



Sur quelques extensions des chaînes de Markov cachées et couples. Applications à la segmentation non-supervisée de signaux radar.

Nicolas Brunel

► To cite this version:

Nicolas Brunel. Sur quelques extensions des chaînes de Markov cachées et couples. Applications à la segmentation non-supervisée de signaux radar.. Mathématiques [math]. Université Pierre et Marie Curie - Paris VI, 2005. Français. NNT: . tel-00011302

HAL Id: tel-00011302

<https://theses.hal.science/tel-00011302>

Submitted on 5 Jan 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



THÈSE DE DOCTORAT DE L'UNIVERSITÉ PARIS 6

Spécialité : MATHÉMATIQUES APPLIQUÉES

PRÉSENTÉE PAR

NICOLAS BRUNEL

EN VUE DE L'OBTENTION DU TITRE DE DOCTEUR DE L'UNIVERSITÉ PARIS 6

SUJET :
SUR QUELQUES EXTENSIONS DES CHAÎNES DE MARKOV CACHÉES ET COUPLES.
APPLICATIONS À LA SEGMENTATION NON-SUPERVISÉE DE SIGNAUX RADAR.

soutenue le 5 décembre 2005

devant le jury composé de :

<i>Rapporteurs</i>	Eric MOULINES	ENST
	Bernard PRUM	Université d'Evry
<i>Directeurs de thèse</i>	Paul DEHEUVELS	Université Paris 6
	Wojciech PIECZYNSKI	INT
<i>Examineurs</i>	Alain HILLION	ENST Bretagne
	Daniel PIERRE-LOTI-VIAUD	Université Paris 6
<i>Invité, co-encadrant</i>	Frédéric BARBARESCO	THALES Air Defence

Table des matières

1	INTRODUCTION	11
2	Segmentation bayésienne d’images et de signaux	17
2.1	Estimation bayésienne	17
2.1.1	Modélisation probabiliste des images	17
2.1.2	Modèles markoviens	19
2.2	Modèles de Markov Couples	22
2.2.1	Définition et propriété fondamentale	22
2.2.2	Chaînes de Markov Couples	23
2.2.3	Calcul des probabilités a posteriori et des estimateurs bayésiens	29
2.3	Modèles triplets	31
3	Estimation statistique des CMCa	35
3.1	Introduction	35
3.2	Maximum de vraisemblance dans les CMCa	37
3.2.1	Le modèle CMCa-BI	37
3.2.2	Les hypothèses de Leroux	38
3.2.3	D’autres résultats de consistance pour les CMCa	40
3.3	Détermination du nombre de classes	41
3.3.1	Critère AIC	42
3.3.2	Critère BIC	43
3.3.3	Vraisemblance pénalisée	43
3.4	Détermination de l’EMV	45
3.4.1	L’algorithme EM	45
3.4.2	Application à l’estimation des mélanges finis	47
3.4.3	Algorithme <i>Stochastic</i> EM	48
3.5	Fonction estimante et Estimation Conditionnelle Itérative	49
3.5.1	Fonction estimante et données manquantes	50
3.5.2	Estimation Conditionnelle Itérative	57
3.5.3	Quelques algorithmes ECI	60
4	Modèles CMCa-BI multivariées	65
4.1	Mélange de lois exponentielles	66
4.1.1	Définition	66

4.1.2	Estimation par EM de modèles exponentiels	67
4.2	Mélange de lois elliptiques	69
4.2.1	Lois elliptiques et Vecteurs Aléatoires Sphériquement Invariants	70
4.2.2	Liens entre loi d'un SIRV et loi de la texture associée	72
4.2.3	Estimation par Maximum de vraisemblance	75
4.2.4	Estimation de mélanges de lois elliptiques	79
4.2.5	Expérimentations	80
4.3	Les copules	84
4.3.1	Définition et propriétés	85
4.3.2	Inférence d'une copule	90
4.3.3	Inférence d'un mélange de copules	92
4.3.4	Expérimentations	94
5	Chaînes de Markov Couples et Copules	105
5.1	Segmentation des processus stationnaires	105
5.1.1	Processus stationnaires et copules	105
5.1.2	Le modèle CMCa stationnaire avec copules	107
5.1.3	Liens avec les processus autorégressifs à changements de régimes markovien	111
5.2	Copule et dépendance temporelle	112
5.2.1	Dépendance dans les CMCa-BI et CMCa	112
5.2.2	Quelques remarques sur la dépendance des processus stationnaires	114
5.3	Algorithme d'estimation	116
5.3.1	Estimation des chaînes de Markov avec copules	117
5.3.2	Estimation des CMCa avec copules	117
5.3.3	Expérimentations	118
5.4	Chaînes de Markov couples	131
5.4.1	Le modèle CMCo stationnaire	132
5.4.2	Estimation des modèles couples	132
5.4.3	Expérimentations	134
5.4.4	Discussion sur la complexité des modèles	136
6	Segmentation de l'environnement radar	139
6.1	Le signal radar	140
6.1.1	Principes et objectifs du radar	140
6.1.2	Chaîne de traitement d'un radar	142
6.1.3	Segmentation des fouillis	147
6.2	Segmentation Doppler	147
6.2.1	Information Doppler et modélisation autorégressive	147
6.2.2	Segmentation bayésienne et cartographie Doppler	149
6.2.3	Le modèle statistique paramétrique utilisé	150
6.2.4	Exemples	151
6.3	Segmentation polarimétrique	161
6.3.1	Principe physique	161
6.3.2	La segmentation polarimétrique	165

6.3.3 Exemple	165
7 CONCLUSION	169
A Vecteurs Gaussiens Complexes	173
A.1 Généralités	173
A.2 Vecteurs gaussiens et sphériquement invariants	174
B Processus stationnaires et coefficients de réflexion	177
B.1 Prédiction linéaire	177
B.2 Relations de Passage	178
B.2.1 Prédiction linéaire et autocovariance : $R_p \leftrightarrow A_p$	178
B.2.2 Prédiction linéaire et coefficients de réflexion : $A_p \leftrightarrow \mu$	178
B.2.3 Factorisation de Choleski de $R_p^{-1} : \mu \leftrightarrow R_p$	179
B.3 Estimation des Coefficients de Réflexion	180
B.3.1 Algorithme de Burg	180
B.3.2 Algorithme de Burg régularisé	181
C Chaînes de Markov	183
C.1 Quelques notations	183
C.2 Définitions et propriétés	183
C.3 Propriétés de mélanges des chaînes de Markov	186
Bibliographie	189

Notations

Symboles

$\mathcal{Y}$	Espace des observations (espace topologique)
$\mathcal{X}$	Espace des états cachés (espace topologique)
$\mathcal{U}$	Espace du processus auxiliaire (espace topologique)
$\mathcal{B}(\mathcal{Y}), \mathcal{B}(\mathcal{X}), \mathcal{B}(\mathcal{U})$	Tribu des boréliens sur $\mathcal{Y}, \mathcal{X}, \mathcal{U}$
$(\Omega, \mathcal{A}, P)$	Espace probabilisé
$\mathbb{S}^{p-1}$	Sphère unité de $\mathbb{R}^p$
$\mathbb{CS}^{p-1}$	Sphère unité de $\mathbb{C}^p$
Y	Variable aléatoire de $(\Omega, \mathcal{A}, P)$ dans $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$
X	Variable aléatoire de $(\Omega, \mathcal{A}, P)$ dans $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$
U	Variable aléatoire de $(\Omega, \mathcal{A}, P)$ dans $(\mathcal{U}, \mathcal{B}(\mathcal{U}))$
μ	Mesure sur $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$
λ	Mesure de Lebesgue sur $\mathbb{R}^d$, $d \geq 1$
N	Nombre d'observations
K	Nombre de classes
S	Ensemble d'indices
$ S $	Cardinal de l'ensemble S
$\mathbf{Y}_S$	Champ markovien observé
$\mathbf{Y}_A$	$(Y_s)_{s \in A}$, lorsque $A \subset S$
$\mathbf{X}_S$	Champ markovien caché
$\mathbf{Y}_n$	Chaîne observée de longueur $n : (Y_i)_{1 \leq i \leq n}$
$\mathbf{X}_n$	Chaîne cachée de longueur $n : (X_i)_{1 \leq i \leq n}$
$\mathbf{U}_n$	Chaîne auxiliaire de longueur $n : (U_i)_{1 \leq i \leq n}$
$\mathbf{y}$	Réalisation de $\mathbf{Y}_S$ (ou $\mathbf{Y}_N$)
$\mathbf{x}$	Réalisation de $\mathbf{X}_S$ (ou $\mathbf{X}_N$)
$\mathbf{y}_n$	Réalisation de $\mathbf{Y}_n : (y_i)_{1 \leq i \leq n}$
$\mathbf{y}_{p:n}$	Réalisation de $\mathbf{Y}_{p:n} : (y_i)_{p \leq i \leq n}$

$p(\cdot)$	Densité d'une variable aléatoire relativement à une mesure de référence
$p(\cdot \cdot)$	Densité conditionnelle d'une variable aléatoire relativement à une mesure de référence
α_n	Probabilité avant pour $1 \leq n \leq N$
β_n	Probabilité arrière pour $1 \leq n \leq N$
ψ_n	Probabilité marginale a posteriori pour $1 \leq n \leq N$
$\mathcal{P}$	Système de Pearson
Θ	Espace de paramètres
$f(y, \theta)$	Densité paramétrique pour les lois d'émission, avec $\theta \in \Theta$
$F(y, \theta)$	Fonction de répartition de la densité $f(y, \theta)$
$\{p(\cdot, \theta), \theta \in \Theta\}$	Modèle paramétrique pour une densité
$N(m, \Sigma)$	Loi normale réelle de moyenne m et de variance Σ
$CN(m, \Sigma)$	Loi normale complexe circulaire de moyenne m et de variance Σ
$\gamma(a, b)$	Loi gamma
$i\gamma(a, b)$	Loi inverse gamma
$Tr(A)$	Trace de la matrice $A = (a_{ij})_{1 \leq i, j \leq p}$
C	Copule
c	Densité de copule
η	Paramètre de copule
A	Matrice de transition de taille $K \times K$
$ M $	Déterminant de la matrice carrée M
ϕ	Paramètre d'un modèle complet (ou d'une CMCo)
L_c	Log-vraisemblance d'un modèle complet
$\mathbf{V}$	Données complètes
p_c	Densité des observations du modèle complet
$\mathbf{U}$	Données incomplètes
L	Log-vraisemblance d'un modèle incomplet
$D(P, Q)$	Divergence de Kullback-Leibler entre deux mesures de probabilités ($D(p, q)$: divergence entre deux densités p et q)
$\text{cov}(X, Y)$	Covariance entre les variables aléatoires X et Y
$\text{cov}_\phi(X, Y)$	Covariance entre les variables aléatoires X et Y lorsque la loi jointe est indexée par le paramètre ϕ
$\alpha(\mathcal{F}, \mathcal{G})$	Coefficient de mélangeance forte entre les tribus $\mathcal{F}$ et $\mathcal{G}$ (où $\mathcal{F}, \mathcal{G} \subset \mathcal{A}$, avec $(\Omega, \mathcal{A}, P)$ espace probabilisé)
$\beta(\mathcal{F}, \mathcal{G})$	Coefficient de β -mélangeance entre les tribus $\mathcal{F}$ et $\mathcal{G}$ (où $\mathcal{F}, \mathcal{G} \subset \mathcal{A}$, avec $(\Omega, \mathcal{A}, P)$ espace probabilisé)
$\phi(\mathcal{F}, \mathcal{G})$	Coefficient de ϕ -mélangeance forte entre les tribus $\mathcal{F}$ et $\mathcal{G}$ (sous-tribus de $\mathcal{A}$, où $(\Omega, \mathcal{A}, P)$ est un espace probabilisé)
$\sigma(U)$	Tribu engendrée par la variable aléatoire U (définie sur l'espace probabilisée $(\Omega, \mathcal{A}, P)$)

Acronymes

v.a.	Variable aléatoire
f.d.r	Fonction de répartition
i.i.d	Indépendant et identiquement distribué
AIC	An Information Criterion (critère d'Akaike)
BIC	Bayesian Information Criterion
ACI	Analyse en Composantes Indépendantes
TFD	Transformée de Fourier Discrète

Modèles

MMCa	Modèle de Markov Caché
MMCa-BI	Modèle de Markov Caché à Bruit Indépendant
CMCa	Chaîne de Markov Caché
CMCa-BI	Chaîne de Markov Caché à Bruit Indépendant
CMCo	Chaîne de Markov Couple
SIRV	Vecteur Aléatoire Sphériquement Invariant (Spherically Invariant Random Vector)

Chapitre 1

INTRODUCTION

La segmentation d'images est le traitement qui consiste à rassembler les pixels d'une image en zones homogènes en utilisant le contenu de ceux-ci. Le résultat d'une segmentation est alors une carte qui constitue un résumé de l'image initiale et qui en facilite l'analyse, bien qu'elle ne corresponde pas toujours à une réalité physique. Cette notion d'homogénéité dépend bien sûr des critères qui ont été choisis pour la construction de la carte, et il existe essentiellement deux familles de méthodes de segmentation [42] : les méthodes par recherche de frontières et les méthodes par recherche de régions.

Les méthodes par recherche de frontières déterminent les contours des zones homogènes en se basant sur les fluctuations de l'intensité de l'image. Par exemple, l'utilisation d'équations aux dérivées partielles permet de faire évoluer ces contours de sorte à déterminer les changements brusques du gradient et donc les zones. Dans le cas de la morphologie mathématique, le concept de ligne de partage des eaux, définie à partir d'une distance, permet de déterminer directement les frontières des zones recherchées.

De manière duale, l'approche par recherche de régions consiste à regrouper des pixels présentant des caractéristiques proches, et à leur attribuer une étiquette. Il s'agit de l'angle d'attaque choisi dans cette thèse. Plus précisément, notre cadre de travail est le cadre probabiliste proposé par Geman et Geman [73] et Besag [20]. L'idée fondamentale de cette approche est de considérer la segmentation d'une image comme l'estimation bayésienne d'un processus aléatoire (à valeurs discrètes) à partir d'un processus observé. La segmentation est alors dans ce contexte un problème de classification dont une particularité est l'importance de la dépendance spatiale des observations. Celle-ci est prise en compte à travers le concept clé de modèle de Markov caché, dans lequel le processus caché est supposé être un champ de Markov : c'est lui qui est la cause de la dépendance observée des pixels. Cette approche s'est révélée très féconde, et a engendré de nombreux développements dont un panorama est dressé dans les récentes monographies de Chalmond [38] et Winkler [156]. Cette problématique est formellement équivalente à celle du lissage en traitement du signal ou en automatique [4], et pour lequel on cherche à estimer une chaîne de Markov non-observée à partir d'un signal bruité. Dans cette optique, les signaux peuvent être considérés comme des images unidimensionnelles mais pour lesquelles leur structure orientée particulière permet le développement de méthodes spécifiques.

Les idées qui ont guidé nos travaux peuvent être alors développées soit dans le contexte du

traitement d'image, soit dans celui du traitement du signal. Dans la thèse, nous nous sommes intéressés à la segmentation de signaux (unidimensionnels) en raison des données que nous avons traitées, mais aussi par la possibilité de transformer une image en un signal par un parcours de l'image selon une courbe de Peano (voir [18] pour une application du parcours de Peano en segmentation d'images). Le "dépliage" de l'image permet alors d'utiliser les algorithmes de chaînes qui sont particulièrement rapides, et évite l'emploi de champs de Markov nécessitant des calculs assez lourds. Enfin, l'importance pratique des chaînes de Markov cachées dans de nombreux domaines (voir la bibliographie "en ligne" de O. Cappé¹ sur les chaînes de Markov cachées) justifie à elle seule l'étude de ce modèle et de ses généralisations. Ce succès dans les applications repose sur les possibilités d'une part de calculer (exactement ou approximativement) les probabilités a posteriori d'intérêt, et d'autre part d'estimer les paramètres du modèle avant de réaliser l'estimation du processus caché. Cette procédure en deux étapes, appelée segmentation non-supervisée, donne toute son importance pratique à l'approche statistique en traitement d'image. Il suffit de se donner un modèle paramétrique adapté aux données à traiter pour avoir un algorithme de segmentation automatique. Cependant, comme pour tout modèle statistique, la modélisation peut se révéler insuffisante et diminuer les performances de segmentation ou fournir des informations erronées sur les caractéristiques des classes. De nombreuses modifications ou extensions du modèle de chaîne de Markov cachée ont été proposées afin de relâcher certaines hypothèses ou d'incorporer des propriétés particulières des phénomènes observés, comme par exemple les chaînes de Markov cachées factorielles (introduisant plusieurs processus cachés [75]) ou les processus cachés semi-markoviens (pour lequel le processus caché n'est plus nécessairement markovien [84, 66, 64]). On pourra consulter le récent ouvrage de Cappé, Moulines et Rydén rassemblant de nombreuses généralisations des chaînes de Markov cachées [32].

Les problématiques que nous avons traitées dans la thèse concernent la généralisation et le raffinement des procédures de segmentation par chaîne de Markov cachée. Cependant, notre but est de fournir des algorithmes de segmentation non-supervisée pouvant être utilisés dans des conditions opérationnelles exigeantes. En collaboration avec la société Thales Air Defence, nous avons développé des techniques économes en termes de complexité ou de temps de calcul, dans l'objectif d'une implémentation matérielle et d'une utilisation en temps réel des algorithmes de segmentation.

Modélisation

Pour la classification bayésienne de N observations $\mathbf{Y} = (Y_n)_{1 \leq n \leq N}$ indépendantes et identiquement distribuées (i.i.d), on suppose qu'il existe des variables aléatoires $(X_n)_{1 \leq n \leq N}$ à valeurs dans un espace discret $\mathcal{X}$ (de cardinal K), où X_n est la classe de l'observation Y_n . Nous notons alors pour tout $k \in \mathcal{X}$, $\pi_k = P(X_n = k)$ la probabilité d'appartenance à la classe k de la variable Y_n , et nous supposons que la loi de Y_n conditionnellement à $X_n = k$ admet une densité paramétrique $f(y_n, \theta_k)$. Toutes les observations Y_n ont donc pour densité le mélange $\sum_{k=1}^K \pi_k f(y_n, \theta_k)$ et la loi a posteriori de la variable X_n que nous utilisons pour classer finalement Y_n se déduit directement de la formule de Bayes grâce à l'indépendance des observations.

