auto-similarité

On donne tout d'abord les définitions de processus auto-similaires et à accroissements stationnaires dont les mouvements browniens fractionnaires sont des cas particuliers.

Définition 5.1.2 (Processus auto-similaire)**.** *Un processus réel à temps continu $Z = (Z_t)_{t \in \mathbb{R}}$ est auto-similaire d'indice $H > 0$ si pour tout $a > 0$, le processus $(Z_{at})_{t \in \mathbb{R}}$ suit la même loi que $(a^H Z_t)_{t \in \mathbb{R}}$.*
L'indice H est appelé indice ou paramètre de Hurst.

Définition 5.1.3 (Processus à accroissements stationnaires)**.** *Un processus réel à temps continu $Z = (Z_t)_{t \in \mathbb{R}}$ est à accroissements stationnaires si pour tout $h \in \mathbb{R}$ les variables aléatoires $Z_{t+h} - Z_t$ et $Z_h - Z_0$ ont mêmes lois.*

Définition 5.1.4 (Mouvement brownien fractionnaire)**.** *Un processus réel à temps continu $B = (B_t)_{t \in \mathbb{R}}$ est un mouvement brownien fractionnaire de paramètre de Hurst $H > 0$ s'il est gaussien, auto-similaire de paramètre H et à accroissements stationnaires.*

Un processus stationnaire du second ordre $Z = (Z_t)_{t \in \mathbb{R}}$ auto-similaire d'indice H et à accroissements stationnaires satisfait les propriétés suivantes :

1. $Z_0 = 0$;
2. Si $H \neq 1$, alors $\mathbb{E}(Z_t) = 0$ pour tout $t \in \mathbb{R}$;
3. Z_{-t} et $-Z_t$ suivent la même loi ;
4. Soit $\Gamma_H(s,t) = \mathbb{E}(Z_t Z_s)$ et $\sigma^2 = \mathbb{E}(Z_0^2)$. Si $H \neq 1$, alors :

$$\Gamma_H(s,t) = \frac{\sigma^2}{2}[|t|^{2H} + |s|^{2H} - |t-s|^{2H}] \ ;$$

5. $H \leq 1$;
6. Si $H = 1$, $Z_t = tZ_1$ pour tout $t \in \mathbb{R}$;
7. Si Z est un mouvement brownien fractionnaire de paramètre de Hurst H, alors la dimension de Hausdorff de la trajectoire $(t, Z_t)_{t\in\mathbb{R}}$ est égale à $2 - H$ (pour la définition de la dimension de Hausdorff, se reporter à l'annexe B).

Bruits gaussiens fractionnaires

Un bruit fractionnaire est le processus d'accroissements d'un processus auto-similaire à accroissements stationnaires, soit :

Définition 5.1.5 (Bruit fractionnaire)**.** *Un processus stationnaire du second ordre $Y = (Y_n)_{n\in\mathbb{Z}}$ est un bruit fractionnaire de moyenne m s'il existe un processus auto-similaire à accroissements stationnaires $Z = (Z_t)_{t\in\mathbb{R}}$ tel que :*

$$\forall n \in \mathbb{Z},\ Y_n = Z_{n+1} - Z_n + m.$$

Lorsque le processus Z est un mouvement brownien fractionnaire, Y est appelé "bruit gaussien fractionnaire".

Soit $Y = (Y_n)_{n\in\mathbb{Z}}$ un bruit fractionnaire et $Z = (Z_t)_{t\in\mathbb{R}}$ le processus auto-similaire de paramètre H à accroissements stationnaires associé à Y. Le bruit fractionnaire Y vérifie les propriétés suivantes :

1. Pour tout $n \in \mathbb{Z}$, $\mathbb{E}(Y_n) = m$;
2. Soit $\sigma^2 = \mathbb{E}(Z_0^2)$. La famille de covariances de Y est donnée par :

$$\forall k \in \mathbb{Z},\ \gamma(k) = \frac{\sigma^2}{2}[|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}] \ ; \tag{5.1}$$

3. On a :
   (a) Si $H = \frac{1}{2}$, alors $\gamma(k) = 0$ pour $k \neq 0$;
   (b) Si $H < \frac{1}{2}$, alors $\gamma(k) < 0$ pour $k \neq 0$;
   (c) Si $H > \frac{1}{2}$, alors $\gamma(k) > 0$ pour $k \neq 0$;
4. Si $H \neq \frac{1}{2}$, alors

$$\gamma(k) \sim_{+\infty} \sigma^2 H(2H-1)k^{2H-2} \ ; \tag{5.2}$$

5. On a :
   (a) Si $H = \frac{1}{2}$, alors Y est un bruit blanc ;
   (b) Si $H < \frac{1}{2}$, alors $2H - 2 < -1$ et Y est à dépendance intermédiaire ;
   (c) Si $H > \frac{1}{2}$, alors $2H - 2 > -1$ et Y est à dépendance longue.

5.1.3 Processus FARIMA

Les "Fractional Autoregressive Integrated Moving Average" (FARIMA) sont, sous certaines conditions, à mémoire longue. Avant d'étudier plus précisement les FARIMA, rappelons quelques notions sur les processus stationnaires du second ordre.

Théorème 5.1.1 (Théorème d'Herglotz)**.** *Soit $Y = (Y_n)_{n\in\mathbb{Z}}$ un processus stationnaire du second ordre, à valeurs réelles, de famille de covariances $(\gamma(k))_{k\in\mathbb{Z}}$.*
Si $(\gamma(k))_{k\in\mathbb{Z}}$ est de type positif (toutes les matrices

$$\begin{pmatrix} \gamma(0) & \gamma(1) & \dots & \dots & \gamma(n) \\ \gamma(1) & \gamma(0) & \gamma(1) & \dots & \gamma(n-1) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma(n) & \gamma(n-1) & \vdots & \vdots & \gamma(0) \end{pmatrix}$$

ont leurs valeurs propres strictement positives), alors il existe une unique mesure μ sur $]-\pi,\pi[$, appelée "mesure spectrale" du processus Y telle que :

$$\gamma(k) = \int_{]-\pi,\pi[} e^{-ik\lambda} d\mu(\lambda) \text{ pour tout } k \in \mathbb{Z}. \tag{5.3}$$

De plus, si la famille de covariances est sommable, μ possède une densité f par rapport à la mesure de Lebesgue appelée "densité spectrale" de Y et donnée par :

$$f(\lambda) = \frac{1}{2\pi} \sum_{k\in\mathbb{Z}} \gamma(k) e^{i\lambda k} \text{ pour tout } \lambda \in]-\pi,\pi[. \tag{5.4}$$

On remarque que la mesure spectrale est la transformée de Fourier de la famille de covariances. Notons que la sommabilité de la famille de covariances est une condition suffisante et non nécessaire d'existence de la densité spectrale. En effet, pour les processus FARIMA, la densité spectrale existe, mais ils peuvent être à dépendance longue, auquel cas la famille de covariances n'est pas sommable.

Exemple 5.1.1 (Bruit blanc)**.** Si Y est un bruit blanc de variance σ^2, alors sa densité spectrale est :

$$f(\lambda) = \frac{\sigma^2}{2\pi}, \text{ pour tout } \lambda \in]-\pi,\pi[.$$

Nous introduisons maintenant la définition d'un filtre linéaire. On notera L^2 l'ensemble des variables aléatoires de carré intégrable. On dira qu'une suite $(Z_n)_{n\in\mathbb{N}}$ d'éléments de L^2 converge au sens L^2 vers la variable aléatoire Z si $\lim_{n\to+\infty} \mathbb{E}\left((Z - Z_n)^2\right) = 0$, on rappelle que la limite Z est dans L^2.

Définition 5.1.6.

1. *Un filtre linéaire est une famille de réels $a = (a_k)_{k\in\mathbb{Z}}$;*

2. *Lorsque* $\sum_{k\in\mathbb{Z}} |a_k|^p$ *converge pour* $p \geq 1$, *on dit que le filtre est dans* l^p *et on note* $a \in l^p$. *Lorsque le filtre est dans* l^1, *on dit que le filtre est stable ;*
3. *Si pour tout* $k < 0$, $a_k = 0$, *le filtre est dit causal ;*
4. *On dit que le filtre* a *admet une transformée en* z *si la série* $\sum_{k\in\mathbb{Z}} a_k z^k$ *converge dans un voisinage du cercle unité* $\{z \in \mathbb{C} : |z| = 1\}$. *On notera alors* $a(z) = \sum_{k\in\mathbb{Z}} a_k z^k$ *la transformée en* z *de* a *;*
5. *Soient* $a = (a_k)_{k\in\mathbb{Z}}$ *et* $b = (b_k)_{k\in\mathbb{Z}}$ *deux familles de réels tels que* $a \in l^p$ *et* $b \in l^q$ *avec* $\frac{1}{p} + \frac{1}{q} \geq 1$. *Le produit de convolution de* a *et* b *noté* $a * b$ *est défini par :*

$$(a * b)_n = \sum_{k\in\mathbb{Z}} a_k b_{n-k} \ ;$$

6. *Soit* $Y^0 = (Y^0_n)_{n\in\mathbb{Z}}$ *un processus stationnaire du second ordre et* a *un filtre.*
   - *Si* a *est un filtre stable, alors pour tout* n, *la série* $\sum_{k\in\mathbb{Z}} a_k Y^0_{n-k}$ *converge presque sûrement vers une variable aléatoire de carré intégrable et converge au sens* L^2 *vers la même variable aléatoire. La sortie* Y *du filtre* a *est définie comme la limite de la série* $\sum_{k\in\mathbb{Z}} a_k Y^0_{n-k}$, *que l'on notera* $Y = a * Y^0$ *;*
   - *si* $a \in l^2$ *et si la famille de covariances de* Y^0 *est sommable, alors pour tout* n, *la série* $\sum_{k\in\mathbb{Z}} a_k Y^0_{n-k}$ *converge au sens* L^2. *La sortie du filtre* a *est définie comme la limite dans* L^2 *de la série* $\sum_{k\in\mathbb{Z}} a_k Y^0_{n-k}$, *on la note également* $Y = a * Y^0$.

Remarque : Si le filtre a est stable, alors la série $\sum_{k\in\mathbb{Z}} a_k z^k$ converge pour $|z| = 1$. Mais ceci n'implique pas que a possède une transformée en z. Lorsque a est stable, on notera de même $a(z)$ la somme $\sum_{k\in\mathbb{Z}} a_k z^k$ pour $|z| = 1$.
La mesure spectrale de la sortie d'un filtre linéaire est donnée par la proposition suivante :

Proposition 5.1.1. *Soit* $Y^0 = (Y^0_n)_{n\in\mathbb{Z}}$ *un processus stationnaire du second ordre de famille de covariances* $\gamma_{Y^0} = (\gamma_{Y^0}(k))_{k\in\mathbb{Z}}$ *sommable et de mesure spectrale* μ_{Y^0} *et soit* a *un filtre linéaire dans* l^2.
Soit $Y = a * Y^0 + m$, *où* m *est un réel, alors on a :*

1. $\gamma_Y = a * \check{a} * \gamma_{Y^0}$, *où* $\check{a}_k = a_{-k}$ *;*
2. *Si de plus* a *est stable, la mesure spectrale* μ_Y *de* Y *a une densité par rapport à la mesure* μ_{Y^0} *donnée par :*

$$d\mu_Y(\lambda) = |a(e^{i\lambda})|^2 d\mu_{Y^0}(\lambda).$$

Remarque : La stabilité du filtre a est une condition suffisante et non nécessaire pour que μ_Y admette une densité par rapport à la mesure μ_{Y^0}. Si le filtre a n'est pas stable et si les singularités de $a(z)$ sur le disque unité sont des pôles, il suffit en fait que $\lambda \to |a(e^{i\lambda})|^2$ soit une fonction intégrable par rapport à la mesure μ_{Y^0} pour que μ_Y admette une densité par rapport à la mesure μ_{Y^0}.
On remarque de plus que si le filtre a est dans l^2, alors la famille de covariances γ_Y est bornée mais n'est pas obligatoirement sommable. Par contre, lorsque le filtre a est stable, la famille de covariances γ_Y est sommable. Ainsi si Y est un processus à dépendance longue, le filtre a ne peut pas être stable.

L'exemple le plus connu des processus s'exprimant comme la sortie d'un filtre linéaire est celui des processus "Auto-Regressive Moving Average" (ARMA) :

Définition 5.1.7 (ARMA(p,q))**.** *Soient $p \geq 0$ et $q \geq 0$ deux entiers. Un processus $Y = (Y_n)_{n\in\mathbb{Z}}$ est un ARMA(p,q) de moyenne nulle s'il existe un bruit blanc $B = (B_n)_{n\in\mathbb{Z}}$ de moyenne nulle, deux polynômes $P(X) = 1 + \sum_{k=1}^{p} \alpha_k X^k$ et $Q(X) = 1 + \sum_{k=1}^{q} \beta_k X^k$ tels que :*

1. *P et Q n'ont pas de racines en commun ;*
2. *P et Q n'ont pas de racines sur le cercle unité ni à l'intérieur du disque unité ;*
3. *$Y_n + \sum_{k=1}^{p} \alpha_k Y_{n-k} = B_n + \sum_{k=1}^{q} \beta_k B_{n-k}$, que l'on note $P * Y = Q * B$.*

Un processus Y est un ARMA de moyenne $m \in \mathbb{R}$ s'il existe Y^0 un ARMA de moyenne nulle tel que $Y = Y^0 + m$.

Remarque : Un ARMA(p,q) est bien la sortie d'un filtre linéaire. En effet, comme P et Q n'ont pas de racines à l'intérieur du disque unité ni sur le cercle unité, alors la fonction $z \to \dfrac{Q(z)}{P(z)}$ est holomorphe dans un voisinage du disque unité et donc dans un voisinage du cercle unité. Cette fonction admet un développement de Laurent au voisinage du cercle unité donné par $\dfrac{Q(z)}{P(z)} = \sum_{k\in\mathbb{N}} a_k z^k$ qui est la transformée en z du filtre causal et stable $a = (a_k)_{k\in\mathbb{N}}$.
On montre que $Y = a * B$ ainsi Y est la sortie du filtre linéaire a. Le filtre a est appelé "filtre constructeur".
La proposition suivante montre que les processus ARMA ne sont pas à dépendance longue :

Proposition 5.1.2. *Soit Y un processus ARMA(p,q) de filtre constructeur a et de famille de covariances $(\gamma(k))_{k\in\mathbb{Z}}$. Il existe $\rho \in [0, 1[$ et $K > 0$ tels que :*

$$|\gamma(k)| \leq K\rho^k \text{ pour tout } k \geq 0.$$

Les processus FARIMA que nous introduisons ci-dessous généralisent les processus ARMA et peuvent être à dépendance longue.
Soit d un réel non nul, considérons a la fonction de $\mathbb{C}$ dans $\mathbb{C}$ définie par $a(z) = (1 - z)^{-d}$.

Cette fonction est holomorphe à l'intérieur du disque unité et admet donc un développement de Laurent pour $|z| < 1$ donné par :

$$a(z) = \sum_{k=0}^{+\infty} a_k z^k,$$

où $a_k = \dfrac{\Gamma(d+k)}{\Gamma(d)\Gamma(k+1)}$ (voir Annexe A pour la définition de la fonction Γ). Le filtre causal $a = (a_k)_{k\in\mathbb{N}}$ est stable si $d < 0$ et dans l^2 si $d < \dfrac{1}{2}$.

Définition 5.1.8. *Soient $p \geq 0$ et $q \geq 0$ deux entiers et soit $d \in \left]-\frac{1}{2}, \frac{1}{2}\right[- \{0\}$. Un processus stationnaire du second ordre $Y = (Y_n)_{n\in\mathbb{Z}}$ est un FARIMA(p,d,q) de moyenne $m \in \mathbb{R}$ s'il existe un ARMA(p,q) de moyenne nulle Y^0 tel que :*

$$Y = a * Y^0 + m, \text{ où pour tout } k \geq 0,\ a_k = \frac{\Gamma(d+k)}{\Gamma(d)\Gamma(k+1)}.$$

On supposera dans la suite que $p = q = 0$; le processus Y^0 est donc un bruit blanc de moyenne nulle. On notera σ^2 la variance de ce bruit blanc.
Le filtre a n'est pas stable si $d > 0$; cependant, conformément à la remarque de la proposition 5.1.1, la fonction $a(z) = (1-z)^{-d}$ admet 1 pour seule singularité sur le disque unité et cette singularité est un pôle. De plus, $\left|a(e^{i\lambda})\right|^2 = \dfrac{1}{\left[4\sin^2\left(\frac{\lambda}{2}\right)\right]^d}$ définit une fonction intégrable sur $]-\pi, \pi[$. On en déduit qu'un FARIMA possède une densité spectrale par rapport à la mesure de Lebesgue donnée par :

$$f(\lambda) = \frac{\sigma^2}{2\pi} \times \frac{1}{\left[4\sin^2\left(\frac{\lambda}{2}\right)\right]^d}. \tag{5.5}$$

La proposition suivante donne la famille de covariances d'un FARIMA.

Proposition 5.1.3. *Soit Y un processus FARIMA(0,d,0) avec $d \in [-\frac{1}{2}, \frac{1}{2}[-\{0\}$, sa covariance est donnée par :*

$$\gamma(k) = \sigma^2 \frac{\Gamma(k+d)\Gamma(1-2d)}{\Gamma(k-d+1)\Gamma(d)\Gamma(1-d)} = \sigma_F^2 \frac{\Gamma(1-d)\Gamma(k+d)}{\Gamma(d)\Gamma(k-d+1)}, \tag{5.6}$$

où

$$\sigma_F^2 = \gamma(0) = \sigma^2 \frac{\Gamma(1-2d)}{[\Gamma(1-d)]^2}.$$

Preuve. Voir [19] pages 522-523. □

On déduit de cette proposition que :

$$\gamma(k) \sim_{+\infty} ck^{2d-1} \text{ où } c = \sigma^2 \frac{\Gamma(1-2d)}{\Gamma(d)\Gamma(1-d)} = \sigma_F^2 \frac{\Gamma(1-d)}{\Gamma(d)}. \tag{5.7}$$

On peut alors donner les conditions de dépendance longue d'un FARIMA(0,d,0) :

– si $d = 0$, le FARIMA(0,0,0) est un bruit blanc ;
– si $d < 0$, le FARIMA(0,d,0) est à dépendance intermédiaire ;
– si $d > 0$, le FARIMA(0,d,0) est à dépendance longue.

Concernant le comportement asymptotique de la famille de covariances lorsque k tend vers l'infini, que ce soit pour les bruits gaussiens fractionnaires ou pour les processus FARIMA, elle décroît en puissance lorsque $k \to +\infty$, soit $\gamma(k) \sim_{+\infty} Ck^{-\alpha}$, où :

– $\alpha = 2 - 2H$ pour les bruits gaussiens fractionnaires ;
– $\alpha = 1 - 2d$ pour les processus FARIMA.

Ainsi la covariance d'un processus FARIMA de paramètre de dépendance d a le même comportement asymptotique que celle d'un bruit gaussien fractionnaire de paramètre de Hurst $H = d + \frac{1}{2}$.

On peut faire les mêmes remarques concernant la densité spectrale. La densité spectrale d'un bruit gaussien fractionnaire est donnée par la proposition suivante :

Proposition 5.1.4. *Soit $Y = (Y_n)_{n\in\mathbb{Z}}$ un bruit gaussien fractionnaire de variance σ^2 et de paramètre de Hurst $H \in]0,1[$. Sa densité spectrale existe et est donnée par :*

$$f(\lambda) = \sigma^2 \frac{2H\Gamma(2H)\sin(\pi H)}{\pi} \times (1 - \cos(\lambda)) \times \sum_{k\in\mathbb{Z}} |2k\pi + \lambda|^{-(2H+1)} . \tag{5.8}$$

Preuve. Voir [113] pages 28 à 34. □

De la formule (5.5), au voisinage de 0, la densité spectrale d'un FARIMA se comporte comme $\frac{\sigma^2}{2\pi}|\lambda|^{-2d}$. De la formule (5.8), la densité spectrale d'un bruit gaussien fractionnaire se comporte comme $\frac{H\Gamma(2H)\sin(\pi H)}{\pi}\sigma^2|\lambda|^{1-2H}$. Ainsi, dans les deux cas, la densité spectrale se comporte comme $C|\lambda|^{-\beta}$, où $\beta = 2H - 1$ pour les bruits gaussiens fractionnaires et $\beta = 2d$ pour les processus FARIMA.

5.1.4 Estimation des processus à dépendance longue

Nous étudions dans cette sous-section l'estimation des paramètres pour trois types de processus. Le premier est un processus stationnaire du second ordre de moyenne m et dont la famille de covariances est donnée par :

$$\forall k \in \mathbb{Z},\ \gamma(k) = \sigma^2(|k| + 1)^{-\alpha},\ \text{où } \alpha \in \mathbb{R}^+.$$

Nous étudierons ensuite l'estimation des FARIMA et des bruits gaussiens fractionnaires. L'estimation que nous proposerons sera semi-paramétrique. Nous estimerons la densité spectrale par une méthode non-paramétrique, puis nous estimerons les paramètres à partir d'un échantillon de ce spectre par une méthode paramétrique. Nous choisirons d'étudier d'abord l'estimation des FARIMA, car comme nous le verrons, l'estimation des bruits gaussiens fractionnaires utilise celle des FARIMA.

1. Cas de la covariance $\gamma(k) = \sigma^2(|k|+1)^{-\alpha}$

Les paramètres du modèle sont la moyenne m, la variance σ^2 et le paramètre de dépendance α. Soit $y_{1:N} = (y_1, \ldots, y_N)$ une réalisation du processus à dépendance longue. Les estimateurs sont donnés par :

$$\hat{m} = \frac{1}{N}\sum_{n=1}^{N} y_n, \tag{5.9}$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{m})^2, \tag{5.10}$$

$$\hat{\alpha} = -\frac{1}{\log(2)}\log\left(\frac{\hat{\gamma(1)}}{\hat{\sigma}^2}\right), \tag{5.11}$$

où

$$\begin{aligned}
\hat{\gamma(1)} &= \frac{1}{N-1}\sum_{n=1}^{N-1}(y_n - \hat{m}_1)(y_{n+1} - \hat{m}_2),\\
\hat{m}_1 &= \frac{1}{N-1}\sum_{n=1}^{N-1} y_n,\\
\hat{m}_2 &= \frac{1}{N-1}\sum_{n=1}^{N-1} y_{n+1}.
\end{aligned}$$

2. Cas des processus FARIMA

On pourrait estimer $\alpha = 1-2d$ (resp. $\alpha = 2-2H$) par la méthode précédente, cependant, l'estimateur proposé précédemment donne de mauvais résultats dans des modèles plus spécifiques comme les FARIMA ou les bruits gaussiens fractionnaire. Nous proposons alors une estimation semi-paramétrique inspirée de celle de E. Moulines et P. Soulier [84]. Les paramètres du modèle sont la moyenne m, la variance du bruit blanc filtré σ^2 (resp. la variance du processus FARIMA σ_F^2) et le paramètre de dépendance d.

Etape non paramétrique : estimation du spectre
Soit $y_{1:N} = (y_1, \ldots, y_N)$ une réalisation d'un processus FARIMA. La densité spectrale est alors estimée pour $\lambda \in [-\pi, \pi]$ par :

$$\hat{f}(\lambda) = \frac{1}{2\pi N}\left|\sum_{n=1}^{N}(y_n - \bar{y})\exp(in\lambda)\right|^2, \tag{5.12}$$

où $\bar{y} = \frac{1}{N}\sum_{n=1}^{N} y_n$.

Etape paramétrique : estimation de d et de σ
Une fois la densité spectrale estimée, nous disposons d'un échantillon $(\hat{f}(\lambda_1), \ldots, \hat{f}(\lambda_M))$, où

$\lambda_j = -\pi + \dfrac{2j\pi}{M}$. D'après (5.5), la densité spectrale d'un FARIMA vérifie :

$$\log(f(\lambda)) = \log\left(\frac{\sigma^2}{2\pi}\right) - d\log\left(4\sin^2\left(\frac{\lambda}{2}\right)\right).$$

Ainsi, posant $\omega_j = \log\left(4\sin^2\left(\dfrac{\lambda_j}{2}\right)\right)$, $z_j = \log(\hat{f}(\lambda_j))$ et $B = \log\left(\dfrac{\sigma^2}{2\pi}\right)$, nous devons chercher B et d minimisant :

$$\sum_{j=1}^{M} \left[z_j - (B - d\omega_j)\right]^2.$$

Les estimateurs au sens des moindres carrés de B et d sont alors :

$$\begin{aligned} \hat{d} &= -\frac{\hat{\sigma}_{\omega,z}}{\hat{\sigma}_{\omega}^2}, \\ \hat{B} &= \bar{z} + \hat{d}\bar{\omega}, \end{aligned} \tag{5.13}$$

où $\bar{\omega} = \dfrac{1}{M}\sum_{j=1}^{M}\omega_j$, $\bar{z} = \dfrac{1}{M}\sum_{j=1}^{M} z_j$, $\hat{\sigma}_{\omega}^2 = \dfrac{1}{M}\sum_{j=1}^{M}(\omega_j - \bar{\omega})^2$ et $\hat{\sigma}_{\omega,z} = \dfrac{1}{M}\sum_{j=1}^{M}(\omega_j - \bar{\omega})(z_j - \bar{z})$.
L'estimation de σ est donnée par :

$$\hat{\sigma}^2 = 2\pi\exp(\hat{B}), \tag{5.14}$$

et donc la variance du FARIMA est estimée par :

$$\hat{\sigma}_F^2 = 2\pi\exp(\hat{B}) \times \frac{\Gamma(1-2\hat{d})}{\left[\Gamma(1-\hat{d})\right]^2}.$$

Estimation de la moyenne m
La moyenne m ne peut être estimée à partir du spectre. Nous utilisons l'estimateur empirique proposé en (5.9).

3. Cas des bruits gaussiens fractionnaires

Comme la densité spectrale (5.8) d'un bruit gaussien fractionnaire est difficile à exploiter directement, on utilise le fait qu'elle a le même comportement asymptotique lorsque λ tend vers 0 que celle d'un processus FARIMA. Les paramètres à estimer sont la moyenne m, la variance du bruit gaussien fractionnaire σ^2 et le paramètre de Hurst H. Soit $y_{1:N} = (y_1, \ldots, y_N)$ la réalisation d'un bruit gaussien fractionnaire.

Estimation de la moyenne m
De la même façon que précédemment, l'estimateur de la moyenne est celui proposé en (5.9).

Estimation de la variance et du paramètre de Hurst
Soit $(\hat{f}(\lambda_1), \ldots, \hat{f}(\lambda_M))$ un échantillon du spectre estimé par (5.12) tel que les λ_j soient les

plus proches possibles de 0, on prendra dans les expérimentations, $\lambda_j = -10^{-2} + \frac{2j10^{-2}}{M}$ et $M = 100$. On commence par estimer d et B en utilisant la méthode des moindres carrés décrite précédemment. On pose :

$$\hat{H} = \hat{d} + \frac{1}{2}. \tag{5.15}$$

Pour l'estimation de la variance, on remarque qu'au voisinage de 0, la densité spectrale d'un FARIMA est équivalente à $\frac{\sigma_0^2}{2\pi} \times |\lambda|^{-2d}$ lorsque σ_0^2 est la variance du bruit blanc filtré et celle d'un bruit gaussien fractionnaire est équivalente à $\frac{H\Gamma(2H)\sin(\pi H)}{\pi}\sigma^2|\lambda|^{1-2H}$, on pose ainsi :

$$\hat{\sigma}^2 = \frac{\pi}{\hat{H}\Gamma(2\hat{H})\sin(\pi\hat{H})}\exp(\hat{B}). \tag{5.16}$$

L'estimateur du paramètre α dans le cas où la covariance est donnée par $\gamma(k) = \sigma^2(|k| + 1)^{-\alpha}$ sera utilisé dans les deux sections suivantes. Nous utiliserons l'estimation semi-paramétrique de bruit gaussien fractionnaire dans la section 5.4.