Le modèle de chaîne de Markov caché permet de prendre en compte une dépendance entre les observations Y_n (et passer ainsi du cas i.i.d au cas stationnaire) en considérant que les variables

¹<http://www.tsi.enst.fr/~cappe/docs/hmmbib.html>

$(X_n)_{1 \leq n \leq N}$ sont une chaîne de Markov de loi stationnaire $\pi = (\pi_1 \dots \pi_K)'$. La loi du processus $\mathbf{Y}$ est alors complètement spécifiée en rajoutant les hypothèses suivantes :

- les observations Y_n sont indépendantes conditionnellement au processus des classes ;
- la loi de Y_n conditionnellement à $(X_n)_{1 \leq n \leq N}$ ne dépend que de X_n .

Malgré l'introduction de cette dépendance, il reste possible de calculer de manière particulièrement simple les probabilités a posteriori $P(X_n|\mathbf{Y})$ grâce aux formules récursives avant-arrière de Baum-Welch (voir [134]), et donc d'exploiter facilement toutes les observations pour la classification. De plus, les probabilités utiles dans le calcul de l'estimateur par maximum de vraisemblance des paramètres du modèle sont toutes calculées par l'algorithme de Baum-Welch.

Le sujet de cette thèse est la modélisation et l'estimation statistique pour la segmentation non-supervisée des signaux, et nos travaux se sont alors développés autour de deux questions :

1. Comment peut-on mieux décrire le processus observé, et en particulier sa structure de dépendance ?
2. Comment pouvons nous estimer le processus caché et les paramètres de manière à conserver des traitements simples et satisfaisants pour permettre l'utilisation pratique de programmes les mettant en oeuvre ?

Ces deux questions sont motivées, entre autres, par le développement de procédures de segmentation prenant en compte les propriétés spécifiques des signaux radar.

Nous remarquons alors que la modélisation de la dépendance des observations peut être faite de manière plus générale en considérant que le processus couple $\mathbf{Z} = (Z_n)_{1 \leq n \leq N} = (X_n, Y_n)_{1 \leq n \leq N}$ est une chaîne de Markov stationnaire et non plus seulement le processus $\mathbf{X}$. Cette modélisation directe du processus joint correspond à l'hypothèse de chaîne de Markov couple, modèle récemment introduit par Pieczynski dans [124]. Pieczynski a montré que la markovianité du processus couple $\mathbf{Z}$ est une condition suffisante pour avoir la markovianité a posteriori du processus $\mathbf{X}$ ainsi que l'existence de procédures récursives de calcul des probabilités a posteriori. Ceci permet de transposer aux chaînes couples tous les traitements "rapides" des chaînes de Markov cachées basées sur ces propriétés. Comme pour les chaînes de Markov cachées, la loi d'un modèle couple stationnaire est décrite par la densité de la loi jointe de Y_1, Y_2 , mais elle n'est plus nécessairement égale à $\sum_{k,l} p_{kl} f(y_1, \theta_k) f(y_2, \theta_l)$ (où p_{kl} est la probabilité $P(X_1 = k, X_2 = l)$). En toute généralité, elle est égale à $\sum_{k,l} p_{kl} f(y_1, y_2, \theta_{kl})$ et l'utilisation de densités bivariées $f(y_1, y_2, \theta_{kl}) \neq f(y_1, \theta_k) f(y_2, \theta_l)$ permet alors de décrire des cas dans lesquelles les observations ne sont pas indépendantes conditionnellement au processus des états. Nous introduisons les copules [114] dans la modélisation de ces densités, ce qui nous permet non seulement de vérifier facilement la contrainte d'égalité des densités marginales dues à la stationnarité du processus $\mathbf{Y}$, mais aussi d'écrire explicitement la densité des observations et du processus (nécessaire dans le cadre bayésien). Ce travail constitue, à notre connaissance, la première utilisation des copules dans le contexte des modèles de Markov cachés. Lorsque nous supposons que le processus $\mathbf{X}$ est markovien, la différence entre chaînes couples et chaînes de Markov cachées consiste alors exactement en l'ajout de cette dépendance conditionnelle, représentée par les copules. L'intérêt des copules est de permettre une meilleure modélisation de la structure de dépendance des données observées (notamment l'autocovariance) et nous montrons les faiblesses, non remarquées jusqu'à présent, des chaînes de Markov cachées sur cet aspect-là.

L'introduction de la dépendance conditionnelle dans les modèles de Markov cachés n'est pas une nouveauté, et il est classique, pour la description de phénomènes possédant des dynamiques complexes, de supposer que le processus des observations est une chaîne de Markov conditionnellement aux observations. De tels processus sont appelés processus autorégressifs à changement de régime markovien et les traitements classiques d'estimation du processus caché et des paramètres se généralisent à ces modèles [59, 32]. Bien que ces modèles soient aussi des chaînes de Markov couple, le modèle que nous développons s'en distingue fondamentalement par l'usage des copules, et par la connaissance de la loi (stationnaire) de Y_n .

Un autre aspect de la modélisation par chaîne de Markov cachée consiste en le choix d'un modèle paramétrique $f(y_n, \theta_k)$ adapté pour représenter la densité de Y_n conditionnellement à $X_n = k$. Dans le cadre de la classification de données multivariées et de la segmentation de signaux et d'images, le modèle gaussien reste le modèle le plus utilisé, même si d'autres modèles paramétriques ont été utilisés dans différentes applications [109]. Nous nous sommes intéressés à l'utilisation (et l'estimation) de modèles plus réalistes ou capables d'intégrer une connaissance sur les densités marginales. Nous étudions les lois de Student et les lois K, qui appartiennent aux modèles elliptiques appelés "Vecteur Aléatoires Sphériquement Invariants" et qui sont des modèles particulièrement pertinents pour représenter les signaux radar. L'utilisation de la loi K dans le contexte de la segmentation non-supervisée est rendue possible par le développement d'une méthode d'estimation par maximum de vraisemblance. Nous introduisons pour la première fois les copules dans le contexte de la segmentation de données multivariées, ce qui nous permet de définir de nouveaux modèles multivariés et nous montrons en particulier l'intérêt des lois gamma multivariées pour la segmentation d'image.

Estimation

L'estimation des paramètres des chaînes de Markov cachées (ainsi que des processus markoviens à changement de régime markovien) est faite le plus souvent par la méthode du maximum de vraisemblance, qui donne un estimateur possédant les propriétés usuelles de convergence presque-sûre, de normalité et d'efficacité asymptotique [102, 58, 21, 59, 32]. Malgré la complexité de la vraisemblance de ces modèles, les points stationnaires de la log-vraisemblance sont calculés grâce à des méthodes récursives : l'algorithme Espérance-Maximisation (*Expectation-Maximization*, EM) de Dempster, Laird et Rubin [52], ou l'une de ses variantes [108].

Cependant, lorsque la log-vraisemblance se complique, notamment par l'utilisation des copules, la recherche du maximum de vraisemblance, même par EM, aboutit à des problèmes d'optimisation numérique qui peuvent fortement limiter l'applicabilité des procédures de segmentation non-supervisée. Nous proposons alors une nouvelle méthode d'estimation des paramètres de modèles à données manquantes, basées sur les fonctions estimantes (introduites par Godambe [79]). L'estimateur est déterminé à partir d'un échantillon $\mathbf{y} = (y_1, \dots, y_N)$ en résolvant une équation de la forme $g(\mathbf{y}, \theta) = 0$. g est une "fonction estimante" dépendant des paramètres et des observations, qui, sous certaines conditions, donne des estimateurs convergents. L'intérêt de ces méthodes est de permettre de choisir la forme de la fonction g , notamment sa complexité, de manière relativement indépendante de la log-vraisemblance du modèle. En développant une idée proposée par Heyde et Morton [90], nous montrons qu'il est possible de construire de manière générale des

fonctions estimantes de modèles à données manquantes (comme les chaînes de Markov cachées ou couples) à partir des fonctions estimantes spécifiées pour un modèle complet. Nous proposons un cadre théorique pour l'étude de leurs propriétés, qui nous permet de rassembler dans le même formalisme plusieurs estimateurs et méthodes de calcul existantes pour les modèles à données incomplètes. En effet, la détermination de l'estimateur à partir de fonctions estimantes aboutit à un problème de recherche de racines, similaire à celui du maximum de vraisemblance, et pour lequel nous proposons un algorithme itératif généralisant EM, appelé Estimation Conditionnelle Itérative. Nous donnons des conditions générales sous lesquelles cet algorithme peut être défini et converge. Nous en déduisons des algorithmes originaux d'estimation de chaînes de Markov cachées multivariées et de chaînes de Markov couples, et restant particulièrement simples (ne nécessitant aucune procédure d'optimisation numérique multidimensionnelle).

Cartographie de l'environnement radar

A partir de l'intensité, du spectre Doppler et de la polarisation du signal reçu, un radar peut fournir des informations respectivement sur la réflectivité électromagnétique, la vitesse et la structure physique de l'environnement. Ces informations étant acquises "ligne par ligne" lors de la rotation de l'antenne d'un radar veille, nous proposons de construire des algorithmes de cartographie automatique dans ce contexte, en utilisant les méthodes développées plus hauts. Nous intégrons notamment la propriété de dépendance spatiale de l'intensité reçue (remarquée et étudiée dans [152, 104]) par des modèles de Markov couples avec copules. De plus, nous proposons des modes de représentation paramétrique de l'information Doppler ou polarimétrique, qui se prête à des méthodes de segmentation bayésienne. Nous utilisons alors des lois de statistique directionnelle pour modéliser les densités $f(y_n, \theta_k)$ et proposer de nouvelles méthodes de segmentation non-supervisée, que nous illustrons sur données simulées et réelles.

Organisation du manuscrit

Dans le chapitre 1, nous présentons les modèles de Markov couples, et démontrons les propriétés qui justifient leur emploi pour la segmentation bayésienne. Nous nous concentrons alors sur les propriétés des chaînes couples, et nous introduisons une typologie de modèles que nous utilisons constamment par la suite : Chaîne de Markov Cachée à Bruit Indépendant (CMCa-BI) pour les chaînes de Markov cachées classiques, Chaîne de Markov Cachée (CMCa) pour les chaînes de Markov Couple pour lesquelles le processus caché est lui aussi markovien, et enfin Chaîne de Markov Couple (CMCo) dans le cas général.

Le chapitre 2 traite de l'inférence des CMCa et peut se décomposer en deux sous-parties. La première traite des méthodes et conditions requises pour la convergence de l'estimateur du maximum de vraisemblance des chaînes de Markov cachées classiques et des processus de Markov à changement de régimes markoviens, ainsi que des estimateurs du nombre de classes. Nous abordons également les aspects calculatoires de la détermination du maximum de la vraisemblance avec les algorithmes EM et *Stochastic-EM*. Dans la seconde sous-partie, nous exposons les concepts de base de la théorie des fonctions estimantes, et nous introduisons la méthode de construction de fonctions estimantes par projection. Nous abordons alors le problème de la détermination des racines des équations ainsi obtenues et nous proposons l'algorithme d'Estimation Conditionnelle

Itérative (ECI). Nous concluons par l'application de l'ECI à différentes fonctions estimantes.

Dans le chapitre 3, nous présentons différents modèles paramétriques de densités multivariées $f(y_n, \theta_k)$ et leurs propriétés les plus importantes. Nous donnons les algorithmes d'estimation obtenus soit par algorithme EM, soit par algorithme ECI dans le cadre des chaînes de Markov cachées classiques, et illustrons les résultats sur données réelles et simulées.

Dans le chapitre 4, nous décrivons l'utilisation des CMCo pour la segmentation de processus stationnaires, et l'introduction des copules pour la représentation de la dépendance conditionnelle. Nous étudions essentiellement le cas des CMCa à observations scalaires, et nous montrons l'influence de la dépendance conditionnelle en modélisation et en segmentation. Nous développons alors l'estimation des paramètres pour les CMCa et CMCo à partir de fonctions estimantes bien choisies, et montrons sur données réelles la construction de cartes de l'environnement radar à partir de l'intensité.

Dans le chapitre 5, nous présentons les principes physiques du radar, les traitements qui sont effectués sur le signal, et la structure des données que nous segmentons effectivement. Nous décrivons le spectre Doppler et l'ellipse de polarisation auxquels nous nous intéressons pour caractériser l'environnement radar, et nous donnons les modes de représentation et les modèles utilisés pour les segmenter. Nous illustrons les procédures de segmentation non-supervisée sur des données réelles. En conclusion, nous proposons des perspectives de recherche dans le prolongement des questions ouvertes dans cette thèse.

Cette thèse est le lieu de rencontre des domaines du traitement d'image, de la statistique et du traitement du signal radar, et constitue un travail de mathématiques appliquées. En effet, des problèmes pratiques rencontrés en image et en radar sont le moteur du développement des nouveaux modèles et des méthodes présentées ici. Ce travail s'est appuyé alors sur de nombreuses simulations et confrontations aux données réelles, dont nous exposons ici les résultats les plus parlants. Notre objectif est alors de fournir un exposé des questions spécifiques que nous avons eues à traiter dans chacun de ces domaines de telle sorte que nos contributions soient exploitables par des praticiens issus de ces différents horizons. Nous rappelons par conséquent explicitement les définitions et les résultats essentiels des méthodes classiques de chacun de ces domaines, soit dans le corps de la thèse lui-même, soit dans les annexes. Dans ces dernières, nous rappelons les concepts classiques en traitement du signal de vecteurs aléatoires complexes circulaires, et rappelons les liens entre les différentes paramétrisations des processus stationnaires et les algorithmes utilisés pour les calculer. Enfin, nous présentons rapidement les propriétés essentielles de chaînes de Markov à espace d'état général, et quelques résultats sur leur comportement asymptotique.

Chapitre 2

Segmentation bayésienne d’images et de signaux

Nous rappelons ici les principes qui fondent l’analyse bayésienne des images et notamment le concept fondamental de champs markoviens, ainsi que celui de champs markoviens cachés. Nous introduisons alors les modèles de Markov couples [131] qui en sont une généralisation, et nous rappelons, dans un cadre général, la propriété qui motive leur étude : la markovianité a posteriori. Par la suite, nous étudions particulièrement les chaînes de Markov couples [124] et nous redémontrons les procédures récursives de calcul des probabilités a posteriori, généralisant les algorithmes qui ont rendus les chaînes de Markov cachées si populaires. Ceci permettra de proposer des traitements “rapides” malgré la complexité accrue du modèle. En dernier lieu, nous rappelons la définition des modèles triplets [123, 129] pour lesquelles les traitements bayésiens sont encore envisageables en exploitant la markovianité d’un processus plus grand : nous donnons des exemples de ces modèles ce qui permet de montrer que “l’approche triplet” réunit dans un même cadre plusieurs généralisations déjà proposées des modèles de Markov cachés classiques.

2.1 Estimation bayésienne

2.1.1 Modélisation probabiliste des images

Nous présentons dans cette section le paradigme de la segmentation bayésienne d’images : le modèle stochastique et les règles de décision bayésienne que nous utilisons pour la construction d’une cartographie des zones homogènes d’une image.

Une image $\mathbf{y}$ est la réalisation d’une variable aléatoire $\mathbf{Y}_S$ définie sur un espace probabilisé $(\Omega, \mathcal{A}, P)$. S est un ensemble fini d’indice, de cardinal $|S|$ et représente l’ensemble des pixels (ou sites) qui constituent l’image. $\mathbf{Y}_S = (Y_s)_{s \in S}$ est une famille de variable aléatoire $Y_s : (\Omega, \mathcal{A}, P) \longrightarrow (\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$, où $\mathcal{Y}$ est un espace topologique (et $\mathcal{B}(\mathcal{Y})$ désigne la tribu des boréliens sur $\mathcal{Y}$). Dans les applications que nous traitons, $\mathcal{Y}$ peut être :

- l’espace euclidien $\mathbb{R}^d$ ou $\mathbb{C}^d$, $d \geq 1$ (mesure multicapteur ou multispectrale) ;
- la sphère unité de $\mathbb{R}^d$ notée $S^{d-1} = \left\{ y = (y_1, \dots, y_d) \in \mathbb{R}^d \mid \|y\| = \left(\sum_{i=1}^d y_i^2 \right)^{1/2} \right\}$ (représen-

tation partielle de l'information Doppler ou polarimétrique).

En toute généralité, $\mathbf{Y}_S$ est appelé un champ aléatoire¹. Si A est un sous-ensemble de S , nous notons $\mathbf{Y}_A = (Y_s)_{s \in A}$ la variable aléatoire marginale et $\mathbf{y}_A = (y_s)_{s \in A}$ sa réalisation. Lorsque $S = [1..N]$ (avec $N \geq 1$), $\mathbf{Y}_S$ (noté alors $\mathbf{Y}_N$) modélise alors un signal unidimensionnel et est appelée une chaîne.

Segmenter l'image $\mathbf{y}$, c'est estimer la réalisation d'une image $\mathbf{X} = (X_s)_{s \in S}$ définie sur $(\Omega, \mathcal{A}, P)$ et dont l'espace des valeurs $\mathcal{X}$ est un espace discret de cardinal $K \geq 2$, que l'on identifie à $[1..K]$. Cette réalisation $\mathbf{x}$ est appelée "réalité terrain". Nous notons μ une mesure de référence définie sur $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$, δ la mesure de comptage sur $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, et nous supposons que la loi de $(\mathbf{X}_S, \mathbf{Y}_S)$ admet une densité $p(\mathbf{x}, \mathbf{y})$ relativement à la mesure produit $(\delta \otimes \mu)^{|S|}$. Les densités des lois de $\mathbf{X}_S$ et $\mathbf{Y}_S$ sont notées respectivement $p(\mathbf{x})$ et $p(\mathbf{y})$. En statistique bayésienne, les densités $p(\mathbf{x})$, $p(\mathbf{y}|\mathbf{x})$ et $p(\mathbf{y})$ sont classiquement appelées (respectivement) loi a priori, vraisemblance et loi d'observation (ou vraisemblance intégrée). Enfin, la loi a posteriori $p(\mathbf{x}|\mathbf{y})$ se déduit des densités précédentes par la formule de Bayes

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})} \quad (2.1)$$

$$\propto p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) \quad (2.2)$$

L'estimation de $\mathbf{X}_S$ à partir de $\mathbf{Y}_S$ est effectuée par l'application d'une règle de décision bayésienne. Celle-ci est construite à partir d'une fonction de coût $L : \mathcal{X}^{|S|} \times \mathcal{X}^{|S|} \rightarrow \mathbb{R}$. L'estimateur bayésien de $\mathbf{X}_S$ correspondant au coût L est la fonction $\mathbf{y} \mapsto \hat{\mathbf{x}}(\mathbf{y})$ qui minimise le coût moyen, i.e.