5.2 Chaînes couples partiellement de Markov

Nous définissons ici le modèle de chaînes couples partiellement de Markov. Nous proposerons ensuite un modèle de chaînes couples partiellement de Markov particulier permettant des traitements bayésiens non supervisés dans le cadre des observations à dépendance longue.

5.2.1 Chaînes couples partiellement de Markov : modèle général

Soient $X = (X_n)_{1\leq n\leq N}$ et $Y = (Y_n)_{1\leq n\leq N}$ deux processus aléatoires tels que chaque X_n prend ses valeurs dans l'ensemble fini $\mathcal{X} = \{\omega_1, \ldots, \omega_K\}$ et chaque Y_n prend ses valeurs dans $\mathbb{R}$.

Définition 5.2.1 (Chaînes couples partiellement de Markov)**.** *Le processus $Z = (X_n, Y_n)_{1\leq n\leq N}$ est une chaîne couple partiellement de Markov (CCPM) si sa distribution $p(z_{1:N})$ vérifie :*

$$p(z_{1:N}) = p(z_1)\prod_{n=1}^{N-1} p(z_{n+1}|x_n, y_{1:n}), \tag{5.17}$$

où $y_{1:n} = (y_1, \ldots, y_n)$.

Remarque : A ce stade, nous disposons des modèles suivants :
- chaînes de Markov cachées à bruit indépendant (CMC-BI) ;
- chaînes de Markov couples (CMCouple) ;
- chaînes couples partiellement de Markov (CCPM).

Nous avions vu au chapitre 3 que le modèle CMCouple était plus général que le modèle CMC-BI. Dans le modèle CMCouple, le processus X n'est plus obligatoirement markovien et les variables aléatoires Y_n ne sont plus obligatoirement indépendantes conditionnellement à

X. Quant au modèle CCPM, il généralise le modèle CMCouple. Dans le modèle CCPM, le processus Z n'est plus obligatoirement une chaîne de Markov, ce qui accorde au modèle une plus grande richesse.
En utilisant l'équivalence entre markovianité et factorisation d'une loi vue au chapitre 2, on montre que si $p(x_{1:N}, y_{1:N})$ est la distribution d'une couple partiellement de Markov, alors $p(x_{1:N}|y_{1:N})$ est la distribution d'une chaîne de Markov. La proposition ci-dessous détaille l'algorithme de Baum-Welsh conditionnel dans le cas du modèle CCPM. Celui-ci permet de calculer les lois a posteriori $p(x_n|y_{1:N})$ et $p(x_{n+1}|x_n, y_{1:N})$, à condition toutefois que la distribution $p(z_{n+1}|x_n, y_{1:n})$ soit calculable.

Proposition 5.2.1 (Algorithme de Baum-Welsh)**.** *Soit $Z = (X_n, Y_n)_{1\leq n\leq N}$ une chaîne couple partiellement de Markov. Les probabilités a posteriori $p(x_n|y_{1:N})$ et $p(x_{n+1}|x_n, y_{1:N})$ sont données par :*

$$p(x_n|y_{1:N}) = \tilde{\alpha}_n(x_n)\tilde{\beta}_n(x_n), \tag{5.18}$$

$$p(x_{n+1}|x_n, y_{1:N}) \propto \frac{\tilde{\beta}_{n+1}(x_{n+1})}{\tilde{\beta}_n(x_n)} p(z_{n+1}|x_n, y_{1:n}), \tag{5.19}$$

$$\text{où} \quad \tilde{\alpha}_1(x_1) = p(x_1|y_1) \quad \text{et} \quad \tilde{\alpha}_{n+1}(x_{n+1}) \propto \sum_{x_n} \tilde{\alpha}_n(x_n) p(z_{n+1}|x_n, y_{1:n}) \text{ pour } n \geq 1,$$

$$\tilde{\beta}_N(x_N) = 1 \quad \text{et} \quad \tilde{\beta}_n(x_n) = \frac{\sum_{x_{n+1}} \tilde{\beta}_{n+1}(x_{n+1}) p(z_{n+1}|x_n, y_{1:n})}{\sum_{x_{n+1}} \sum_{x_n} \tilde{\alpha}_n(x_n) p(z_{n+1}|x_n, y_{1:n})} \text{ pour } n \leq N-1.$$

5.2.2 Observations gaussiennes à dépendance longue

Nous proposons dans cette sous-section un modèle de chaînes couples partiellement de Markov pour lequel la quantité $p(z_{n+1}|x_n, y_{1:n})$ est calculable. Nous verrons par la suite comment ce modèle permet de considérer des observations à dépendance longue.
Le modèle particulier est défini par les trois hypothèses suivantes :

1. $p(x_{n+1}|x_n, y_{1:n}) = p(x_{n+1}|x_n)$;
2. $p(y_{n+1}|x_n, x_{n+1}, y_{1:n}) = p(y_{n+1}|x_{n+1}, y_{1:n})$;
3. Les distributions $p(y_{n+1}|x_{n+1}, y_{1:n})$ sont des lois normales définies par :
   - $p(y_1|x_1)$ est une distribution gaussienne de moyenne m_{x_1} et de variance $\gamma_{x_1}(0)$;
   - $p(y_{n+1}|x_{n+1}, y_{1:n})$ est une distribution gaussienne de moyenne $\tilde{m}_{x_{n+1}}$ et de variance $\tilde{\gamma}_{x_{n+1}}$ données par :
   $$\begin{aligned} \tilde{m}_{x_{n+1}} &= m_{x_{n+1}} + \Gamma^{2,1}_{x_{n+1}} (\Gamma^n_{x_{n+1}})^{-1} (y_{1:n} - m^n_{x_{n+1}}), \\ \tilde{\gamma}_{x_{n+1}} &= \gamma_{x_{n+1}}(0) - \Gamma^{2,1}_{x_{n+1}} (\Gamma^n_{x_{n+1}})^{-1} \Gamma^{1,2}_{x_{n+1}}, \end{aligned}$$

où $m^n_{x_{n+1}} = (\underbrace{m_{x_{n+1}}, \ldots, m_{x_{n+1}}}_{\text{n fois}})$,

$$\Gamma^n_{x_{n+1}} = \begin{pmatrix} \gamma_{x_{n+1}}(0) & \gamma_{x_{n+1}}(1) & \ldots & \gamma_{x_{n+1}}(n-2) & \gamma_{x_{n+1}}(n-1) \\ \gamma_{x_{n+1}}(1) & \gamma_{x_{n+1}}(0) & \ldots & \gamma_{x_{n+1}}(n-3) & \gamma_{x_{n+1}}(n-2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma_{x_{n+1}}(n-1) & \gamma_{x_{n+1}}(n-2) & \ldots & \gamma_{x_{n+1}}(1) & \gamma_{x_{n+1}}(0) \end{pmatrix}$$

est une matrice symétrique définie positive et $\Gamma^{2,1}_{x_{n+1}} = \big(\gamma_{x_{n+1}}(n), \ldots, \gamma_{x_{n+1}}(1)\big) = (\Gamma^{1,2}_{x_{n+1}})^T$ est un vecteur de taille $n \times 1$;

La première hypothèse implique que X est une chaîne de Markov. Selon le point 3., les distributions $p(z_{n+1}|x_n, y_{1:n})$ sont calculables, il en résulte que les probabilités a posteriori $p(x_n|y_{1:N})$ et $p(x_{n+1}|x_n, y_{1:N})$ sont calculables par la proposition 5.2.1. Lorsque l'une au moins des familles de covariance γ_{ω_j} est à mémoire longue, ce modèle sera appelé "chaîne de Markov cachée à mémoire longue" (CMC-ML).
La proposition suivante résume les propriétés classiques des vecteurs gaussiens.

Proposition 5.2.2.

– *Soient W^1 et W^2 deux vecteurs réels gaussiens de dimensions respectives d^1 et d^2, de moyennes respectives M^1 et M^2, de matrices de covariance respectives Γ^1 et Γ^2 et tels que le vecteur joint $W = (W^1, W^2)$, de densité $p(w^1, w^2)$, soit un vecteur gaussien de covariance :*

$$\Gamma = \begin{pmatrix} \Gamma^1 & \Gamma^{1,2} \\ \Gamma^{2,1} & \Gamma^2 \end{pmatrix}.$$

Alors la densité $p(w^2|w^1)$ est celle d'une loi normale de moyenne :

$$M^{2|1} = M^2 + \Gamma^{2,1}(\Gamma^1)^{-1}(w^1 - M^1),$$

et de matrice de covariance :

$$\Gamma^{2|1} = \Gamma^2 - \Gamma^{2,1}(\Gamma^1)^{-1}\Gamma^{1,2}.$$

– *Réciproquement, soient W^1 et $(W^2|W^1 = w^1)$ deux vecteurs réels gaussiens de dimensions respectives d^1 et d^2 et de densités $p(w^1)$ et $p(w^2|w^1)$. Si W^1 a pour moyenne M^1 et covariance Γ^1 et si $(W^2|W^1 = w^1)$ a pour moyenne $M^{2|1} = Aw^1 + B$ et pour covariance $\Gamma^{2|1}$, alors le vecteur joint $W = (W^1, W^2)$ de densité $p(w^1, w^2)$ est gaussien de moyenne :*

$$M = \begin{pmatrix} M^1 \\ AM^1 + B \end{pmatrix},$$

et de matrice de covariance :

$$\Gamma = \begin{pmatrix} \Gamma^1 & \Gamma^1 A^T \\ A\Gamma^1 & \Gamma^{2|1} + A\Gamma^1 A^T \end{pmatrix}.$$

La proposition 5.2.2 nous permettra également de calculer les lois gaussiennes $p(y_{1:n}|x_{1:n})$, qui seront utiles dans l'estimation des paramètres. Plus précisément :

Proposition 5.2.3. *Si $p(y_1|x_1)$ suit la loi normale $\mathcal{N}_{\mathbb{R}}\left(m_{x_1}, \gamma_{x_1}(0)\right)$ et si $p(y_{n+1}|x_{n+1}, y_{1:n})$ suit la loi normale :*

- *de moyenne $m_{x_{n+1}} + \Gamma^{2,1}_{x_{n+1}}(\Gamma^n_{x_{n+1}})^{-1}(y_{1:n} - m^n_{x_{n+1}})$;*
- *de variance $\gamma_{x_{n+1}}(0) - \Gamma^{2,1}_{x_{n+1}}(\Gamma^n_{x_{n+1}})^{-1}\Gamma^{1,2}_{x_{n+1}}$.*

Alors $p(y_{1:n}|x_{1:n})$ suit la loi normale de $\mathbb{R}^n$, $\mathcal{N}_{\mathbb{R}^n}\left(M^{x_{1:n}}, \Gamma^{x_{1:n}}\right)$, où $M^{x_{1:n}}$ et $\Gamma^{x_{1:n}}$ sont calculés par les récursions suivantes :

1. *Initialisation :*

$$M^{x_1} = m_{x_1} \text{ et } \Gamma^{x_1} = \gamma_{x_1}(0) \ ;$$

2. *Itération :*

$$M^{x_{1:n+1}} = \begin{pmatrix} M^{x_{1:n}} \\ m_{x_{n+1}} + \Gamma^{2,1}_{x_{n+1}}(\Gamma^n_{x_{n+1}})^{-1}\left[M^{x_{1:n}} - m^n_{x_{n+1}}\right] \end{pmatrix},$$

$$\Gamma^{x_{1:n+1}} = \begin{pmatrix} \Gamma^{x_{1:n}} & \Gamma^{x_{1:n}}(\Gamma^n_{x_{n+1}})^{-1}\Gamma^{1,2}_{x_{n+1}} \\ \Gamma^{2,1}_{x_{n+1}}(\Gamma^n_{x_{n+1}})^{-1}\Gamma^{x_{1:n}} & \gamma_{x_{n+1}}(0) - \Gamma^{2,1}_{x_{n+1}}(\Gamma^n_{x_{n+1}})^{-1}\left[\Gamma^{1,2}_{x_{n+1}} - \Gamma^{x_{1:n}}(\Gamma^n_{x_{n+1}})^{-1}\Gamma^{1,2}_{x_{n+1}}\right] \end{pmatrix}.$$

Preuve. Voir [66]. □

Notons que le modèle est relativement complexe car $p(y_n|x_{1:n})$ dépend de tous les x_k pour $k \leq n$. Nous reviendrons sur cette remarque lors de l'estimation des paramètres.

5.2.3 Estimation des paramètres

Nous détaillons dans cette sous-section l'algorithme ICE dans le cas d'un modèle CMC-ML gaussien. Lorsque chaque X_n prend ses valeurs dans l'ensemble fini $\mathcal{X} = \{\omega_1, \ldots, \omega_K\}$, les paramètres à estimer sont les K moyennes m_{ω_j}, les K variances $\sigma^2_{\omega_j} = \gamma_{\omega_j}(0)$ et les K paramètres de dépendance qui selon le modèle sont :

- α_{ω_j} lorsque $\gamma_{\omega_j}(k) = \sigma^2_{\omega_j}(k+1)^{-\alpha_{\omega_j}}$;
- H_{ω_j} lorsque $(\gamma_{\omega_j}(k))_{k\in\mathbb{N}}$ est la famille de covariances d'un bruit gaussien fractionnaire ;
- d_{ω_j} lorsque $(\gamma_{\omega_j}(k))_{k\in\mathbb{N}}$ est la famille de covariances d'un processus FARIMA ;

et les K^2 paramètres $p(x_n = \omega_i, x_{n+1} = \omega_j)$. On suppose que la chaîne X est une chaîne de Markov stationnaire et réversible.

Afin d'utiliser l'algorithme ICE, nous devons nous donner un estimateur à partir des données complètes. Concernant les paramètres $p_{i,j}$, nous considérerons l'estimateur classique :

$$\hat{p}_{i,j}(x_{1:N}, y_{1:N}) = \frac{1}{2(N-1)}\sum_{n=1}^{N-1}\left[I(x_n = \omega_i, x_{n+1} = \omega_j) + I(x_n = \omega_j, x_{n+1} = \omega_i)\right]. \quad (5.20)$$

L'espérance conditionnelle $\mathbb{E}\left(\hat{p}_{i,j}(X, y_{1:N})|y_{1:N}; \theta_q\right)$ est calculable et la re-estimation $p^{q+1}_{i,j}$ de $p_{i,j}$ donne :

$$p^{q+1}_{i,j} = \frac{1}{2(N-1)}\sum_{n=1}^{N-1}\left[p(x_n = \omega_i, x_{n+1} = \omega_j|y_{1:N}; \theta_q) + p(x_n = \omega_j, x_{n+1} = \omega_i|y_{1:N}; \theta_q)\right]. \quad (5.21)$$

L'estimation des paramètres m_{ω_j}, $\sigma^2_{\omega_j}$ et des paramètres de dépendance est plus délicate car si $x_{n+1} \neq x_n$, la covariance de $p(y_{1:n+1}|x_{1:n+1})$ n'est pas égale à $\Gamma^{n+1}_{x_{n+1}}$.
Pour voir les difficultés liées à l'estimation de ces paramètres, considérons l'exemple suivant :

Exemple 5.2.1. Supposons $K = 2$ et $N = 10$ et considérons le modèle où la covariance est donnée par $\gamma(k) = \sigma^2 (k+1)^{-\alpha}$. On observe un échantillon $(x_{1:10}, y_{1:10})$ et on cherche à estimer les moyennes m_{ω_1} et m_{ω_2}, les variances $\sigma^2_{\omega_1}$ et $\sigma^2_{\omega_2}$ et les paramètres de dépendance α_{ω_1} et α_{ω_2}.
Supposons que l'on ait $x_{1:10} = (\omega_1, \omega_1, \omega_2, \omega_2, \omega_1, \omega_1, \omega_2, \omega_2, \omega_2, \omega_2)$. S'il s'agissait du modèle classique de chaînes de Markov cachées à bruit corrélé, nous aurions $p(y_n|x_{1:10}) = p(y_n|x_n)$. Sous une telle hypothèse, les estimateurs des moyennes et des covariances seraient donnés par :

$$\begin{aligned}
\hat{m}_{\omega_1}(x_{1:10}, y_{1:10}) &= \frac{y_1 + y_2 + y_5 + y_6}{4}, \\
\hat{m}_{\omega_2}(x_{1:10}, y_{1:10}) &= \frac{y_3 + y_4 + y_7 + y_8 + y_9 + y_{10}}{6}, \\
\hat{\gamma}_{\omega_1}(1) &= \frac{(y_1 - \hat{m}_{\omega_1}, y_2 - \hat{m}_{\omega_1}) \begin{pmatrix} y_1 - \hat{m}_{\omega_1} \\ y_2 - \hat{m}_{\omega_1} \end{pmatrix} + (y_5 - \hat{m}_{\omega_1}, y_6 - \hat{m}_{\omega_1}) \begin{pmatrix} y_5 - \hat{m}_{\omega_1} \\ y_6 - \hat{m}_{\omega_1} \end{pmatrix}}{2},
\end{aligned}$$

et une formule similaire pour $\hat{\gamma}_{\omega_2}(1)$. L'utilisation de la dernière formule est possible dans le cas du modèle classique de chaînes de Markov cachées car $p(y_n, y_{n+1}|x_{1:n+1}) = p(y_n, y_{n+1}|x_n, x_{n+1})$. Cependant, cette égalité n'est plus vraie dans le cas des chaînes de Markov cachées dont les observations sont à dépendance longue. Considérons les échantillons $x_{1:4} = (\omega_1, \omega_1, \omega_2, \omega_2)$ et $y_{1:4}$ extraits de $x_{1:10}$ et de $y_{1:10}$. La matrice

$$\Gamma^2_{\omega_1} = \begin{pmatrix} \gamma_{\omega_1}(0) & \gamma_{\omega_1}(1) \\ \gamma_{\omega_1}(1) & \gamma_{\omega_1}(0) \end{pmatrix} \tag{5.22}$$

est bien la matrice de covariance de $p(y_1, y_2|x_1, x_2)$ tandis que la matrice

$$\Gamma^2_{\omega_2} = \begin{pmatrix} \gamma_{\omega_2}(0) & \gamma_{\omega_2}(1) \\ \gamma_{\omega_2}(1) & \gamma_{\omega_2}(0) \end{pmatrix} \tag{5.23}$$

n'est plus la matrice de covariance de $p(y_3, y_4|x_{1:4})$.
Il en est de même si le modèle considéré est un modèle FARIMA ou bruit gaussien fractionnaire. La densité spectrale de $p(y_1, y_2|x_1, x_2)$ est bien la transformée de Fourier de la famille de covariances définie par (5.22) mais celle de $p(y_3, y_4|x_{1:4})$ n'est pas la transformée de Fourier de la famille de covariances définie par (5.23).

Revenons au problème général. Notons $(m^{(q)}_{\omega_j})_{1\leq j\leq K}$, $(\Gamma^{n,(q)}_{\omega_j})_{1\leq j\leq K}$ les moyennes et les matrices de covariance correspondant aux paramètres estimés lors de l'étape q de ICE. D'après ce qui précède, si $x_{n_1} = \ldots = x_{n_1+n_2}$, sous le paramètre θ_q, la distribution $p(y_{n_1:n_1+n_2}|x_{1:n_1+n_2})$ n'est pas la distribution normale de moyenne $m^{n_2+1,(q)}_{x_{n_1}} = (\underbrace{m^{(q)}_{x_{n_1}}, \ldots, m^{(q)}_{x_{n_1}}}_{n_2+1 \text{ fois}})$ et de covariance $\Gamma^{n_2+1,(q)}_{x_{n_1}}$. En fait, la distribution $p(y_{n_1:n_1+n_2}|x_{1:n_1+n_2})$ est la distribution marginale de la loi normale $p(y_{1:n_1+n_2}|x_{1:n_1+n_2})$ de moyenne $M^{x_{1:n_1+n_2},(q)}$ et de covariance $\Gamma^{x_{1:n_1+n_2},(q)}$ calculées

par la procédure décrite dans la proposition 5.2.3 avec $m_{\omega_j} = m_{\omega_j}^{(q)}$ et $\Gamma_{\omega_j}^{n} = \Gamma_{\omega_j}^{n,(q)}$. Notons $M^{x_{n_1:n_1+n_2},(q)}$ et $\Gamma^{x_{n_1:n_1+n_2},(q)}$ la moyenne et la covariance de la distribution $p(y_{n_1:n_1+n_2}|x_{1:n_1+n_2})$ sous θ_q. Afin de pouvoir re-estimer les paramètres par l'algorithme ICE, nous devons effectuer une transformation de l'échantillon, soit $\tilde{y}_{n_1:n_1+n_2} = Ay_{n_1:n_1+n_2}$, de façon à ce que $p(\tilde{y}_{n_1:n_1+n_2}|x_{1:n_1+n_2})$ suive une loi normale de moyenne $m_{x_{n_1}}^{n_2+1,(q)}$ et de covariance $\Gamma_{x_{n_1}}^{n_2+1,(q)}$. Notons $\Gamma^{x_{n_1:n_1+n_2},(q)} = CC^T$ et $\Gamma_{x_{n_1}}^{n_2+1,(q)} = DD^T$ les transformées de Choleski des matrices $\Gamma^{x_{n_1:n_1+n_2},(q)}$ et $\Gamma_{x_{n_1}}^{n_2+1,(q)}$. Alors la transformation qui convient est :

$$\tilde{y}_{n_1:n_1+n_2} = DC^{-1}\left(y_{n_1:n_1+n_2} - M^{x_{n_1:n_1+n_2},(q)}\right)(C^T)^{-1}D^T + m_{x_{n_1}}^{n_2+1,(q)}. \tag{5.24}$$

Finalement, l'algorithme ICE proposé fonctionne de la manière suivante :

1. Sous θ_q, calculer les distributions $p(y_{1:n}|x_{1:n})$ en utilisant la proposition 5.2.3 avec $m_{\omega_i} = m_{\omega_i}^{(q)}$ et $\Gamma_{\omega_i}^{n} = \Gamma_{\omega_i}^{n,(q)}$ pour tout $1 \leq i \leq K$;
2. considérer $J(i) = \{n_1, \ldots, n_r\}$ sous-ensemble de $\{1, \ldots, N\}$, tel que pour tout $1 \leq j \leq r$, $x_{n_j-1} \neq x_{n_j}$ et il existe $m_j \geq 0$ tel que $x_{n_j} = x_{n_j+1} = \ldots = x_{n_j+m_j} = \omega_i$ et $x_{n_j} \neq x_{n_j+m_j+1}$. Considérer les r échantillons correspondants $(y_{n_1}, \ldots, y_{n_1+m_1})$,..., $(y_{n_r}, \ldots, y_{n_r+m_r})$;
3. soient, pour tout $1 \leq j \leq r$, $M^{x_{n_j:n_j+m_j},(q)}$ et $\Gamma^{x_{n_j:n_j+m_j},(q)}$ les moyennes et variances de $p(y_{n_j}, \ldots, y_{n_j+m_j}|x_{1:n_j+m_j})$ calculés à l'étape 1. Calculer les transformées de Choleski $\Gamma^{x_{n_j:n_j+m_j},(q)} = C_jC_j^T$ et $\Gamma_{\omega_i}^{m_j+1,(q)} = D_jD_j^T$ et poser :

$$\begin{pmatrix} \tilde{y}_{n_j} \\ \tilde{y}_{n_j+1} \\ \vdots \\ \tilde{y}_{n_j+m_j} \end{pmatrix} = D_jC_j^{-1}\left(\begin{pmatrix} y_{n_j} \\ y_{n_j+1} \\ \vdots \\ y_{n_j+m_j} \end{pmatrix} - M^{x_{n_j:n_j+m_j},(q)}\right)(C_j^T)^{-1}D_j^T + m_{\omega_i}^{m_j+1,q} ;$$

4. estimer m_i, $\gamma_i(0)$ et le paramètre de dépendance correspondant à la classe i à partir des échantillons $(\tilde{y}_{n_j}, \ldots, \tilde{y}_{n_j+m_j})$ par les estimateurs de la sous-section 5.1.4.

Remarque : Lorsque la covariance est de la forme $\gamma(k) = \sigma^2(k+1)^{-\alpha}$, l'estimation de α utilise seulement les estimations de $\gamma(0)$ et $\gamma(1)$, on peut donc se contenter d'échantillons $(\tilde{y}_{n_j}, \tilde{y}_{n_j+1})$ de taille $m_j + 1 = 2$. L'algorithme ICE s'écrit alors :

1. Sous θ_q, calculer les distributions $p(y_{1:n}|x_{1:n})$ en utilisant la proposition 5.2.3 avec $m_{\omega_i} = m_{\omega_i}^{(q)}$ et $\Gamma_{\omega_i}^{n} = \Gamma_{\omega_i}^{n,(q)}$ pour tout $1 \leq i \leq K$;
2. considérer $J(i) = \{n_1, \ldots, n_r\}$ sous-ensemble de $\{1, \ldots, N\}$, tel que pour tout $1 \leq j \leq r, x_{n_j} = x_{n_j+1} = \omega_i$. Considérer les r échantillons correspondants (y_{n_1}, y_{n_1+1}),..., (y_{n_r}, y_{n_r+1}) ;
3. soient, pour tout $1 \leq j \leq r$, $M^{x_{n_j:n_j+1},(q)}$ et $\Gamma^{x_{n_j:n_j+1},(q)}$ les moyennes et variances de $p(y_{n_j}, y_{n_j+1}|x_{1:n_j+1})$ calculées à l'étape 1. Calculer les transformées de Choleski $\Gamma^{x_{n_j:n_j+1},(q)} = C_jC_j^T$ et $\Gamma_{\omega_i}^{2,(q)} = D_jD_j^T$ et poser :

$$\begin{pmatrix} \tilde{y}_{n_j} \\ \tilde{y}_{n_j+1} \end{pmatrix} = D_jC_j^{-1}\left(\begin{pmatrix} y_{n_j} \\ y_{n_j+1} \end{pmatrix} - M^{x_{n_j:n_j+1},(q)}\right)(C_j^T)^{-1}D_j^T + m_{\omega_i}^{2,q} ;$$

4. estimer les paramètres m_{ω_i}, $\sigma^2_{\omega_i}$ et α_{ω_i} par :

$$m^{q+1}_{\omega_i} = \frac{1}{2}\left(m^{q+1}_{\omega_i,1} + m^{q+1}_{\omega_i,2}\right) \ ;$$

$$(\sigma^{q+1}_{\omega_i})^2 = \frac{1}{2}\left(\hat{\gamma}_1(0) + \hat{\gamma}_2(0)\right) \ ;$$

$$\alpha^{q+1}_{\omega_i} = -\frac{\log\left(\dfrac{\hat{\gamma}(1)}{(\sigma^{q+1}_{\omega_i})^2}\right)}{\log(2)},$$

où

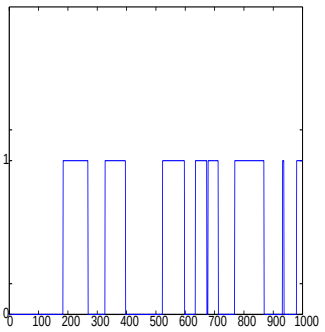
$$\begin{pmatrix} m^{q+1}_{\omega_i,1} \\ m^{q+1}_{\omega_i,2} \end{pmatrix} = \frac{1}{r}\sum_{j=1}^{r}\begin{pmatrix} \tilde{y}_{n_j} \\ \tilde{y}_{n_j+1} \end{pmatrix},$$

et

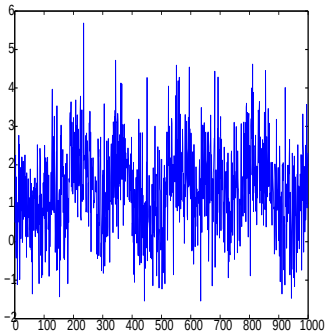
$$\begin{aligned} \Gamma^{2,(q+1)}_{\omega_i} &= \frac{1}{r}\sum_{j=1}^{r}\begin{pmatrix} \tilde{y}_{n_j} - m^{q+1}_{\omega_i,1} \\ \tilde{y}_{n_j+1} - m^{q+1}_{\omega_i,2} \end{pmatrix}\left(\tilde{y}_{n_j} - m^{q+1}_{\omega_i,1}, \tilde{y}_{n_j+1} - m^{q+1}_{\omega_i,2}\right) \\ &= \begin{pmatrix} \hat{\gamma}_1(0) & \hat{\gamma}(1) \\ \hat{\gamma}(1) & \hat{\gamma}_2(0) \end{pmatrix}. \end{aligned}$$

5.2.4 Expérimentations

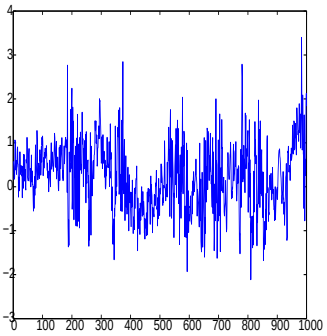
Nous proposons trois séries d'expériences. Pour les trois séries, la chaîne de Markov X est stationnaire et réversible à valeurs dans $\mathcal{X} = \{\omega_1, \omega_2\}$ avec $p(x_1 = \omega_1, x_2 = \omega_1) = p(x_1 = \omega_2, x_2 = \omega_2) = 0.495$ et $p(x_1 = \omega_1, x_2 = \omega_2) = p(x_1 = \omega_2, x_2 = \omega_1) = 0.005$. La covariance est de la forme $\gamma(k) = \sigma^2\,(k+1)^{-\alpha}$, la taille des échantillons est $N = 1000$. Les simulations sont présentées dans la figure 5.1. Dans la première expérience, on considère une chaîne de Markov cachée à bruit indépendant dont les moyennes sont égales respectivement à 1 et 2 et dont les variances sont égales à 1. La réalisation observée $y_{1:N}$ est ensuite segmentée en utilisant trois méthodes. La première utilise les vrais paramètres du modèle. La seconde méthode utilise ICE pour estimer les paramètres et nous supposons que les données observées sont issues du vrai modèle de chaînes de Markov cachées à bruit indépendant (CMC-BI). Quant à la dernière méthode, on suppose que les données sont issues du modèle de chaînes de Markov cachées à mémoire longue (CMC-ML), les paramètres étant estimés par ICE proposé ci-dessus. Pour les trois méthodes, les états cachés sont estimés par MPM.



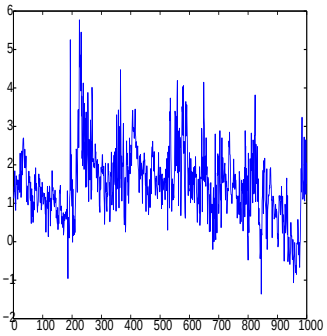
(a)



(b)



(c)

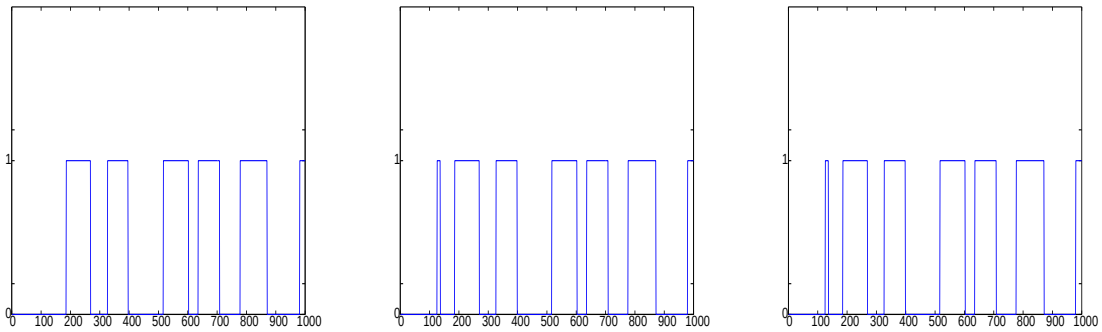


(d)

FIG. 5.1 – (a) Chaîne de Markov à deux états, (b) Version bruitée avec bruit indépendant, (c) Bruit à dépendance longue, mêmes moyennes, mêmes variances, $\alpha_{\omega_1} = 0.1$ et $\alpha_{\omega_2} = 1$, (d) Bruit à dépendance longue, $m_{\omega_1} = 1$ et $m_{\omega_2} = 2$, variance commune égale à 1, $\alpha_{\omega_1} = 0.1$ et $\alpha_{\omega_2} = 0.9$.

Paramètres	CMC-BI		CMC-ML		Vraies valeurs	
	ω_1	ω_2	ω_1	ω_2	ω_1	ω_2
m	0.92	1.99	0.89	1.96	1	2
σ^2	1	1	0.98	1.05	1	1
α	-	-	> 100	> 100	-	-
Taux d'erreur	5.2%		5.2%		4.1%	

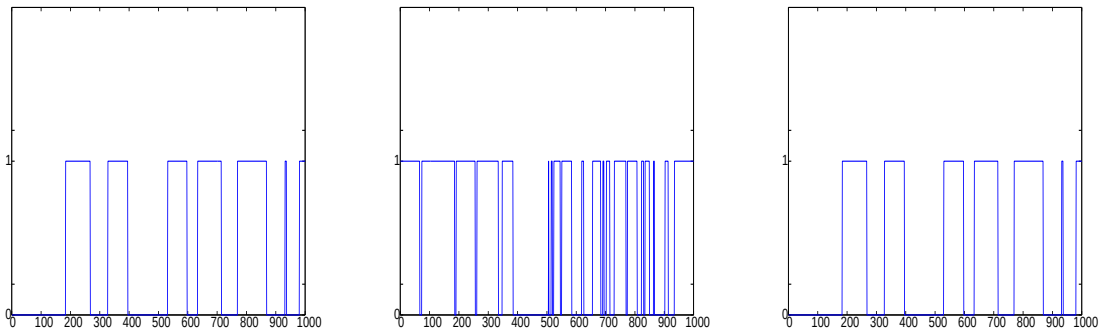
TAB. 5.1 – Estimation de la loi d'observation en utilisant les modèles CMC-BI et CMC-ML. La réalisation $y_{1:N}$ est issue du modèle CMC-BI.



(a) (b) (c)

FIG. 5.2 – Segmentation de $y_{1:N}$ issue du modèle CMC-BI : (a) Fondée sur les vrais paramètres, 4.1% d'erreur, (b) Modèle CMC-BI : 5.2% d'erreur, (c) Modèle CMC-ML : 5.2% d'erreur.

D'après les figures 5.2,(b) et 5.2,(c), les résultats obtenus avec CMC-BI et CMC-ML sont comparables, ce qui montre la bonne robustesse, dans le cadre de l'expérience, du modèle à dépendance longue. On peut remarquer également que les moyennes et variances sont bien estimées et le paramètre de dépendance α est supérieur à 100 (voir tableau 5.1), ainsi le modèle CMC-ML est capable de reconnaître des situations où la covariance décroît rapidement.



(a) (b) (c)

FIG. 5.3 – Segmentation de $y_{1:N}$ issue du modèle CMC-ML (deuxième expérience) : (a) Fondée sur les vrais paramètres, 2.1% d'erreur, (b) Modèle CMC-BI : 48% d'erreur, (c) Modèle CMC-ML : 1.9% d'erreur.

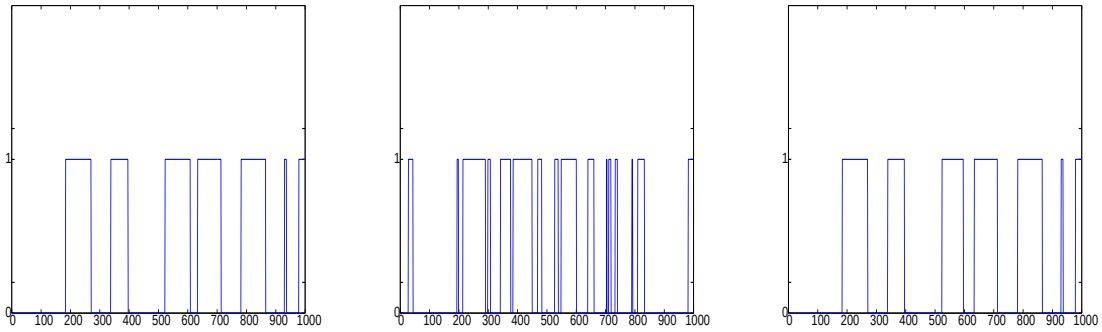
La deuxième expérience est complémentaire de la précédente. Les deux moyennes sont égales à 0, les deux variances à 1 tandis que les paramètres de dépendance valent respectivement 0.1 et 1 (figure 5.1,(c)). Les données simulées ne peuvent alors être issues du modèle CMC-BI. Il est alors intéressant de savoir si le modèle CMC-BI peut tout de même bien estimer les états cachés.

Paramètres	CMC-BI		CMC-ML		Vraies valeurs	
	ω_1	ω_2	ω_1	ω_2	ω_1	ω_2
m	0.61	-0.41	0.24	0.22	0	0
σ^2	0.35	0.26	0.37	0.81	1	1
α	-	-	0.28	1.1	0.1	1
Taux d'erreur	48%		1.9%		2.1%	

TAB. 5.2 – Estimation de la loi d'observation en utilisant les modèles CMC-BI et CMC-ML. La réalisation $y_{1:N}$ est issue du modèle CMC-ML (deuxième expérience).

On peut remarquer sur la figure 5.3 que les résultats obtenus en utilisant le modèle CMC-ML sont très bons tandis que les résultats obtenus en utilisant le modèle CMC-BI sont très médiocres. Il en résulte qu'il existe des situations dans lesquelles une CMC-ML ne peut être approchée par une CMC-BI. Concernant l'estimation des paramètres présentée dans le tableau 5.2, nous voyons que les paramètres sont mal estimés lorsque l'on utilise le modèle CMC-BI. Cependant, les paramètres ne sont pas non plus très bien estimés lorsque l'on utilise le modèle CMC-ML.

Dans le dernier exemple, les moyennes des deux classes sont respectivement égales à 1 et 2, les paramètres de corrélation sont égaux à 0.1 et 0.9, la variance est égale à 1.



(a) (b) (c)

FIG. 5.4 – Segmentation de $y_{1:N}$ lorsque (X, Y) est une chaîne de Markov cachée à dépendance longue (troisième expérience) : (a) Basée sur les vrais paramètres, 4.9% d'erreur, (b) Modèle CMC-BI : 32.4% d'erreur, (c) Modèle CMC-ML : 4.1% d'erreur.

Paramètres	CMC-BI		CMC-ML		Vraies valeurs	
	ω_1	ω_2	ω_1	ω_2	ω_1	ω_2
m	1.06	2.25	1.33	1.73	1	2
σ^2	0.42	0.67	0.53	1.17	1	1
α	-	-	0.22	0.66	0.1	0.9
Taux d'erreur	32.4%		4.1%		4.9%	

TAB. 5.3 – Estimation de la loi d'observation en utilisant les modèles CMC-BI et CMC-ML. La réalisation $y_{1:N}$ est issue du modèle CMC-ML (troisième expérience).

Selon les résultats de la figure 5.4, nous constatons que l'estimation des états cachés par le modèle CMC-ML est nettement meilleure que celle par le modèle CMC-BI. Cependant, on peut également constater que l'estimation des paramètres du modèle (Tableau 5.3) n'est pas très bonne en utilisant le modèle de chaînes de Markov cachées à dépendance longue. De plus, on peut remarquer que même avec des paramètres mal estimés, les états cachés restent bien estimés par le modèle CMC-ML. Ces différents résultats montrent que le modèle CMC-ML est très différent du modèle CMC-BI. Par ailleurs, la méthode ICE proposée semble bien adaptée à la segmentation non supervisée.

5.3 Chaînes semi-markoviennes cachées à dépendance longue

Nous proposons dans cette section un autre modèle partiellement de Markov permettant de modéliser des observations à dépendance longue. Dans ce nouveau modèle, chaque y_n ne dépend plus de tous les y_k tels que $k \leq n$ mais seulement des $y_{n_1}, \ldots, y_n$ tels que $x_{n_1} = \ldots = x_n$ et $x_{n_1-1} \neq x_{n_1}$. Ce nouveau modèle nous permettra de traiter des processus de taille plus grande. En effet, dans le modèle précédent, le calcul de la transformée de Choleski $\Gamma^{x_{n_1:n_1+n_2}} = CC^T$ nécessite le stockage des matrices $\Gamma^{x_{1:n}}$, ce qui rend l'algorithme d'estimation ICE inopérant lorsque le processus est de grande taille ($N \simeq 10^4$), en raison de débordement de mémoire. En effet, avec un ordinateur disposant d'une mémoire de 3 giga-octets, le logiciel MATLAB ne peut gérer des matrices de plus de 25×10^7 éléments réels. Cependant, le calcul de la transformée de Choleski nécessite également le stockage de C, ainsi les matrices ne peuvent avoir plus de 12.5×10^7 éléments réels et donc les processus ne peuvent être de taille plus grande que $N = 11025$. Par ailleurs, ce nouveau modèle est étendu au cas où la chaîne cachée est semi-markovienne.

5.3.1 Chaînes triplets partiellement de Markov et dépendance longue

On considère un processus triplet $Z = (X_n, U_n, Y_n)_{1 \leq n \leq N}$ tel que chaque X_n prend ses valeurs dans $\mathcal{X} = \{\omega_1, \ldots, \omega_K\}$, chaque U_n prend ses valeurs dans un ensemble fini $\Lambda = \{0, 1, \ldots, L\}$ et chaque Y_n prend ses valeurs dans $\mathbb{R}$.

Définition 5.3.1 (Chaînes triplets partiellement de Markov)**.**
Le processus $Z = (X_n, U_n, Y_n)_{1 \leq n \leq N}$ est une chaîne triplet partiellement de Markov si le processus couple (V, Y) où $V = (X, U)$ est une chaîne couple partiellement de Markov.

Afin de modéliser des observations à dépendance longue, nous proposons la distribution $p(x_{1:N}, u_{1:N}, y_{1:N})$ suivante :

$$p(x_{1:N}, u_{1:N}, y_{1:N}) = p(x_1)\delta_0(u_1)p(y_1|x_1)\prod_{n=1}^{N-1} p(x_{n+1}|x_n)p(u_{n+1}|x_n, x_{n+1}, u_n)p(y_{n+1}|x_{n+1}, u_{n+1}, y_{1:n}), \tag{5.25}$$

où

$$p(u_{n+1}|x_n, x_{n+1}, u_n) = \begin{cases} \delta_{u_n+1}(u_{n+1}) \text{ si } x_n = x_{n+1} \text{ et } u_n < L\,; \\ \delta_0(u_{n+1}) \text{ si } x_n \neq x_{n+1} \text{ ou } u_n = L\,; \end{cases} \tag{5.26}$$

$$p(y_{n+1}|x_{n+1}, u_{n+1}, y_{1:n}) = \begin{cases} p(y_{n+1}|x_{n+1}) \text{ si } u_{n+1} = 0\,; \\ p(y_{n+1}|x_{n+1}, y_{n-u_{n+1}+1:n}) \text{ si } u_{n+1} > 0\,; \end{cases} \tag{5.27}$$

et $p(y_{n+1}|x_{n+1}, y_{n-u_{n+1}+1:n})$ est tel que $p(y_{n-u_{n+1}+1:n+1}|x_{n+1})$ suive une loi normale de moyenne

$$m_{x_{n+1}}^{u_{n+1}+1} = (\underbrace{m_{x_{n+1}}, \ldots, m_{x_{n+1}}}_{u_{n+1}+1 \text{ fois}}),$$

et de matrice de covariance

$$\Gamma_{x_{n+1}}^{u_{n+1}+1} = \begin{pmatrix} \gamma_{x_{n+1}}(0) & \gamma_{x_{n+1}}(1) & \ldots & \gamma_{x_{n+1}}(u_{n+1}-1) & \gamma_{x_{n+1}}(u_{n+1}) \\ \gamma_{x_{n+1}}(1) & \gamma_{x_{n+1}}(0) & \ldots & \gamma_{x_{n+1}}(u_{n+1}-2) & \gamma_{x_{n+1}}(u_{n+1}-1) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma_{x_{n+1}}(u_{n+1}) & \gamma_{x_{n+1}}(u_{n+1}-1) & \ldots & \gamma_{x_{n+1}}(1) & \gamma_{x_{n+1}}(0) \end{pmatrix}.$$

On remarque que X est une chaîne de Markov, de plus u_n représente le temps de séjour minimal écoulé dans l'état x_n depuis le changement d'état. Si $L = N$, u_n est le temps de séjour exact écoulé depuis le changement d'état. Ce modèle ressemble à celui des chaînes semi-markoviennes, il en est cependant très différent. Dans le modèle des chaînes semi-markoviennes, le processus U permet de définir la loi de X, tandis que dans notre modèle triplet partiellement de Markov, la loi de X est déjà définie comme celle d'une chaîne de Markov et le processus U est défini à partir de X.
Comme pour le modèle couple partiellement de Markov décrit dans la section précédente, l'algorithme de Baum-Welsh est utilisable si les quantités $p(y_{n+1}|x_{n+1}, u_{n+1}, y_{1:n})$ sont calculables. En effet, lorsque $u_{n+1} = 0$, $p(y_{n+1}|x_{n+1}, u_{n+1}, y_{1:n}) = p(y_{n+1}|x_{n+1})$ est une loi normale dans $\mathbb{R}$ de moyenne $m_{x_{n+1}}$ et de variance $\gamma_{x_{n+1}}(0)$; et lorsque $u_{n+1} > 0$, $p(y_{n+1}|x_{n+1}, u_{n+1}, y_{1:n}) = p(y_{n+1}|x_{n+1}, y_{n-u_{n+1}+1:n})$ est une loi normale de moyenne

$$M_{x_{n+1}} = m_{x_{n+1}} + (\gamma_{x_{n+1}}(u_{n+1}), \ldots, \gamma_{x_{n+1}}(1)) \left(\Gamma_{x_{n+1}}^{u_{n+1}}\right)^{-1} \left(y_{n-u_{n+1}+1:n} - m_{x_{n+1}}^{u_{n+1}}\right),$$

et de variance

$$\tilde{\gamma}_{x_{n+1}}(0) = \gamma_{x_{n+1}}(0) - (\gamma_{x_{n+1}}(u_{n+1}), \ldots, \gamma_{x_{n+1}}(1)) \left(\Gamma_{x_{n+1}}^{u_{n+1}}\right)^{-1} \begin{pmatrix} \gamma_{x_{n+1}}(u_{n+1}) \\ \vdots \\ \gamma_{x_{n+1}}(1) \end{pmatrix}.$$

Pour calculer les quantités $p(y_{n+1}|x_{n+1}, y_{n-u_{n+1}+1:n})$, il n'est pas nécessaire de stocker les matrices $\Gamma_{x_{n+1}}^{u_{n+1}}$, celles-ci pouvant être de grande taille dans le cas des applications en segmentation d'image. En effet, les matrices $\Gamma_{x_{n+1}}^{u_{n+1}}$ sont de Toeplitz (la valeur de l'élément (i,j) de la matrice ne dépend que de la différence $|i-j|$) et de plus le vecteur $(\gamma_{x_{n+1}}(u_{n+1}), \ldots, \gamma_{x_{n+1}}(1))$ est extrait de la matrice $\Gamma_{x_{n+1}}^{u_{n+1}+1}$. Ainsi nous pouvons utiliser l'algorithme de Durbin-Levinson décrit dans la sous-section 6.3.1 du chapitre 6 pour calculer $(\gamma_{x_{n+1}}(u_{n+1}), \ldots, \gamma_{x_{n+1}}(1)) \left(\Gamma_{x_{n+1}}^{u_{n+1}}\right)^{-1}$.

5.3.2 Le modèle semi-markovien

Le modèle précédent est enrichi en introduisant un autre processus auxiliaire modélisant la semi-markovianité. Considérons le processus $Z = (X_n, U_n^1, U_n^2, Y_n)_{1\leq n\leq N}$, où chaque X_n prend ses valeurs dans $\mathcal{X} = \{\omega_1, \ldots, \omega_K\}$, chaque U_n^1 prend ses valeurs dans $\Lambda^1 = \{1, \ldots, L_1\}$, chaque U_n^2 prend ses valeurs dans $\Lambda^2 = \{0, \ldots, L_2\}$ et chaque Y_n prend ses valeurs dans $\mathbb{R}$. Dans ce modèle, on considère que X est une chaîne semi-markovienne telle que la loi de (X, U^1) soit définie par les formules (3.15), (3.16) et (3.17). La loi de (X, U^1, U^2, Y) est alors définie par :

$$p(x_{1:N}, u_{1:N}^1, u_{1:N}^2, y_{1:N}) = p(x_1)p(u_1^1|x_1)\delta_0(u_1^2)p(y_1|x_1)$$
$$\times \prod_{n=1}^{N-1} p(x_{n+1}|x_n, u_n^1)p(u_{n+1}^1|x_{n+1}, u_n^1)p(u_{n+1}^2|x_n, x_{n+1}, u_n^2)p(y_{n+1}|x_{n+1}, u_{n+1}^2, y_{1:n}), \quad (5.28)$$

avec

$$p(u_1^1|x_1) = \bar{d}(x_1, u_1^1); \quad (5.29)$$

$$p(x_{n+1}|x_n, u_n^1) = \begin{cases} \delta_{x_n}(x_{n+1}) \text{ si } u_n^1 > 1; \\ r(x_{n+1}|x_n) \text{ si } u_n^1 = 1; \end{cases} \quad (5.30)$$

$$p(u_{n+1}^1|x_{n+1}, u_n^1) = \begin{cases} \delta_{u_n^1-1}(u_{n+1}^1) \text{ si } u_n^1 > 1; \\ \bar{d}(x_{n+1}, u_{n+1}^1) \text{ si } u_n^1 = 1; \end{cases} \quad (5.31)$$

$$p(u_{n+1}^2|x_n, x_{n+1}, u_n^2) = \begin{cases} \delta_{u_n^2+1}(u_{n+1}^2) \text{ si } x_n = x_{n+1} \text{ et } u_n < L_2\,; \\ \delta_0(u_{n+1}^2) \text{ si } x_n \neq x_{n+1} \text{ ou } u_n = L_2\,; \end{cases} \quad (5.32)$$

$$p(y_{n+1}|x_{n+1}, u_{n+1}^2, y_{1:n}) = \begin{cases} p(y_{n+1}|x_{n+1}) \text{ si } u_{n+1}^2 = 0\,; \\ p(y_{n+1}|x_{n+1}, y_{n-u_{n+1}^2+1:n}) \text{ si } u_{n+1}^2 > 0\,; \end{cases} \quad (5.33)$$

où pour tout ω_j, $\bar{d}(\omega_j, .)$ est une densité de probabilité sur Λ^1 et $r(.|.)$ est un noyau de transition sur $\mathcal{X}^2$ conformément aux notations de la sous-section 3.3.3 du chapitre 3.
Dans la suite, nous appelerons ce modèle "chaînes semi-markoviennes cachées à mémoire longue" (CSMC-ML). Lorsque le processus X est une chaîne de Markov, on l'appelera "chaînes triplets partiellement de Markov à mémoire longue" (CTPM-ML) pour le différencier du modèle étudié dans la section 5.2.

5.3.3 Estimation des paramètres

Nous ne traiterons ici que de l'estimation de la loi d'observation, l'estimation de la loi $p(x_{1:N}, u_{1:N}^1, u_{1:N}^2)$ ayant été traitée au chapitre 3. Si θ_q désigne le vecteur paramètre obtenu à l'étape q de ICE, on procède de la manière suivante :

1. simuler un échantillon $(x_{1:N}, u_{1:N}^1, u_{1:N}^2)$ selon la loi a posteriori $p(x_{1:N}, u_{1:N}^1, u_{1:N}^2|y_{1:N}; \theta_q)$;

2. pour chaque ω_j, considérer l'échantillon $(y_m^{\omega_j})_{1\leq m\leq N_j}$, où pour tout m, $x_m = \omega_j$ et N_j est le nombre de x_n égaux à ω_j ;
3. estimer les paramètres correspondant à l'état ω_j à partir de $(y_m^{\omega_j})_{1\leq m\leq N_j}$ en utilisant les estimateurs présentés à la sous-section 5.1.4.

5.3.4 Expérimentations

Le modèle de chaînes semi-markoviennes cachées à dépendance longue généralise à la fois le modèle de chaînes semi-markoviennes cachées (CSMC) et le nouveau modèle de chaînes triplets partiellement de Markov à dépendance longue (CTPM-ML). Le but de cette sous-section est de regarder les améliorations apportées par ces deux généralisations. Les expérimentations seront faites sur des processus de taille $N = 128 \times 128$. Les réalisations de ces processus seront représentées comme des images bi-dimensionnelles grâce au parcours d'Hilbert-Peano.
Nous présentons trois séries d'expériences. Dans la première série, les données sont issues du modèle CSMC et la question est de savoir si le modèle CSMC-ML, plus complexe, donne des résultats similaires à ceux obtenus par le modèle CSMC. Dans la seconde série, les données sont issues du modèle CSMC-ML et nous estimons les paramètres et les états cachés en utilisant les modèles CSMC, CTPM-ML et CSMC-ML. Nous conclurons cette sous-section par la segmentation d'une image réelle, les données n'étant alors probablement issues d'aucun des 3 modèles. Dans toutes les expériences proposées dans cette sous-section, la covariance est de la forme $\gamma(k) = \sigma^2 (k+1)^{-\alpha}$.
Dans la première expérience, on considère une chaîne semi-markovienne cachée (CSMC) $(X, U^1, Y) = (X_n, U_n^1, Y_n)_{1\leq n\leq N}$ telle que chaque X_n prend ses valeurs dans $\mathcal{X} = \{\omega_1, \omega_2\}$ et chaque U_n^1 prend ses valeurs dans $\Lambda^1 = \{1, \ldots, 10\}$. La loi $p(x_{1:N}, u_{1:N}^1)$ est définie par (3.15), (3.16) et (3.17) et la loi $p(y_{1:N}|x_{1:N}, u_{1:N}^1)$ est donnée par :

$$p(y_{1:N}|x_{1:N}, u_{1:N}^1) = \prod_{n=1}^{N} p(y_n|x_n).$$

Lorsque $x_n = \omega_1$, $p(y_n|x_n)$ est une loi normale de moyenne 1 et de variance 20 et lorsque $x_n = \omega_2$, c'est une loi normale de moyenne 2 et de variance 20. La distribution $\bar{d}(\omega_j, .)$ est celle d'une loi uniforme sur Λ^1 pour tout $j \in \{1, 2\}$. De plus, $r(x_{n+1}|x_n) = 0.999$ si $x_n = x_{n+1}$ et $r(x_{n+1}|x_n) = 0.001$ si $x_n \neq x_{n+1}$. La réalisation $y_{1:N}$ est ensuite segmentée en utilisant trois méthodes. La première suppose que les données sont issues du modèle CSMC et utilise les vrais paramètres. La seconde méthode suppose que les données sont issues du modèle CSMC et on estime les paramètres par ICE. Quant à la troisième méthode, on suppose que les données sont issues du modèle plus général CSMC-ML et on estime également les paramètres par ICE. Dans le modèle CSMC-ML, on supposera que $L_2 = 50$. Les résultats sont présentés dans la figure 5.5 et dans le tableau 5.4.

(a) (b) (c) (d) (e)

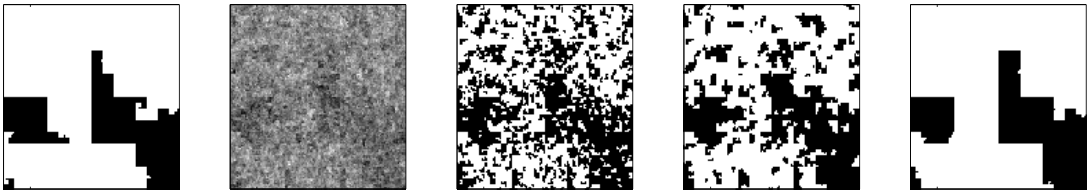
FIG. 5.5 – Segmentation de $y_{1:N}$ lorsque (X, Y) est une chaîne semi-markovienne cachée à bruit indépendant : (a) Réalisation de X, (b) Réalisation de Y, (c) Modèle CSMC avec les vrais paramètres, 3.74% d'erreur, (d) Modèle CSMC en non supervisé : 4.55% d'erreur, (e) Modèle CSMC-ML en non supervisé : 4.57% d'erreur.