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{T}} E[L(\mathbf{X}_S, \mathbf{T}(\mathbf{Y}_S))]$$

où $\mathbf{T}$ décrit l'ensemble des fonctions mesurables et intégrables de $\mathcal{Y}^{|S|}$ dans $\mathcal{X}^{|S|}$. En traitement de l'image et du signal, les fonctions de perte les plus couramment utilisées sont

$$\begin{aligned} \forall \mathbf{x} = (x_s)_{s \in S}, \mathbf{x}' = (x'_s)_{s \in S} \in \mathcal{X}^{|S|}, \quad L_1(\mathbf{x}, \mathbf{x}') &= 1_{\{\mathbf{x} \neq \mathbf{x}'\}} \\ L_2(\mathbf{x}, \mathbf{x}') &= \sum_{s \in S} 1_{x_s \neq x'_s} \end{aligned}$$

L'estimateur obtenu avec la fonction L_1 est le Maximum A Posteriori (MAP) et a pour expression

$$\hat{\mathbf{x}}_{\text{MAP}}(\mathbf{y}) = \arg \max_{\mathbf{x} \in \mathcal{X}^{|S|}} p(\mathbf{x}|\mathbf{y}) \quad (2.3)$$

L'estimateur obtenu avec L_2 est le Maximum a Posteriori des Marges (MPM) et a pour expression

$$\begin{aligned} \hat{\mathbf{x}}_{\text{MPM}}(\mathbf{y}) &= (\hat{x}_{s, \text{MPM}}(\mathbf{y}))_{s \in S} \\ \forall s \in S, \quad \hat{x}_{s, \text{MPM}}(\mathbf{y}) &= \arg \max_{x_s \in \mathcal{X}} p(x_s|\mathbf{y}) \end{aligned} \quad (2.4)$$

Par la suite, nous utiliserons plus particulièrement l'estimateur MPM qui minimise, en moyenne,

¹ S est souvent un sous-ensemble de $\mathbb{Z}^2$, mais l'imagerie tridimensionnelle ou la vidéo amène à considérer des sous-ensembles de $\mathbb{Z}^3$, ou $\mathbb{Z}^d$ avec d quelconque.

le nombre d'observations mal classées, encore appelé taux d'erreur. L'inférence bayésienne de l'image cachée dépend alors uniquement de la loi a posteriori, et il suffit, par la formule de Bayes (2.2), de connaître la loi a priori et la vraisemblance pour pouvoir calculer $\hat{\mathbf{x}}_{\text{MAP}}$ ou $\hat{\mathbf{x}}_{\text{MPM}}$.

Les lois a priori et d'observation représentent deux types de connaissance sur le phénomène physique que nous observons. Dans le cas du traitement d'images, nous voulons intégrer une des caractéristiques essentielles de l'image qui est la dépendance spatiale des observations : ceci est fait par le biais de la densité $p(\mathbf{x})$ qui va représenter la dépendance entre sites. La loi d'observation $p(\mathbf{y}|\mathbf{x})$ est en général phénoménologique : elle représente notre connaissance sur la formation de l'image observée et elle sera différente selon que nous avons une image optique, multispectrale, ... Ainsi, classiquement, nous choisissons d'abord une loi a priori en accord avec l'information spatiale que nous voulons utiliser (processus 1D ou 2D, fortement dépendant spatialement ou non), puis nous choisissons une loi $p(\mathbf{y}|\mathbf{x})$ correspondante au phénomène observé. Les modèles de Markov couples proposent une rupture par rapport à cette approche classique de la modélisation bayésienne : il s'agit de proposer directement la loi jointe $p(\mathbf{x}, \mathbf{y})$ de telle sorte qu'elle intègre une structure de dépendance entre les sites plus riche que dans les modèles markoviens classiquement utilisés, bien que les procédures d'estimation bayésienne soient toujours possibles.

2.1.2 Modèles markoviens

Le problème de l'estimation bayésienne est formellement résolu par les équations (2.2) et l'expression de l'estimateur MPM (2.4). Cependant, la nécessité de calculer les probabilités $p(\mathbf{x}|\mathbf{y})$ alors que $|S|$ est grand devient très difficile dès lors que l'on ne suppose plus l'indépendance des variables (X_s, Y_s) . L'hypothèse de markovianité de $\mathbf{X}_S$ est l'extension la plus simple du cas indépendant au cas dépendant, mais elle permet de modéliser la dépendance spatiale de manière assez riche et interprétable, tout en ayant une loi de densité $p(\mathbf{x})$ calculable ou aisément simulable, par des méthodes de Monte Carlo par Chaînes de Markov (*Markov Chain Monte Carlo*, MCMC), [138, 156]. Nous rappelons ici la définition de champ markovien, fondé sur les notions de cliques et de voisinages.

Définition 2.1.1. Cliques et Voisinages

Soit S un ensemble fini. On appelle un système de voisinage $\mathcal{V} = \{\mathcal{V}_s, \mathcal{V}_s \subset S, s \in S\}$ une collection de sous-ensembles de S telle que $s \notin \mathcal{V}_s$ et $(s \in \mathcal{V}_t \Leftrightarrow t \in \mathcal{V}_s)$. Les sites $t \in \mathcal{V}_s$ sont appelés voisins de s .

Un sous-ensemble c de S est une clique si c'est un singleton ou si tous les sites de c sont voisins. L'ensemble des cliques de S est noté $\mathcal{C}$.

Définition 2.1.2. Champs Markoviens

Le champ aléatoire $\mathbf{X}_S$ est un champ markovien relativement au système de voisinage $\mathcal{V}$ si sa loi admet une densité (relativement à une mesure de référence) vérifiant :

$$\forall s \in S, \quad p(x_s | (x_t)_{t \neq s}) = p(x_s | \mathbf{x}_{\mathcal{V}_s}) \quad (2.5)$$

Les densités conditionnelles $p(x_s | \mathbf{x}_{\mathcal{V}_s})$ sont appelées caractéristiques locales du champ. La propriété de markovianité locale (2.5) s'étend à un ensemble $A \subset S$ quelconque, si on introduit la

notion de fermeture $\overline{A} = \cup_{t \in A} \mathcal{V}_t$ et de frontière $\partial A = \overline{A} \setminus A$. Nous avons alors :

$$p(\mathbf{x}_A | \mathbf{x}_{S \setminus A}) = p(\mathbf{x}_A | \mathbf{x}_{\partial A}) \quad (2.6)$$

Cette propriété permet entre autres d'affirmer que pour tout ensemble $A \subset S$, $\mathbf{x}_A$ et $\mathbf{x}_{S \setminus \overline{A}}$ sont indépendants conditionnellement à ∂A , i.e.

$$\forall A \subset S, p(\mathbf{x}_A, \mathbf{x}_{S \setminus \overline{A}} | \mathbf{x}_{\partial A}) = p(\mathbf{x}_A | \mathbf{x}_{\partial A}) p(\mathbf{x}_{S \setminus \overline{A}} | \mathbf{x}_{\partial A}) \quad (2.7)$$

L'importance des cliques pour l'écriture des caractéristiques locales d'un champ markovien provient du théorème de Hammersley-Clifford qui affirme l'équivalence, lorsque $\forall \mathbf{x}, p(\mathbf{x}) > 0$, entre la markovianité de $\mathbf{X}_S$ et la possibilité d'écrire la densité de $\mathbf{X}_S$ sous la forme :

$$p(\mathbf{x}) = \frac{1}{Z} \exp \left(- \sum_{c \in \mathcal{C}} h_c(\mathbf{x}_c) \right) \quad (2.8)$$

où Z est la constante de normalisation, appelée aussi fonction de partition, et les fonctions h_c sont à valeurs dans $\mathbb{R}$. $p(\mathbf{x})$ est alors appelée une mesure de Gibbs et la fonction $H(\mathbf{x}) = \sum_{c \in \mathcal{C}} h_c(\mathbf{x}_c)$ est son énergie [133]. Nous pouvons dire aussi que la densité (2.8) se factorise sur les cliques, autrement dit il existe des fonctions ψ_c définies pour toute clique $c \in \mathcal{C}$ telles que $p(\mathbf{x}) = \prod_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c)$. Les lois conditionnelles s'écrivent facilement avec les fonctions h_c

$$p(\mathbf{x}_A | \mathbf{x}_{S \setminus A}) = \frac{1}{Z_A} \exp \left(- \sum_{c \cap A \neq \emptyset} h_c(\mathbf{x}_c) \right) \quad (2.9)$$

avec Z_A constante de normalisation.

Une chaîne de Markov $\mathbf{X}_N$ est un champ de Markov relativement au système de voisinages $\mathcal{V}_{\{n\}} = \{n-1, n+1\}$ pour $2 \leq n \leq N-1$, et $\mathcal{V}_{\{1\}} = \{2\}$ et $\mathcal{V}_{\{N\}} = \{N-1\}$. Le système de cliques $\mathcal{C}$ correspondant est constitué des singletons $(\{n\})_{1 \leq n \leq N}$ et des paires $(\{n, n+1\})_{1 \leq n \leq N-1}$. Dans ce cas-là, lorsque A est de la forme $A = [p..n]$, avec $1 \leq p \leq n \leq N$, nous notons $\mathbf{X}_A = \mathbf{X}_{p:n}$, et simplement $\mathbf{X}_n$ pour $\mathbf{X}_{1:n}$. La propriété (2.7) apparaît comme une généralisation de la propriété "d'indépendance du passé et du futur conditionnellement au présent" des chaînes de Markov :

$$p(\mathbf{x}_n, \mathbf{x}_{n+1:N} | x_n) = p(\mathbf{x}_n | x_n) p(\mathbf{x}_{n+1:N} | x_n) \quad (2.10)$$

L'expression sous forme d'une mesure de Gibbs redonne la factorisation classique de la densité d'une chaîne de Markov (lorsque $\forall \mathbf{x}, p(\mathbf{x}) > 0$) :

$$p(\mathbf{x}) = \prod_{n=1}^N \psi_{\{n\}}(x_n) \prod_{n=2}^N \psi_{\{n-1, n\}}(x_{n-1}, x_n)$$

Cette formulation d'une chaîne en tant que champ de Markov occulte l'interprétation des chaînes de Markov comme des processus temporels vérifiant la propriété $p(x_{n+1} | \mathbf{x}_n) = p(x_{n+1} | x_n)$, qui donne un rôle central aux densités de transition $p(x_{n+1} | x_n)$. Les spécifications locales du champ

markovien et les densités de transitions sont liées par l'expression

$$p(x_n | x_{n-1}, x_{n+1}) = \frac{p(x_{n+1} | x_n) p(x_n | x_{n-1})}{\int_{\mathcal{X}} p(x_{n+1} | x'_n) p(x'_n | x_{n-1}) \mu(dx'_n)}$$

Les données radar que nous allons traiter par la suite sont des processus spatiaux à une dimension : l'hypothèse de markovianité (avec une mémoire de taille 1) est donc le modèle le plus simple que l'on puisse faire pour tenir compte de la dépendance spatiale. Pour la réalisation des étapes calculatoires (estimation et segmentation), nous utilisons la description classique en tant que processus temporel, ce qui permet l'utilisation des méthodes récursives rapides que nous décrivons dans la section suivante.

Finalement, les modèles utilisés classiquement pour la modélisation des images et des signaux sont les modèles de Markov cachés, fondés sur les 3 hypothèses suivantes :

- l'image $\mathbf{X}_S$ est un champ markovien avec $p(\mathbf{x}) = \prod_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c)$;
- les observations $(Y_s)_{s \in S}$ sont indépendantes conditionnellement à $\mathbf{X}_S$;
- $\forall s \in S, p(y_s | \mathbf{x}) = p(y_s | x_s)$.

En raison de l'hypothèse d'indépendance, nous appellerons ces modèles "Modèles de Markov Cachés à Bruit Indépendant" (MMCa-BI). La densité du processus couple $\mathbf{Z}_S = (X_s, Y_s)_{s \in S}$ est

$$\begin{aligned} p(\mathbf{z}) &= p(\mathbf{x}) \prod_{s \in S} p(y_s | \mathbf{x}) \\ &= \prod_{c \in \mathcal{C}} \tilde{\psi}_c(\mathbf{z}_c) \end{aligned}$$

Les fonctions $\tilde{\psi}_c(\mathbf{z})$ sont égales à $\psi_c(\mathbf{x})$ si c n'est pas un singleton, sinon nous avons $\tilde{\psi}_{\{s\}}(\mathbf{z}) = \psi_{\{s\}}(x_s) p(y_s | x_s)$. Par conséquent, le processus joint $\mathbf{Z}_S$ est encore un champ de Markov relativement au même système de voisinage que celui de $\mathbf{X}_S$. Le succès dans les applications des modèles MMCa-BI repose sur le fait que l'a priori markovien est "conjugué" au processus d'observation $p(\mathbf{y} | \mathbf{x}) = \prod_{s \in S} p(y_s | x_s)$: la densité a posteriori $p(\mathbf{x} | \mathbf{y})$ se factorise relativement au même système de voisinages $\mathcal{V}$ que $p(\mathbf{x})$. Ceci permet de connaître les spécifications locales de $p(\mathbf{x} | \mathbf{y})$, et donc d'appliquer les techniques de simulation de type MCMC, notamment l'échantillonneur de Gibbs. Il est possible de simuler les champs a posteriori, et partant, d'estimer les probabilités marginales a posteriori² pour le calcul du MPM, ou d'explorer la probabilité a posteriori, notamment pour la détermination de son maximum, [147, 156]. De plus, les méthodes MCMC peuvent être utilisées pour l'estimation de Modèles Markoviens Cachés soit pour une estimation pleinement bayésienne, soit pour le maximum de vraisemblance, [76].

Lorsque $\mathbf{X}_S$ est une chaîne de Markov, nous obtenons une chaîne de Markov cachée à bruit indépendant (CMCa-BI), et l'algorithme "avant-arrière" (ou *forward-backward*, voir section 2.2.2) permet de calculer rapidement et exactement la probabilité a posteriori $p(\mathbf{x} | \mathbf{y})$, contrairement aux champs markoviens généraux. De plus, il est possible de calculer exactement les probabilités marginales $p(x_n | \mathbf{y})$ ainsi que $\arg \max_{\mathbf{x}} p(\mathbf{x} | \mathbf{y})$. Cela fournit donc des algorithmes de calcul rapide des estimateurs bayésiens classiques (dans le cas où $\mathcal{X}$ est fini). Nous redémontrons, dans la section suivante, ces récursions dans le cadre plus général des chaînes de Markov couple (CMCo).

²voire de les calculer exactement pour certaines structures de voisinages

2.2 Modèles de Markov Couples

Nous rassemblons dans cette section les propriétés fondamentales des modèles couples que l'on peut trouver, sous des formes parfois différentes dans les articles [131, 124, 128, 53, 125], et nous en donnons les démonstrations. Nous mettons en particulier en évidence le lien entre markovianité du processus caché et loi conditionnelle des observations de manière plus précise que dans [124]. Nous donnons aussi les démonstrations (non existantes dans les articles précédemment cités) de l'algorithme de Viterbi pour les chaînes couples ainsi que de la propriété 2.2.1.

2.2.1 Définition et propriété fondamentale

Lorsque les champs aléatoires observé $\mathbf{Y}_S$ et caché $\mathbf{X}_S$ sont tels que le processus joint $\mathbf{Z}_S = (X_s, Y_s)_{s \in S}$ soit un champ de Markov, nous dirons que $\mathbf{X}_S, \mathbf{Y}_S$ forment un champ de Markov couple. Dans le cadre de la modélisation, si nous faisons uniquement l'hypothèse de markovianité de $\mathbf{Z}_S$, nous dirons que nous avons un modèle de Markov couple.

Proposition 2.2.1. Markovianité a posteriori des champs couples

Soit $\mathbf{Z}_S = (X_s, Y_s)_{s \in S}$ un champ de Markov relativement à $\mathcal{V}$ alors $\mathbf{X}_S$ est un champ de Markov relativement à $\mathcal{V}$ conditionnellement à $\mathbf{Y}_S$.

Démonstration.

$$\begin{aligned} p(x_s | \mathbf{x}_{S \setminus s}, \mathbf{y}) &= \frac{p(x_s, y_s | \mathbf{x}_{S \setminus s}, \mathbf{y}_{S \setminus s})}{p(y_s | \mathbf{x}_{S \setminus s}, \mathbf{y}_{S \setminus s})} \\ &\propto p(x_s, y_s | \mathbf{x}_{\mathcal{V}_s}, \mathbf{y}_{\mathcal{V}_s}) \\ &\propto p(x_s | \mathbf{x}_{\mathcal{V}_s}, y_s, \mathbf{y}_{\mathcal{V}_s}) \end{aligned}$$

□

En revanche, $\mathbf{X}_S$ n'est plus nécessairement lui-même markovien (non-conditionnellement à $\mathbf{Y}_S$), parce qu'il n'est que le processus marginal d'un processus markovien. De plus, en toute généralité nous avons $p(y_s | \mathbf{x}) \neq p(y_s | x_s)$.

Nous pouvons construire des champs couples en affaiblissant les hypothèses des MMCa-BI. Nous définissons les Modèles Markoviens Cachés (MMCc) $(\mathbf{X}_S, \mathbf{Y}_S)$ comme étant des processus pour lesquelles $\mathbf{X}_S$ est un champ de Markov relativement à un système de voisinage $\mathcal{V}$. Les observations ne sont pas forcément indépendantes conditionnellement à $\mathbf{X}_S$ et une généralisation simple de la loi conditionnelle de $\mathbf{Y}_S$ est de supposer que c'est aussi un champ markovien relativement à $\mathcal{V}$ conditionnellement à $\mathbf{X}_S$ (dont la densité s'écrit sous forme factorisée $\prod_{c \in \mathcal{C}} \psi_{c, \mathbf{x}}(\mathbf{y}_c)$). Dans ce cas-là, nous avons :

$$\begin{aligned} p(\mathbf{x}, \mathbf{y}) &= p(\mathbf{y} | \mathbf{x}) p(\mathbf{x}) \\ &= \prod_{c \in \mathcal{C}} \psi_{c, \mathbf{x}}(\mathbf{y}_c) \prod_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c) \end{aligned}$$

Si chaque fonction $\psi_{c, \mathbf{x}}(\mathbf{y})$ ne dépend de $\mathbf{x}$ qu'à travers les coordonnées appartenant à c nous

aurons alors

$$p(\mathbf{z}) \propto \prod_{c \in \mathcal{C}} \psi_c(\mathbf{z}_c)$$

avec $\psi_c(\mathbf{z}_c) = \psi_{c,\mathbf{x}}(\mathbf{y}_c)\psi_c(\mathbf{x}_c)$. Comme la loi de $\mathbf{Z}_S$ se factorise selon le système de cliques $\mathcal{C}$, c'est donc un champ de Markov relativement à $\mathcal{V}$.

Une méthode générale de construction de MMCa proposée dans [15] (généralisant une méthode proposée dans [83]) consiste à considérer un champ markovien $\mathbf{W}_S$ relativement à $\mathcal{V}$, un ensemble de cliques $(c_i)_{1 \leq i \leq Q}$ formant une partition de S et une fonction f . Si la fonction f se décompose selon les cliques sous la forme $f(\mathbf{x}, \mathbf{w}) = (f_i(\mathbf{x}_{c_i}, \mathbf{w}_{c_i}))_{1 \leq i \leq Q}$, alors le champ $(\mathbf{X}_S, \mathbf{Y}_S = f(\mathbf{X}_S, \mathbf{W}_S))$ est un MMCa. Un moyen plus direct est alors de supposer que $p(\mathbf{z})$ est une mesure de Gibbs selon un système de voisinages adapté. Dans ce cas-là, nous n'avons plus la markovianité de $\mathbf{X}_S$, mais nous pouvons choisir $p(\mathbf{y}|\mathbf{x})$ tel qu'il représente bien la réalité. Dans [131, 14], la densité $p(\mathbf{z})$ du couple a été choisie de telle sorte que la loi conditionnelle soit un champ gaussien ; or les champs gaussiens (en particulier les modèles conditionnels auto-régressifs) restituent bien certaines texture des images, voir Winkler [156].

2.2.2 Chaînes de Markov Couples

Nous traitons dans cette partie des CMCo stationnaires et de leurs propriétés essentielles, notamment les procédures de calcul des probabilités a posteriori. La stationnarité n'est pas nécessaire pour établir ces dernières, mais elle permet de simplifier la présentation du modèle et de mettre en évidence les différences majeures avec les CMCa. Nous utilisons ces propriétés pour mettre en évidence les liens entre markovianité du processus $\mathbf{X}_N$, indépendance conditionnelle des observations et markovianité du processus couple. Ces propriétés ont été présentées dans le cas où l'espace des états $\mathcal{X}$ est fini dans les articles [124, 53] ; elles restent également valables lorsque $\mathcal{X}$ est un espace quelconque avec δ mesure de référence sur $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. Par conséquent, les algorithmes de filtrage pour l'estimation d'un processus $\mathbf{X}_N$ à espace d'état continu sont possibles (dans un contexte gaussien [128], ou général [54] en utilisant le filtrage particulière).

Si $\mathbf{Z}_N = (\mathbf{X}_N, \mathbf{Y}_N)$ est une CMCo stationnaire, elle est entièrement définie par sa densité de transition et sa loi stationnaire, ou de manière équivalente par la densité $p(z_1, z_2) = p(x_1, x_2, y_1, y_2)$ que nous décomposons en $p(x_1, x_2)p(y_1, y_2 | x_1, x_2)$. Nous supposons par la suite que nous connaissons la loi du processus $\mathbf{Z}_N$ par la connaissance de la probabilité jointe $p(x_1, x_2)$ et de la densité conditionnelle $p(y_1, y_2 | x_1, x_2)$ ³.

La densité de transition $p(z_2 | z_1)$, relativement à la mesure produit $\delta \otimes \mu$ s'écrit

$$p(z_2 | z_1) = \frac{p(x_1, x_2)p(y_1, y_2 | x_1, x_2)}{\int_{\mathcal{X}} p(x_1, x_2)p(y_1 | x_1, x_2)\delta(dx_2)}$$

où $p(y_1 | x_1, x_2)$ est la densité marginale "gauche" de $p(y_1, y_2 | x_1, x_2)$ (notée parfois $p_g(y_1 | x_1, x_2)$). De manière symétrique, nous notons $p(y_2 | x_1, x_2)$ la marginale "droite" de $p(y_1, y_2 | x_1, x_2)$ (notée parfois $p_d(y_2 | x_1, x_2)$). La loi stationnaire a pour densité $p(z_1) = \int_{\mathcal{X}} p(x_1, x_2)p_g(y_1 | x_1, x_2)\delta(dx_2)$

³Ceci est une extension de la description habituellement faite des CMCa-BI, pour lesquelles nous connaissons la densité de transition $p(x_2 | x_1)$ et la loi d'émission $p(y_1 | x_1)$.

qui est aussi égale à $p(z_2) = \int_{\mathcal{X}} p(x_1, x_2) p_d(y_2 | x_1, x_2) \delta(dx_1)$: ceci montre que les densités marginales gauche et droite sont liées. Ainsi, lors de la spécification d'un modèle paramétrique pour les chaînes couples stationnaires, nous devons choisir une loi jointe $p(y_1, y_2 | x_1, x_2)$ respectant cette contrainte. Nous aborderons cet aspect dans le chapitre 4.

La simulation des CMCo est plus compliquée que celle des CMCa en raison de l'intrication des processus $\mathbf{X}_N$ et $\mathbf{Y}_N$. En effet, la densité de transition de $\mathbf{Z}_N$ se réécrit encore :

$$p(x_2, y_2 | x_1, y_1) = p(y_2 | x_2, y_1, x_1) p(x_2 | x_1, y_1) \quad (2.11)$$

Remarquons que les CMCa-BI sont un cas particulier de CMCo pour lesquelles nous avons $p(y_1, y_2 | x_1, x_2) = p(y_1 | x_1) p(y_2 | x_2)$ et donc telles que $p(y_1 | x_1, x_2) = p(y_1 | x_1)$ et $p(y_2 | y_1, x_1, x_2) = p(y_2 | x_2)$.

Une réalisation $\mathbf{z} = (z_1, \dots, z_N)$ d'une chaîne couple est obtenue de la façon suivante :

1. x_1 est tiré selon $p(x_1) = \int_{\mathcal{X}} p(x_1, x_2) \delta(dx_1)$ et y_1 est tiré selon le mélange $\int_{\mathcal{X}} p(x_1, x'_2) p(y_1 | x_1, x'_2) \delta(dx'_2)$;
2. Pour tout $n \leq N$, x_n est tiré selon la loi

$$p(x_n | x_{n-1}, y_{n-1}) \propto p(x_{n-1}, x_n) p(y_{n-1} | x_{n-1}, x_n) = p(x_1, x_2) p(y_1 | x_1, x_2) \quad (2.12)$$

et y_n est tiré selon la densité

$$p(y_n | x_n, y_{n-1}, x_{n-1}) = \frac{p(y_1, y_2 | x_1, x_2)}{p(y_1 | x_1, x_2)} \quad (2.13)$$

En toute généralité, il n'est donc pas possible de simuler d'abord la chaîne $(x_n)_{1 \leq n \leq N}$ dans sa totalité puis de tirer les observations y_n selon l'état caché, ou une densité de transition dépendant des valeurs x_{n-1}, x_n , comme cela peut être le cas pour les CMCa.

Malgré une plus grande complexité dans la description des caractéristiques locales des CMCo, nous gardons les propriétés les plus connues des CMCa-BI, à savoir les formules de récurrence avant-arrière.

Propriété 2.2.1. Récurrence avant

Soit $\mathbf{Z}_N$ une CMCo. Les densités $p(y_n | \mathbf{x}_n)$ et $p(x_n | \mathbf{y}_n)$ peuvent être calculées par récursion avant.

$$\begin{aligned} p(y_n | \mathbf{x}_n) &= \frac{\int_{\mathcal{Y}} p(z_n | z_{n-1}) p(y_{n-1} | \mathbf{x}_{n-1}) \mu(dy_{n-1})}{p(x_n | \mathbf{x}_{n-1})} \\ \forall n \geq 2, \\ p(x_n | \mathbf{y}_n) &= \frac{\int_{\mathcal{X}} p(z_n | z_{n-1}) p(x_{n-1} | \mathbf{y}_{n-1}) \delta(dx_{n-1})}{p(y_n | \mathbf{y}_{n-1})} \end{aligned} \quad (2.14)$$

Démonstration.

$$\begin{aligned} p(y_n | \mathbf{x}_n) &= \frac{p(y_n, x_n | \mathbf{x}_{n-1})}{p(x_n | \mathbf{x}_{n-1})} \\ &= \frac{\int_{\mathcal{Y}} p(y_n, x_n | y_{n-1}, \mathbf{x}_{n-1}) p(y_{n-1} | \mathbf{x}_{n-1}) \mu(dy_{n-1})}{p(x_n | \mathbf{x}_{n-1})} \\ &= \frac{\int_{\mathcal{Y}} p(z_n | z_{n-1}) p(y_{n-1} | \mathbf{x}_{n-1}) \mu(dy_{n-1})}{p(x_n | \mathbf{x}_{n-1})} \end{aligned}$$

Par le même calcul, nous obtenons la récurrence pour $p(x_n | \mathbf{y}_n)$.

□

Propriété 2.2.2. Récurrence arrière

Soit $\mathbf{Z}_N$ une CMCo. Les densités $p(\mathbf{y}_{n+1:N} | x_n, y_n)$ et $p(\mathbf{x}_{n+1:N} | x_n, y_n)$ peuvent être calculées par récursion arrière.

$$\begin{aligned} p(\mathbf{y}_{n+1:N} | x_n, y_n) &= \int_{\mathcal{X}} p(\mathbf{y}_{n+2:N} | x_{n+1}, y_{n+1}) p(z_{n+1} | z_n) \delta(dx_{n+1}) \\ \forall n \leq N-2, \\ p(\mathbf{x}_{n+1:N} | x_n, y_n) &= \int_{\mathcal{X}} p(\mathbf{x}_{n+2:N} | x_{n+1}, y_{n+1}) p(z_{n+1} | z_n) \mu(dy_{n+1}) \end{aligned} \quad (2.15)$$

Démonstration.

$$\begin{aligned} p(\mathbf{y}_{n+1:N} | x_n, y_n) &= \int_{\mathcal{X}} p(\mathbf{y}_{n+1:N}, x_{n+1} | x_n, y_n) \delta(dx_{n+1}) \\ &= \int_{\mathcal{X}} p(\mathbf{y}_{n+2:N} | x_{n+1}, y_{n+1}) p(y_{n+1}, x_{n+1} | x_n, y_n) \delta(dx_{n+1}) \end{aligned}$$

Nous avons la même récurrence pour les probabilités $p(\mathbf{y}_{n+1:N} | x_n, y_n)$.

□

Ces formules de récurrence permettent de retrouver les grandeurs classiques introduites pour le filtrage et le lissage des chaînes de Markov cachées. Les “probabilités avant” $(\alpha_n)_{1 \leq n \leq N}$ sont définies sur $\mathcal{X}$ par

$$\begin{cases} \alpha_1(x_1) &= p(x_1 | y_1) \\ \forall n \leq N-1 \\ \alpha_{n+1}(x_{n+1}) &= \int p(z_{n+1} | z_n) \alpha_n(x_n) \delta(dx_n) \end{cases} \quad (2.16)$$

et les constantes de normalisation permettent de calculer la vraisemblance intégrée $p(\mathbf{y})$. Les “probabilités arrière” $(\beta_n)_{1 \leq n \leq N}$, définies dans le cas CMCa-BI par $p(\mathbf{y}_{n+1:N} | x_n)$, sont égales à $p(\mathbf{y}_{n+1:N} | x_n, y_n)$ et sont calculées récursivement sur $\mathcal{X}$ par

$$\begin{cases} \beta_N(x_N) &= 1 \\ \forall n \leq N-1 \\ \beta_n(x_n) &= \int p(z_{n+1} | z_n) \beta_{n+1}(x_{n+1}) \delta(dx_{n+1}) \end{cases} \quad (2.17)$$

Remarque 2.2.1. Relation entre récurrences avant et arrière

La récurrence arrière pour le calcul des probabilités β_n peut se voir comme une récurrence avant pour la chaîne $\tilde{\mathbf{Z}}_N = (\tilde{Z}_n)_{1 \leq n \leq N}$, obtenue en inversant le sens de $\mathbf{Z}_N$. Le processus inversé est toujours une chaîne de Markov stationnaire, qui est complètement connue par la densité $\tilde{p}(\tilde{z}_1, \tilde{z}_2) = p(\tilde{x}_2, \tilde{x}_1) p(\tilde{y}_2, \tilde{y}_1 | \tilde{x}_2, \tilde{x}_1)$. Nous avons alors $\tilde{p}_g(\tilde{y}_1 | \tilde{x}_1, \tilde{x}_2) = p_d(\tilde{y}_1 | \tilde{x}_2, \tilde{x}_1)$ et $\tilde{p}_d(\tilde{y}_2 | \tilde{x}_1, \tilde{x}_2) = p_g(\tilde{y}_2 | \tilde{x}_2, \tilde{x}_1)$. De plus,

$$\begin{aligned}
p(\mathbf{y}_{n+1:N} | x_n, y_n) &= \frac{p(\mathbf{y}_{n+1:N}, x_n, y_n)}{p(x_n, y_n)} \\
&= \frac{\tilde{p}(\tilde{x}_{N-n+1}, \tilde{\mathbf{y}}_{N-n+1})}{\tilde{p}(\tilde{x}_{N-n+1}, \tilde{y}_{N-n+1})}
\end{aligned}$$

Pour calculer les probabilités arrière β_n , il suffit donc de parcourir la chaîne $\mathbf{Z}_N$ dans le sens inverse pour calculer les probabilités avant $\tilde{\alpha}_n$ (correspondant à la probabilité de transition $\tilde{p}(\tilde{z}_2 | \tilde{z}_1)$) et de diviser par la loi stationnaire de $\mathbf{Z}_N$ (identique à celle de $\tilde{\mathbf{Z}}_N$). Nous avons donc

$$\beta_n(x_n) = \frac{\tilde{\alpha}_{N-n+1}(x_n)}{p(x_n, y_n)}$$

Comme dans le cas des CMCa-BI, les densités marginales a posteriori des processus $\mathbf{X}_N$ se calculent directement à partir des probabilités avant et arrière.

Propriété 2.2.3. *Soit $\mathbf{Z}_N$ une CMCo. Nous avons*

$$p(x_n | \mathbf{y}) \propto \alpha_n(x_n) \beta_n(x_n) \quad (2.18)$$

Démonstration.

$$\begin{aligned}
p(x_n | \mathbf{y}) &\propto p(x_n, \mathbf{y}) \\
&\propto p(y_{1:n-1}, z_n, y_{n+1:N}) \\
&\propto p(y_{1:n-1} | z_n) p(y_{n+1:N} | z_n)
\end{aligned}$$

La dernière égalité provient de la propriété d'indépendance conditionnelle du passé et du futur de la chaîne $\mathbf{Z}_N$. □

Nous montrons maintenant le lien entre la markovianité du processus $\mathbf{X}_N$ et la densité $p(y_1, y_2 | x_1, x_2)$, afin d'obtenir une meilleure compréhension du passage du modèle CMCa-BI au modèle CMCo. Pour cela nous démontrons que la markovianité peut découler de deux conditions proches, mais par des mécanismes différents. Pour cette raison, nous explicitons les démonstrations dans les deux cas.

Propriété 2.2.4. Condition 1 de Markovianité

Soit $\mathbf{Z}_N$ une CMCo stationnaire. Si $p(y_1 | x_1, x_2) = p(y_1 | x_1)$ alors $\mathbf{X}_N$ est une chaîne de Markov.

Démonstration. Nous avons par hypothèse $\forall n, p(x_{n+1} | x_n, y_n) = \frac{p(x_{n+1}, x_n, y_n)}{p(x_n, y_n)} = \frac{p(y_n | x_n) p(x_n, x_{n+1})}{p(x_n, y_n)}$, soit encore $p(x_{n+1} | x_n, y_n) = p(x_{n+1} | x_n)$. Ainsi, la densité de transition de $\mathbf{Z}_N$ s'écrit

$$p(x_{n+1}, y_{n+1} | x_n, y_n) = p(y_{n+1} | x_{n+1}, y_n, x_n) p(x_{n+1} | x_n)$$

Par conséquent, nous pouvons factoriser la densité de $\mathbf{X}_N$ en intégrant celle de $\mathbf{Z}_N$:

$$\begin{aligned}
p(\mathbf{x}) &= \int_{\mathcal{Y}^N} p(z_1) \prod_{n=2}^N p(z_n | z_{n-1}) \mu(d\mathbf{y}) \\
&= p(x_1) \prod_{n=2}^N p(x_n | x_{n-1}) \int_{\mathcal{Y}^N} p(y_1 | x_1) \prod_{n=2}^N p(y_n | y_{n-1}, x_{n-1}, x_n) \mu(d\mathbf{y}) \\
&= p(x_1) \prod_{n=2}^N p(x_n | x_{n-1})
\end{aligned}$$

□

Propriété 2.2.5. Condition 2 de Markovianité

Soit $\mathbf{Z}_N$ une CMCo stationnaire. Si $p(y_2 | x_2, x_1) = p(y_2 | x_2)$ alors $\mathbf{X}_N$ est une chaîne de Markov.

Démonstration. Il suffit de considérer la chaîne de Markov inverse $\tilde{\mathbf{Z}}_N = (\tilde{Z}_n)_{1 \leq n \leq N}$ introduite dans la remarque 2.2.1. Si la loi de $\mathbf{Z}_N$ vérifie la condition 2, cela implique que la loi de $\tilde{\mathbf{Z}}_N$ vérifie la condition 1, d'où la markovianité de $\mathbf{X}_N$.

□

Remarque 2.2.2. Une récurrence immédiate permet de montrer que la condition $p(y_2 | x_2, x_1) = p(y_2 | x_2)$ implique que $\forall n \leq N$, $p(y_n | \mathbf{x}_n) = p(y_n | x_n)$.