Paramètres	CSMC		CSMC-ML		Vraies valeurs	
	ω_1	ω_2	ω_1	ω_2	ω_1	ω_2
m	1.06	2.04	0.98	1.97	1	2
σ^2	19.81	20.71	19.84	20.46	20	20
α	-	-	15.28	5.95	-	-
Taux d'erreur	4.55%		4.57%		3.74%	

TAB. 5.4 – Estimation des paramètres à partir de données issues d'une chaîne semi-markovienne à bruit indépendant.

Nous constatons que les résultats sont similaires en utilisant les modèles CSMC et CSMC-ML, ce qui montre la capacité du modèle CSMC-ML à traiter des données issues du modèle CSMC. De plus l'estimation du paramètre α prouve que le modèle CSMC-ML est capable de reconnaître les situations où la covariance décroît rapidement.

Dans la seconde expérience, nous segmentons des données issues du modèle CSMC-ML. Le but de cette expérience est de savoir lequel des deux modèles CSMC ou CTPM-ML donne de meilleurs résultats. La chaîne semi-markovienne (X, U^1) considérée suit la même loi que dans l'expérience précédente. Concernant le processus U^2, on prendra $L_2 = 50$. Les lois $p(y_n|x_n)$ seront respectivement la loi normale de moyenne 1 et de variance 1 lorsque $x_n = \omega_1$ et la loi normale de moyenne 2 et de variance 1 lorsque $x_n = \omega_2$. Le paramètre de dépendance est égal pour les deux classes à 0.5.



(a) (b) (c) (d) (e)

FIG. 5.6 – Segmentation de $y_{1:N}$ lorsque (X, Y) est une chaîne semi-markovienne cachée à mémoire longue : (a) Réalisation de X, (b) Réalisation de Y, (c) Modèle CSMC en non supervisé : 31.15% d'erreur, (d) Modèle CTPM-ML en non supervisé : 21.53% d'erreur, (e) Modèle CSMC-ML en non supervisé : 3.16% d'erreur.

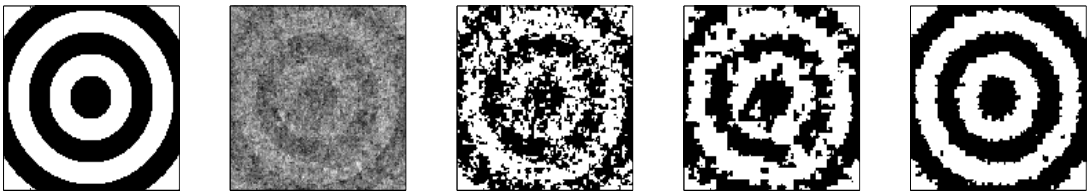
Paramètres	CSMC		CTPM-ML		CSMC-ML		Vraies valeurs	
	ω_1	ω_2	ω_1	ω_2	ω_1	ω_2	ω_1	ω_2
m	0.97	2.44	1.08	2.22	0.98	1.97	1	2
σ^2	0.59	0.56	0.83	0.78	0.96	0.93	1	1
α	-	-	0.69	0.72	0.62	0.61	0.5	0.5
Taux d'erreur	31.15%		21.53%		3.16%		2.85%	

TAB. 5.5 – Estimation des paramètres à partir de données issues de chaînes semi-markoviennes cachées à dépendance longue.

Nous constatons d'après la figure 5.6 et le tableau 5.5, que négliger la dépendance longue donne de plus mauvais résultats que négliger la semi-markovianité. Cependant, d'après la figure 5.6,(e), la considération de la semi-markovianité est importante et améliore nettement les résultats comparés à ceux obtenus à la figure 5.6,(d). Ces remarques sont encore justifiées dans l'estimation des paramètres présentée dans le tableau 5.5. Cependant les qualités d'estimation des paramètres par CTPM-ML et CSMC-ML sont quasi similaires. Globalement, l'expérience montre que le modèle CSMC-ML est plus riche, de manière significative, que chacun des deux modèles CTMP-ML et CSMC.

Dans la dernière expérience, on choisit de segmenter à l'aide des trois modèles CSMC, CTPM-ML et CSMC-ML une image dessinée. De l'image réelle représentant des cercles concentriques, on obtient la réalisation $x_{1:N}$ par le parcours d'Hilbert-Peano. La réalisation $x_{1:N}$ est ensuite bruitée à l'aide du modèle défini par les formules (5.26) et (5.27). Dans cette expérience, on

prendra $L_2 = 50$, les moyennes des deux classes seront respectivement égales à 1 et 2, la variance sera égale à 1, et le paramètre de dépendance sera égal à 0.9.



(a) (b) (c) (d) (e)

FIG. 5.7 – Segmentation d'une image réelle bruitée avec de la dépendance longue : (a) Réalisation de X, (b) Réalisation de Y, (c) Modèle CSMC en non supervisé : 24.70% d'erreur, (d) Modèle CTPM-ML en non supervisé : 18.27% d'erreur, (e) Modèle CSMC-ML en non supervisé : 6.31% d'erreur.

Paramètres	CSMC		CTPM-ML		CSMC-ML		Vraies valeurs	
	ω_1	ω_2	ω_1	ω_2	ω_1	ω_2	ω_1	ω_2
m	0.75	2.26	0.91	2.04	0.99	1.99	1	2
σ^2	0.69	0.66	0.92	0.93	1.01	1.02	1	1
α	-	-	1.07	1.07	0.93	0.92	0.9	0.9
Taux d'erreur	24.70%		18.27%		6.31%		5.95%	

TAB. 5.6 – Estimation des paramètres à partir de données issues d'une image réelle

Des résultats présentés dans la figure 5.7 et dans le tableau 5.6, nous voyons que le modèle semi-markovien est suffisamment général pour prendre en compte des propriétés statistiques de l'image que le modèle markovien ne prend pas en compte. Ainsi, on peut constater une meilleure segmentation en utilisant le modèle CSMC-ML (Figure 5.7,(e)) qu'en utisant le modèle CTPM-ML (Figure 5.7,(d)). Cependant, si on néglige la dépendance longue, les paramètres de la loi d'observation sont mal estimés (Tableau 5.6), ce qui entraîne une mauvaise estimation des états cachés (Figure 5.7,(c)).

5.4 Observations non gaussiennes à dépendance longue

Le dernier modèle que nous présentons dans ce chapitre permet de modéliser des observations non gaussiennes. Nous avons vu au chapitre 4 qu'il est possible d'écrire la loi jointe d'un vecteur aléatoire à partir de ses lois marginales et d'une fonction d'agrégation appelée "copule". Dans ce chapitre, les copules vont nous permettre ainsi de modéliser des observations non gaussiennes à dépendance longue.

5.4.1 Dépendance longue et copules

On considère un processus triplet $(X,U,Y) = (X_n,U_n,Y_n)_{1\leq n\leq N}$ tel que chaque X_n prend ses valeurs dans un ensemble fini $\mathcal{X} = \{\omega_1,\ldots,\omega_K\}$, chaque U_n prend ses valeurs dans l'ensemble fini $\Lambda = \{0,\ldots,L\}$ et chaque Y_n prend ses valeurs dans $\mathbb{R}$. La distribution $p(x_{1:N},u_{1:N},y_{1:N})$ est celle d'une chaîne triplet partiellement de Markov définie par (5.25), (5.26) et (5.27). A la différence avec l'ancien modèle où la distribution $p(y_{n-u_{n+1}+1:n+1}|x_{n+1})$ était celle d'une loi normale, dans le modèle proposé, elle s'écrit :

$$\begin{aligned} p(y_{n-u_{n+1}+1:n+1}|x_{n+1}) &= \prod_{k=n-u_{n+1}+1}^{n+1} p(y_k|x_{n+1}) \\ &\times\ c_{x_{n+1}}^{u_{n+1}+1}\left(F_{x_{n+1}}(y_{n-u_{n+1}+1}),\ldots,F_{x_{n+1}}(y_{n+1})\right), \end{aligned} \tag{5.34}$$

où pour chaque $\omega_j \in \mathcal{X}$, F_{ω_j} est une fonction de répartition marginale, $c_{\omega_j}^{u_{n+1}+1}$ est la copule gaussienne de matrice de corrélation :

$$R_{\omega_j}^{u_{n+1}+1} = \begin{pmatrix} 1 & \gamma_{\omega_j}(1) & \ldots & \gamma_{\omega_j}(u_{n+1}-1) & \gamma_{\omega_j}(u_{n+1}) \\ \gamma_{\omega_j}(1) & 1 & \gamma_{\omega_j}(1) & \ldots & \gamma_{\omega_j}(u_{n+1}-1) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma_{\omega_j}(u_{n+1}) & \gamma_{\omega_j}(u_{n+1}-1) & \ldots & \gamma_{\omega_j}(1) & 1 \end{pmatrix},$$

et chaque $p(y_k|x_{n+1}=\omega_j)$ est une distribution de fonction de répartition F_{ω_j}. Ainsi, si $x_n = x_{n+1} = \ldots = x_{n+k}$, alors $(y_n,\ldots,y_{n+k})$ est la réalisation d'un vecteur aléatoire dont les lois marginales sont de fonction de répartition F_{x_n} et dont la copule est une copule gaussienne.
Posons $z_n = \phi^{-1}\circ F_{x_n}(y_n)$, où ϕ est la fonction de répartition de la loi normale centrée et réduite. La distribution $p(x_{1:N},u_{1:N},z_{1:N})$ est également définie par (5.25), (5.26) et (5.27). Dans ce cas il s'agit du modèle triplet partiellement de Markov tel que $p(z_{n-u_{n+1}+1:n+1}|x_{n+1})$ soit la distribution d'une loi normale à marginales centrées et réduites et de matrice de corrélation égale à $R_{x_{n+1}}^{u_{n+1}+1}$. Si pour tout $1\leq j\leq K$, les familles de covariance $(\gamma_{\omega_j}(k))_{k\in\mathbb{N}}$ satisfont les propriétés de dépendance longue, l'échantillon $z_{1:N}$ est alors à dépendance longue conditionnellement aux classes ; en d'autres termes, si $x_n = x_{n+1} = \ldots = x_{n+k}$, l'échantillon $(z_n,\ldots,z_{n+k})$ est la réalisation d'un processus à dépendance longue. Selon la définition 5.4.1, nous dirons que $(y_n,\ldots,y_{n+k})$ est également un processus à dépendance longue. Dans cette définition, nous introduirons également la notion de stationnarité du second ordre en copule.

Définition 5.4.1. *Soit $Y = (Y_n)_{n\in\mathbb{Z}}$ un processus réel tel que pour tout $n\in\mathbb{Z}$, Y_n ait pour fonction de répartition F.*

- *On dit que le processus Y est stationnaire du second ordre en copule si le processus $U = (F(Y_n))_{n\in\mathbb{Z}}$ est stationnaire du second ordre ;*
- *Lorsque Y est stationnaire du second ordre en copule, on définit les corrélations de Spearman $(\rho_S(k))_{k\in\mathbb{N}}$ par :*

$$\rho_S(k) = \frac{\mathbb{E}\left[(F(Y_n)-\mathbb{E}(F(Y_n)))(F(Y_{n-k})-\mathbb{E}(F(Y_{n-k})))\right]}{\sqrt{\mathbb{E}\left((F(Y_n)-\mathbb{E}(F(Y_n)))^2\right)\mathbb{E}\left((F(Y_{n-k})-\mathbb{E}(F(Y_{n-k})))^2\right)}}.$$

- *On dit que le processus Y est à dépendance longue si :*

$$\rho_S(k) \sim_{+\infty} Ck^{-\alpha} \text{ où } \alpha\in]0,1] \text{ et } C\in\mathbb{R}.$$

Remarque : D'après ce qu'on a précisé au chapitre 4, les coefficients de mélangeance ne dépendent que de la copule. Ainsi si f est un $\mathcal{C}^1$-difféomorphisme croissant de $\mathbb{R}$ dans $\mathbb{R}$, les processus $(Y_n)_{n\in\mathbb{Z}}$ et $(f(Y_n))_{n\in\mathbb{Z}}$ ont la même mélangeance. La dépendance longue peut se définir de manière plus générale : un processus est à dépendance longue s'il est ρ-mélangeant et son coefficient de ρ-mélangeance converge vers 0 en $n^{-\alpha}$ où $\alpha \in]0,1]$.

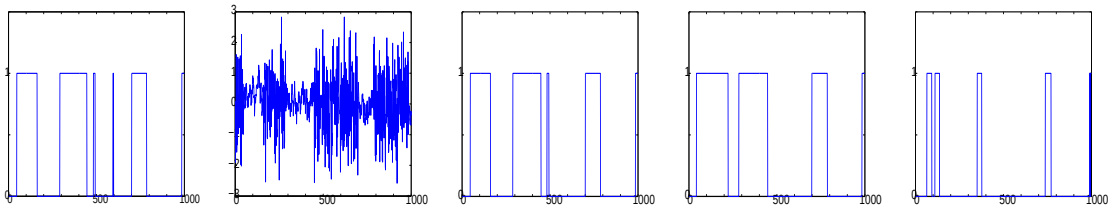
5.4.2 Estimation des paramètres

Soit ϕ la fonction de répartition d'une loi normale centrée-réduite. Nous détaillons ici l'estimation de la loi d'observation $p(y_{1:N}|x_{1:N})$ par l'algorithme ICE. Si θ_q désigne le vecteur paramètre obtenu à l'étape q, on procède de la manière suivante :

1. simuler un échantillon $(x_{1:N}, u_{1:N})$ selon la loi a posteriori $p(x_{1:N}, u_{1:N}|y_{1:N}; \theta_q)$;
2. pour chaque ω_j, considérer l'échantillon $(y_m^{\omega_j})_{1\leq m\leq N_j}$, où pour tout m, $x_m = \omega_j$ et N_j est le nombre de x_n égaux à ω_j ;
3. estimer la loi marginale $p(y_n|x_n = \omega_j)$ à partir de l'échantillon $(y_m^{\omega_j})_{1\leq m\leq N_j}$. Soit $F_{\omega_j}^{q+1}$ la fonction de répartition correspondant aux paramètres re-estimés ;
4. estimer les paramètres de corrélation à partir de l'échantillon $(\phi^{-1} \circ F_{\omega_j}^{q+1}(y_m^{\omega_j}))_{1\leq m\leq N_j}$ avec les estimateurs présentés à la sous-section 5.1.4.

5.4.3 Expérimentations

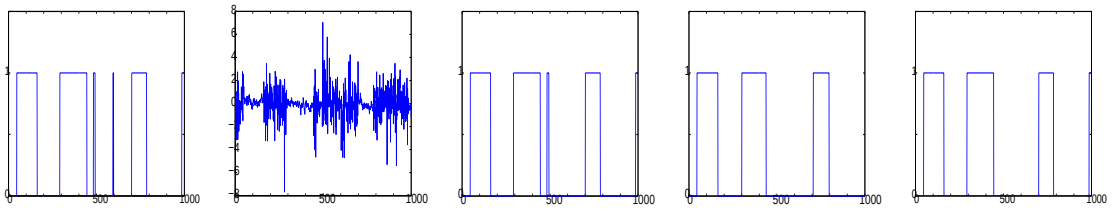
Nous proposons dans cette sous-section deux expériences. Les expériences présentées permettent de savoir si le modèle à dépendance longue gaussien peut être utilisé dans le cas où les données ne sont pas issues d'un modèle gaussien. Elles permettent également de tester la robustesse lorsque l'on utilise différent modèle de bruit. Dans les deux expériences, la taille des échantillons est $N = 1000$, la chaîne de Markov X à valeurs dans $\mathcal{X} = \{\omega_1, \omega_2\}$ est stationnaire réversible de loi donnée par $p(x_n = \omega_1, x_{n+1} = \omega_1) = p(x_n = \omega_2, x_{n+1} = \omega_2) = 0.495$ et $p(x_n = \omega_1, x_{n+1} = \omega_2) = p(x_n = \omega_2, x_{n+1} = \omega_1) = 0.005$ et la famille de covariance $(\gamma_{\omega_j}(k))_{k\in\mathbb{N}}$ est celle d'un bruit gaussien fractionnaire. Les paramètres de Hurst pour les deux expériences sont respectivement égaux à $H_{\omega_1} = 0.5$ et $H_{\omega_2} = 0.99$. On définit pour la suite la loi Γ^- dite "Gamma signée" comme la loi de la variable aléatoire ϵW, où ϵ suit une loi uniforme sur $\{-1, 1\}$ et W suit une loi Γ de paramètres a et b.
Dans la première expérience, les lois marginales sont gaussiennes de moyenne nulle et de variance égale à 1. Les données sont ensuite segmentées par trois méthodes. La première utilise les vrais paramètres du modèle et utilise MPM pour estimer les états cachés. La deuxième suppose que les données sont issues du modèle triplet partiellement de Markov à observations gaussiennes à dépendance longue. La troisième méthode suppose que les données sont à dépendance longue mais les lois marginales sont des lois Γ^-. Le but de cette expérience est de savoir si choisir un autre modèle que le modèle gaussien peut dégrader les résultats si les données sont réellement gaussienne.



(a) (b) (c) (d) (e)

FIG. 5.8 – Segmentation de données issues du modèle triplet partiellement de Markov à bruit gaussien à dépendance longue : (a) Réalisation de X, (b) Réalisation de Y, (c) Avec les vrais paramètres, 1.3% d'erreur, (d) Modèle gaussien à dépendance longue, 9.3% d'erreur, (e) Modèle Γ^- à dépendance longue, 26.1% d'erreur.

Dans la seconde expérience, nous considérons le problème inverse. Les données sont simulées selon le modèle à dépendance longue dont les lois marginales sont des lois Γ^- de paramètres $a_{\omega_1} = a_{\omega_2} = 1$ et $b_{\omega_1} = b_{\omega_2} = 1$. Le but de cette expérience est de savoir si le modèle gaussien peut être suffisamment robuste pour être utilisé même si les données ne sont probablement pas gaussiennes. Notons que les copules sont gaussiennes dans les duex modèles, qui ne se différencient que par les lois marginales.



(a) (b) (c) (d) (e)

FIG. 5.9 – Segmentation de données issues du modèle triplet partiellement de Markov à bruit Γ^- à dépendance longue : (a) Réalisation de X, (b) Réalisation de Y, (c) Avec les vrais paramètres, 0.8% d'erreur, (d) Modèle gaussien à dépendance longue, 6.5% d'erreur, (e) Modèle Γ^- à dépendance longue, 1.9% d'erreur.

Ces expériences montrent qu'il est important de bien choisir la loi d'observation. En effet, si nous segmentons des données en utilisant une loi différente de celle du vrai modèle, les résultats peuvent se dégrader de manière non négligeable.

Conclusion

Nous avons proposé dans ce chapitre plusieurs modèles dans lesquels la chaîne inobservable est cachée par du bruit à dépendance longue. Dans tous les modèles proposés, les probabilités marginales a posteriori sont calculables, ce qui permet l'estimation de la chaîne inobservable par la méthode bayésienne MPM. Dans le premier modèle "chaînes de Markov cachée avec du bruit à mémoire longue gaussien" (CMC-ML), la loi de chacune des marginales Y_n (conditionnellement aux classes) dépend de toutes les marginales Y_k passées ; cependant, les calculs se font grâce au caractère gaussien du bruit. Afin d'estimer les paramètres de ce modèle, nous avons proposé un algorithme ICE original et nous avons montré son bon comportement au travers des simulations. Nous avons également montré la bonne robustesse du modèle CMC-ML lorsque les données sont issues du modèle classique des chaînes de Markov cachées avec du bruit indépendant (CMC-BI). Le second modèle que nous avons proposé, qui est une "chaîne triplet partiellement de Markov cachée par du bruit à mémoire longue" (CTPM-ML), permet de palier aux difficultés algorithmiques du premier modèle et peut ainsi être utilisé en traitement d'image où les données traitées sont, en général, très volumineuses. Ce modèle a été ensuite étendu, en introduisant un processus auxiliaire, de façon à considérer la semi-markovianité éventuelle du processus caché. Nous avons montré aux travers des expérimentations que ni le modèle de chaînes semi-markoviennes cachées à bruit indépendant, ni le modèle CTPM-ML dans le lequel le processus caché est une chaîne de Markov, ne parvient à donner d'aussi bons résultats que le modèle général, où la chaîne cachée est semi-markovienne (les données sont issues de ce dernier). Enfin, tous les modèles précédents peuvent être généralisés avec l'introduction des lois marginales non nécessairement gaussiennes. En effet, en gardant les copules gaussiennes il est possible d'étendre les calculs faisables dans le cas gaussien aux cas où les marginales du bruit sont quelconques. Finalement, en utilisant tous les résultats exposés jusqu'à présent, il est possible de proposer un modèle général dans lequel la chaîne cachée est semi-markovienne éventuellement non stationnaire, et le bruit est à mémoire longue et à marginales quelconques. Un tel modèle peut être utilisé à des fins de segmentations non supervisées, les paramètres pouvant être estimés par une méthode ICE adéquate. En guise de perspectives, il serait intéressant d'étudier le cas de lois marginales à queues lourdes telles que les lois stables de Lévy. Ces lois sont couramment utilisées pour modéliser les phénomènes atypiques dans les données financières. Il devrait ainsi être envisageable de segmenter des données financières en considérant simultanément deux propriétés de celles-ci : la dépendance longue et l'apparition de phénomènes atypiques tels que les " cracks' ".

Chapitre 6

Application au traitement du signal radar

Dans les chapitres précédents, nous nous sommes consacré à l'aspect modélisation et nous avons étendu le modèle classique de chaînes de Markov cachées à bruit indépendant à différents modèles prenant en compte diverses propriétés du signal. La validation des résultats a été faite au travers d'expérimentations et les modèles étaient utilisés pour obtenir des segmentations d'image et de signaux. Dans ce chapitre, nous allons voir comment la segmentation peut être utilisée, comme résumé du signal, par une autre application : celle de la détection de cibles dans les signaux radar. Le détecteur original que nous allons présenté utilise les propriétés statistiques du signal reçu ; plus exactement, il compare le signal reçu d'une distance et d'un angle de visée donnés aux signaux provenant du voisinage. Ce détecteur devra connaître les propriétés statistiques du voisinage, l'intérêt de la segmentation sera alors d'obtenir des voisinages homogènes. Nous commençons à rappeler dans ce chapitre les principes du traitement du signal radar.

6.1 Prérequis en traitement du signal radar

Un radar est constitué de deux antennes, une antenne émettrice et une antenne réceptrice. Des signaux reçus par cette dernière, on peut déterminer trois attribus : l'azimut, la distance et la vitesse du milieu réflechissant. Dans le cas d'antennes tournantes, l'azimut correspond à la direction de pointage du radar. Quant à la distance et à la vitesse, des traitements supplémentaires que nous détaillerons, sont nécessaires pour les déterminer. D'autres détails techniques du radar, tels que la formation de faisceaux pour les radars à antennes fixes, sont présentés dans [28, 72, 90].

6.1.1 Radar à impulsions

Un radar à impulsions émet des impulsions ayant, chacune, une certaine durée τ. Le train d'impulsions est émis pendant une durée appelée "durée de cohérence", qui sera notée T_{coh}. Le temps écoulé entre deux impulsions, appelé "période" ou durée de récurrence, sera noté T_{rec}.

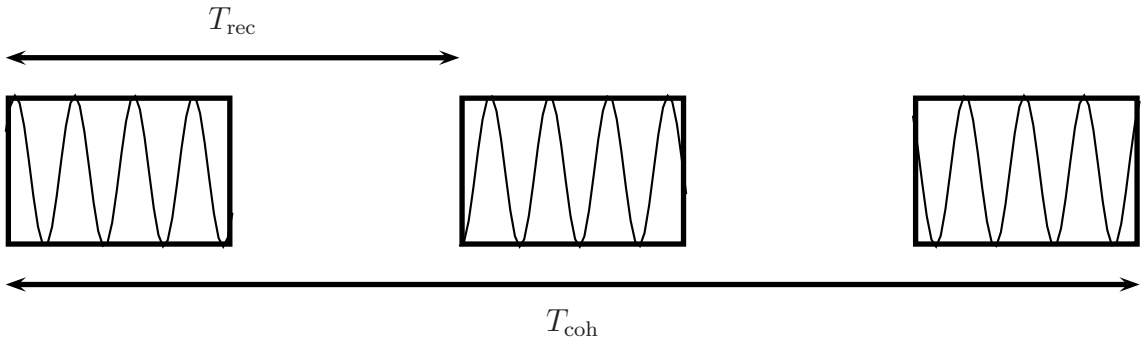


FIG. 6.1 – Schéma d'émission d'un radar à impulsions

Dans le cas du radar à impulsion classique, l'impulsion est portée par un signal sinusoïdal de fréquence f_0 appelée fréquence porteuse. Ainsi, le signal émis complexe S_e en fonction du temps est donné par :

$$S_e(t) = s_e(t)\exp\left(2i\pi f_0 t\right), \text{ où } s_e(t) = \begin{cases} 1 & \text{si} \quad t \in \bigcup_{k=0}^{\frac{T_{\text{coh}}}{T_{\text{rec}}}} \left[kT_{\text{rec}}, kT_{\text{rec}} + \tau\right], \\ 0 & \text{sinon.} \end{cases}$$

Chaque signal émis réfléchit sur un certain milieu situé à une distance D du radar et de vitesse radiale v par rapport au radar qui sera de signe positif lorsque le milieu se rapproche du radar. On notera que ce milieu peut posséder plusieurs composantes cinétiques ; c'est le cas notamment si le milieu est un fluide. De part l'éloignement au radar, il s'ensuit un décalage en temps entre le signal émis et le signal reçu. Si t_0 est l'instant d'émission du signal, le signal réfléchi correspondant sera reçu à l'instant $t = t_0 + \Delta t$, où :

$$\Delta t = \frac{2D}{c},$$

c désignant la vitesse de la lumière.
Le signal réfléchi subit également un décalage en fréquence dû à la vitesse du milieu. Si f_0 est la fréquence porteuse du signal émis, la fréquence porteuse du signal réfléchi est $f = f_0 + \Delta f$, où :

$$\Delta f = \frac{2v}{\lambda},$$

λ étant la longueur d'onde correspondant à la fréquence f_0. Ce phénomène physique est appelé "effet Doppler".
Ainsi, le signal reçu à l'instant t sera la sommation de signaux émis à des instants antérieurs et réfléchis sur des milieux plus ou moins éloignés du radar. La composante du signal reçu correspondante au signal émis à l'instant t_0 a pour expression :

$$A\exp(i\varphi)s_e(t_0)\exp\left(2i\pi(f_0 + \Delta f)t\right) + b(t),$$

où $A\exp(i\varphi)$ est un terme d'atténuation dû au milieu réfléchissant et à la propagation de l'onde et $b(t)$ est un bruit additif, souvent supposé gaussien. Cette composante a été réfléchie sur un milieu situé à une distance $D = \dfrac{c\Delta t}{2}$ et se déplaçant à une vitesse $v = \dfrac{\lambda \Delta f}{2}$.
Du signal reçu, nous avons besoin de connaître la nature du milieu situé à l'instant D. Pour cela, nous devons isoler la composante correspondante. Ceci se fait grâce à un filtrage appelé "filtrage adaptatif". Notons $S_r(t)$ le signal complexe reçu à l'instant t, il est montré dans [72] que la contribution du signal réfléchi par le milieu situé à une distance $D = \dfrac{c\Delta t}{2}$ est approximée par chacune des intégrales :

$$S_r(\Delta t, k) = \int_{kT_{\text{rec}}}^{(k+1)T_{\text{rec}}} S_r(u)\bar{S}_e(u - \Delta t, f_0)du, \tag{6.1}$$

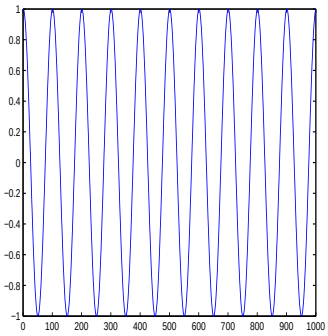
où $k \in \left\{0, \ldots, \dfrac{T_{\text{coh}}}{T_{\text{rec}}} - 1\right\}$. A k fixé, la fonction $\Delta t \to S_r(\Delta t, k)$ est couramment appelée "chirp". Le calcul de toutes les intégrales $S_r(\Delta t, k)$ fournit un échantillonnage en temps du signal reçu provenant de la distance D. Nous l'appelerons "échantillon In-Phase Quadrature" (IQ). Cet échantillon nous est utile pour déterminer le spectre de ce signal. Ce spectre appelé "spectre Doppler", nous permet ensuite de déterminer les composantes cinétiques du milieu. Conformément au théorème d'échantillonnage de Shannon, la largeur de bande du spectre Doppler est égale à l'inverse du pas d'échantillonnage, ainsi toute fréquence située hors de la bande $\left[f_0 - \dfrac{1}{2T_{\text{rec}}}, f_0 + \dfrac{1}{2T_{\text{rec}}}\right]$ ne peut être détectée. Soit $(z_1, \ldots, z_m)$ l'échantillon IQ, le spectre Doppler complexe est donné par :

$$\forall f \in \left[f_0 - \frac{1}{2T_{\text{rec}}}, f_0 + \frac{1}{2T_{\text{rec}}}\right],\ S_{\mathbb{C}}(f) = \frac{1}{\sqrt{2\pi m}} \sum_{k=1}^{m} z_k \exp\left(2i\pi k T_{\text{rec}}(f - f_0)\right), \tag{6.2}$$

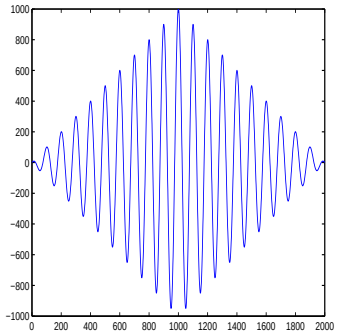
et la puissance spectrale Doppler correspondante par :

$$\forall f \in \left[f_0 - \frac{1}{2T_{\text{rec}}}, f_0 + \frac{1}{2T_{\text{rec}}}\right],\ S(f) = \frac{1}{2\pi m} \left| \sum_{k=1}^{m} z_k \exp\left(2i\pi k T_{\text{rec}}(f - f_0)\right) \right|^2. \tag{6.3}$$

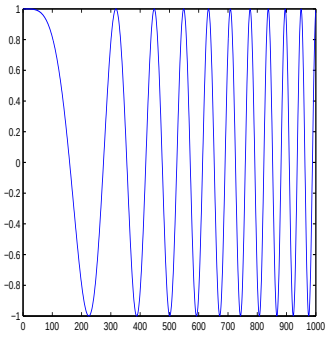
Les fréquences f pour lesquelles on détermine cette puissance seront appelées "fréquences Doppler".
Les radars utilisés pour les applications sont des radar à impulsions particuliers appelés "radar à compression d'impulsions". Dans un radar à compression d'impulsion, chaque impulsion émise est modulée en fréquence au lieu d'être un simple signal sinusoïdal. Les éléments physiques vus précédemment restent applicables, cependant, le filtrage adaptatif pour un k fixé produit un "chirp" qui décroît plus rapidement lorsqu'on s'éloigne de son maximum que dans le cas du radar à impulsion classique. Il s'ensuit un gain en résolution distance. La figure 6.2 présente le filtrage adaptatif d'une impulsion classique et d'une impulsion modulée en fréquence. Dans cet exemple, on suppose que $S_r(t) = S_e(t - \Delta t)$, ce qui correspond à la condition idéale d'un unique corps réfléchissant situé à la distance $D = \dfrac{c\Delta t}{2}$. Nous représentons uniquement les parties réelles des chirps.