Lorsque l'une ou l'autre des conditions des propriétés 2.2.4 et 2.2.5 est vérifiée, nous dirons que la marginale concernée de la densité $p(y_1, y_2 | x_1, x_2)$ ne dépend que de l'état courant. La propriété 2.2.4 a déjà été montré dans [124], alors que la propriété 2.2.5 (bien qu'immédiate conséquence de 2.2.4) n'a jamais été remarquée. Nous mettons ici en évidence la différence entre les marges droite et gauche en donnant deux raisons distinctes pour lesquelles le processus caché d'une CMCo peut être une chaîne de Markov. Par conséquent, si $\mathbf{X}$ est markovien, nous ne pouvons pas conclure quant aux propriétés de $p(y_1, y_2 | x_1, x_2)$ et de ses marginales. Cependant, si nous supposons que $p_d(y_2 | x_1, x_2)$ et $p_g(y_1 | x_1, x_2)$ sont égales (les conditions 1 et 2 se confondent), nous pouvons affirmer alors que la markovianité de $\mathbf{X}$ implique que $p(y_2 | x_1, x_2) = p(y_2 | x_2)$ lorsque $\mathcal{X}$ est fini⁴ (proposition (2.3) de [124]). L'égalité des marges droite et gauche a aussi une conséquence sur les lois marginales conditionnelles des observations.

Théorème 2.2.1. Soit $\mathbf{Z}_N$ PMC stationnaire.

$p(y_2 | x_2, x_1) = p(y_1 | x_2, x_1) = p(y_i | x_i)$ pour $i = 1, 2$ si et seulement si $p(y_n | \mathbf{x}) = p(y_n | x_n)$.

Démonstration. – Nous montrons d'abord l'implication directe.

Nous savons déjà par la remarque 2.2.2 que

$$p(y_2 | x_2, x_1) = p(y_2 | x_2) \implies \forall n, p(y_n | \mathbf{x}_{1:n}) = p(y_n | x_n)$$

La même démonstration s'applique à la chaîne inverse pour laquelle nous montrons que $\tilde{\alpha}_n(\tilde{x}_n) =$

⁴Un résultat similaire a été montré dans le cas où $\mathcal{X}$ est continu mais tel que $\mathbf{Z}_N$ soit un processus gaussien [128]. Dans ce cas-là, l'égalité des marges droite et gauche n'est plus nécessaire.

$\tilde{p}(\tilde{y}_n | \tilde{x}_n)$ car nous avons $\tilde{p}(\tilde{y}_2 | \tilde{x}_2, \tilde{x}_1) = \tilde{p}(\tilde{y}_2 | \tilde{x}_2)$. D'après la propriété (2.2.3), nous avons

$$\begin{aligned} p(y_n | \mathbf{x}) &\propto \alpha_n(y_n) \beta_n(y_n) \\ &\propto \frac{p(y_n | x_n) p(y_n | x_n)}{p(y_n, x_n)} \end{aligned}$$

Donc

$$p(y_n | \mathbf{x}) = p(y_n | x_n)$$

– La réciproque est immédiate.

En effet, nous remarquons que $p(y_2 | x_2, x_1) = \int p(y_2 | \mathbf{x}) p(\mathbf{x}_{3:n} | x_2, x_1) \mu(d\mathbf{x}_{3:n}) = p(y_2 | x_2)$, et donc $\forall n, p(y_n | \mathbf{x}) = p(y_n | x_n) \implies p(y_2 | x_2, x_1) = p(y_1 | x_2, x_1) = p(y_i | x_i)$. □

Nous avons donc $p(y_n | \mathbf{x}) = p(y_n | x_n) \implies \mathbf{X}_N$ chaîne de Markov, ce qui montre la redondance des hypothèses du modèle CMCa dans le cadre des modèles CMCo. Cependant, nous n'avons pas équivalence entre ces deux hypothèses, à moins de supposer la réversibilité de $\mathbf{Z}_N$.

Nous pouvons conclure que malgré la complexité accrue des modèles couples, le calcul du MPM d'une CMCo a la même complexité algorithmique que pour une CMCa. La différence réside uniquement dans le calcul des probabilités de transition $p(z_{n+1} | z_n)$, qui est effectivement plus lourd pour les CMCo que pour les CMCa-BI.

Les CMCo qui ne sont pas des CMCa-BI présentent deux innovations par rapport à ces dernières :

1. la dépendance des observations conditionnellement aux états cachés ;
2. les lois marginales des observations peuvent dépendre de tout le processus caché, et plus seulement de l'état courant.

La première propriété est reliée à la dépendance induite par la densité $p(y_1, y_2 | x_1, x_2)$, alors que l'égalité $p(y_n | \mathbf{x}) = p(y_n | x_n)$ est reliée aux marges de $p(y_1, y_2 | x_1, x_2)$.

Ce changement suppose une modification dans la modélisation avec les CMCo. En effet, le choix des densités $p(y_n | x_n)$, appelées lois d'émission, est guidé par la connaissance physique ou empirique du phénomène observé, par exemple une loi de Rayleigh pour l'intensité rétrodiffusée par le sol en radar (voir le chapitre 3 pour la présentation de plusieurs modèles paramétriques de lois d'émission adaptés à différents types d'information disponibles sur l'environnement radar). Un modèle de Markov couple nécessite (et permet) une modélisation plus précise du phénomène observé car la loi d'émission est décomposé comme le mélange de plusieurs lois :

$$p(y_n | x_n) = \sum_{x'_{n+1}=1}^K p(x'_{n+1} | x_n) p(y_n | x_n, x'_{n+1})$$

et chaque composante représente différent comportement décrivant l'interaction entre les différentes classes. Pour la modélisation d'une image, ceci permet d'intégrer des modèles différents pour les zones frontières et pour les zones homogènes.

2.2.3 Calcul des probabilités a posteriori et des estimateurs bayésiens

Pour la simulation de la chaîne a posteriori ou le calcul du Maximum de $p(\mathbf{x}|\mathbf{y})$, nous avons besoin de connaître la loi de $\mathbf{X}_N$ conditionnellement à $\mathbf{Y}_N$. Nous savons que c'est une chaîne de Markov non-homogène dont nous pouvons calculer directement la densité de transition :

$$\begin{aligned}
 \forall n \leq N-1, p(x_{n+1}|\mathbf{x}_n, \mathbf{y}) &= \frac{p(\mathbf{y}_{n+2:N}, \mathbf{z}_{n+1})}{p(\mathbf{y}_{n+1:N}, \mathbf{z}_n)} \\
 &= \frac{p(\mathbf{y}_{n+2:N}|\mathbf{z}_{n+1})p(\mathbf{z}_{n+1})}{p(\mathbf{y}_{n+1:N}|\mathbf{z}_n)p(\mathbf{z}_n)} \\
 &= \frac{p(\mathbf{y}_{n+2:N}|z_{n+1})}{p(\mathbf{y}_{n+1:N}|z_n)}p(z_{n+1}|z_n) \\
 &\propto p(\mathbf{y}_{n+2:N}|x_{n+1}, y_{n+1})p(x_n, x_{n+1})p(y_n, y_{n+1}|x_n, x_{n+1})
 \end{aligned}$$

Soit en normalisant

$$\forall x_n, x_{n+1}, p(x_{n+1}|x_n, \mathbf{y}) = \frac{\beta_{n+1}(x_{n+1})p(x_{n+1}, x_n)p(y_{n+1}, y_n|x_n, x_{n+1})}{\int \beta_{n+1}(x'_{n+1})p(x_{n+1}, x'_{n+1})p(y_{n+1}, y_n|x_n, x'_{n+1})\delta(dx'_{n+1})} \quad (2.19)$$

Par le même calcul, nous pouvons obtenir une expression explicite pour les probabilités jointes a posteriori :

$$\begin{aligned}
 p(x_n, x_{n+1}|\mathbf{y}) &\propto p(\mathbf{y}_{n+2:N}|y_{n+1}, x_{n+1}, y_n, x_n, \mathbf{y}_{n-1})p(x_{n+1}, y_{n+1}|x_n, y_n, \mathbf{y}_{n-1})p(x_n, \mathbf{y}_n) \\
 &\propto \beta_{n+1}(x_{n+1})p(x_{n+1}, y_{n+1}|x_n, y_n)p(x_n, \mathbf{y}_n)
 \end{aligned}$$

soit en normalisant

$$\forall x_n, x_{n+1}, p(x_n, x_{n+1}|\mathbf{y}) = \frac{\alpha_n(x_n)p(y_{n+1}, x_{n+1}|y_n, x_n)\beta_{n+1}(x_{n+1})}{\int \int \alpha_n(x'_n)p(y_{n+1}, x'_{n+1}|y_n, x'_n)\beta_{n+1}(x'_{n+1})\delta(dx'_n)\delta(dx'_{n+1})} \quad (2.20)$$

Ainsi, de même que les probabilités marginales a posteriori se calcule directement par la propriété (2.2.3) :

$$\forall n, \hat{x}_{n,\text{MPM}}(\mathbf{y}) = \arg \max_x \alpha_n(x)\beta_n(x) \quad (2.21)$$

les expressions (2.19) et (2.20) permettront de calculer directement les probabilités jointes intervenant dans la simulation du processus a posteriori (ou l'estimation des paramètres, voir section 3.4.2).

Comme l'estimateur MPM, l'estimateur MAP d'une CMCo peut se calculer par une extension de l'algorithme de Viterbi [134]. Nous décrivons cet algorithme, déjà utilisé dans [53] mais sans justification, qui permet de calculer avec une complexité polynômiale la séquence des états ayant la probabilité d'occurrence maximum. Initialement, cet algorithme a été proposé pour les CMCa-BI (cette propriété est perdue pour les champs de Markov cachés, et la recherche du MAP doit être réalisée par des méthodes optimisation telles que le recuit-simulé [6]). Nous recherchons la suite

définie par

$$\hat{x}_{\text{MAP}} = (\hat{x}_{n,\text{MAP}})_{n \leq N} = \arg \max_{\mathcal{X}^N} p(\mathbf{x} | \mathbf{y})$$

La solution recherchée est aussi le maximum de la probabilité jointe $p(\mathbf{x}, \mathbf{y})$. Nous introduisons alors la quantité

$$\forall n \geq 2, \delta_n(x_n) = \max_{\mathbf{x}_{n-1}} p(x_n, \mathbf{x}_{n-1}, \mathbf{y}_n)$$

et nous notons $\delta_1(x_1) = p(x_1, y_1)$.

Proposition 2.2.2. Relation de récurrence des δ_n

$$\forall n \leq N-1, \delta_{n+1}(x_{n+1}) = \max_{x_n} p(x_{n+1}, y_{n+1} | x_n, y_n) \delta_n(x_n) \quad (2.22)$$

Démonstration. Nous montrons que la maximisation globale sur $\mathcal{X}^N$ peut être décomposée en une succession de “petites maximisations”, en utilisant la factorisation de la fonction de vraisemblance. Comme nous avons

$$\forall x_n, x_{n+1}, p(x_{n+1}, \mathbf{x}_n, \mathbf{y}_{n+1}) = p(x_{n+1}, y_{n+1} | x_n, y_n) p(x_n, \mathbf{x}_{n-1}, \mathbf{y}_n)$$

La maximisation de $p(x_{n+1}, \mathbf{x}_n, \mathbf{y}_{n+1})$ en $\mathbf{x}_n$ se décompose en :

$$\max_{\mathbf{x}_n} p(x_{n+1}, \mathbf{x}_n, \mathbf{y}_{n+1}) = \max_{x_n, \mathbf{x}_{n-1}} p(x_{n+1}, y_{n+1} | x_n, y_n) p(x_n, \mathbf{x}_{n-1}, \mathbf{y}_n)$$

soit encore

$$\max_{\mathbf{x}_n} p(x_{n+1}, \mathbf{x}_n, \mathbf{y}_{n+1}) = \max_{x_n} p(x_{n+1}, y_{n+1} | x_n, y_n) \times \delta_n(x_n)$$

□

Remarque 2.2.3. Les δ_n sont des bornes supérieures pour les probabilités jointes

$$\forall x_{n+1}, \forall \mathbf{x}_n, p(x_{n+1}, \mathbf{x}_n, \mathbf{y}_{n+1}) \leq \delta_{n+1}(x_{n+1})$$

La seconde propriété sur laquelle est basée l'algorithme de Viterbi est que l'on peut déterminer successivement les états $\hat{x}_{n,\text{MAP}}(\mathbf{y})$ en parcourant la chaîne de manière rétrograde.

Proposition 2.2.3. Si $\mathbf{y}$ est observé et si nous notons $(\hat{x}_{n,\text{MAP}}(\mathbf{y}))_{1 \leq n \leq N} = \arg \max_{\mathbf{x} \in \mathcal{X}^N} p(\mathbf{x} | \mathbf{y})$, alors nous avons

$$\hat{x}_{N,\text{MAP}}(\mathbf{y}) = \arg \max_x \delta_N(x)$$

et

$$\hat{x}_{N-1,\text{MAP}}(\mathbf{y}) = \arg \max_{x_{N-1}} p(\hat{x}_{N,\text{MAP}}(\mathbf{y}), y_N | x_{N-1}, y_{N-1}) \delta_{N-1}(x_{N-1})$$

Démonstration. La première affirmation est vraie par définition de δ_N . La seconde affirmation provient de la relation de récurrence entre les vraisemblances du processus $Z_n : p(\mathbf{z}) = p(z_N | z_{N-1}) p(\mathbf{z}_{N-1})$. Si $\hat{x}_{N,\text{MAP}}(\mathbf{y})$ est connu, la recherche de la valeur à l'instant $N-1$ est obtenue par la maximisation

suiivante (en notant $z_N = (\hat{x}_{N,\text{MAP}}(\mathbf{y}), y_N)$)

$$\arg \max_{x_{N-1}, \mathbf{x}_{N-2}} p(z_N | z_{N-1}) p(\mathbf{z}_{N-1}) = \arg \max_{x_{N-1}} \{p(\hat{x}_{N,\text{MAP}}(\mathbf{y}), y_N | x_{N-1}, y_{N-1}) \delta_{N-1}(x_{N-1})\}$$

□

L'algorithme de Viterbi se déroule en deux temps : le calcul récursif des bornes supérieures δ_n , puis la détermination récursive (rétrograde) des $\hat{x}_{n,\text{MAP}}$.

Algorithme 2.2.1. *Algorithme de Viterbi*

$$\forall x_1 \in \mathcal{X}, \delta_1(x_1) = p(x_1, y_1)$$

$$\forall n \leq N, \forall x_n \in \mathcal{X}, \delta_n(x_n) = \max_{x_{n-1}} \{\delta_{n-1}(x_{n-1}) \times p(x_n, y_n | x_{n-1}, y_{n-1})\}$$

Nous avons alors

$$\hat{x}_{N,\text{MAP}} = \arg \max_x \delta_N(x)$$

et

$$\forall n \leq N-1, \hat{x}_{n,\text{MAP}} = \arg \max_{x_n} \{\delta_n(x_n) \times p(\hat{x}_{n+1,\text{MAP}}, y_{n+1} | x_n, y_n)\}$$

Comme dans sa version classique, la complexité de cet algorithme est $O(NK^2)$: le calcul des N fonctions δ_n nécessite K^2 multiplications. Il suffit ensuite de chercher N fois un maximum parmi K valeurs. Nous pouvons proposer des procédures d'estimation bayésienne MPM et MAP très générales⁵, la différence entre les CMCa et les CMCo portent uniquement sur le calcul de la matrice de transition du processus complet.

2.3 Modèles triplets

L'inférence bayésienne dans les modèles à données manquantes est facilitée lorsque le processus a posteriori est markovien : il est possible soit de calculer exactement la loi a posteriori, soit de l'échantillonner de telle sorte que l'on puisse approcher par Monte Carlo n'importe quel estimateur. Cependant la markovianité du couple peut se révéler être une hypothèse trop forte pour $\mathbf{Z}_N$ (ou pour $\mathbf{X}_N$ conditionnellement à $\mathbf{Y}_N$). Les modèles triplet sont une extension des modèles couples, pour lesquelles il est possible de calculer "facilement" les probabilités $p(\mathbf{x}|\mathbf{y})$ ou de les échantillonner [16, 123, 129].

⁵et donc conserver les mêmes programmes informatiques pour le calcul des probabilités α_n, β_n , et de $\hat{\mathbf{x}}_{\text{MPM}}$ et $\hat{\mathbf{x}}_{\text{MAP}}$ dans les modèles CMCa et CMCo. Il suffit de modifier les procédures de calcul des matrices $p(z_{n+1} | z_n)$ pour $n = 1..N-1$.

Définition 2.3.1. Champ de Markov Triplet

Soit $\mathbf{Y}_S$ un champ aléatoire observé et $\mathbf{X}_S$ le champ aléatoire des états cachés. Nous dirons que $\mathbf{Y}_S$ est un champ de Markov Triplet (CMT) si il existe un processus auxiliaire $\mathbf{U}_S = (U_s)_{s \in S}$ à valeur dans $\mathcal{U}$ (fini ou continu) tel que le processus joint $\mathbf{Z}_S = (X_s, U_s, Y_s)_{s \in S}$ soit un champ de Markov.

Si $S = [1..N]$, nous dirons que $\mathbf{Y}_N$ est une chaîne de Markov triplet. Le processus $\mathbf{U}_N$ est un processus latent inobservé similaire à $\mathbf{X}_N$, et tous les processus marginaux sont donc markoviens conditionnellement aux autres. En particulier, $(\mathbf{U}_N, \mathbf{X}_N)$ est markovien conditionnellement à $\mathbf{Y}_N$: nous pouvons donc calculer les probabilités $p(x_n, u_n | \mathbf{y})$ et $p(x_n, u_n | \mathbf{y}, x_{n-1}, u_{n-1})$ par les récursions avant et arrière. La densité $p(x_n | \mathbf{y})$ s'obtient alors en intégrant la densité $p(x_n, u_n | \mathbf{y})$, ce qui est fait facilement lorsque l'espace $\mathcal{U}$ est fini, mais devient beaucoup plus difficile lorsque $\mathcal{U}$ est continu. Cela nécessite alors l'utilisation de méthodes de simulation et d'approximation telles que les méthodes MCMC ou Monte Carlo séquentiel [60, 138].

Un modèle triplet peut donc se voir comme un modèle couple $((\mathbf{U}_N, \mathbf{X}_N), \mathbf{Y}_N)$ dont le processus caché $(\mathbf{U}_N, \mathbf{X}_N)$ ne nous intéresse que partiellement. Nous pouvons dire de manière équivalente qu'un modèle triplet est un processus markovien partiellement (ou imparfaitement⁶) observé i.e. il existe un processus markovien $\mathbf{Z}_N$ à valeurs dans un espace $\mathcal{Z}$ (muni d'une tribu $\mathcal{B}(\mathcal{Z})$) et une fonctions φ mesurable et déterministe telles que nous observons $\forall n, Y_n = \varphi(Z_n)$. Le modèle triplet correspond au cas où nous pouvons faire une décomposition "cartésienne" des variables $Z_n = (U_n, X_n, Y_n)$ et où la fonction φ correspond à la projection π_Y selon la dernière coordonnée. Notre objectif est d'estimer la fonction $\pi_X(Z_n)$, où π_X est la projection selon la deuxième coordonnée. Inversement, un processus markovien Z_n à valeurs dans $\mathcal{Z}$ telle que nous n'observons que $\varphi(Z_n)$, avec $\varphi : \mathcal{Z} \rightarrow \mathcal{Y}$ (surjective) peut être considéré comme un processus triplet. Si nous choisissons une fonction $\varphi^\perp : \mathcal{Z} \rightarrow \mathcal{V}$ (surjective), telle que la fonction $z \mapsto (\varphi(z), \varphi^\perp(z))$ (de $\mathcal{Z} \rightarrow \mathcal{Y} \times \mathcal{V}$) soit une bijection⁷ alors $Y_n = \varphi(Z_n)$ sera une chaîne de Markov Triplet. Si nous voulons estimer un processus $h(Z_n)$, nous posons $X_n = h(Z_n)$, et alors $(\varphi^\perp(Z_n), X_n, Y_n)$ est une chaîne de Markov.

Les modèles triplets (ou modèles markoviens partiellement observés) rassemblent de multiples extensions des MMCa introduits pour affaiblir les hypothèses des modèles classiques (MMCa-BI). Ainsi, selon la définition du processus auxiliaire $\mathbf{U}_N$, nous sommes en mesure de retrouver de nombreux modèles visant à relâcher la stationnarité (la matrice de transition du processus $\mathbf{X}_N$ évolue en fonction de U_n [99]), la markovianité de $\mathbf{X}_N$ (U_n est le temps de séjour dans chaque état et ne suit plus nécessairement une loi exponentielle [127, 66, 84]) ou encore l'indépendance conditionnelle (le processus $(\mathbf{Y}_N, \mathbf{U}_N)$ est une CMCa-BI conditionnellement à $\mathbf{X}_N$ [27]). Une liste des différentes généralisations des CMCa-BI et leur réécriture sous forme de modèle triplet est faite dans [129].

Un cas particulièrement intéressant est celui où le processus auxiliaire est continu, car cela permet de formuler des modèles statistiques complexes en considérant que ce sont des observations partielles de processus simples. Ainsi, si le couple $(X_n, Y_n)_{n \geq 1}$ est une chaîne de Markov, alors le processus $(Y_n)_{n \geq 1}$ conditionnellement à $(X_n)_{n \geq 1}$ est un processus markovien. Un modèle simple

⁶pour reprendre la terminologie employée par L. Younes dans [158], concernant l'estimation par maximum de vraisemblance de modèles plus généraux que les champs de Markov cachés.

⁷c'est un changement de coordonnées non-linéaire

de processus markovien est le processus autorégressif linéaire d'ordre 1 que nous pouvons écrire $Y_{n+1} = a(X_n)Y_n + \sigma(X_n)\epsilon_n$ conditionnellement à $\mathbf{X}_N$, $(\epsilon_n)_n$ étant un bruit blanc indépendant de $(X_n)_{n \geq 1}$, de variance 1. Nous pouvons, afin d'améliorer le modèle, utiliser des modèles AR d'ordre p , que nous pouvons toujours écrire sous forme matricielle $\check{Y}_{n+1} = a(X_n)\check{Y}_n + \sigma(X_n)\epsilon_n$ (les observations que nous considérons sont alors les vecteurs $(\check{Y}_n)_{n \geq 1}$, et $a(X_n)$ et $\sigma(X_n)$ sont des matrices dépendantes des états). Cependant la modélisation auto-régressive peut s'avérer insuffisante pour décrire la richesse du processus et nous pouvons considérer qu'il est plutôt de type ARMA, i.e.

$$\forall n, Y_n - \sum_{k=1}^p a_k(X_n)Y_{n-k} = \epsilon_n - \sum_{l=1}^q b_l(X_n)\epsilon_{n-l} \quad (2.23)$$

avec $(\epsilon_n)_{n \in \mathbb{Z}}$ bruit blanc. Or, tout processus ARMA admet alors une représentation "espace d'états" :

$$\begin{cases} \check{U}_{n+1,p} &= A(X_n)\check{U}_{n,p} + C(X_n)\epsilon_{n+1} \\ Y_n &= B(X_n)\check{U}_{n,p} \end{cases} \quad (2.24)$$

avec $\check{U}_{n,p} = (U_n \dots U_{n-p})$. Ainsi si $\mathbf{Y}$ conditionnellement à $\mathbf{X}$ est un processus ARMA (de degrés p, q constants), alors c'est aussi un processus triplet en considérant le processus $\mathbf{U}$ stationnaire vérifiant la représentation (2.24). Le processus $\mathbf{U}$ est un moyen de varier et d'augmenter la complexité des modèles utilisées pour décrire la dynamique de $\mathbf{Y}$ conditionnellement à $\mathbf{X}$: l'utilisation de modèle MMCA plus généraux (non-linéaires) que (2.24) est donc envisageable :

$$\begin{cases} \check{U}_{n+1,p} = f(\check{U}_{n,p}, \epsilon_{n+1}, A(X_n)) \\ Y_n = h(\check{U}_{n,p}, \zeta_n, B(X_n)) \end{cases} \quad (2.25)$$

avec A, B des paramètres dépendant des états cachés $\mathbf{X}$, et (ζ_n) un bruit blanc normalisé.

Les modèles triplets permettent aussi une modélisation plus réaliste du signal radar, présentée dans [27, 29]. Nous savons que la famille des vecteurs aléatoires sphériquement invariants (voir section 4.2.1) donne une modélisation des trains d'onde en accord avec les données [44]. Il est donc souhaitable qu'un modèle markovien pour la segmentation de signaux soit en accord avec cette propriété. Soit $(\mathbf{X}_N, \mathbf{Y}_N)$ le processus joint, tel que les observations soient des vecteurs aléatoires sphériquement invariants d'émission de moyenne nulle. Nous savons qu'il existe alors un troisième processus $\mathbf{U}_N$, appelé texture, tel que

$$\forall n \leq N, Y_n = U_n^{1/2} \epsilon_n \quad (2.26)$$

avec $\epsilon_n \sim N(0, \Sigma_{x_n})$ (le speckle), et U_n de densité $g(u, \theta_{x_n})$. Si nous supposons que les couples $(U_n, \epsilon_n)_n$ sont indépendants les uns des autres conditionnellement à $\mathbf{X}$, et que $\mathbf{X}$ est markovien, alors $(\mathbf{X}, \mathbf{Y})$ est une CMCA-BI. Cependant, nous savons que les fouillis radar (i.e. les signaux réfléchies par les obstacles naturelles) sont corrélés spatialement (même en environnement homogène), et que cette dépendance est due au processus de texture. Pour modéliser cette dernière, nous faisons l'hypothèse que $\mathbf{U}$ est une chaîne de Markov conditionnellement à $\mathbf{X}$ (une hypothèse de Markovianité similaire a déjà été proposé pour la simulation d'une texture corrélée avec marges

gamma dans [104], ou encore dans [153]). Ainsi, la texture n'est pas observable directement, mais l'est par le biais de la transformation (2.26).

L'ajout de la dépendance dans la texture complique les procédures de segmentation et d'estimation, en raison de son caractère continu. En effet, il n'est pas possible (hormis dans le cas gaussien) de calculer analytiquement les probabilités avant et arrière de (x_n, u_n) . Malgré cela, nous avons montré dans [27] grâce à un algorithme proposé par Perez et Vermaak [120], et par un algorithme d'échantillonnage Monte-Carlo séquentiel que nous pouvons calculer une approximation des probabilités avant $\alpha_n(x, u)$ et arrière $\beta_n(x, u)$ et donc de la densité de lissage $p(x_n | \mathbf{y}_N)$. De plus, ces modèles sont plus fidèles à la réalité, ce qui permet d'évaluer les impacts de la non-prise en compte de la corrélation dans les traitements [29].

Les deux exemples précédents montrent que les modèles triplets permettent de décrire des situations complexes, tout en conservant l'existence de procédures récursives pour le calcul des probabilités a posteriori. La recherche de modèles encore plus généraux possédant cette propriété a conduit à l'introduction des modèles partiellement markoviens, pour lesquelles la seule propriété requise est la markovianité a posteriori du processus caché, [126].

Plus généralement, la théorie des modèles graphiques (voir Lauritzen [100]) montre que la capacité à calculer les probabilités d'intérêt de lois multivariées spécifiées par des lois conditionnelles tient au fait qu'il soit possible de factoriser la probabilité jointe de variables observées ou cachées selon un graphe. Il est alors possible dans certains cas (lorsque le graphe ne possède pas de boucle) d'étendre les récursions avant-arrière pour effectuer la marginalisation (i.e. le calcul des densités marginales d'intérêt) : c'est l'algorithme de propagation des croyances de Pearl (*belief propagation*) qui permet de calculer exactement les probabilités a posteriori. Cet algorithme a été étendu aux graphes possédant des boucles sous le nom de *loopy belief propagation*, et permet de donner une estimation des probabilités recherchées en itérant le processus de propagation jusqu'à convergence [155]. La compréhension du succès en pratique de cet algorithme est le sujet actuel de nombreux développements, [157, 92].

Chapitre 3

Estimation statistique des CMCa

3.1 Introduction

Nous avons présenté dans la partie précédente les règles de décision que nous utilisons pour segmenter une image ou un signal. Celles-ci sont basées sur une modélisation probabiliste des observations et des états cachés et utilisent la loi de $\mathbf{X}_N$ conditionnellement aux observations $\mathbf{Y}_N$. Dans le cadre dans lequel nous travaillons, les densités a posteriori $p(\mathbf{x}|\mathbf{y})$ sont inconnues et nous devons d'abord les estimer avant de pouvoir segmenter une série d'observations $\mathbf{y}$. Pour cette raison, nous faisons l'hypothèse que le processus $\mathbf{Y}_N$ (ou $\mathbf{Z}_N$) est stationnaire, et que les densités $p(y_1, y_2 | x_1, x_2)$ et $p(x_1, x_2)$ appartiennent à des modèles paramétriques. Il est alors possible, et c'est le but de cette partie, d'estimer ces paramètres à partir des seules observations $\mathbf{y}$ et de calculer un estimateur *plug-in* (par injection) des probabilités a posteriori d'appartenance aux classes $\hat{p}(x_n | \mathbf{y})$ en utilisant ces estimations des paramètres. La segmentation réalisée avec les probabilités $(\hat{p}(x_n | \mathbf{y}))_{1 \leq n \leq N}$ est appelée segmentation non-supervisée, par opposition à la segmentation supervisée effectuée avec les vrais probabilités.

Il peut paraître détourné dans le cadre de la segmentation d'une image, d'estimer les paramètres des lois d'émission, alors que nous voulons simplement avoir un estimateur des probabilités a posteriori. Cette approche est motivée par la possibilité d'utiliser les nombreux outils de la statistique paramétrique, et de bénéficier de résultats théoriques puissants (pour l'estimation, notamment au sujet du maximum de vraisemblance) [45]. De plus, la modélisation paramétrique donne des résumés compacts des classes, permettant de les caractériser. Cependant, les résultats des procédures d'estimation sont à manipuler avec précaution, parce que l'estimation correcte des paramètres n'assurent pas forcément une estimation correcte des probabilités a posteriori. Il est difficile de relier la qualité et la variabilité des estimateurs des paramètres des lois d'émission, à celles des estimateurs $\hat{p}(x_n | \mathbf{y})$ et des règles de segmentation qui leur sont associées. Pour cette raison, nous nous intéresserons dans les chapitres suivants, à l'écart entre taux d'erreur en segmentation supervisée et non-supervisée.

Si nous revenons au problème statistique de l'estimation d'une CMCa-BI (ou de manière générale d'une CMCo), celui-ci sort du cadre classique de l'estimation paramétrique, et s'inscrit dans la problématique de l'estimation des modèles à données manquantes (ou à variable latente), qui n'est pas sans poser des problèmes théoriques et pratiques importants. Si depuis les travaux de

Baum et Petrie [13], la méthode du maximum de vraisemblance est largement utilisée pour l'estimation des nombreux modèles de Markov cachés proposés depuis les années 60¹, ce n'est qu'assez récemment qu'ont été montrées les bonnes propriétés de cet estimateur dans un contexte général (voir [32] pour un panorama des récents résultats en inférence des CMCA). Il n'en reste pas moins que la recherche du maximum global de la vraisemblance constitue une réelle difficulté à laquelle des solutions partielles ont été proposées, parmi lesquelles l'algorithme EM et ses nombreuses variantes.

Parallèlement, le développement de méthodes d'échantillonnage performantes et flexibles : MCMC et échantillonnage d'importance (*Importance Sampling*) [138], Monte Carlo séquentiel [60], permet à l'approche bayésienne de fournir un estimateur dans le cadre de modèles très complexes, et se prête particulièrement bien à l'inférence des MMCA. De plus, le nombre de classes peut être estimé conjointement aux paramètres du modèle grâce aux chaînes de Markov Monte Carlo à saut réversible (*Reversible Jump Markov Chain Monte Carlo* ou RJMCMC, introduit par Green dans [80]), en considérant plusieurs modèles et en "sautant" de l'un à l'autre [136]. Finalement, l'estimation bayésienne en segmentation statistique donne des méthodes globales que l'on appelle pleinement bayésienne, dans lesquelles processus caché et paramètres sont estimés conjointement. Ce traitement permet d'éviter les reproches faits à une segmentation non-supervisée en deux étapes, mais la complexité d'implémentation et la lourdeur des simulations ne permettent pas toujours d'envisager cette approche, notamment en traitement d'images où la taille des données peut être très grande, ou en traitement radar sous la contrainte d'applications temps réel.

L'une comme l'autre de ces méthodes butent sur la complexité de la vraisemblance des CMCA : alors que l'algorithme EM et ses variantes ne cherchent que le maximum de la vraisemblance, les méthodes bayésiennes explorent la vraisemblance dans sa totalité. La difficulté de ce problème incite à utilisation des fonctions plus faciles à explorer, choisies pour la complexité des calculs qu'ils nécessitent. Pour contourner ce problème de maximisation, le principe d'Estimation Conditionnelle Itérative introduit dans [122], propose de calculer l'espérance d'un estimateur défini sur les données complètes conditionnellement aux données observées. Cette opération est répétée jusqu'à la stabilisation de la suite des itérés, en s'inspirant de la dynamique de l'algorithme EM avec lequel il est équivalent sous certaines conditions, [50]. L'avantage de cette méthode est de permettre l'emploi d'estimateur ad hoc, ce qui évite d'avoir à calculer les vraisemblances, et à les maximiser.

L'objectif de cette partie est de présenter les méthodes d'estimation que nous avons utilisé par la suite dans le cadre des CMCA-BI (chapitre 3) et des CMCA (Chapitre 4), et nous mettons en évidence les ressorts théoriques sur lesquelles elles reposent. Nous exposons d'abord les résultats théoriques existants pour l'EMV, qui motivent son emploi pour l'estimation des CMCA-BI, ainsi que pour les processus autorégressifs à changements de régime markovien (ou *Switching Autoregressive process*, SAR) qui sont un cas particulier de CMCA. Nous rappelons alors l'algorithme EM et quelques variantes pour le calcul de l'EMV des modèles à données manquantes et nous donnons son expression particulière pour les mélanges finis.

Nous étudions ensuite une méthode alternative et originale d'estimation des modèles à données manquantes à l'aide de la théorie des fonctions estimantes. Cette approche a été proposée de manière très générale, par Heyde et Morton dans [90]. Nous développons particulièrement dans

¹Pour un aperçu des travaux théoriques ou appliqués relatifs aux HMM sur les 10 dernières années, voir la bibliographie "10 years of HMM", accessible sur internet <http://www.tsi.enst.fr/~cappe/docs/hmmbib.html>

cette partie la projection, au sens de l'espérance conditionnelle, des fonctions estimantes. Ceci nous permet de définir clairement une nouvelle de familles de fonctions estimantes pour les modèles à données manquantes, et d'en donner les premières propriétés de base. Nous donnons aussi de premiers éléments pour la démonstration de la consistance des estimateurs que l'on peut déduire.

Nous proposons alors une méthode de résolution de ces fonctions estimantes, appelée Estimation Conditionnelle Itérative (ECI), qui permet de retrouver en particulier l'algorithme EM. Nous donnons certains résultats sur le comportement asymptotique de cette méthode, similaires à ceux existants pour l'algorithme EM. Nous montrons aussi comment l'algorithme ECI permet de reformuler le principe d'Estimation Conditionnelle Itérative, introduit par Pieczynski dans [122], et de retrouver d'autres algorithmes précédemment proposés pour l'estimation de modèles à données manquantes.

3.2 Maximum de vraisemblance dans les CMCA

Dans le contexte de la segmentation non-supervisée, nous sommes en premier lieu intéressés par l'obtention d'un estimateur ponctuel des paramètres. Pour cette raison, nous rappelons des conditions de consistance forte de l'EMV, obtenues pour les CMCA-BI par Leroux ([102]), dont l'avantage est d'être assez peu exigeantes. Nous signalons aussi la consistance forte et faible de l'EMV pour les SAR démontrée récemment par Douc, Moulines et Ryden [59] dans un cadre assez général. L'intérêt de ces résultats est de justifier dans un grand nombre de situations pratiques le "bon comportement" des procédures utilisées. Outre l'intérêt propre de la connaissance de la vitesse de convergence de l'EMV par un résultat de normalité asymptotique, celle-ci est reliée à l'écart observée entre taux d'erreur de segmentation supervisée et non-supervisée. Cependant, la complexité du lien entre la variance des estimateurs utilisés et le taux d'erreur est tel qu'il est souvent nécessaire pour le praticien d'avoir recours à des simulations pour s'assurer en dernier lieu de la fiabilité et de l'intérêt des procédures proposées. Pour cette raison, l'étude de la qualité des estimateurs est complétée par une étude de la variabilité des taux d'erreur en segmentation non-supervisée.

Nous décrivons tout d'abord la structure générale du modèle paramétrique de CMCA-BI que nous utilisons dans les applications (voir Chapitre 3) et nous commentons les hypothèses de Leroux qui garantissent la consistance de l'EMV pour ce type de modèle. Dans un second temps, nous décrivons le cas particulier de CMCA pour lequel Douc et al ont montré la convergence de l'EMV.