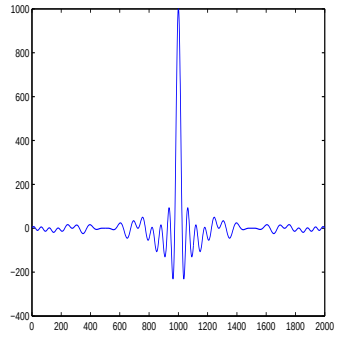
(a) Impulsion non modulée en fréquence.



(b) Impulsion non modulée en fréquence après filtrage.



(c) Impulsion modulée en fréquence.



(d) Impulsion modulée en fréquence après filtrage.

FIG. 6.2 – Filtrage adaptatif de deux types d'impulsion.

6.1.2 Principe de la détection TFAC

Le filtrage adaptatif présenté précédemment nous permet de déterminer les puissances reçues en fonction de la distance, ainsi que la répartition de cette puissance dans le spectre Doppler. Cependant, l'intérêt premier d'un radar est de pouvoir détecter automatiquement certains corps réfléchissant appelés "cibles". Ces cibles sont des objets matériels tels que des bateaux ou des êtres humains, ils ont la propriété de réfléchir plus le signal que le milieu environnant. Nous présentons dans cette sous-section, la technique permettant de détecter automatiquement ces cibles. Celle-ci est basée sur les tests d'hypothèses et est appelée "détection à Taux de Fausses Alarmes Constant" (TFAC). Notons (H0) l'hypothèse "absence de cible" et (H1) l'hypothèse "présence de cible". On appelera "case" le triplet (distance,azimut,fréquence Doppler), le couple (distance,azimut) sera appelé case distance-azimut. La case pour laquelle on cherche à déterminer la présence d'une cible sera appelée "case sous test". Le signal reçu d'une case donnée est la quantité $S_{\mathbb{C}}(f)$ obtenue après filtrage adaptatif et transformée de Fourier pour un azimut, un Δt et une fréquence f donnés. Soit r sa partie réelle. Cette quantité suit une certaine loi de probabilité qui diffère selon qu'il y ait une cible ou non. Notons $p(r|H0)$ la densité de probabilité de r en absence de cible et $p(r|H1)$ sa densité de probabilité en présence de cible. La probabilité de l'hypothèse Hi conditionnellement à l'observation r

est calculée par la règle de Bayes :

$$p(Hi|r) \propto p(r|Hi)p(Hi).$$

Nous décidons la présence d'une cible si :

$$\frac{p(H1|r)}{p(H0|r)} > 1,$$

ce qui est équivalent à :

$$\frac{p(r|H1)}{p(r|H0)} > \frac{p(H0)}{p(H1)}.$$

L'a priori $\lambda = \dfrac{p(H0)}{p(H1)}$ est déterminé de façon à maintenir la probabilité de fausses alarmes constante. Cette probabilité de fausses alarmes est égale à :

$$\mathbb{P}_{\text{fa}} = \int_{\frac{p(r|H1)}{p(r|H0)}>\lambda} p(r|H0)dr.$$

On définit également la probabilité de détection par :

$$\mathbb{P}_{\text{d}} = \int_{\frac{p(r|H1)}{p(r|H0)}>\lambda} p(r|H1)dr.$$

On remarque que plus λ est petit et mieux on détecte mais plus on a de fausses alarmes. Il nous faut donc trouver un compromis entre taux de fausses alarmes bas et taux de détection élevé. La qualité d'un détecteur est mesuré par la probabilité de détection en fonction de la probabilité de fausses alarmes. Dans la pratique, on exige un taux de fausses alarmes égal à 10^{-6} et le taux de détection s'échelonne entre 0.5 et 0.9 selon les applications.
Examinons, pour illustrer, le cas où le signal complexe reçu en absence de cible suit une loi normale complexe circulaire centrée de variance $\sigma^2_{\mathbb{C}}$, hypothèse faite dans la plupart des cas. La quantité $\sigma^2_{\mathbb{C}}$ est la puissance du signal reçu en absence de cible. Sous cette condition, r suit une loi normale réelle centrée de variance $\sigma^2 = \dfrac{\sigma^2_{\mathbb{C}}}{2}$. Ainsi :

- sous (H0) :

  $$r = b,$$

  où b suit une loi normale réelle centrée et de variance σ^2 ;
- sous (H1) :

  $$r = a + b,$$

  où $a \in \mathbb{R}^+$.

Le rapport $SNR = \dfrac{a^2}{2\sigma^2}$ est appelé rapport signal sur bruit (SNR). Dans ce cas, la décision "présence d'une cible" est équivalente à :

$$\frac{r}{\sigma} > \frac{\sigma}{a}\log(\lambda) + \frac{a}{2\sigma}.$$

Par conséquent, pour une probabilité de fausses alarmes et un rapport signal sur bruit donnés, on a :

$$\log \lambda = \sqrt{2SNR}\phi^{-1}\left(1 - \mathbb{P}_{\text{fa}}\right) - SNR,$$

où ϕ est la fonction de répartition d'une loi normale centrée réduite.
On montre alors que $\mathbb{P}_{\text{d}}$ est donné en fonction de $\mathbb{P}_{\text{fa}}$ par :

$$\mathbb{P}_{\text{d}} = 1 - \phi\left(\phi^{-1}\left(1 - \mathbb{P}_{\text{fa}}\right) - \sqrt{2SNR}\right).$$

Dans la pratique, l'écart-type σ est estimé à partir des signaux reçus de cases ayant même azimut et même fréquence Doppler et dont les distances sont voisines de celle de la case sous test. On y estime également la moyenne m et on retranche au signal reçu de la case sous test cette moyenne. On décide alors la présence d'une cible si $\frac{|r-m|}{\sigma} >^{H1} \alpha$, où $\alpha = \frac{\sigma}{a}\log(\lambda) + \frac{a}{2\sigma}$. Le détecteur ainsi construit est optimal lorsque le bruit est gaussien. Ainsi, les courbes de performances présentées dans la figure 6.3 sont les meilleures que l'on puissent avoir. Cependant, on remarque que pour un rapport signal sur bruit égal à 2 et une probabilité de fausses alarmes égale à 10^{-4}, la probabilité de détection est inférieure à 0.1. Ainsi, même dans des conditions idéales, les performances de ce détecteur ne sont pas si bonnes. Cette remarque nous a motivé à concevoir un détecteur TFAC original. Dans le détecteur que nous proposons, la détection s'effectue directement sur les échantillons IQ, ces derniers contenant toute la connaissance disponible sur le spectre Doppler.

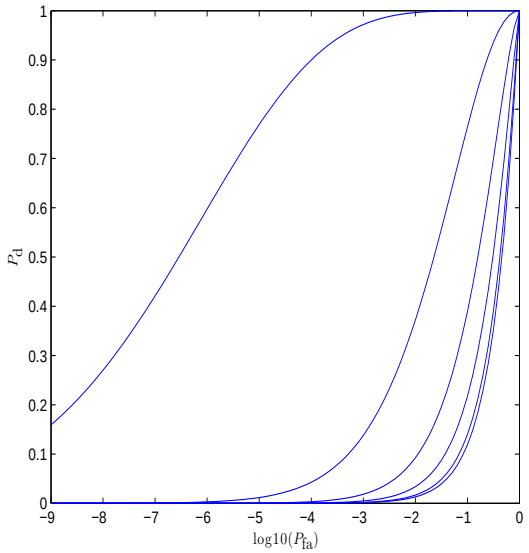


FIG. 6.3 – Courbes de performance $\mathbb{P}_{\text{d}}$ fonction de $\log_{10}\mathbb{P}_{\text{fa}}$ pour différents rapport signal sur bruit : 0.005, 0.02, 0.125, 0.5, 2 et 12.5 de bas en haut.

6.2 Détection à partir des données IQ

Nous présentons dans cette section un détecteur TFAC original. Contrairement au détecteur TFAC classique qui n'utilise uniquement l'intensité rétrodiffusée, celui-ci utilise tout un échantillon IQ obtenu à une distance et un azimut donné. Pour définir notre détecteur, nous nous inspirons de la détection TFAC classique. Le détecteur classique utilise la quantité $\frac{|r-m|}{\sigma}$ qui s'interprète comme une distance entre le signal r de la case sous test et la moyenne m des cases environnantes. Dans notre nouveau détecteur, nous définirons la distance entre l'échantillon IQ observé et la moyenne des échantillons IQ environnants. La notion de distance entre échantillon IQ n'est pas triviale. En effet, la distance entre deux échantillons IQ ayant même densité spectrale devra être nulle. On supposera dans la suite que les échantillons IQ sont réalisations de processus gaussiens complexes circulaires stationnaires du second ordre et dont la famille de covariance est sommable. Ainsi la densité spectrale est la transformée de Fourier de la matrice de covariance de l'échantillon IQ. La distance entre deux échantillons IQ sera donc définie comme la distance entre leur matrice de covariance respective. Il s'agira donc d'une distance entre distributions de probabilité, qui, comme nous le verrons, n'est pas euclidienne. Une fois la notion de distance définie, on définira la notion de moyenne d'une famille de matrices de covariance, nous verrons en particulier que la notion de moyenne dépend de la distance que l'on a définie.

6.2.1 Distance entre distributions d'une même famille paramétrique

Considérons une famille de densité de probabilité $\Lambda = \{y \in \mathcal{Y} \to p(y;\theta) : \theta \in \Theta\}$ par rapport à une mesure de référence ν sur $\mathcal{Y}$. L'ensemble des paramètres $\Theta \subset \mathbb{R}^k$ sera supposé ouvert non vide de $\mathbb{R}^k$. Nous allons détailler ici la construction de la distance entre deux distributions $p(.;\theta_1)$ et $p(.;\theta_2)$ telle qu'elle a été faite par C. R. Rao [3].
Tout d'abord, il est courant de considérer l'information de Kullback :

$$\bar{K}(\theta_1,\theta_2) = \int \log\left(\frac{p(y;\theta_1)}{p(y;\theta_2)}\right) p(y;\theta_1) d\nu(y),$$

comme une "métrique" entre distributions. Cependant, celle-ci n'est pas symétrique. Afin de définir la distance de Rao, considérons le développement limité de la fonction $\tilde{\theta} \to \bar{K}(\theta,\tilde{\theta})$ au voisinage de θ :

$$\bar{K}(\theta,\theta+d\theta) = \frac{1}{2}(d\theta)^T I_Y(\theta) d\theta + \circ(\|d\theta\|^2),$$

où $I_Y(\theta)$ est la matrice d'information de Fisher. Ainsi, pour deux valeurs de paramètres θ et $\theta + d\theta$ suffisamment proches l'une de l'autre, à une constante près, la distance entre les distributions $p(.;\theta)$ et $p(.;\theta+d\theta)$ peut être approximée par $(d\theta)^T I_Y(\theta) d\theta$. On remarque que cette forme quadratique est symétrique et elle permet de définir la métrique différentielle :

$$dl = \sqrt{(d\theta)^T I_Y(\theta) d\theta}, \tag{6.4}$$

appelée "distance de Rao". On remarquera en reprenant le raisonnement de la sous-section 1.2.1 du chapitre 1 que cette métrique est invariante par changement de paramétrage $\eta = g(\theta)$ et ne dépend donc que de l'espace fonctionnel Λ. L'élément dl sera interprété comme l'élément

de longueur de Λ. Considèrons une courbe de dimension 1 dans Λ d'extrémités $p(.;\theta_1)$ et $p(.;\theta_2)$, celle-ci est paramétrée par :

$$t \in [t_1, t_2] \rightarrow \gamma_\Lambda(t) \in \Lambda,$$

où $\gamma_\Lambda(t_j) = p(.;\theta_j)$. A cette courbe, on peut lui faire correspondre une courbe θ de Θ par $\gamma_\Lambda = \varphi \circ \theta$ où φ est l'application qui à $\theta \in \Theta$ associe $p(.;\theta)$. La longueur de cette courbe est égale à :

$$L_\theta(\theta_1, \theta_2) = \int_{t_1}^{t_2} \sqrt{(\theta'(t))^T I_Y(\theta(t))(\theta'(t))} dt. \tag{6.5}$$

Toute courbe minimisant l'intégrale (6.5) est appelée géodésique et la distance entre $p(.;\theta_1)$ et $p(.;\theta_2)$ est par définition la longueur d'une géodésique, elle est définie par :

$$D(\theta_1, \theta_2) = \min_\theta \int_{t_1}^{t_2} \sqrt{(\theta'(t))^T I_Y(\theta(t))(\theta'(t))} dt. \tag{6.6}$$

On remarquera que nous avons utilisé l'information de Kullback entre deux distributions afin de définir la métrique différentielle de Rao. Dans [9], il y figure différentes mesures de divergence entre lois dont la différentielle seconde donne la même métrique différentielle.
Expliquons maintenant le lien entre la métrique de Rao et les mesures de Jeffreys abordées au chapitre 1, ceci peut justifier l'usage de la métrique de Rao plutôt qu'une autre divergence symétrique entre lois. Lorsque l'espace Λ est muni de l'élément de longueur dl exprimé dans le paramétrage Θ par (6.4), l'élément vectoriel de longueur $\overrightarrow{dl}$ a pour expression :

$$\overrightarrow{dl} = \begin{pmatrix} dl_1 \\ \vdots \\ dl_k \end{pmatrix} = (I_Y(\theta))^{\frac{1}{2}} \begin{pmatrix} d\theta_1 \\ \vdots \\ d\theta_k \end{pmatrix}.$$

On en déduit alors que l'élément de volume de Λ a pour expression :

$$d\tau = \sqrt{\det I_Y(\theta)} d\theta_1 \dots d\theta_k.$$

On reconnaît alors la mesure de Jeffreys. Plus exactement, choisir θ selon une loi de Jeffreys sur Θ est équivalent à choisir la fonctionnelle $p(.;\theta)$ selon une loi uniforme sur Λ. On retrouve ainsi le caractère non informatif de la loi uniforme. On remarquera d'ailleurs, si Θ est discret, l'ensemble Λ l'est aussi et peut être identifié à Θ.

Nous allons calculer dans la section suivante la distance entre deux lois gaussiennes centrées et circulaires. On définira la distance entre deux matrices de covariance comme la distance entre les lois gaussiennes centrées circulaires correspondantes.

6.2.2 Distance entre lois normales complexes centrées circulaires

Considérons la famille des lois gaussiennes complexes circulaires centrées paramétrées par la matrice de covariance Σ. Dans la suite, cette famille sera confondue avec celle des matrices hermitiennes définies positives. La densité de telles lois est :

$$z \in \mathbb{C}^n \quad \rightarrow \quad \frac{1}{\pi^n \det \Sigma} \exp\left(-\bar{z}^T \Sigma^{-1} z\right).$$

En utilisant le développement limité de l'information de Kullback, on montre que la forme quadratique différentielle associée à la métrique de Rao est :

$$dl^2 = \mathrm{tr}\left(\Sigma^{-1} d\Sigma \Sigma^{-1} d\Sigma\right).$$

Soient $\Sigma^{(1)}$ et $\Sigma^{(2)}$ deux matrices hermitiennes définies positives. Considérons une courbe $t \in [0,1] \rightarrow \Sigma(t)$ dans l'espace des paramètres telle que $\Sigma(0) = \Sigma^{(1)}$ et $\Sigma(1) = \Sigma^{(2)}$. La longueur de cette courbe est donnée par :

$$\int_0^1 \sqrt{\mathrm{tr}\left(\Sigma^{-1}(t)\Sigma'(t)\Sigma^{-1}(t)\Sigma'(t)\right)} dt,$$

et la distance entre deux lois gaussiennes circulaires de covariances respectives $\Sigma^{(1)}$ et $\Sigma^{(2)}$ est alors donnée par :

$$D(\Sigma^{(1)}, \Sigma^{(2)}) = \min_{\Sigma} \int_0^1 \sqrt{\mathrm{tr}\left(\Sigma^{-1}(t)\Sigma'(t)\Sigma^{-1}(t)\Sigma'(t)\right)} dt.$$

Calculons cette distance. Pour cela, on effectue le changement de variable $\tilde{\Sigma} = (\Sigma^{(1)})^{-\frac{1}{2}} \Sigma (\Sigma^{(1)})^{-\frac{1}{2}}$ où $\Sigma^{\alpha} = \exp\left(\alpha \log(\Sigma)\right)$. La métrique différentielle s'écrit alors :

$$dl^2 = \mathrm{tr}\left(\tilde{\Sigma}^{-1} d\tilde{\Sigma} \tilde{\Sigma}^{-1} d\tilde{\Sigma}\right),$$

et donc $D(\Sigma^{(1)}, \Sigma^{(2)}) = D(\mathrm{Id}, (\Sigma^{(1)})^{-\frac{1}{2}} \Sigma^{(2)} (\Sigma^{(1)})^{-\frac{1}{2}})$ et $t \rightarrow \Sigma(t)$ est une géodésique entre $\Sigma^{(1)}$ et $\Sigma^{(2)}$ si et seulement si $t \rightarrow (\Sigma^{(1)})^{-\frac{1}{2}} \Sigma(t) (\Sigma^{(1)})^{-\frac{1}{2}}$ est une géodésique entre $\tilde{\Sigma}^{(1)} = \mathrm{Id}$ et $\tilde{\Sigma}^{(2)} = (\Sigma^{(1)})^{-\frac{1}{2}} \Sigma^{(2)} (\Sigma^{(1)})^{-\frac{1}{2}}$.
Par le changement de variable $\tilde{\Sigma} = P \Delta \bar{P}^T$, où Δ est diagonale à valeurs propres réelles positives et P est une matrice orthogonale, on a de même :

$$dl^2 = \mathrm{tr}\left(\Delta^{-1} d\Delta \Delta^{-1} d\Delta\right),$$

qui est indépendant de $t \rightarrow P(t)$. Soient λ_j les éléments diagonaux de Δ, valeurs propres de $(\Sigma^{(1)})^{-\frac{1}{2}} \Sigma (\Sigma^{(1)})^{-\frac{1}{2}}$ et $x_j = \log(\lambda_j)$. On a :

$$dl^2 = \sum_{j=1}^{n} \left(\frac{d\lambda_j}{\lambda_j}\right)^2 = \sum_{j=1}^{n} dx_j^2.$$

La métrique différentielle correspond donc à une métrique euclidienne dans l'espace des logarithmes des valeurs propres de Δ. On en déduit que la géodésique dans l'espace des valeurs propres est unique et donnée par $t \rightarrow t \log(\mu_j)$, où les μ_j sont les valeurs propres de $(\Sigma^{(1)})^{-\frac{1}{2}} \Sigma^{(2)} (\Sigma^{(1)})^{-\frac{1}{2}}$.
La distance entre les matrices hermitiennes définies positives $\Sigma^{(1)}$ et $\Sigma^{(2)}$ est finalement donnée par :

$$D(\Sigma^{(1)}, \Sigma^{(2)}) = \sqrt{\sum_{j=1}^{n} (\log(\mu_j))^2}. \tag{6.7}$$

Cette distance entre matrices hermitiennes définies positives a été définie dans un contexte différent par C. L. Siegel [111], nous l'appelerons donc distance de Siegel. De l'expression de la géodésique dans l'espace des logarithmes des valeurs propres, on en déduit que $\log(\Delta(t)) = t\log(\Delta^{(2)})$ et donc $\Delta(t) = \left(\Delta^{(2)}\right)^t$. Ainsi toute géodésique $t \to \tilde{\Sigma}(t)$ a pour expression :

$$\tilde{\Sigma}(t) = P(t)\bar{P}(1)^T \left(\tilde{\Sigma}^{(2)}\right)^t P(1)\bar{P}(t)^T.$$

On en déduit finalement que toute géodésique entre $\Sigma^{(1)}$ et $\Sigma^{(2)}$ dans l'espace des matrices hermitiennes définies positives a pour expression :

$$t \in [0,1] \to (\Sigma^{(1)})^{\frac{1}{2}} P(t)\bar{P}(1)^T \left[(\Sigma^{(1)})^{-\frac{1}{2}}\Sigma^{(2)}(\Sigma^{(1)})^{-\frac{1}{2}}\right]^t P(1)\bar{P}(t)^T(\Sigma^{(1)})^{\frac{1}{2}},$$

où $t \in [0,1] \to P(t)$ est une fonction différentiable à valeurs dans l'ensemble des matrices orthogonales telle que les colonnes de $P(1)$ soient les vecteurs propres de $(\Sigma^{(1)})^{-\frac{1}{2}}\Sigma^{(2)}(\Sigma^{(1)})^{-\frac{1}{2}}$. Dans la suite, nous choisirons la géodésique telle que pour tout t, $P(t) = P(1)$, ainsi son expression est :

$$t \in [0,1] \to (\Sigma^{(1)})^{\frac{1}{2}} \left[(\Sigma^{(1)})^{-\frac{1}{2}}\Sigma^{(2)}(\Sigma^{(1)})^{-\frac{1}{2}}\right]^t (\Sigma^{(1)})^{\frac{1}{2}}. \tag{6.8}$$

6.2.3 Moyenne de matrices hermitiennes définies positives

Nous définissons la moyenne de matrices hermitiennes définies positives, également utile dans notre détecteur TFAC. La moyenne dépend de la distance et est donnée par la définition suivante :

Définition 6.2.1 (Moyenne de deux éléments d'une variété différentielle)**.** *Soit Λ un espace métrique muni de la distance D. Une moyenne de deux éléments f et g est l'élément $m = m(f,g)$ d'une géodésique vérifiant :*

$$D(f,m) = D(m,g) = \frac{1}{2}D(f,g).$$

Si la géodésique est unique, la moyenne est unique. Dans le cas contraire, on doit se fixer une géodésique pour définir la moyenne. Reprenons le cas des lois gaussiennes complexes circulaires centrées qui est identifié à l'espace des matrices hermitiennes définies positives. La moyenne de $\Sigma^{(1)}$ et de $\Sigma^{(2)}$ relative à la géodésique (6.8) est égale à :

$$M = (\Sigma^{(1)})^{\frac{1}{2}} \left[(\Sigma^{(1)})^{-\frac{1}{2}}\Sigma^{(2)}(\Sigma^{(1)})^{-\frac{1}{2}}\right]^{\frac{1}{2}} (\Sigma^{(1)})^{\frac{1}{2}}.$$

Dans le cas monovarié, on retrouve la moyenne géométrique. Nous appelerons alors cette moyenne "moyenne géométrique" de deux matrices hermitiennes définies positives.
Pour définir la moyenne de plusieurs éléments, nous généralisons la notion d'isobarycentre. Il y a plusieurs manières de généraliser la notion d'isobarycentre. La première est celle de D. Petz [93] qui utilise la propriété suivante :

Proposition 6.2.1. *Soient $M_1, \ldots, M_n$, n points d'un espace vectoriel et soit G leur isobarycentre. Notant G_j l'isobarycentre des points $(M_i)_{i\neq j}$, alors G est également isobarycentre des points $G_1, \ldots, G_n$.*

Bien que l'ensemble des matrices hermitiennes définies positives ne soit pas un espace vectoriel, il est possible d'étendre la définition de barycentre en utilisant cette propriété. Ainsi, il propose la méthode récursive suivante pour le calcul de la moyenne de $N+1$ matrices hermitiennes définies positives $\Sigma^{(1)}, \ldots, \Sigma^{(N+1)}$:

1. soit M_N l'opérateur qui à N matrices hermitiennes définies positives associe sa moyenne ;
2. on construit la suite de polygône T_k suivante :
   (a) $T_0 = \left(\Sigma^{(1)}, \ldots, \Sigma^{(N+1)}\right)$;
   (b) $T_{n+1} = \left(G_{n+1}^{(1)}, \ldots, G_{n+1}^{(N+1)}\right)$, où $G_{n+1}^{(j)} = M_N\left(G_n^{(1)}, \ldots, G_n^{(j-1)}, G_n^{(j+1)}, \ldots, G_n^{(N+1)}\right)$.

La suite des polygônes T_n converge vers un singleton. La moyenne M_{N+1} est alors définie comme la limite de cette suite.
Cependant, dans la pratique nous ne calculerons pas la moyenne de cette manière, l'algorithme ainsi construit étant de complexité $\mathcal{O}(N!)$. T. Ando et R. Mathias [4] propose une autre définition qui utilise l'associativité du barycentre et qui permet de calculer la moyenne pour 2^n matrices. Pour cela, si on sait calculer la moyenne de 2 matrices, la moyenne de 4 matrices est définie comme la moyenne de la moyenne des 2 premières et de la moyenne des 2 dernières et ainsi de suite. Cependant, la moyenne obtenue est différente de la précédente et ne généralise pas convenablement la notion de barycentre.
Nous allons plutôt utiliser la définition de H. Kärcher [62]. Cette définition généralise la propriété suivante :

$$\overrightarrow{GM_1} + \ldots + \overrightarrow{GM_N} = 0, \tag{6.9}$$

où G est l'isobarycentre des points $M_1, \ldots, M_N$.
Considérons un espace muni d'une métrique différentielle. Un point G est un barycentre de Kärcher de N points $M_1, \ldots, M_N$ s'il satisfait la relation :

$$\sum_{k=1}^{N} \left.\frac{d\gamma_k(t)}{dt}\right|_{t=0} = 0, \tag{6.10}$$

où γ_k est la géodésique de G à M_k.
Dans le cas de l'ensemble des matrices hermitiennes définies positives muni de la géodésique définie par (6.8), soit G l'isobarycentre des N matrices $\Sigma^{(1)}, \ldots, \Sigma^{(N)}$, la géodésique γ_k est donnée par :

$$\gamma_k(t) = G^{\frac{1}{2}}\left[G^{-\frac{1}{2}}\Sigma^{(k)}G^{-\frac{1}{2}}\right]^t G^{\frac{1}{2}}.$$

Ainsi :

$$\left.\frac{d\gamma_k(t)}{dt}\right|_{t=0} = G^{\frac{1}{2}}\log\left(G^{-\frac{1}{2}}\Sigma^{(k)}G^{-\frac{1}{2}}\right)G^{\frac{1}{2}}.$$

On en déduit que la moyenne G de Kärcher de $\Sigma^{(1)}, \ldots, \Sigma^{(N)}$ vérifie :

$$\sum_{k=1}^{N} \log\left(G^{-\frac{1}{2}}\Sigma^{(k)}G^{-\frac{1}{2}}\right) = 0.$$

Cette moyenne peut se calculer récursivement de la manière suivante [7] :

1. initialisation de la suite G_0 ;
2. à partir de G_m, on calcule $G_{m+1} = G_m^{\frac{1}{2}} \exp\left(-\epsilon \sum_{k=1}^{N} \log\left(G_m^{-\frac{1}{2}} \Sigma^{(k)} G_m^{-\frac{1}{2}}\right)\right) G_m^{\frac{1}{2}}$, avec $-1 < \epsilon < 0$.

La limite de cette suite est la moyenne de Kärcher.

6.2.4 Principe du détecteur TFAC

Dans cette sous-section, nous décrivons notre détecteur TFAC. En chaque case distance-azimut, nous observons des échantillons IQ qui sont des vecteurs complexes. Avant de tester la présence de cible, nous commençons par estimer les matrices de covariances des échantillons. L'estimation de ces matrices de covariance sera expliquée dans la section suivante. Une fois ces matrices estimées, nous calculons les moyennes des matrices des cases environnantes, puis nous calculons la distance de la matrice de la case sous test à cette moyenne. Le test s'écrit alors :

$$\text{Si } D(M, \Sigma_{\text{test}}) > \lambda, \text{ nous décidons la présence d'une cible,}$$

où M est la moyenne, Σ_{test} la matrice de covariance sous test et λ un seuil déterminé pour maintenir le taux de fausses alarmes.
Ce détecteur TFAC pourra également utiliser une segmentation du signal obtenue à l'aide d'un modèle de chaînes de Markov cachées. La segmentation bayésienne apparaît ainsi comme un traitement en amont du signal. Une fois cette segmentation obtenue, on considèrera uniquement les cases environnantes dans la même classe que la case sous test, les échantillons IQ considérés pour le calcul de la moyenne auront alors des propriétés statistiques voisines.
La section suivante présente la segmentation bayésienne à partir des données observées qui sont les échantillons IQ. Ces derniers étant de grande taille (256 valeurs complexes), nous aurons besoin de réduire l'espace d'observation tout en conservant le maximum d'information. La technique de réduction de l'espace d'observation est également présentée dans la section suivante.

6.3 Segmentation et prétraitement des données radar

L'algorithme de Burg présenté dans la sous-section suivante permet de réduire l'espace d'observation afin de pouvoir utiliser les modèles de chaînes de Markov cachées pour la segmentation des données radar. Les vecteurs complexes de taille réduite obtenus par l'algorithme de Burg seront ensuite utilisés pour l'estimation des matrices de covariance des échantillons IQ par l'algorithme de Gohberg-Semencul [35].

6.3.1 Algorithme de Burg et estimation des covariances

Algorithme de Burg

Soit $Y = (Y_n)_{n\in\mathbb{Z}}$ un processus centré et stationnaire du second ordre à valeurs dans $\mathbb{C}$ de covariance complexe $\gamma(k) = \mathbb{E}\left(Y_n \bar{Y}_{n-k}\right)$. L'ensemble des variables aléatoires de carré

intégrable est muni du produit scalaire $\langle X, Z\rangle = \mathbb{E}\left(X\bar{Z}\right)$, on notera $\|.\|$ la norme associée. On a :

$$Y_n = \underbrace{\left(Y_n + \sum_{k=1}^{p} a_k^{(p)} Y_{n-k}\right)}_{\text{Orthogonal à } Y_{n-1},\ldots,Y_{n-p}} + \underbrace{\left(-\sum_{k=1}^{p} a_k^{(p)} Y_{n-k}\right)}_{\text{Proj}(Y_n|Y_{n-1},\ldots,Y_{n-p})},$$

où $\text{Proj}(Y_n|Y_{n-1},\ldots,Y_{n-p})$ est la projection orthogonale de Y_n sur le sous-espace vectoriel engendré par $Y_{n-1},\ldots,Y_{n-p}$. Les coefficients $a_k^{(p)}$ sont appelés coefficients auto-régressifs et les coefficients $\mu_p = a_p^{(p)}$ sont appelés coefficients d'auto-corrélation partielle ou coefficients de réflexion. Ce sont les coefficients de réflexion estimés par l'algorithme de Burg que l'on observera lors de la segmentation, la réalisation de Y correspond alors aux données IQ d'une case distance-azimut.
Les coefficients de réflexion sont solutions de l'équation matricielle de Yule-Walker suivante :

$$\begin{pmatrix} \gamma(0) & \gamma(1) & \ldots & \ldots & \gamma(p-1) \\ \bar{\gamma}(1) & \gamma(0) & \gamma(1) & \ldots & \gamma(p-2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \ldots & \ldots & \ldots & \bar{\gamma}(1) & \gamma(0) \end{pmatrix} \times \begin{pmatrix} a_p^{(p)} \\ \vdots \\ a_1^{(p)} \end{pmatrix} = - \begin{pmatrix} \gamma(p) \\ \vdots \\ \gamma(1) \end{pmatrix}. \tag{6.11}$$

Ces équations peuvent se résoudre par l'algorithme de Durbin-Levinson suivant :

1. initialisation :

$$\mu_1 = a_1^{(1)} = -\frac{\gamma(1)}{\gamma(0)},$$

$$\epsilon_1 =^{\text{def}} \|Y_2 - \text{Proj}(Y_2|Y_1)\|^2 = \gamma(0)\left(1 - |\mu_1|^2\right) \ ;$$

2. itération :

$$a_{p+1}^{(p+1)} = \mu_{p+1} = -\frac{\gamma(p+1) + \sum_{k=1}^{p} a_k^{(p)} \gamma(p+1-k)}{\epsilon_p},$$

$$\epsilon_{p+1} =^{\text{def}} \|Y_{p+2} - \text{Proj}(Y_{p+2}|Y_1,\ldots,Y_{p+1})\|^2 = \epsilon_p\left(1 - |\mu_{p+1}|^2\right),$$

et pour $k \in \{1,\ldots,p\}$,

$$a_k^{(p+1)} = a_k^{(p)} + \mu_{p+1}\bar{a}_{p+1-k}^{(p)}. \tag{6.12}$$

Contrairement à l'algorithme de Durbin-Levinson, l'algorithme de Burg ne nécessite pas de connaître les covariances. De plus, ces dernières seraient mal estimées à partir des échantillons IQ, ceux-ci n'étant pas assez grands. Cependant, l'algorithme de Burg ne calcule pas les coefficients de réflexion de manière exacte mais les estime par une méthode semblable à celle des moindres carrés.

Détaillons l'algorithme de Burg. Soit $(y_1, \ldots, y_M)$ un échantillon IQ en une case distance-azimut donnée. On définit les erreurs de filtrage (directes) et de lissage (rétrogrades) pour $1 \leq p \leq M-1$ par :

$$\begin{aligned} f_p(n) &= y_n - \mathrm{Proj}(y_n | y_{n-1}, \ldots, y_{n-p}) = y_n + \sum_{k=1}^{p} a_k^{(p)} y_{n-k}, \\ b_p(n) &= y_{n-p} - \mathrm{Proj}(y_{n-p} | y_{n-p+1}, \ldots, y_n) = y_{n-p} + \sum_{k=1}^{p} \bar{a}_k^{(p)} y_{n-p+k}. \end{aligned}$$

L'algorithme de Burg consiste à minimiser :

$$U^{(p)} = \sum_{n=p+1}^{M} \left[|f_p(n)|^2 + |b_p(n)|^2 \right].$$

Pour $p = 1$, on doit chercher $\mu_1 = a_1^{(1)}$ minimisant :

$$U^{(1)} = \sum_{n=2}^{M} \left[|y_n + \mu_1 y_{n-1}|^2 + |y_{n-1} + \bar{\mu}_1 y_n|^2 \right],$$

μ_1 est alors donné par :

$$\mu_1 = -2 \frac{\sum_{n=2}^{M} y_n \bar{y}_{n-1}}{\sum_{n=2}^{M} \left(|y_{n-1}|^2 + |y_n|^2 \right)}. \tag{6.13}$$

En utilisant la formule (6.12), on montre que :

$$\begin{aligned} f_{p+1}(n) &= f_p(n) + \mu_{p+1} b_p(n-1), \\ b_{p+1}(n) &= b_p(n-1) + \bar{\mu}_{p+1} f_p(n), \end{aligned}$$

ainsi μ_{p+1} doit minimiser :

$$U^{(p+1)} = \sum_{n=p+2}^{M} \left[|f_p(n) + \mu_{p+1} b_p(n-1)|^2 + |b_p(n-1) + \bar{\mu}_{p+1} f_p(n)|^2 \right],$$

il est alors donné par :

$$\mu_{p+1} = -2 \frac{\sum_{n=p+2}^{M} f_p(n) \bar{b}_p(n-1)}{\sum_{n=p+2}^{M} \left(|f_p(n)|^2 + |b_p(n-1)|^2 \right)}. \tag{6.14}$$

Ainsi l'algorithme de Burg fonctionne de la manière suivante :

1. initialisation :
   - calcul de μ_1 en utilisant la formule (6.13) ;
   - calcul des erreurs de filtrage et de lissage pour $n \geq 2$:
   $$\begin{aligned} f_1(n) &= y_n + \mu_1 y_{n-1}, \\ b_1(n) &= y_{n-1} + \bar{\mu}_1 y_n \ ; \end{aligned}$$
2. itération :
   - calcul de μ_{p+1} en utilisant (6.14) ;
   - calcul des erreurs de filtrage et de lissage pour $n \geq p+2$:
   $$\begin{aligned} f_{p+1}(n) &= f_p(n) + \mu_{p+1} b_p(n-1), \\ b_{p+1}(n) &= b_p(n-1) + \bar{\mu}_{p+1} f_p(n) \ ; \end{aligned}$$
3. estimation de la variance $\gamma(0)$ puis calcul des erreurs de prédictions :
$$\begin{aligned} \epsilon_1 &= \gamma(0)\left(1 - |\mu_1|^2\right), \\ \epsilon_{p+1} &= \epsilon_p\left(1 - |\mu_{p+1}|^2\right) ; \end{aligned}$$
4. calcul des $a_k^{(p)}$ par (6.12).

La variance $\gamma(0)$ correspond à la puissance reçue de la case distance-azimut et les coefficients de réflexion caractérisent la forme du spectre Doppler.
On peut montrer que si $\mu_{p+1} = 0$, alors pour tout $k \geq p+1$, $\mu_k = 0$, le processus Y est alors un processus auto-régressif d'ordre p (AR(p)). Dans la pratique, on imposera un ordre p et on considérera que pour $k \geq p+1$, $\mu_k = 0$. Dans ce cas, pour tout $n \geq p$, le processus Y satisfait pour $n \geq p$:

$$\mathrm{Proj}(Y_n | Y_{n-1}, \ldots, Y_{n-p}) = \mathrm{Proj}(Y_n | (Y_m)_{m \leq n-1}),$$

et $B_n = Y_n - \mathrm{Proj}(Y_n | (Y_m)_{m \leq n-1})$ est une séquence indépendante centrée et de variance ϵ_p appelée bruit d'innovation. Comme :

$$B_n = Y_n + \sum_{k=1}^{p} a_k^{(p)} Y_{n-k},$$

alors la densité spectrale de Y correspondant au spectre Doppler s'écrit :

$$\forall f \in \left[f_0 - \frac{1}{2T_{\mathrm{rec}}}, f_0 + \frac{1}{2T_{\mathrm{rec}}}\right], \ S(f) = \frac{\epsilon_p}{2\pi} \times \frac{1}{\left|1 + \sum_{k=1}^{p} a_k^{(p)} \exp\left(2ik\pi T_{\mathrm{rec}}(f - f_0)\right)\right|^2}. \quad (6.15)$$

Dans la pratique, il se peut que l'ordre p imposé soit trop petit et que μ_{p+1} ait une valeur non négligeable. Ainsi, dans [6], il est proposé un algorithme de Burg régularisé de façon à imposer une décroissance rapide des μ_p. Dans cette nouvelle version de Burg, la quantité $U^{(p)}$ à minimiser est remplacée par :

$$E^{(p)} = U^{(p)} + \frac{1}{N-p}\left[\gamma_0 \int_{-\frac{1}{2}}^{\frac{1}{2}} \left|A^{(p)}(\lambda)\right|^2 d\lambda + \gamma_1 \int_{-\frac{1}{2}}^{\frac{1}{2}} \left|\frac{dA^{(p)}}{d\lambda}(\lambda)\right|^2 d\lambda\right],$$

où γ_0 et γ_1 sont deux coefficients de régularisation et

$$A^{(p)}(\lambda) = 1 + \sum_{k=1}^{p} a_k^{(p)} e^{2i\pi k\lambda}.$$

Le calcul des coefficients de réflexion s'effectue de manière similaire et est détaillé dans [6]. Dans les expérimentations, nous prendrons $p = 10$, $\gamma_0 = 0$ et $\gamma_1 = 10000$.
Nous voyons ainsi que l'algorithme de Burg permet de transformer les vecteurs IQ de chaque case distance-azimut en vecteurs complexes de taille réduite égale à p.
A partir des coefficients de réflexion, il est possible de calculer la famille de covariance grâce à l'algorithme de Gohberg-Semencul que nous présentons dans le paragraphe suivant.

Algorithme de Gohberg-Semencul

Nous présentons maintenant le calcul de la covariance $\gamma(k) = \mathbb{E}\left(Y_n \bar{Y}_{n-k}\right)$ à partir des coefficients de réflexion $(\mu_1, \ldots, \mu_p)$, de la puissance spectrale $\gamma(0)$, des erreurs de prédictions $(\epsilon_1, \ldots, \epsilon_{p-1})$ et des coefficients auto-régressifs. Ce calcul s'effectue récursivement de la manière suivante :

1. initialisation :

$$\gamma(1) = -\gamma(0)\mu_1 \ ;$$

2. itération :
pour $m \leq p-1$:

$$\gamma(m+1) = -\left[\epsilon_m \mu_{m+1} + \sum_{k=1}^{m} a_k^{(m)} \gamma(m+1-k)\right],$$

et pour $m \geq p$:

$$\gamma(m+1) = -\sum_{k=1}^{p} a_k^{(p)} \gamma(m+1-k).$$

La connaissance de $(\gamma(0), \ldots, \gamma(p))$ suffit pour déterminer les coefficients auto-régressifs jusqu'à l'ordre p grâce aux formules de Durbin-Levinson. Remarquant alors que pour $m \geq p$, le calcul de $\gamma(m+1)$ n'utilise uniquement les coefficients auto-régressifs jusqu'à l'ordre p, ainsi la connaissance de $(\gamma(0), \ldots, \gamma(p))$ suffit pour déterminer toute la covariance. Ainsi, dans les expérimentations, lorsque l'on devra calculer les moyennes de matrices et la distance entre matrices, nous n'utiliserons uniquement les matrices de covariance suivantes :

$$\Sigma = \begin{pmatrix} \gamma(0) & \ldots & \gamma(p) \\ \vdots & \vdots & \vdots \\ \bar{\gamma}(p) & \ldots & \gamma(0) \end{pmatrix}.$$

6.3.2 Modèles CMC utilisés

Nous précisons ici le modèle que nous allons utiliser pour la segmentation. En chaque case distance-azimut (m,n), nous observons un vecteur de coefficients de réflexion $\mu^{(m,n)} = (\mu_1^{(m,n)}, \ldots, \mu_p^{(m,n)})$ ainsi que la puissance $r^{(m,n)} = \gamma^{(m,n)}(0)$. L'observation est donc la réalisation $(\mu^{(m,n)}, r^{(m,n)})_{(m,n)\in\mathcal{S}}$ d'un champ aléatoire bi-dimensionnel indexé sur $\mathcal{S} = \{1,\ldots,N_D\} \times \{1,\ldots,N_A\}$, où N_D est le nombre de distances et N_A le nombre d'azimuts considérés. Ce processus est transformé en processus mono-dimensionnel de taille $N = N_D \times N_A$ en utilisant un parcours azimut par azimut. La réalisation du processus mono-dimensionnel est donc :

$$((\mu^{(1,1)}, r^{(1,1)}), (\mu^{(2,1)}, r^{(2,1)}), \ldots, (\mu^{(N_D,1)}, r^{(N_D,1)}), (\mu^{(1,2)}, r^{(1,2)}), \ldots, (\mu^{(N_D,N_A)}, r^{(N_D,N_A)})).$$

On notera désormais $(\mu^{(n)}, r^{(n)})$ la $n^{\text{ième}}$ marginale de cette réalisation mono-dimensionnelle. Cette réalisation est celle du processus observé que l'on notera $Z = (\mu^{(n)}, r^{(n)})_{1\leq n\leq N}$. Le processus caché sera noté $X = (X_n)_{1\leq n\leq N}$ et chaque X_n prendra ses valeurs dans un ensemble fini $\mathcal{X} = \{\omega_1, \ldots, \omega_K\}$. Le modèle considéré est celui des chaînes de Markov cachées à bruit indépendant. La distribution de (X,Z) sera alors donnée par :

$$p(x_{1:N}, z_{1:N}) = p(x_1)p(z_1|x_1) \prod_{n=1}^{N-1} p(x_{n+1}|x_n)p(z_{n+1}|x_{n+1}).$$

Quant à la loi d'observation $p(z_n|x_n)$, elle sera donnée par :

$$p(z_n|x_n) = p(r^{(n)}|x_n) \times \prod_{k=1}^{p} p\left(\left|\mu_k^{(n)}\right| \mid x_n\right) p\left(\arg\left(\mu_k^{(n)}\right) \mid x_n\right),$$

où $p(r^{(n)}|x_n = \omega_j)$ est la distribution de l'inverse d'une loi gamma de paramètres a_j et b_j :

$$p(r^{(n)}|x_n = \omega_j) = \frac{1}{\Gamma(a_j)b_j^{a_j}} \frac{1}{y^{a_j+1}} \exp\left(-\frac{1}{b_j y}\right), \tag{6.16}$$

$p\left(\left|\mu_k^{(n)}\right| \mid x_n = \omega_j\right)$ est la distribution d'une loi gamma de paramètres α_j et β_j et
$p\left(\arg\left(\mu_k^{(n)}\right) \mid x_n = \omega_j\right)$ est la distribution d'une loi de Von Mises Fisher de paramètre de direction $\mu_{0,j}$ et de concentration κ_j.

6.4 Expérimentations

Nous commençons cette section par comparer notre détecteur TFAC avec le détecteur TFAC classique. Dans un second temps, nous utiliserons la segmentation obtenue par le modèle de chaînes de Markov cachées dans notre algorithme de détection. Les données traitées sont issues du radar HF Wera utilisé par la société Actimar (Brest) dans le cadre du PREI Decimall en collaboration avec Thalès et financé par la DGA. Ce radar a une résolution de 1 kilomètres par case distance et nous disposons de 120 distances. La résolution en azimut est de 0.02 radians et nous disposons de 32 azimuts consécutifs.

6.4.1 Comparaison qualitative des deux détecteurs

Pour comparer les deux détecteurs, nous n'utiliserons pas de courbes $(\mathbb{P}_{\text{fa}}, \mathbb{P}_{\text{d}})$, étant donné que pour estimer la probabilité de détection pour une probabilité de fausses alarmes égale à 10^{-6}, il faudrait plus de 1000000 cases distance-azimut et les algorithmes de détection ne pourraient fonctionner à cause du manque de mémoire vive. On va plutôt regarder le gain en terme de distance pour les cases où nous savons qu'il y a une cible. Pour les deux détecteurs, nous calculerons les spectres complexes et la densité spectrale pour 512 fréquences consécutives dans $\left[f_0 - \dfrac{1}{2T_{\text{rec}}}, f_0 + \dfrac{1}{2T_{\text{rec}}}\right]$.
Détaillons comment nous procédons pour les deux détecteurs.

- Cas du TFAC classique :
  1. soit $r(n, \theta, k)$ le signal réel reçu de la case de distance $n \in \{1, \ldots, 120\}$, d'azimut $\theta \in \{1, \ldots, 32\}$ et de fréquence $k \in \{1, \ldots, 512\}$. Calculer les moyennes $m(n, \theta, k)$ et écart type $\sigma(n, \theta, k)$ des signaux reçus des cases, dites environnantes, de distance $n' \in \{n-5, \ldots, n-10\} \cup \{n+5, \ldots, n+10\}$, d'azimut θ et de fréquence k. Puis calculer les distances $\dfrac{|r(n, \theta, k) - m(n, \theta, k)|}{\sigma(n, \theta, k)}$;
  2. calculer, pour chaque case distance-azimut, la quantité

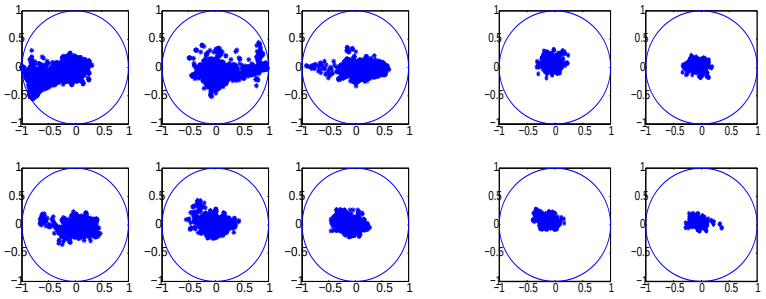
$$D_{TFAC}(n, \theta) = \max_{1 \leq k \leq 512} \frac{|r(n, \theta, k) - m(n, \theta, k))|}{\sigma(n, \theta, k)},$$

     qui sera appelée par la suite "distance". Le test de présence d'une cible s'écrit :

$$D_{TFAC}(n, \theta) > \alpha.$$

- Cas du TFAC amélioré :
  1. pour chaque case de distance n et d'azimut θ, calculer les moyennes des matrices de covariance des échantillons IQ provenant des cases, dites environnantes, d'azimut θ et de distance $n' \in \{n-5, \ldots, n-10\} \cup \{n+5, \ldots, n+10\}$;
  2. calculer ensuite la distance de Siegel $D^{\text{Siegel}}_{TFAC}(n, \theta)$ de la matrice de covariance issue de la case (n, θ, k) à la moyenne correspondante.

Afin de pouvoir comparer correctement les deux détecteurs, nous nous donnons une case de référence de distance n_0 et d'azimut θ_0. Nous divisons ensuite chaque distance $D_{TFAC}(n, \theta)$ (resp. $D^{\text{Siegel}}_{TFAC}(n, \theta)$) par $D_{TFAC}(n_0, \theta_0)$ (resp. $D^{\text{Siegel}}_{TFAC}(n_0, \theta_0)$). Nous choisissons pour case de référence celle située à 57 kilomètres du radar et ayant un azimut de 13.13° avec l'axe de visée du radar. De la figure 6.6, nous voyons qu'en utilisant le détecteur classique, de nombreuses valeurs atypiques apparaissent ce qui peut générer des fausses alarmes. Nous avons pu mettre en évidence grâce aux vérités terrain la présence de deux cibles, la première est située à une distance de 54 kilomètres et un azimut de $-9.78°$ et la deuxième est située à une distance de 53 kilomètres et un azimut de $-67.08°$. Les figures 6.7 et 6.8 représentent les distances de Siegel et les distances obtenues par le détecteur TFAC classique après normalisation pour les azimuts $-9.78°$ et $-67.08°$.



(a) Six premières composantes. (b) Quatres dernières composantes.

FIG. 6.4 – Représentation des 120×32 coefficients de réflexion dans le plan complexe.

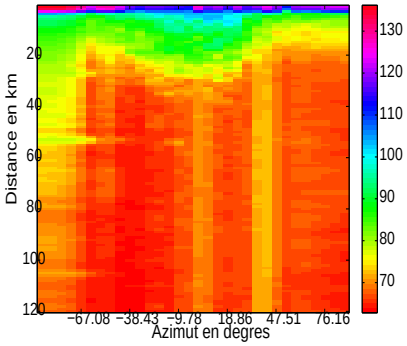
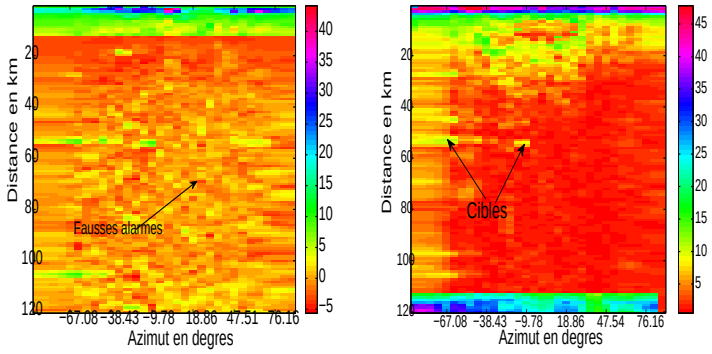
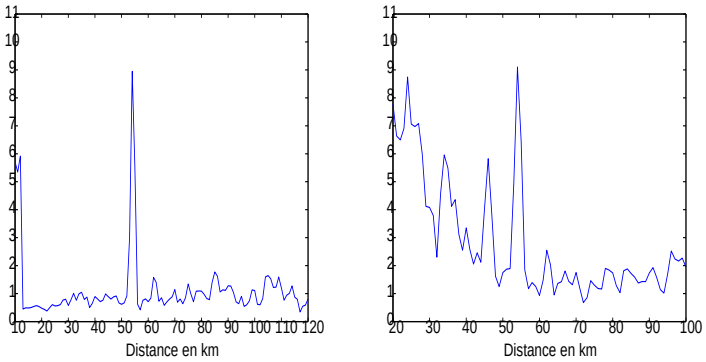


FIG. 6.5 – Intensité rétrodiffusée en chaque case distance-azimut, l'intensité est mesurée en décibels, soit $10\log(\gamma(0))$ (labels des couleurs à droite des graphiques).



(a) TFAC classique. (b) Nouveau TFAC.