3.2.1 Le modèle CMCA-BI

Le processus caché $\mathbf{X} = (X_n)_{n \geq 1}$ est une chaîne de Markov stationnaire, dont la matrice de transition est notée A . Nous supposons qu'elle admet une unique loi stationnaire $\pi \in \mathcal{S}$ où $\mathcal{S} = \left\{ (\pi_i)_{1 \leq i \leq K} \in \mathbb{R}^K \mid \sum_{i=1}^K \pi_i = 1, 0 \leq \pi_i \leq 1 \right\}$.

Le processus des observations est noté $\mathbf{Y} = (Y_n)_{n \geq 1}$ et nous supposons que chaque densité $p(y_1 | x_1 = k)$ appartient à un modèle paramétrique $\{f(y, \theta_k), \theta_k \in \Theta_k\}$, $k = 1..K$. Le modèle CMCA-BI est paramétré par le vecteur de paramètres $\phi = (A, \theta = (\theta_k)_{1 \leq k \leq K})$ appartenant à l'espace $\Phi = \mathcal{S}^K \times \Theta_1 \times \dots \times \Theta_K$. Nous supposons que les espaces Θ_k sont des parties de $\mathbb{R}^{d_k}$, $d_k \geq 1$. Dans la quasi-totalité des applications traitées par la suite, toutes les lois appartiennent au même

modèle statistique indexé par un ensemble Θ .

Cette restriction sur les types des lois de chaque classe est très largement répandue dans les applications et a pour justification de simplifier la mise en oeuvre informatique et de faciliter l'interprétation des classes (il suffit de comparer les paramètres estimés). Cependant, il existe des cas où l'utilisation de types de loi différents a une interprétation précise. Par exemple dans le cadre de la classification de données directionnelles, McLachlan (voir [109], chapitre 11 et les références à l'intérieur) décrit les classes d'intérêt par des lois de Von Mises-Fisher (caractérisées par une direction moyenne et un coefficient de dispersion) et introduit une classe avec une loi uniforme sur la sphère correspondant à une classe "bruit de fond". Il s'agit de situations où une réflexion sur le comportement physique (aléatoire) des classes est traduite en terme de choix de modèles. Le même travail a été effectué en télédétection, pour décrire les différences de fluctuations de l'intensité rétrodiffusée selon les zones géographiques dans des images SAR (Delignon, [49]). Les mélanges généralisés, étudiés par Giordana et Pieczynski dans [78], sont des modèles dans lesquelles l'a priori sur le type de loi pertinent pour chaque classe est beaucoup moins fort : plusieurs modèles paramétriques de lois sont en concurrence, et c'est lors de l'étape d'estimation qu'est choisi le modèle paramétrique collant le mieux aux données, grâce à un critère de décision (basé sur la distance de Kolmogorov par exemple). La même idée a été développée dans [48], en utilisant le système de Pearson² $\mathcal{P}$, ce qui permet de sélectionner le modèle le plus adapté conjointement à la phase d'estimation. Finalement, $\mathbf{Y}$ est un processus stationnaire tel que la densité de Y_n soit le mélange paramétrique

$$p(y_n, \phi) = \sum_{k=1}^K \pi_k f(y_n, \theta_k) \quad (3.1)$$

et dont la structure de dépendance est décrite par la matrice A . Nous notons $p(\mathbf{y}_n, \phi)$ la densité de $\mathbf{Y}_n = (Y_1, \dots, Y_n)$ égale à

$$p(\mathbf{y}_n, \phi) = \sum_{x_1, \dots, x_N} \pi_{x_1} f(y_1, \theta_{x_1}) \prod_{i=2}^N a_{x_{i-1}x_i} f(y_i, \theta_{x_i}) \quad (3.2)$$

Le modèle CMCA-BI est donc un prolongement au cas dépendant du modèle statistique de mélange utilisé pour la classification de données indépendantes.

3.2.2 Les hypothèses de Leroux

La démonstration de Leroux de la consistance forte³ de l'EMV suit la stratégie en trois étapes de Wald, [150] :

1. Convergence de la log-vraisemblance normalisée $\frac{1}{n} \log(p(\mathbf{Y}_n, \phi))$ vers une fonction $\lambda(\phi)$, P_{ϕ_0} presque-sûrement (ϕ_0 est le vrai paramètre) ;
2. Vérification que cette fonction limite λ est un contraste, i.e. vérifie

$$\lambda(\phi) \leq \lambda(\phi_0) \text{ et } (\lambda(\phi) = \lambda(\phi_0) \implies \phi = \phi_0) ;$$

²L'intérêt de ce modèle réside dans la réunion, dans le même formalisme, de familles paramétriques ayant des expressions et des supports différents (les lois normales, gamma et beta appartiennent à $\mathcal{P}$). Ceci interdit par contre une estimation par maximum de vraisemblance et l'estimation se fait alors par les moments.

³La consistance forte désigne la convergence presque-sûre, et la consistance faible consiste en la convergence en loi.

3. Convergence du maximum de $\frac{1}{n} \log(p(\mathbf{Y}_n, \phi))$ vers celui de λ , i.e.

$$\lim_{n \rightarrow \infty} \arg \max_{\phi} \left(\frac{1}{n} \log(p(\mathbf{Y}_n, \phi)) \right) = \arg \max_{\phi} \lambda(\phi)$$

Nous développons dans cette section les hypothèses de Leroux qui permettent l'aboutissement de ce plan, en donnant, lorsque c'est le cas, une traduction pratique de la pathologie que les hypothèses cherchent à éviter pour garantir le bon comportement de l'EMV.

L'accomplissement de l'étape 1 repose sur la propriété d'ergodicité⁴ du processus $\mathbf{Y}$, ainsi que sur l'existence de certains moments qui permettent d'étendre la propriété de convergence au processus de log-vraisemblance $\log p(\mathbf{Y}_n)$. L'ergodicité du processus $\mathbf{Y}$ découle de l'ergodicité du processus caché $\mathbf{X}$: pour cela, on suppose que la matrice A est irréductible (cf. annexe C) ce qui implique que le processus $\mathbf{X}$ est positif récurrent. En particulier, cela impose que la loi stationnaire π charge tous les états ($\forall k, \pi_k > 0$), et donc que le nombre d'états K soit correct. Pour avoir la convergence presque-sûre de la log-vraisemblance normalisée, Leroux suppose aussi que :

- $\forall k \leq K, E_{\phi_0} [|\log(f(Y_1, \theta_k))|] < \infty$, où ϕ_0 est le vrai paramètre ;
- $\forall \theta_k \in \Theta_k, \exists \delta > 0, E_{\phi_0} \left[\sup_{\|\theta_k - \theta'_k\| < \delta} \left(\log(f(Y_1, \theta'_k)) \right)^+ \right] < \infty$, où $x^+ = \max(0, x)$.

L'étape 2 consiste à montrer que la fonction limite λ est une divergence de Kullback-Leibler entre des lois de processus. La possibilité de conclure que $\lambda(\phi) = \lambda(\phi_0) \implies \phi = \phi_0$ repose alors sur l'identifiabilité du modèle paramétrique, i.e. l'injectivité de la paramétrisation : si $p(\mathbf{y}, \phi) = p(\mathbf{y}, \phi_0)$ presque-sûrement alors $\phi = \phi_0$ ($p(\mathbf{y}, \phi_0)$ est la densité du processus $\mathbf{Y}$). L'identifiabilité de la loi du processus $\mathbf{Y}$ découle de l'identifiabilité du mélange (3.1), grâce à un résultat de Teicher sur l'identifiabilité des produits de mélange de lois [148]. Cependant l'identifiabilité d'un mélange ne peut être définie qu'à une permutation près des classes, mais cette difficulté peut être contournée en considérant la convergence de l'EMV modulo la relation d'équivalence de permutations d'indices des paramètres. Il est donc supposé l'identifiabilité forte du modèle, définie par

$$\sum_{k=1}^K \pi_k f(y, \theta_k) = \sum_{k=1}^K \pi'_k f(y, \theta'_k) \mu - \text{ps} \implies \forall k \leq K, \pi_k = \pi'_k \text{ et } \theta_k = \theta'_k \quad (3.3)$$

Remarque 3.2.1. *L'invariance par permutation des indices de la loi du processus pose aussi des problèmes dans la recherche numérique du maximum de vraisemblance. L'invariance signifie que nous avons $K!$ maxima globaux de la log-vraisemblance et non plus un seul (lorsqu'il en existe un), ce qui en complique la recherche en pratique. En effet, pour résoudre les équations normales définissant l'EMV, nous sommes amenés à construire des suites d'estimation des paramètres qui amélioreraient en moyenne la vraisemblance de l'échantillon (par exemple avec l'échantillonnage MCMC pour l'estimation bayésienne), et l'on essaye souvent d'améliorer la précision de l'estimateur en moyennant sur les dernières itérations considérées proches d'un maximum. La multiplication des maxima rend cette opération potentiellement dangereuse, car la moyennisation peut donner alors un estimateur final sous-optimal situé entre deux maxima de la log-vraisemblance. Ce phénomène*

⁴ Soit $\mathbf{Y} = (Y_n)_{n \in \mathbb{Z}} : (\Omega, \mathcal{A}, P) \longrightarrow (\mathcal{Y}^{\mathbb{Z}}, \mathcal{B}(\mathcal{Y}^{\mathbb{Z}}))$ un processus stationnaire et T l'opérateur de décalage sur $(\mathcal{Y}^{\mathbb{Z}}, \mathcal{B}(\mathcal{Y}^{\mathbb{Z}}))$, tel que si $\mathbf{y} = (y_n)_{n \in \mathbb{Z}} \in \mathcal{Y}^{\mathbb{Z}}$, nous ayons $T(\mathbf{y}) = (y_{n+1})_{n \in \mathbb{Z}}$. Nous notons $\mathcal{I} = \{A \mid T^{-1}(A) = A\}$ la tribu des invariants de T , et $\mathcal{J} = \mathbf{Y}^{-1}(\mathcal{I})$ la tribu image. Le processus $\mathbf{Y}$ est ergodique si la tribu $\mathcal{J}$ est telle que $B \in \mathcal{J} \implies P(B) = 0 \text{ ou } 1$. Ceci implique alors le théorème ergodique de Birkoff qui affirme que $\frac{1}{n} \sum_{k=1}^n f(Y_k) \longrightarrow E[f(Y_1)]$ P -presque-sûrement (et aussi dans L^1), pour tout fonction f mesurable et intégrable.

impose donc de vérifier les zones dans lesquels évolue l'algorithme d'estimation. Ce problème est intrinsèque aux modèles de mélange, et apparaît même lorsqu'il n'y a pas de maxima locaux.

Enfin, la démonstration de la troisième étape repose de manière essentielle sur les propriétés de continuité en θ des densités $f(y, \theta)$, et de compacité de l'espace des paramètres Φ . En effet, l'existence d'un maximum de la log-vraisemblance $\hat{\phi}_n$ pour tout n est garantie si les densités des lois d'émission (et donc la vraisemblance) sont continues et que l'espace des paramètres est compact⁵. La nécessité de l'hypothèse de compacité est illustrée par l'exemple classique (donné par Robert dans [137]) d'un mélange de lois normales, paramétrées par les moyennes et les variances (m_i, σ_i^2) . Si nous prenons $\Theta = \mathbb{R} \times \mathbb{R}^{+*}$, la vraisemblance calculée avec un échantillon $(y_1, \dots, y_N)$ n'est pas bornée supérieurement, ce que l'on peut voir en prenant pour $m_1 = y_1$ et en faisant tendre σ_1^2 vers 0. Il est donc nécessaire dans ce cas de borner et de fermer l'espace des variances en excluant 0. De même pour les proportions, il faut se limiter à un compact "loin" des bords du simplexe $\mathcal{S}$ puisque nous considérons des mélanges dont les proportions sont toutes strictement positives. Ainsi, en toute rigueur, la recherche du maximum de la log-vraisemblance doit se faire sous contrainte d'appartenance à un compact convenablement défini selon le type de modèle. En pratique, cette maximisation est faite sans contrainte, et les "solutions divergentes" sont écartées.

3.2.3 D'autres résultats de consistance pour les CMCa

Les techniques de démonstration utilisées par Mevel [112], et généralisées par Douc et Matias ont permis de traiter le cas des modèles CMCa-BI non-stationnaires (i.e. $P(X_1) \neq \pi$), ainsi que celui où l'espace d'états est seulement compact. Cette généralisation justifie donc l'utilisation de l'estimateur du maximum de vraisemblance dans les modèles à espaces d'états continus, fermés et bornés (si nous sommes dans $\mathbb{R}^d$) et pas seulement finis. Cependant, les modèles couramment utilisés en statistique et traitement du signal (notamment dans le contexte du filtrage de Kalman) ont des espaces d'états non-bornés, ce qui limite la portée pratique de ces résultats.

Les consistances forte et faible de l'EMV ont pu être montrées pour les processus autorégressifs à basculement (voir Douc, Moulines et Ryden, [59]), qui n'appartiennent plus à la classe des processus CMCa-BI : le processus des observations est une chaîne de Markov conditionnellement au processus des états. Ces modèles sont utilisés entre autres en économétrie [85], ainsi qu'en contrôle adaptatif [61], et en suivi de mouvement [36]. Ils sont paramétrés de la manière suivante : le processus $\mathbf{X}$ est une chaîne de Markov homogène, de densité de transition q_ϕ , et $\mathbf{Y}$ est une chaîne de Markov (de taille de mémoire égale à s) conditionnellement à $\mathbf{X}$, de densité de transition (homogène) $g_\phi(y_{s+1} | \mathbf{y}_s, x_{s+1})$ ne dépendant que de l'état courant. La densité de $\mathbf{Y}_N$ conditionnellement aux valeurs initiales $\mathbf{y}_{-s+1:0} = (y_0, \dots, y_{-s+1})$ s'écrit :

$$p_\phi(\mathbf{y} | \mathbf{y}_{-s+1:0}) = \int \cdots \int p(x_1) g_\phi(y_1 | \mathbf{y}_{-s+1:0}, x_1) \prod q_\phi(x_{N-1}, x_N) \prod g_\phi(y_N | \mathbf{y}_{N-s:N-1}, x_N) \delta(d\mathbf{x})$$

Une représentation stochastique de ce type de modèle existe pour le modèle autorégressif linéaire

⁵ou bien comme le fait Leroux, l'ajout de l'hypothèse de nullité à l'infini des densités $\theta \mapsto f(y, \theta)$ permet de compactifier l'espace des paramètres en rajoutant un point θ_∞ , et de traiter ainsi des espaces de paramètres non-nécessairement bornés.

à saut pour lequel nous avons :

$$Y_n = \sum_{k=1}^s a_k(X_n, \phi) Y_{n-k} + \sigma(X_n) \epsilon_n \quad (3.4)$$

La démonstration de la consistance et de la normalité asymptotique de l'EMV suit toujours la stratégie de Wald et utilise les propriétés de Markovianité du processus joint $\mathbf{Z}$ et du processus a posteriori $(X_n | \mathbf{Y}_N)_{n \geq 1}$. L'argument clé est la possibilité d'approcher la log-vraisemblance (ou le Hessien) par une fonctionnelle additive d'une suite stationnaire pour laquelle une loi forte des grands nombres existe. Ceci permet de montrer que la log-vraisemblance converge presque-sûrement et uniformément vers une fonction contraste. Les hypothèses formulées impliquent entre autres l'ergodicité uniforme (voir Annexe C) de la chaîne de Markov complète $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$.

3.3 Détermination du nombre de classes

Les résultats de convergence de l'EMV, et les procédures de calcul que lui sont associées, supposent que le modèle utilisé $\{p(\cdot, \phi), \phi \in \Phi\}$ est "suffisamment grand" pour qu'il existe un paramètre ϕ^* tel que $\mathbf{y}$ soit une réalisation d'une variable aléatoire ayant pour densité $p(\cdot, \phi^*)$. Or le modèle $\{p(\cdot, \phi), \phi \in \Phi\}$ repose sur 4 hypothèses faites sur la loi du processus $\mathbf{Y}$:

1. Le processus caché a K états ;
2. Les densités de chaque classe appartiennent aux modèles $\{f(\cdot, \theta), \theta \in \Theta\}$;
3. La dépendance entre les observations est le fait d'un processus caché markovien stationnaire ;
4. Les observations sont indépendantes conditionnellement aux états.

Il est alors souhaitable après l'estimation, de vérifier a posteriori si ces 4 hypothèses sont vérifiées, pour évaluer la validité des conclusions que nous pouvons tirer du modèle estimé. Nous traitons ici uniquement de l'adéquation de l'hypothèse 1. La vérification des hypothèses sur la structure de dépendance de $\mathbf{Y}$ (hypothèses 3 et 4) est abordée dans le chapitre 5. L'adéquation des modèles des lois d'émission est traitée entre autres par MacKay, [3]. Au-delà de l'aspect "contrôle" des résultats statistiques, la vérification de l'hypothèse 1 participe à une meilleure compréhension de la structure cachée. En effet, dans le cas de la segmentation non-supervisée, cette vérification a un rôle inférentiel et est appelée "détermination (voire estimation) du nombre de classes". Nous sommes alors dans une optique de choix de modèles et à cette fin, différents outils ont été développés : les tests statistiques, les critères informationnelles et de vraisemblance pénalisée.

L'approche classique, basée sur un test du rapport du maximum de vraisemblance, doit faire face à des problèmes de détermination de la loi asymptotique de la statistique du test, en raison des problèmes d'identifiabilité du modèle (dus à des composantes du mélange qui sont nulles). Gassiat et Kéribin [72] ont proposé un test séquentiel de détermination de K pour les CMCa-BI, en utilisant de test du rapport de vraisemblance. Cependant, cette méthode n'est pas satisfaisante en pratique, car le test se présente sous la forme

$$\begin{cases} H_0 : & K = p \\ H_1 : & K = p + 1 \end{cases}$$

où p est un entier. Le test est alors effectué de manière croissante et a tendance à s'arrêter trop tôt, si bien que l'on ne peut tester des nombres "grands" de composantes (voir [109]). Une seconde difficulté consiste en l'apparition du niveau de confiance du test, qui devient un paramètre à choisir et dont l'influence sur la qualité de la segmentation n'est pas toujours clair.

De nombreuses méthodes existent pour le choix de modèles statistiques [109, 63], et nous nous concentrons sur les critères informationnels AIC [30] et BIC [65], largement employés dans le contexte des chaînes de Markov cachées.