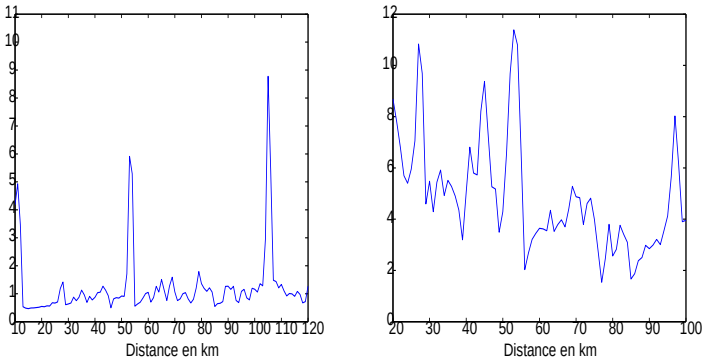
FIG. 6.6 – Cartes distance-azimut représentant les distances après normalisation obtenues par le détecteur TFAC classique et celles obtenues par le nouveau détecteur TFAC.



(a) Détecteur TFAC classique. (b) Nouveau détecteur TFAC.

FIG. 6.7 – Comparaison de la détection pour l'azimut de $-9.78°$. A gauche, on représente les distances obtenues par le TFAC classique fonction de la case distance et à droite les distances de Siegel. Les résultats sont présentés après normalisation.

Dans la figure 6.7, on voit que pour les deux types de détecteurs, les résultats sont similaires. La cible est bien détectée dans les deux cas, car le rapport signal sur bruit est élevé.
Cependant, ce qui différencie un bon détecteur d'un mauvais détecteur est qu'il parvient à détecter pour des faibles rapports signal sur bruit.



(a) Détecteur TFAC classique. (b) Nouveau détecteur TFAC.

FIG. 6.8 – Comparaison de la détection pour l'azimut de $-67.08°$.

Les résultats présentés dans la figure 6.8 sont ceux relatifs à la deuxième cible qui est plongée dans un fort bruit ambiant. On peut constater une nette amélioration en utilisant le nouveau détecteur. La distance obtenue passe de la valeur de 5.92 avec l'ancien détecteur à 11.38 avec le nouveau détecteur. Ainsi, si on avait utilisé dans la détection un seuil de 10, le nouveau détecteur aurait détecté alors que l'ancien n'aurait pas détecté la cible.

6.4.2 Détection utilisant une segmentation bayésienne

Nous proposons d'utiliser une segmentation pour affiner la détection. Plus exactement, lorsque l'on calculera la moyenne des matrices, nous n'utiliserons uniquement celles appartenant à la même classe que la case sous test. Nous comparerons la détection en utilisant notre nouveau détecteur sans segmentation et avec segmentation. Nous utiliserons le modèle de chaîne de Markov cachée à bruit indépendant décrit dans la sous-section 6.3.2. La figure 6.9 présente la segmentation de la carte distance-azimut en 4 classes. La figure 6.10 présente l'appartenance des différents coefficients de réflexion à une classe. On peut alors remarquer que les coefficients de réflexion de la classe "jaune" tendent rapidement vers 0 lorsque l'on se rapproche de l'ordre $p = 10$. Ainsi, la classe "jaune" correspond à du bruit blanc. Pour finir, nous avons représenté dans la figure 6.11 les densités de lois gamma correspondant aux densités estimées de l'inverse de la puissance pour chacune des classes. Nous y avons également superposé les histogrammes de l'inverse de la puissance en ne considérant que les cases d'appartenance à une même classe. Ceci permet de mettre en évidence le bon comportement de l'algorithme ICE utilisé pour l'estimation.

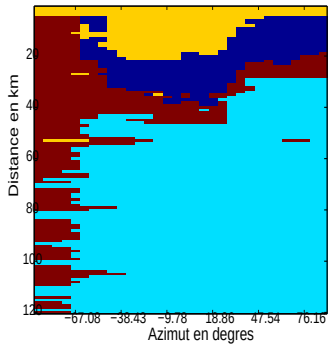
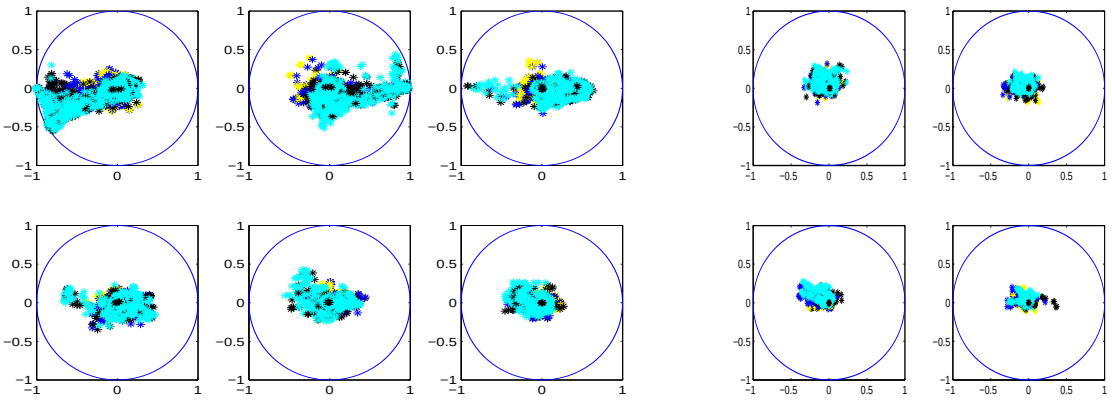


FIG. 6.9 – Segmentation de la carte distance-azimut.



(a) Six premières composantes. (b) Quatre dernières composantes.

FIG. 6.10 – Classification des différents coefficients de réflexion, la couleur “noire” correspond à la classe brune sur la segmentation.

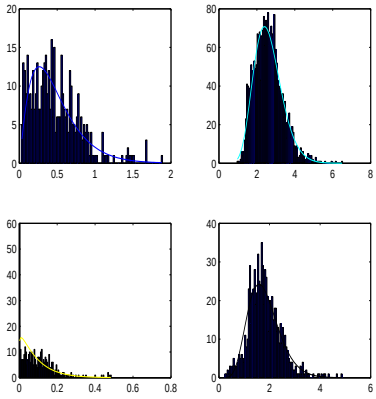
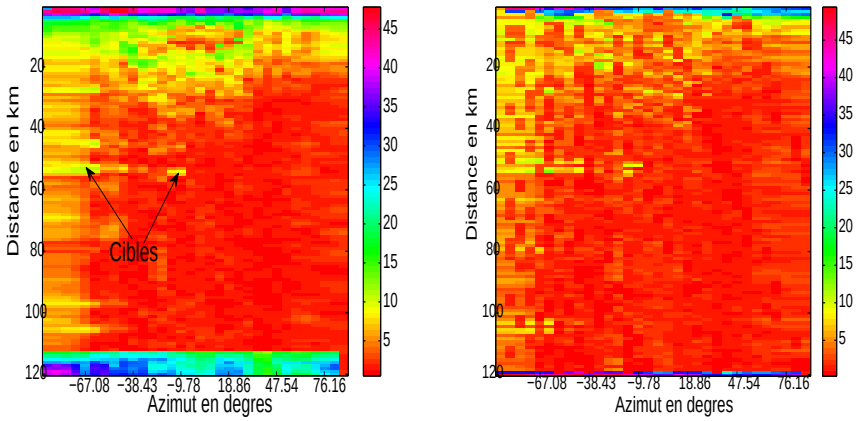
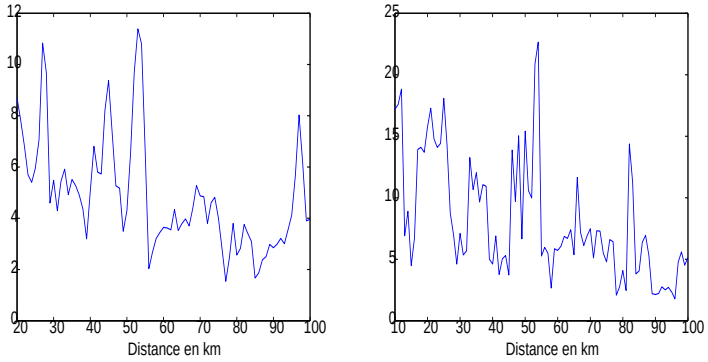


FIG. 6.11 – Comparaison des histogrammes des inverses des puissances pour chaque classe avec les densités estimées (loi gamma).



(a) Sans segmentation. (b) Avec segmentation.

FIG. 6.12 – Distances de Siegel sans et avec segmentation. Avec segmentation, la moyenne est calculée pour les cases dans la même classe que la case sous test.



(a) Sans segmentation. (b) Avec segmentation.

FIG. 6.13 – Distance de Siegel sans et avec segmentation pour l'azimut −67.08°.

De la figure 6.12 on constate une amélioration du contraste autour de la cible située à 53 kilomètres et $-67.08°$. La figure 6.13 présente la distance de Siegel pour l'azimut de $-67.08°$ sans segmentation et avec segmentation. De cette dernière figure, on constate que l'apport de la segmentation est important. D'une part, la distance de Siegel passe de 11.38 à 22.65 mais d'autre part certains maximas certainement responsables de fausses alarmes sont atténués. La segmentation permet ainsi de rendre plus robuste notre détecteur. Sans la segmentation, le signal d'une des cases environnantes pourrait avoir un comportement statistique très différent des signaux des autres cases environnantes. La moyenne des matrices de covariance serait alors non représentative de l'environnement de la case sous test et la détection en serait affectée. La segmentation permet ainsi d'éliminer du calcul de la moyenne les cases pour lesquelles la statistique du signal serait trop différente de celle de l'ambiance. Ainsi, lors de la détection, on évite de comparer la statistique du signal sous test avec celle du signal provenant d'une cible.

Conclusion

Dans ce chapitre, nous avons vu comment la segmentation bayésienne pouvait être utilisée comme un pré-traitement des données par d'autres applications. L'application concernée dans ce chapitre est celle de la détection de cible à partir d'un signal radar. Le détecteur que nous avons implémenté utilise les propriétés statistiques du signal reçu. Pour cela, nous avons défini une distance entre lois de probabilité. Cette distance a été ensuite calculée dans le cas des lois normales complexes centrées et circulaires. Parmi les perspectives à envisager, on peut étendre la distance entre matrices de covariance à celle entre lois normales complexes circulaires non centrées. On peut également envisagé le cas non gaussien en utilisant notamment les travaux de M. Calvo et J. M. Oller [26] dans lesquels il généralise au cas de lois elliptiques. Le cas des lois Γ a également été traité dans l'article [107] de F. Reverter et J. M. Oller. Parmi les autres perspectives de ces travaux, on peut également prévoir l'extension du détecteur TFAC médian classique, qui au lieu de calculer les moyennes des cases environnantes, calcule les médianes des cases environnantes. Ce détecteur est alors plus robuste aux valeurs atypiques. En effet, une valeur très grande dans un échantillon ne modifie pas la médiane. Ainsi l'idée serait d'étendre la notion de médiane pour des matrices hermitiennes définies positives en utilisant une propriété de la médiane. Concernant le modèle utilisé pour la segmentation, on peut envisager l'utilisation de lois K [17] pour l'intensité et l'utilisation de lois sur le polydisque unité $\left\{\mu \in \mathbb{C}^p : |\mu_1|^2 + \ldots + |\mu_p|^2 = 1\right\}$ au lieu d'un produit de lois de Von Mises-Fisher. L'introduction de la dépendance longue peut être envisagée également lorsque les données sont par exemple issues de radar de haute résolution, auquel cas cette dépendance longue due à la structure du fouilli de mer (pseudo-périodicité due aux vagues) ne peut être négligée.

Conclusion et perspectives

Le contexte général de notre étude était l'estimation automatique, à partir des variables observées $Y = (Y_s)_{s\in\mathcal{S}}$, des réalisations des variables inobservées $X = (X_s)_{s\in\mathcal{S}}$. Le modèle classique par chaînes de Markov cachées (CMC) est parmi les modèles les plus utilisés et les plus efficaces pour accomplir une telle estimation lorsque le nombre de variables est trop grand pour que l'on puisse utiliser la loi $p(x, y)$ du couple (X, Y) dans toute sa généralité. Cependant, ce modèle a ses limites à cause de la simplicité de la loi $p(y|x)$, qui modélise le "bruit". Ce qui le rend difficilement justifiable dans un certain nombre de situations réelles.
Les différentes contributions originales de notre mémoire concernent différentes extensions du modèle CMC. Ces différentes extensions s'appuient sur les modèles génériques que sont les "chaînes de Markov triplets" (CMT, [103]) et les "chaînes triplets partiellement de Markov" (CTPM, [97]). Concernant les CMT, leur pouvoir modélisant a commencé à être exploité récemment dans la thèse de P. Lanchantin [65]. En particulier, les CMT permettent de modéliser les CMC non stationnaires. Nous avons proposé un certain nombre de modélisations particulières se rapportant à cette problématique. Nous avons montré que les modèles classiques par chaînes semi-markoviennes cachées (CSMC) sont des CMT particulières. Dans ce contexte, nous avons proposé une CSMC originale, autorisant des temps de calcul bien moindres que ceux nécessaires à l'utilisation des CSMC classiques. Ensuite, les CSMC ont été étendues au cas où la chaîne cachée n'est pas stationnaire. Enfin, en développant les idées proposées dans la thèse de N. Brunel [21, 23], nous avons introduits dans les nouveaux modèles les copules, ce qui autorise une grande richesse de la modélisation de la loi $p(y|x)$. Pour chacun des nouveaux modèles nous avons proposé une méthode adéquate d'estimation des paramètres, de type ICE, et l'intérêt des segmentations bayésiennes non supervisées associées a été validé par des expérimentations informatiques.
Concernant les CTPM, leur potentiel modélisant a été décrit dans [97, 98] ; cependant, contrairement aux CMT, poser le modèle général ne conduit pas immédiatement aux méthodes exploitables. Afin de pouvoir traiter des bruits $p(y|x)$ à mémoire longue, nous avons proposé, en collaboration avec P. Lanchantin, un modèle particulier dans lequel la chaîne cachée est markovienne, et le bruit est gaussien. De telles "chaînes de Markov cachées par du bruit à mémoire longue" (CMC-ML) sont alors exploitables et permettent d'estimer la chaîne cachée par les méthodes bayésiennes classiques [66]. Nous avons ensuite étendu ces CMC-ML aux modèles dans lesquels seule la copule reste gaussienne, les lois marginales du bruit pouvant être quelconques. Enfin, une extension non triviale du principe général de la méthode d'estimation des paramètres ICE nous a permis de proposer des méthodes de segmentation non supervisées, dont le bon comportement a été constaté par des expérimentations. Notons qu'il y a plusieurs types de bruit à mémoire longue comme les processus FARIMA ou les bruits gaussiens fractionnaires, et les modèles proposés sont utilisables dès que l'on connaît la forme de la fonction de covariance. Nos modèles s'appliquent ainsi, de manière automatique, dans

différents problèmes de détection de changements de régime dans des phénomènes aléatoires à mémoire longue.

En guise de perspective, on peut envisager de généraliser les modèles considérés dans cette thèse à des modèles de dépendance plus complexes tels que les arbres ou autres modèles graphiques. De telles généralisations peuvent se faire facilement dans les arbres de Markov, qui peuvent être vus comme des chaînes "hiérarchisées". Notons également certains résultats déjà existant, concernant la non stationnarité ou l'utilisation des copules, dans le cadre des champs de Markov [11, 12, 13].

Annexe A

Fonctions eulériennes et fonctions de Bessel

A.1 Fonctions eulériennes Gamma et Beta

Définition A.1.1 (Fonction Gamma et Beta)**.** *La fonction* Γ *est définie sur* $\mathbb{C}^+ = \{z \in \mathbb{C} : Re(z) > 0\}$ *par :*

$$\Gamma(z) = \int_0^{+\infty} t^{z-1} \exp(-t) dt. \tag{A.1}$$

La fonction β *est définie sur* $\mathbb{C}^{+2}$ *par :*

$$\beta(z_1, z_2) = \frac{\Gamma(z_1)\Gamma(z_2)}{\Gamma(z_1 + z_2)} \tag{A.2}$$

Propriétés 1. *On a les propriétés suivantes :*

1. *Pour tout entier* $n \geq 1$, $\Gamma(n) = (n-1)!$ *et* $\Gamma(z) = (z-1)\Gamma(z-1)$.
2. *On a la formule de Stirling pour tout* $z \in \mathbb{C}^+$ *:*

$$\Gamma(z) \sim^{+\infty} \sqrt{2\pi}(z-1)^{z-\frac{1}{2}} e^{-(z-1)}.$$

3. *Pour tous réels* $a > 0$ *et* $b > 0$*, on a :*

$$\beta(a, b) = \int_0^1 t^{a-1}(1-t)^{b-1} dt.$$

4. *On a propriété asymptotique :*

$$\Gamma(z) = \lim_{n \to +\infty} \frac{n! n^z}{z(z+1)\dots(z+n)}.$$

5. Γ *se prolonge à une fonction méromorphe sur* $\mathbb{C}$ *dont les pôles sont les éléments de* $\mathbb{Z}^-$.

6. *Le produit infini de cette fonction méromorphe est donné par :*

$$\frac{1}{\Gamma(z)} = z\exp(\gamma z)\prod_{k=1}^{+\infty}\left(1+\frac{z}{k}\right)\exp\left(-\frac{z}{k}\right),$$

où $\gamma = \lim_{n\to+\infty}\sum_{k=1}^{n}\frac{1}{k} - \ln(n)$ *est la constante d'Euler.*

7. *On a* $\Gamma(z)\Gamma(1-z) = \frac{\pi}{\sin(\pi z)}$, *pour tout* $z \in \mathbb{C}^+$.

8. *Pour tout* $z \in \mathbb{C}^+$, *la formule des compléments est donnée par :*

$$\Gamma(z)\Gamma\left(z+\frac{1}{n}\right)\ldots\Gamma\left(z+\frac{n-1}{n}\right) = (2\pi)^{\frac{n-1}{2}}n^{\frac{1}{2}-nz}\Gamma(nz).$$

Corollaire A.1.1. *La fonction "digamma" est définie sur* $\mathbb{C}-\mathbb{Z}^-$ *par* $\psi(z) = \frac{\Gamma'(z)}{\Gamma(z)}$, *celle-ci est développable en série de Laurent sur* $\mathbb{C}-\mathbb{Z}^-$ *et :*

$$\psi(z) = -\frac{1}{z} - \gamma + \sum_{k=1}^{+\infty}\frac{z}{k(z+k)}. \tag{A.3}$$

A.2 Fonctions de Bessel modifiées

Les deux fonctions de Bessel modifiées de première espèce et de seconde espèce constituent une base de l'espace vectoriel des solutions d'une équation différentielle linéaire du second ordre appelée "équation de Bessel modifiée".

Définition A.2.1. *Soit* $\nu \in \mathbb{R}$, *on appelle équation de Bessel modifiée l'équation différentielle :*

$$x^2y'' + xy' - (x^2+\nu^2)y = 0, \tag{A.4}$$

où y *est une fonction de* $\mathbb{R}$ *dans* $\mathbb{R}$.

Une base de l'ensemble des solutions est donnée par les deux fonctions de Bessel modifiées :

Définition A.2.2. *On appelle fonction de Bessel modifiée de première espèce notée* I_ν *l'unique solution de (A.4) telle que* $x^{-\nu}I_\nu$ *soit développable en série entière et*

$$\lim_{x\to 0} x^{-\nu}I_\nu(x) = \frac{1}{2^\nu\Gamma(\nu+1)}.$$

Celle-ci est donnée par le développement de Laurent :

$$I_\nu(x) = \left(\frac{x}{2}\right)^\nu\sum_{p=0}^{+\infty}\frac{\left(\frac{x}{2}\right)^{2p}}{p!\Gamma(\nu+p+1)}.$$

On appelle fonction de Bessel modifiée de seconde espèce (notée K_ν*) la fonction définie par :*

$$K_\nu(x) = \frac{\pi}{2}\frac{I_{-\nu}(x) - I_\nu(x)}{sin(\pi\nu)}.$$

De plus c'est l'unique solution de l'équation de Bessel modifiée telle que :

- $\left[\dfrac{2\sin(\pi\nu)}{\pi}K_\nu + I_\nu\right] x^\nu$ *soit développable en série entière.*
- $\lim\limits_{x\to 0}\left[\dfrac{2\sin(\pi\nu)}{\pi}K_\nu(x) + I_\nu(x)\right] x^\nu = \dfrac{2^\nu}{\Gamma(1-\nu)}$.

Les deux propositions suivantes nous permettent d'écrire les fonctions de Bessel sous forme intégrale :

Proposition A.2.1. *Les deux fonctions de Bessel I_ν et K_ν s'écrivent sous les formes intégrales suivantes :*

$$I_\nu(x) = \frac{x^\nu}{2^\nu\Gamma(\nu+1)\displaystyle\int_0^\pi \sin^{2\nu}(\theta)d\theta} \times \int_0^\pi \sin^{2\nu}(\theta)\exp\left(x\cos(\theta)\right)d\theta, \tag{A.5}$$

$$K_\nu(x) = \frac{1}{2}\int_0^{+\infty} u^{\nu-1}\exp(-\frac{x}{2}(u+\frac{1}{u}))du = \frac{1}{2}\int_{\mathbb{R}} \cosh(\nu t)\exp(-x\cosh t)dt. \tag{A.6}$$

Preuve. Utiliser les caractérisations des fonctions de Bessel. □

Annexe B

Eléments de géométrie fractale

B.1 Dimension topologique, dimension de Hausdorff et ensembles fractals

Définition B.1.1. *Soient $(E, \mathcal{T}_E)$ et $(F, \mathcal{T}_F)$ deux espaces topologiques. Une application φ de E dans F est un homéomorphisme si elle est bijective, continue et si son application inverse φ^{-1} est continue.*

Définition B.1.2 (Dimension topologique)**.** *Soit $(E, \mathcal{T}_E)$ un espace topologique. Sa dimension topologique est finie et égale à d s'il existe un homéomorphisme de $\mathbb{R}^d$ dans E.*

Définition B.1.3 (Mesure et dimension de Hausdorff)**.** *Soit (E, d) un espace métrique et soit $\nu > 0$ un réel strictement positif, la mesure de Hausdorff μ_ν de paramètre ν est définie sur la tribu borélienne de E par :*

$$\mu_\nu(A) = \lim_{\epsilon \to 0} \sum_{n \in \mathbb{N}} diam(\mathcal{O}_n^\epsilon)^\nu, \tag{B.1}$$

où diam est le diamètre d'un ensemble et $(\mathcal{O}_n^\epsilon)_{n\in\mathbb{N}}$ est un recouvrement dénombrable de A par des ouverts de diamètre inférieur ou égal à ϵ.
La dimension de Hausdorff de E est définie par :

$$d_H(E) = \inf\left(\nu : \mu_\nu(E) = 0\right). \tag{B.2}$$

Remarque : Lorsque E est un $\mathbb{R}$-espace vectoriel et que $\nu = n$ est un entier, alors μ_n est la mesure de Lebesgue $\lambda_{\mathbb{R}^n}$. Ainsi, la dimension de Hausdorff généralise la dimension algébrique. De plus, si d désigne la dimension de Hausdorff de E, la mesure μ_d est invariante par translation et homogène.
On a :

Définition et proposition B.1.1 (Ensembles fractals)**.** *La dimension de Hausdorff d'un espace métrique est supérieure ou égale à la dimension topologique.*
Un ensemble est dit fractal si sa dimension de Hausdorff est strictement supérieure à sa dimension topologique.

Parmi l'ensemble des fractals on peut citer les ensembles auto-similaires pour lesquels la dimension de Hausdorff est simple à calculer.

Définition B.1.4 (Ensembles auto-similaires)**.** *Un sous-ensemble E d'un $\mathbb{R}$-espace vectoriel est auto-similaire s'il existe une sous-partie F de E et une homothétie de rapport $\lambda > 0$ telles que $\lambda F = E$.*

La section suivante présente trois exemples d'ensembles auto-similaires.

B.2 Exemples d'ensembles auto-similaires et leurs propriétés

Exemple B.2.1 (Ensemble triadique de Cantor)**.** L'ensemble triadique de Cantor noté $\mathcal{C}$ est l'ensemble des nombres de $[0, 1]$ s'écrivant en base 3 uniquement avec les chiffres 0 et 2. Posant $I_0 = [0, 1]$, I_{n+1} s'obtient à partir de I_n en enlevant le segment du milieu de chaque composante connexe de I_n.

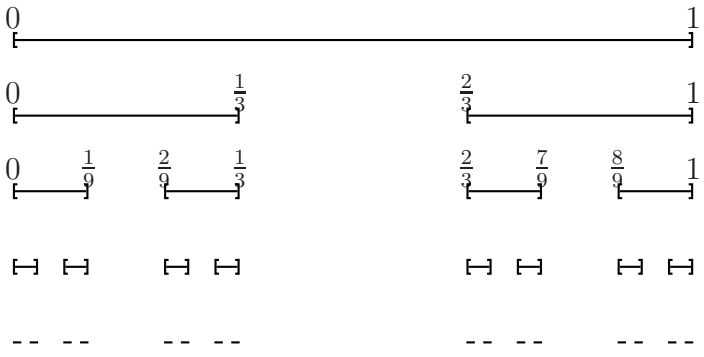


Fig. B.1 – Premières itérations de l'ensemble de Cantor.

L'image de $\mathcal{C} \cap \left[0, \frac{1}{3}\right]$ par une homothétie de rapport 3 est $\mathcal{C}$ et l'image de $\mathcal{C}$ par cette même homothétie de rapport 3 est composée de deux copies de $\mathcal{C}$. Ainsi si d_H désigne la dimension de Hausdorff de $\mathcal{C}$, on a $\mu_{d_H}(3\mathcal{C}) = 3^{d_H}\mu_{d_H}(\mathcal{C})$ par homogénéité de la mesure et $\mu_{d_H}(3\mathcal{C}) = 2\mu_{d_H}(\mathcal{C})$ par invariance par translation. La dimension de Hausdorff de l'ensemble triadique de Cantor est alors égale à $d_H = \dfrac{\log(2)}{\log(3)}$. La dimension topologique ne pouvant être qu'entière, sa dimension topologique est alors égale à 0. L'ensemble triadique de Cantor est donc fractal.

Exemple B.2.2 (Triangle de Sierpinski)**.** Le triangle de Sierpinski s'obtient à partir d'un triangle en enlevant le triangle inversé du milieu, puis on itère la méthode pour les trois triangles restant.
L'image du triangle de Sierpinski par une homothétie de rapport 2 est constituée de trois

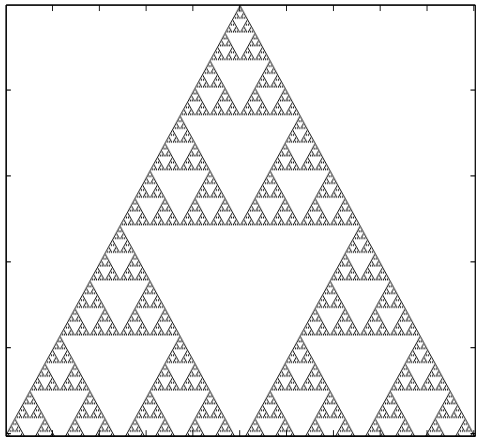
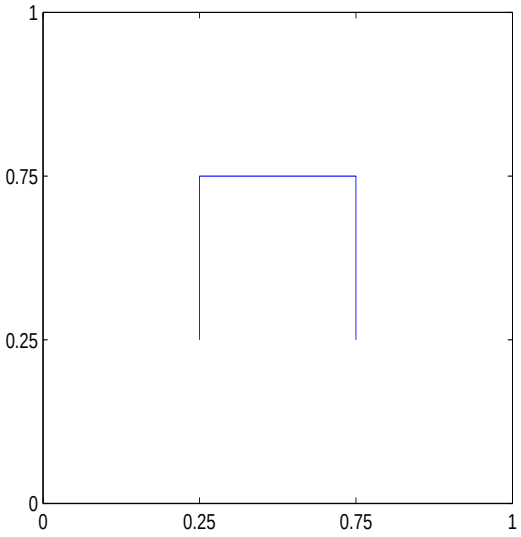


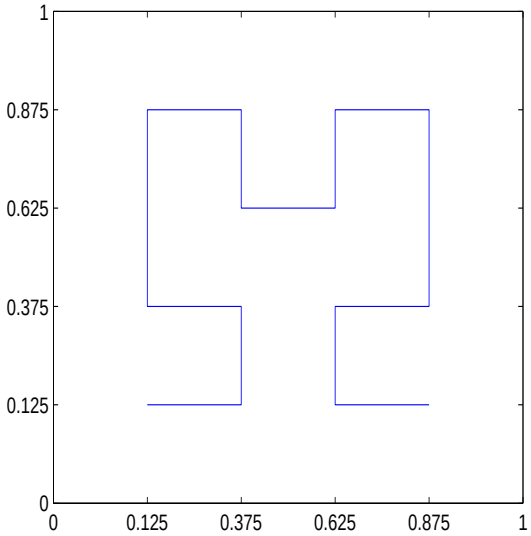
FIG. B.2 – Triangle de Sierpinski.

copies du triangle de Sierpinski. Ainsi, sa dimension de Hausdorff est égale à $d_H = \frac{\log(3)}{\log(2)}$. Sa dimension topologique est égale à 1, le triangle de Sierpinski est donc une courbe.

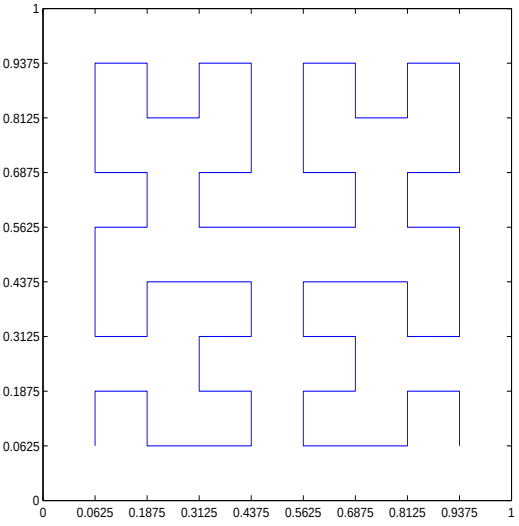
Exemple B.2.3 (Parcours d'Hilbert-Peano)**.** Le parcours d'Hilbert-Peano est la limite de la suite de courbes $\mathcal{C}_n$ obtenues de la manière suivante :



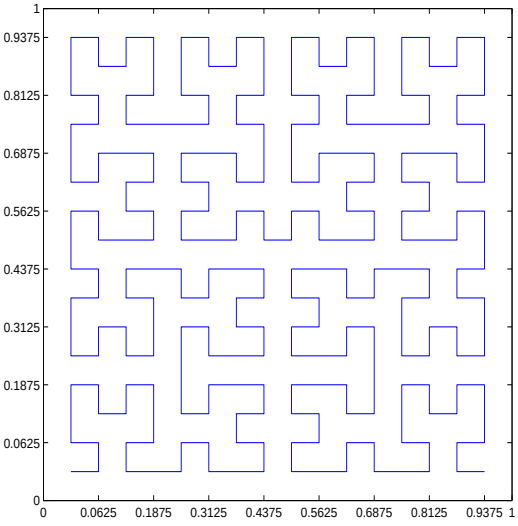
(a) $\mathcal{C}_1$



(b) $\mathcal{C}_2$



(c) $\mathcal{C}_3$



(d) $\mathcal{C}_4$

Fig. B.3 – Quatre premières itérations du parcours d'Hilbert-Peano.

Le parcours d'Hilbert-Peano étant une courbe, sa dimension topologique est égale à 1. Une homothétie de rapport 2 transforme le parcours d'Hilbert-Peano en 4 copies du parcours d'Hilbert-Peano. Sa dimension de Hausdorff est égale à $d_H = 2$. Le parcours d'Hilbert-Peano est donc une courbe remplissant tout le carré $[0, 1]^2$.

Bibliographie

[1] B. Ait-el-Fquih. *Estimation bayésienne non supervisée dans les chaînes de Markov triplets continues*. PhD thesis, Institut National des Télécommunications, Evry, France, 2007.

[2] B. Ait-el-Fquih and F. Desbouvries. Kalman filtering in triplet Markov chains. *IEEE Trans. on Signal Processing*, 54(8) :2957–2963, 2006.

[3] S.-I. Amari, O. E. Barndorff-Nielsen, R. E. Kass, S. L. Lauritzen, and C. R. Rao. *Differential geometry in statistical inference*. Monograph series, 1987.

[4] T. Ando, C.-K. Li, and R. Mathias. Geometric means. *Linear Algebra Appl.*, 385 :305–334, 2004.

[5] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra. Clustering on the Unit Hypersphere using Von Mises-Fisher Distributions. *Journal of Machine Learning Research*, 6 :1345–1382, 2005.

[6] F. Barbaresco. Recursive eigendecomposition via autoregressive analysis and ago-antagonistic regularization. Munich, Germany, 1997. ICASSP 97.

[7] F. Barbaresco. Interactions between symmetric cone and information geometries : Bruhat-Tits and Siegel spaces models for high resolution autoregressive Doppler imagery. Rome, Italy, May 2008. IEEE Radar Conference 2008.

[8] V. Barbu and N. Limnios. Maximum likelihood estimation for hidden semi-Markov models. *Comptes Rendus de l'Académie des Sciences*, 342 :201–205, 2006.

[9] M. Basseville. Information : entropies, divergences et moyennes. Technical report, IRISA, Rennes, France, May 1996.

[10] E. Bellone, J. Hugues, and P. Guttorp. A hidden Markov model for downscaling synoptic athmospheric patterns to precipitation amounts. *Climate Research*, 15(1) :1–12, 2000.

[11] D. Benboudjema and W. Pieczynski. Unsupervised image segmentation using triplet Markov fields. *Computer Vision and Image Understanding*, 99(3) :476–498, 2005.

[12] D. Benboudjema and W. Pieczynski. Unsupervised segmentation of non stationary images with non Gaussian correlated noise using triplet Markov fields and the Pearson system. Toulouse, France, May 2006. ICASSP 2006.

[13] D. Benboudjema and W. Pieczynski. Unsupervised statistical segmentation of nonstationary images using triplet Markov field. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(8) :1367–1378, 2007.

[14] B. Benmiloud and W. Pieczynski. Estimation des paramètres dans les chaînes de Markov cachées et segmentation d'images. *Traitement du Signal*, 12(5) :433–454, 1995.

[15] J. M. Bernardo and A. F. M. Smith. *Bayesian theory*. Wiley, 1994.

[16] J. Besag. On the statistical analysis of dirty picture. *Journal of the Royal Statistical Society*, 48 :259–302, 1986.

[17] N. Bon, A. Khenchaf, J-M. Quellec, M. Chabat, and R. Garello. Détection MV-TFAC dynamique dans un fouillis de mer hétérogène. Louvain-la-Neuve, France, September 2005. GRETSI 2005.

[18] J. V. Braun and H.-G. Müller. Statistical methods for DNA sequence segmentation. *Statistical Science*, 13(2) :142–162, 1998.

[19] P. J. Brockwell and R. A. Davis. *Time series : Theory and Methods*. Springer-Verlag, 1986.

[20] M. Broniatowski, G. Celeux, and J. Diebolt. Reconnaissance de mélange de densités par un algorithme d'apprentissage probabiliste. *Data Analysis and Informatics*, 3 :359–373, 1984.

[21] N. Brunel. *Sur quelques extensions des chaînes de Markov cachées et couples. Application à la segmentation non supervisée de signaux radar*. PhD thesis, Université Paris VI-INT, Paris , France, 2005.

[22] N. Brunel and W. Pieczynski. Modeling temporal dependence of Spherically Invariant Random Vectors with triplet Markov chains. Bordeaux, France, 2005. IEEE Workshop on Statistical Signal Processing.

[23] N. Brunel and W. Pieczynski. Unsupervised signal restoration using hidden Markov chains with copulas. *Signal Processing*, 85(12) :2304–2315, 2005.

[24] N. Brunel, W. Pieczynski, and F. Barbaresco. Chaînes de Markov cachées multivariées à bruit corrélé non gaussien, avec applications à la segmentation du signal radar. Louvain-la-Neuve, France, September 2005. GRETSI 2005.

[25] C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, 268 :78–94, 1997.

[26] M. Calvo and J. M. Oller. A distance between elliptical distributions based in an embedding into the Siegel group. *Journal of Computational and Applied Mathematics*, 145(2) :319–334, 2002.

[27] O. Cappé, E. Moulines, and T. Ryden. *Inference in Hidden Markov Models*. Springer-Verlag, 2005.

[28] M. Carpentier. *Radar, Bases modernes*. Masson, 1977.

[29] G. Celeux and J. Diebolt. The SEM algorithm : A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistic Quarterly*, 2 :73–82, 1985.

[30] M. Chaouachi. La détection de la mémoire longue dans la chronique de pétrole. Technical report, Université Paris I, December 2005.

[31] M. Chen, A. Kundu, and J. Zhou. Off-Line Handwritten Work Recognition Using a Hidden Markov Model Type Stochastic Network. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 16(5) :481–496, 1994.

[32] O. Chryssaphinou, M. Karaliopoulou, and N. Limnios. On Discrete Time Semi-Markov Chains and Applications in Words Occurences. *Communications in Statistics, Theory and Methods*, 37 :1306–1322, 2008.

[33] G. A. Churchill. Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology*, 51(1) :79–94, 1989.

[34] J.-P. Cocquerez and S. Phillip. *Analyse d'images : filtrage et segmentation*. Masson, 1995.

[35] P. Debajyoti. Gohberg-Semencul type formulas via embedding of Lyapunov equations. *IEEE Trans. on Signal Processing*, 41(6) :2208–2215, 1993.

[36] C. Decoux. Estimation de chaînes semi-markoviennes par échantillonnage de Gibbs. *Revue de Statistique Appliquée*, 45(2) :71–88, 1997.

[37] P. Deheuvels. La fonction de dépendance empirique et ses propriétés–un test non paramétrique d'indépendance. *Bulletin de l'Académie Royale de Belgique–Classe des sciences*, 65 :274–292, 1979.

[38] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, 39(1) :1–38, 1977.

[39] S. Derrode and W. Pieczynski. Signal and image segmentation using Pairwise Markov chains. *IEEE Trans. on Signal Processing*, 52(9) :2477–2489, 2004.

[40] P. A. Devijver. Baum's forward-backward algorithm revisited. *Pattern Recognition*, 3(6) :369–373, 1985.

[41] S. Faisan, L. Thoraval, J.-P. Armspach, M.-N. Metz-Lutz, and F. Heitz. Unsupervised learning and mapping of active brain functional MRI signals based on hidden semi-Markov event sequence models. *IEEE Trans. on Medical Imaging*, 24(2) :263–276, 2005.

[42] J. D. Ferguson. Variable duration models for speech. Princeton, New Jersey, 1980. Proceedings of the Symposium on the Applications of Hidden Markov Models to Text and Speech.

[43] J. C. Fodor and M. Roubens. *Fuzzy Preference Modelling and Multicriteria Decision Support*. Kluwer Academic Publishers, 1994.

[44] C. Fourgeaud and A. Fuchs. *Statistiques*. Dunod, 1967.

[45] M. Fréchet. Sur les tableaux de corrélation dont les marges sont données. *Ann. Univ. Lyon*, 3(1) :53–77, 1951.

[46] C. Genest and R. J. MacKay. Copules archimédiennes et familles de lois bidimensionnelles dont les marges sont données. *The Canadian Journal of Statistics*, 14(2) :145–159, 1986.

[47] J. K. Ghosh, M. Delampady, and T. Samanta. *An introduction to Bayesian analysis. Theory and methods*. Springer texts in Statistics, 2006.

[48] N. Giordana and W. Pieczynski. Estimation of generalized multisensor hidden Markov chains and unsupervised image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(5) :465–475, 1997.

[49] P. Gonçalvès and J.-B. Durand. Statistical inference for hidden Markov tree models and application to wavelet tree. Technical report, INRIA, Septembre 2001. Rapport de recherche 4248.

[50] M. Grabisch and C. Labreuche. A decade of application of the Choquet and Sugeno integrals in multi-criteria decision aid. *A Quarterly Journal of Operations Research*, 6(1) :1–44, 2008.

[51] S. Grégoir and F. Lenglart. Measuring the probability of a business cycle turning point by using a multivariate qualitative hidden Markov model. *Journal of Forecasting*, 19(2) :81–102, 2000.

[52] Y. Guédon. Estimating hidden semi-Markov chains from discrete sequences. *Journal of Computational and Graphical Statistics*, 12(3) :604–639, 2003.

[53] D. Guégan. La persistance dans les marchés financiers. *Banque et Marchés*, 90 :34–43, 2007.

[54] X. Guyon. *Champs aléatoires sur un réseau : modélisation, statistique et applications*. Masson, 1993.

[55] S. Haykin, R. Bakker, and B. W. Currie. Uncovering nonlinear dynamics : the case of sea clutter. *Proceedings of the IEEE, Special issue on Applications of Nonlinear Dynamics to Electronic and Information Engineering*, 90(5) :860–881, 2002.

[56] M. Henry and P. Zaffaroni. *Theory and applications of long-range dependence*, chapter The long-range dependence paradigm for macroeconomics and finance. Birkhäuser, 2002.

[57] C. C. Heyde. *Quasi-likelihood and its application : a general approach to optimal parameter estimation*. Springer-Verlag, 1997.

[58] G. L. Jones. On the Markov chain central limit theorem. *Probability Surveys*, 1 :299–320, 2004.

[59] M. I. Jordan. *Learning in graphical models*. Kluwer Academic Publishers, 1998.

[60] M. I. Jordan and F. R. Bach. Analyse en composantes indépendantes et réseaux bayésiens. Paris, France, September 2003. GRETSI 2003.

[61] I. Kaj, Lasse Leskelä, I. Norros, and V. Schmidt. Scaling limits for random fields with long-range dependence. *Annals of Probability*, 35(2) :528–550, 2007.

[62] H. Karcher. Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics*, 30(5) :509–541, 1977.

[63] E. P. Klement, R. Mesiar, and E. Pap. Measure-based aggregation operators. *Fuzzy Sets and Systems*, 142(1) :3–14, 2004.

[64] T. Koski. *Hidden Markov models for bioinformatics*. Kluwer Academic Publishers, 2001.

[65] P. Lanchantin. *Chaînes de Markov triplets et segmentation non supervisée de signaux*. PhD thesis, Institut National des Télécommunications, Evry, France, 2006.

[66] P. Lanchantin, J. Lapuyade-Lahorgue, and W. Pieczynski. Unsupervised segmentation of triplet Markov chains hidden with long memory noise. *Signal Processing*, 88(5) :1134–1151, 2008.

[67] P. Lanchantin and W. Pieczynski. Chaînes et arbres de Markov évidentiels avec applications à la segmentation des processus non stationnaires. *Traitement du Signal*, 22(1) :15–26, 2005.

[68] J. Lapuyade-Lahorgue and W. Pieczynski. Unsupervised segmentation of hidden semi-Markov non stationary chains. Paris, France, July 2006. Twenty sixth International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MaxEnt 2006).

[69] J. Lapuyade-Lahorgue and W. Pieczynski. Partially Markov models and unsupervised segmentation of semi-Markov chains hidden with long dependence noise. Chania, Crete, Greece, May 2007. International Symposium on Applied Stochastic Models and Data Analysis (ASMDA 2007).

[70] S. L. Lauritzen. *Graphical models*. Oxford Science Publications, 1996.

[71] F. Le Ber, M. Benoît, C. Scott, J.-F. Mari, and C. Mignolet. Studying crop sequences with Carrotage, a HMM-based data mixing software. *Ecological Modelling*, 191(1) :170–185, 2006.

[72] F. Le Chevalier. *Principles of Radar and Sonar Signal Processing*. Artech House, 2003.

[73] S. E. Levinson. Continuously variable duration hidden Markov models for automatic speech recognition. *Computer Speech and Language*, 1(1) :29–45, 1986.

[74] N. Limnios. Estimation of the Stationary Distribution of Semi-Markov Processes with Borel State Space. *Statistics and Probability Letters*, 76(14) :1536–1542, 2006.

[75] N. Limnios and G. Oprişan. *Semi-Markov Processes and Reliability*. Birkhäuser, 2001.

[76] C. H. Ling. Representation of associative functions. *Publicationes Mathematicae Debrecen*, 12 :189–212, 1965.

[77] H. Maître. *Traitement des images de Radar à Synthèse d'Ouverture*. Hermes, 2001.

[78] B. B. Mandelbrot and J. W. Van Ness. Fractional brownian motions, fractional noises and applications. *Revue of the Society for Industrial and Applied Mathematics*, 10(4) :422–437, 1968.

[79] G. McLachlan and T. Krishnan. *The EM algorithm and extensions*. Wiley, 1997.

[80] K. Menger. Statistical metrics. *Proceedings of the National Acadademy of Sciences*, 28 :535–537, 1942.

[81] E. Monfrini, J. Lecomte, F. Desbouvries, and W. Pieczynski. Image and signal restoration using pairwise Markov trees. Saint-Louis, Missouri, USA, 2003. IEEE Workshop on Statistical Signal Processing 2003.

[82] A. Montarani. *Theory and Applications of Long-Range Dependence*, chapter Long-range dependence in hydrology. Birkhäuser, 2002.

[83] M. D. Moore and M. I. Savic. Speech reconstruction using a generalized HSMM (GHSMM). *Digital Signal Processing*, 14(1) :37–53, 2004.

[84] E. Moulines and P. Soulier. *Theory and applications of long-range dependence*, chapter Semiparametric spectral estimation for fractional processes. Birkhäuser, 2002.

[85] R. B. Nelsen. *An introduction to Copulas*. Springer-Verlag, 1999.

[86] P. Nicolas, L. Bize, F. Muri-Majoube, M. Hoebeke, F. Rodolphe, S. Dusko, Ehrlich, B. Prum, and P. Bessières. Mining Bacillus subtilis chromosome heterogeneities using hidden Markov models. *Nucleic Acid Research*, 30(6) :1418–1426, 2002.

[87] I. Norros. *Theory and applications of long-range dependence*, chapter Large deviations of queues with long-range dependent input. Birkhäuser, 2002.

[88] K. Nour, R. David, and C. Raffalli. *Introduction à la logique–théorie de la démonstration*. Dunod, 2001.

[89] G. Nuel and B. Prum. *Analyse statistique des séquences biologiques : modélisation markovienne, alignements et motifs*. Hermes, 2007.

[90] M. Oudin. *Etude d'algorithmes de traitement d'antenne sur signaux large bande et signaux bande étroite à antenne tournante*. PhD thesis, Université Paris VI-INT, Paris, France, 2008.

[91] M. Parizeau and G. Lorette. A comparative analysis of regional correlation, dynamic time warping and skeletal tree matching for signature verification. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 12(7) :710–717, 1990.

[92] K. S. Pedersen and M. Nielsen. The Hausdorf dimension and scale-space normalisation of natural images. *Lecture notes in computer science*, 1682 :271–282, 1999.

[93] D. Petz. Means of positive matrices : Geometry and a conjecture. *Annales Mathematicae et Informaticae*, (32) :129–139, 2005.

[94] W. Pieczynski. Arbres de Markov couples. *Comptes Rendus de l'Académie des Sciences*, 335(1) :79–82, 2002.

[95] W. Pieczynski. Chaînes de Markov triplet. *Comptes Rendus de l'Académie des Sciences*, 335(3) :275–278, 2002.

[96] W. Pieczynski. Pairwise Markov Chains. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(5) :634–639, 2003.

[97] W. Pieczynski. Triplet partially Markov chains and trees. Brest, France, July 2004. 2^{nd} International Symposium on Image/Video Communications over fixed and mobile networks (ISIVC'04).

[98] W. Pieczynski. Copules gaussiennes dans les chaînes triplets partiellement de Markov. *Comptes Rendus de l'Académie des Sciences*, 341(3) :189–194, 2005.

[99] W. Pieczynski. Multisensor triplet Markov chains and theory of evidence. *Journal of Approximate Reasoning*, 45(1) :1–16, 2007.

[100] W. Pieczynski. *Problèmes inverses en imagerie et en vision*, chapter Chaînes de Markov triplets et segmentation d'images. Hermes, 2008.

[101] W. Pieczynski. Sur la convergence de l'estimation conditionnelle itérative. *Comptes Rendus de l'Académie des Sciences*, 346(7) :457–460, 2008.

[102] W. Pieczynski and F. Desbouvries. On triplet Markov chains. Brest, France, May 2005. International Symposium on Applied Stochastic Models and Data Analysis (ASMDA 2005).

[103] W. Pieczynski, C. Hulard, and T. Veit. Triplet Markov Chains in hidden signal restoration. Crete, Greece, September 2002. SPIE's International Symposium on Remote Sensing.

[104] W. Pieczynski and A.-N. Tebbache. Champs aléatoires de Markov couple et segmentation des images texturées. Paris, France, February 2000. Actes du Congrès Reconnaissance des Formes et Intelligence Artificielle.

[105] M. Rangaswamy, D. Weiner, and A. Öztürk. Non-gaussian random vector identification using spherically invariant random process. *IEEE Trans. on Aerospace and Electronic Systems*, 29(1) :111–124, 1993.

[106] C. Raphael. Automatic segmentation of acoustic musical signals using hidden Markov models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(4) :360–370, 1999.

[107] F. Reverter and J. M. Oller. Computing the Rao distance for Gamma distributions. *Journal of Computational and Applied Mathematics*, 157(1) :155–167, 2003.

[108] M. J. Russel and R. K. Moore. Explicit modelling of state occupancy in hidden Markov models for automatic speech recognition. Tampa, Florida, USA, 1985. Proceedings of the International Conference on Acoustics, Speech and Signal Processing.

[109] Y. Sato and K. Kogure. Online signature verification based on shape, motion and writing. Munich, Germany, October 1982. IEEE International Conference on Pattern Recognition.

[110] B. Schweizer and A. Sklar. *Probabilistic metric Spaces*. Elsevier Science Publishing Company, 1983.

[111] C. L. Siegel. *Symplectic Geometry*. Academic Press, New York, 1964.

[112] A. Sklar. Fonction de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistiques de Paris*, 8 :229–231, 1959.

[113] M. S. Taqqu. *Theory and applications of long-range dependence*, chapter Fractional Brownian motion and long-range dependence. Birkhäuser, 2002.

[114] G. Teyssière and A. P. Kirman. *Long Memory in Economics*. Springer, 2007.

[115] L. Thomas, D. Allen, and N. Morkel-Kingsbury. A hidden Markov chain model for the term structure of bond credit risk spreads. *International Review of Financial Analysis*, 11(3) :311–329, 2002.

[116] R. L. Tweedie and S. P. Meyne. *Markov chains and stochastic stability*. Springer, 1993.

[117] M. Vrac. *Analyse et modélisation de données probabilistes par décomposition de mélange de copules et application à une base de données climatologiques*. PhD thesis, Université Paris IX-Dauphine, Paris, France, 2002.

[118] S. Watts. The performance of cell-averaging CFAR systems in sea clutter. IEEE Radar Conference 2000, July 2000.

[119] S. Watts. Radar clutter and CFAR detection. Technical report, Thales Airborne Systems UK, 2004.

[120] W. Willinger, V. Paxson, R. H. Riedi, and M. S. Taqqu. *Theory and Applications of Long-Range Dependence*, chapter Long-range dependence and data network traffic. Birkhäuser, 2002.

[121] L. Yang, B. K. Widjaja, and R. Prasad. Application of hidden Markov models for signature verification. *Pattern Recognition*, 28(2) :161–170, 1995.

Résumé

L'objectif de cette thèse est de proposer différents modèles généralisant le modèle classique des chaînes de Markov cachées à bruit indépendant couramment utilisé en inférence bayésienne de signaux. Les diverses extensions de ce modèle visent à l'enrichir et à prendre en compte différentes propriétés du signal, comme le caractère non gaussien du bruit, ou la nature semi-markovienne du signal caché. Dans un problème d'inférence bayésienne, nous disposons de deux processus aléatoires X et Y, on observe la réalisation y de Y et nous estimons la réalisation cachée x de X. Le lien existant entre les deux processus est modélisé par la distribution de probabilité $p(x, y)$. Dans le modèle classique des chaînes de Markov cachées à bruit indépendant, la distribution $p(x)$ est celle d'une chaîne de Markov et la distribution $p(y|x)$ est celle de marginales indépendantes conditionnellement à x. Bien que ce modèle puisse être utilisé dans de nombreuses applications, il ne parvient pas à modéliser toutes les situations de dépendance. Le premier modèle que nous proposons est de type "chaînes de Markov triplet", on considère ainsi un troisième processus U tel que le triplet (X, U, Y) soit une chaîne de Markov. Dans le modèle proposé, ce processus auxiliaire modélise la semi-markovianité de X ; on parvient ainsi à prendre en compte la non markovianité éventuelle du processus caché. Dans un deuxième modèle, nous considérons des observations à dépendance longue et nous proposons un algorithme d'estimation original des paramètres de ce modèle. Nous étudions par ailleurs différents modèles prenant en compte simultanément la semi-markovianité des données cachées, la dépendance longue dans les observations ou la non stationnarité des données cachées. Enfin, la nature non nécessairement gaussienne du bruit est prise en compte via l'introduction des copules. L'intérêt des différents modèles proposés est également validé au travers d'expérimentations.
Dans la dernière partie de cette thèse, nous étudions également comment la segmentation obtenue par une méthode bayésienne peut être utilisée dans la détection de cibles dans le signal radar. Le détecteur original que nous implémentons utilise la différence de statistiques entre un signal reçu et les signaux reçus de son voisinage. Le détecteur ainsi implémenté s'avère donner de meilleurs résultats en présence de fort bruit que le détecteur habituellement utilisé en traitement radar.

Mots clés Inférence bayésienne, chaînes de Markov cachées, chaînes de Markov couples et triplets, chaînes semi-markoviennes, dépendance longue, copules, Espérance Maximisation (EM), Estimation Conditionnelle Itérative (ICE), distance de Rao, mesures de Jeffreys, information de Kullback, entropie de Shannon, détection à Taux de Fausses Alarmes Constant (TFAC).

$\star$

Abstract

The objective of this thesis is to propose different - more general - models than the classical hidden Markov chain, which is often used in Bayesian inference of signals or images. The different extensions of this model aim to take into account different properties of the signal, such as its non gaussian behaviour or semi-markovianity of the hidden process. In a Bayesian inference context, we have two random processes X and Y, where the realisation y of Y is observed and the realisation x of X has to be estimated. The relationship between the two

processes X and Y is modeled by the probability distribution $p(x, y)$. In the classical model of hidden Markov chains with independent noise, the distribution $p(x)$ is that of a Markov chain and the distribution $p(y|x)$ is that of a vector whose marginal distributions are independent conditionally to x. Although this model can be used in several applications, it fails to model all situations of dependence. The first model we propose belongs to the Markov triplet models, in which we consider a third process U such that the triplet (X, U, Y) is a Markov chain. In the proposed model, this auxiliary process models the semi-markovianity of X, and so we are able to take into account the possible non-markovianity of the hidden process. In a second model, we consider long dependence within the observations and we propose an original algorithm to estimate the parameters of this model. We also study different models taking simultaneously into account semi-markovianity of the hidden data, long dependence and non-stationarity. Finally, the non-Gaussian properties of noise are taken into account through the introduction of copulas. The viability of different models is also confirmed through experimentation.
In the last part of this thesis, we are studying how the segmentation obtained by a Bayesian method can be used in detecting targets in the radar signal. The original detector that we implement uses statistical difference between a signal and the signals received from its surroundings. To achieve this, we define a distance between distributions that is used in detection. Through a series of tests, we show how the new detector thus implemented produces better results than the classical detector in a very noisy environment.

Key words Bayesian inference, hidden Markov chains, pairwise and triplet Markov chains, semi-Markov chains, long dependence, copulas, Expectation Maximisation (EM), Iterative Conditional Estimation (ICE), Rao metric, Jeffreys' measure, Kullback information, Shannon entropy, Constant False Alarm Ratio detection (CFAR).