3.3.1 Critère AIC

Introduit par Akaike en 1973, le critère AIC (*An Information Criterion*) consiste en une correction du biais dans l'estimation de la divergence de Kullback-Leibler entre la loi obtenue par l'EMV et la vraie loi P_0 dont est issue l'échantillon $\mathbf{y}$. La divergence de Kullback-Leibler est une mesure d'écart entre mesures de probabilités ou entre densités (ce n'est pas une distance parce qu'elle ne possède pas la propriété de symétrie).

Définition 3.3.1. Divergence de Kullback-Leibler

Soient P, Q deux mesures de probabilités sur $(\Omega, \mathcal{A})$ tel que P admette une densité (dérivée de Radon-Nikodym) relativement à Q , notée $\frac{dP}{dQ}$. La divergence de Kullback-Leibler entre P et Q est

$$D(P, Q) = \int_{\Omega} \log \left(\frac{dP}{dQ}(\omega) \right) P(d\omega)$$

Si p et q sont deux densités relativement à une mesure de référence μ , on note

$$D(p, q) = \int \log \left(\frac{p}{q}(y) \right) p(y) \mu(dy)$$

Etant donné des observations $\mathbf{y}$, l'objectif du critère AIC est de choisir parmi M modèles paramétriques $\mathcal{P}(\Phi_i) = \{P_{\phi}, \phi \in \Phi_i\}$, $i = 1..M$, le plus proche de P_0 au sens de la divergence de Kullback-Leibler. Pour chaque modèle $\mathcal{P}(\Phi_i)$, l'EMV $\phi_i(\mathbf{y})$ est le point qui minimise la divergence de Kullback observée. Pour faire le choix du modèle au vu des données, nous comparons donc les divergences entre P_0 et le point le plus proche $P_{\phi_i(\mathbf{y})}$ au sein de chaque modèle $\mathcal{P}(\Phi_i)$. La quantité d'intérêt est alors la moyenne de la divergence entre P_0 et $P_{\phi_i(\mathbf{y})}$. Celle-ci a pour expression $E_{P_0, \mathbf{Y}} [D(P_0, P_{\phi_i(\mathbf{Y})})]$

$$E_{P_0, \tilde{Y}} [\log(P_0(\tilde{Y}))] - E_{P_0, \mathbf{Y}} [E_{P_0, \tilde{Y}} [\log(P_{\phi_i(\mathbf{Y})}(\tilde{Y}))]] \quad (3.5)$$

Le premier terme de l'Eq. (3.5) est une constante, et il suffit de comparer le second terme pour chaque modèle $\mathcal{P}(\Phi_i)$. Nous pouvons estimer ce terme à partir de $\mathbf{y}$ par $\log(\frac{P_0(\mathbf{y})}{P_{\phi_i(\mathbf{y})}(\mathbf{y})})$ dont la moyenne est $E_{P_0, \mathbf{Y}} [\log(\frac{P_0(\mathbf{Y})}{P_{\phi_i(\mathbf{Y})}(\mathbf{Y})})]$. Cet estimateur est biaisé parce que les quantités $E_{P_0, \mathbf{Y}} [\log(P_{\phi(\mathbf{Y})}(\mathbf{Y}))]$ et $E_{P_0, \mathbf{Y}} [E_{P_0, \tilde{Y}} [\log(P_{\phi(\mathbf{Y})}(\tilde{Y}))]]$ sont différentes. Akaike a mis en évidence ce biais asymptotique B :

$$E_{P_0, \mathbf{Y}} [\log(P_{\phi(\mathbf{Y})}(\mathbf{Y}))] = E_{P_0, \mathbf{Y}} [E_{P_0, \tilde{Y}} [\log(P_{\phi(\mathbf{Y})}(\tilde{Y}))]] + B$$

Ainsi la quantité à maximiser pour minimiser $E_{P_0, \mathbf{Y}} [D(P_0, P_{\phi_i(\mathbf{Y})})]$ n'est pas la quantité $\log(P_{\phi(\mathbf{y})}(\mathbf{y}))$, mais la quantité $\log(P_{\phi(\mathbf{y})}(\mathbf{y})) - B$.

Une approximation asymptotique au premier ordre de B est $|\Phi_i|$, le nombre de paramètres libres estimés dans le modèle Φ_i . Le critère d'Akaike à maximiser est donc

$$AIC_i = \log(P_{\phi_i(\mathbf{y})}(\mathbf{y})) - |\Phi_i| \quad (3.6)$$

3.3.2 Critère BIC

Le critère BIC (*Bayesian Information Criterion*) apparaît lors de l'étude du comportement asymptotique du rapport de Bayes pour la détermination du meilleur modèle $\mathcal{P}(\Phi_i)$, $i = 1..M$. Le point de départ, bayésien, est de considérer que le type de modèle est lui aussi une variable aléatoire qui est estimée par MAP. Sous l'hypothèse d'uniformité de la loi a priori sur les modèles, cela conduit à comparer les probabilités $P(\mathbf{y} | \mathcal{P}(\Phi_i))$. Les intégrales définissant les vraisemblances intégrées

$$P(\mathbf{y} | \mathcal{P}(\Phi_i)) = \int P(\mathbf{y} | \phi_i, \mathcal{P}(\Phi_i)) P(\phi_i | \mathcal{P}(\Phi_i)) d\phi_i$$

sont rarement calculables et doivent être approchées par la méthode de Laplace. Sous des conditions techniques, on obtient un développement asymptotique du logarithme du rapport de Bayes. Ce dernier est défini entre deux modèles $\mathcal{P}(\Phi_i)$ et $\mathcal{P}(\Phi_j)$ par :

$$B_{i,j} = \frac{P(\mathbf{y} | \mathcal{P}(\Phi_i))}{P(\mathbf{y} | \mathcal{P}(\Phi_j))}$$

Le développement asymptotique du logarithme (au premier ordre en le nombre d'observations N) est :

$$\log B_{i,j} \simeq L_i(\mathbf{y}, \phi_i(\mathbf{y})) - L_j(\mathbf{y}, \phi_j(\mathbf{y})) - \frac{1}{2}(|\Phi_i| - |\Phi_j|) \log N$$

où $\phi_i(\mathbf{y})$ est l'estimateur MAP du paramètre ϕ_i , et $L_i(\mathbf{y}, \phi_i)$ est la log-vraisemblance du modèle $\mathcal{P}(\Phi_i)$. Ce résultat est doublement remarquable :

- pour comparer 2 modèles, il suffit de calculer la quantité $L_i(\mathbf{y}, \phi_i(\mathbf{y})) - \frac{1}{2} |\Phi_i| \log N$ (le BIC) et de prendre le modèle qui la maximise.
- L'approximation au premier ordre ne dépend pas des loi a priori des modèles $P(\phi_i | \mathcal{P}(\Phi_i))$. Ainsi, même si le BIC est déduit par des arguments bayésiens, il est utilisable en l'état dans la démarche fréquentiste : c'est ce qui "autorise" son utilisation notamment dans le cadre de la sélection du nombre de classe d'une CMCA-BI, lorsque l'estimateur des paramètres est obtenu par maximum de vraisemblance.

3.3.3 Vraisemblance pénalisée

Les critères AIC et BIC posent des problèmes dans leur applications aux mélanges, et en particulier aux CMCA-BI. Le critère AIC se révèle être un mauvais estimateur du nombre de classes [109, 22]. En effet, AIC "ne pénalise pas assez" les modèles ayant beaucoup de classes car

le terme de complexité ne tient pas compte de la taille de l'échantillon $\mathbf{y}$, contrairement à BIC. Ce dernier se révèle satisfaisant pour sélectionner le bon nombre de classes. Cependant, son utilisation rigoureuse repose sur des hypothèses dites de Laplace-régularité (conditions données par Kass et Raftery [96] permettant le développement asymptotique évoqué dans le paragraphe précédent) dont la justesse est soit invérifiable, soit irréalisable dans les modèles à données manquantes.

Ces critères informationnels aboutissent à des critères de vraisemblance pénalisée et leur mise en pratique met clairement en évidence la caractère décisif de la vitesse de croissance de la pénalisation (en fonction de la taille de l'échantillon). Dans [71], Gassiat propose des conditions suffisantes de vitesse de croissance pour garantir la sélection du bon nombre de classes presque-sûrement (sans se préoccuper de la manière dont ces pénalisations ont été construites). Le processus $\mathbf{Y}$ est considéré uniquement comme un processus stationnaire tel que la densité de Y_n soit un mélange de K densités. Pour tout N , on introduit la fonction de pénalisation $K \mapsto a_N(K)$ qui permettra de construire la vraisemblance (marginale) pénalisée du processus, dont la maximisation (sur K) donnera un estimateur $\hat{K}$ du nombre de composantes. Le critère étudié est :

$$L_{p,K} = \max_{\phi_i \in \Phi_i} \sum_{n=1}^N \log p(y_n, \phi_i) - a_N(K) \quad (3.7)$$

ce qui donne pour l'estimateur du nombre de classe :

$$\hat{K} = \arg \max_K \left\{ \max_{\phi_i \in \Phi_i} \sum_{n=1}^N \log p(y_n, \phi_i) - a_N(K) \right\}$$

Les conditions suffisantes pour assurer la consistance forte de l'estimateur $\hat{K}$ sont :

1. $K \mapsto a_N(K)$ est une fonction croissante
2. $K_1 > K_2 \implies a_N(K_1) - a_N(K_2) \xrightarrow{N \rightarrow \infty} +\infty$
3. $\forall K_1, K_2, \frac{a_N(K_1) - a_N(K_2)}{N} \xrightarrow{N \rightarrow \infty} 0$

La fonction de pénalisation de BIC, de la forme $a_N(K) = \frac{|\Phi_i|}{2} \log N$, vérifie ces conditions. Ceci vient a posteriori expliquer le bon comportement de l'usage de BIC pour la détermination du nombre de classes dans les CMCA. Cependant, ce résultat permet juste de déterminer la bonne vitesse pour $a_N(K)$, et pose alors le problème de la détermination pratique du coefficient multiplicatif modulo lequel est défini une fonction de pénalisation vérifiant les hypothèses 1, 2 et 3.

Remarque 3.3.1. Hypothèses de mélangeance

Le théorème 2.1 de [71] affirmant la convergence presque-sûre de l'estimateur $\hat{K}$ utilise l'hypothèse CMCA-BI dans sa démonstration uniquement pour affirmer la β -mélangeance⁶ du processus observé $(Y_n)_{n \geq 1}$ (celle-ci est transmise à $\mathbf{Y}$ par la chaîne de Markov cachée $\mathbf{X}$). Le même résultat de consistance s'applique donc si l'on suppose directement que le processus $\mathbf{Y}$ est β -mélangeant, sans pour autant être une CMCA-BI. Ainsi la même pénalisation fournit un estimateur fortement consistant du nombre de classes pour les CMCA et CMCo β -mélangeantes.

⁶voir section 4.2

3.4 Détermination de l'EMV

La détermination pratique de l'EMV pose un difficile problème d'optimisation, ce qui provient de la forme particulière de la densité (3.2). La complexité de l'optimisation provient essentiellement de la modélisation par mélange de lois, et les mêmes difficultés apparaissent dans le cas des mélange de lois indépendantes⁷, des champs markoviens ou des réseaux bayésiens.

Nous rappelons dans cette partie l'algorithme Espérance-Maximisation (*Expectation-Maximization*, EM) qui est une des méthodes les plus couramment utilisées pour rechercher les maxima de la vraisemblance d'un mélange de lois. Nous le présentons tout d'abord dans le contexte général des modèles à données manquantes, ainsi que sa variante EM Stochastique (*Stochastic-EM*, SEM).

De par la paramétrisation que nous avons utilisée, les paramètres des lois d'émission et de la loi du processus $\mathbf{X}_N$ ne sont pas liés, si bien que la loi du processus caché $\mathbf{X}_N$ (i.i.d, chaîne ou champ de Markov) n'a pas d'impact sur l'expression des formules d'estimation des θ_k : en effet, le processus caché n'intervient dans l'algorithme EM qu'à travers les probabilités marginales a posteriori. Ainsi les algorithmes développés dans cette section pour les CMCa-BI (ainsi que ceux dans le chapitre 3), peuvent être utilisés pour différentes structures cachées, lorsque les observations restent indépendantes conditionnellement à $\mathbf{X}_N$ (MMCa-BI).

3.4.1 L'algorithme EM

Dempster, Laird et Rubin, ont, dans leur article de 1977 [52], formalisé l'algorithme EM et démontré ses principales propriétés. Il s'agit d'un algorithme général de recherche de maxima de la log-vraisemblance, au même titre que les méthodes d'optimisation numérique (itérative) de type Newton-Raphson ou descente du gradient (voir la monographie de McLachlan et Krishnan [108]). A la différence des habituelles méthodes d'optimisation, l'approche EM a une interprétation statistique reposant sur la notion de modèles complet et incomplet. Nous décrivons de manière générale le mécanisme de l'algorithme.

Nous supposons que nous avons observé $\mathbf{u}$, réalisation d'une variable aléatoire $\mathbf{U}$, dont la densité $p(\mathbf{u}, \phi)$ (relativement à une mesure μ) appartient à un modèle paramétrique indexé par un ensemble Φ . L'objectif est de déterminer $\arg \max_{\phi \in \Phi} \log(p(\mathbf{u}, \phi))$. Nous notons $L(\mathbf{u}, \phi)$ la log-vraisemblance de $\mathbf{u}$, appelée aussi log-vraisemblance incomplète ou observée. Nous supposons qu'il existe une variable aléatoire $\mathbf{V}$ de densité $p_c(\mathbf{v}, \phi)$ et une fonction déterministe φ tel que $\mathbf{U} = \varphi(\mathbf{V})$ ($\mathbf{U}$ est une observation partielle de $\mathbf{V}$) : la variable $\mathbf{V}$ est appelée donnée complète et $\mathbf{U}$ est appelée donnée incomplète. Si $\mathbf{v}$ est une réalisation de $\mathbf{V}$, $L_c(\mathbf{v}, \phi) = \log(p_c(\mathbf{v}, \phi))$ est la log-vraisemblance complète ou complétée. Enfin, nous écrirons simplement $E_\phi[h(\mathbf{V})|\mathbf{u}]$ (si h est une fonction) pour désigner l'espérance conditionnelle $E_\phi[h(\mathbf{V})|\mathbf{U}]$ évaluée au point $\mathbf{u}$.

Partant d'un paramètre initial ϕ_0 , l'objectif de l'algorithme EM est de construire une suite de paramètre $(\phi_n)_{n \geq 1}$ telle que $(L(\mathbf{u}, \phi_n))_{n \geq 1}$ atteigne un maximum de la log-vraisemblance $L(\mathbf{u}, \phi)$, noté L^* . La construction de cette suite est basée sur l'alternance de deux phases, pour tout $n \geq 1$:

⁷Si le problème d'optimisation est commun aux 2 types de modèles, nous rappelons que pour le modèle CMCa-BI se rajoute le problème théorique de la consistance de l'EMV due à l'estimation avec des données dépendantes, cf. section 3.2.1.

Espérance Calcul de la fonction $Q(\phi, \phi_n) = E_{\phi_n} [L_c(\mathbf{V}, \phi) | \mathbf{u}]$

Maximisation $\phi_{n+1} = \arg \max_{\phi \in \Phi} Q(\phi, \phi_n)$

La suite ainsi construite a la propriété d'augmenter la vraisemblance incomplète. Pour le voir, nous exploitons la décomposition suivante :

$$L(\mathbf{u}, \phi) = L_c(\mathbf{v}, \phi) - \log(p(\mathbf{v} | \mathbf{u}, \phi)) \quad (3.8)$$

$p(\mathbf{v} | \phi, \mathbf{u})$ désigne la densité de $\mathbf{V}$ conditionnellement $\mathbf{U}$ pour le paramètre ϕ . Nous pouvons décomposer alors un accroissement de la log-vraisemblance incomplète en deux parties :

$$L(\mathbf{u}, \phi) - L(\mathbf{u}, \phi_n) = L_c(\mathbf{v}, \phi) - L_c(\mathbf{v}, \phi_n) - \log \left(\frac{p(\mathbf{v} | \mathbf{u}, \phi)}{p(\mathbf{v} | \mathbf{u}, \phi_n)} \right) \quad (3.9)$$

Nous prenons l'espérance conditionnelle à $\mathbf{U} = \mathbf{u}$ (lorsque le paramètre vaut ϕ_n) à droite et gauche de cette égalité :

$$L(\mathbf{u}, \phi) - L(\mathbf{u}, \phi_n) = E_{\phi_n} [L_c(\mathbf{V}, \phi) - L_c(\mathbf{V}, \phi_n) | \mathbf{u}] + E_{\phi_n} \left[\log \left(\frac{p(\mathbf{V} | \mathbf{u}, \phi)}{p(\mathbf{V} | \mathbf{u}, \phi_n)} \right) | \mathbf{u} \right] \quad (3.10)$$

Le second terme du membre de droite est la divergence de Kullback-Leibler entre les densités (conditionnelles) $p(\mathbf{v} | \mathbf{u}, \phi_n)$ et $p(\mathbf{v} | \mathbf{u}, \phi)$ qui est toujours positive. Nous avons donc la minoration suivante :

$$L(\mathbf{u}, \phi) - L(\mathbf{u}, \phi_n) \geq E_{\phi_n} [L_c(\mathbf{V}, \phi) - L_c(\mathbf{V}, \phi_n) | \mathbf{u}] \quad (3.11)$$

L'équation (3.11) montre qu'une étape EM maximise une borne inférieure de l'accroissement de la log-vraisemblance incomplète. Cette inégalité montre aussi qu'il suffit de prendre ϕ_{n+1} tel que $Q(\phi_{n+1}, \phi_n) > 0$ pour assurer la croissance de la vraisemblance observée, sans nécessairement prendre le maximum : de tels algorithmes sont appelés "Generalized Expectation - Maximization" (GEM), et sont utilisés lorsque la recherche du maximum de $Q(\cdot, \phi_n)$ est une opération trop difficile ou coûteuse en termes de temps de calcul. Wu [108] a montré sous des conditions assez générales la convergence de la suite des log-vraisemblances $(L(\phi_n))_{n \geq 0}$ obtenue par un algorithme GEM vers un maximum local de la log-vraisemblance L^* , ainsi que la convergence de $(\phi_n)_{n \geq 0}$ vers un point ϕ^* atteignant ce maximum $L^* = L^*(\mathbf{u}, \phi^*)$.

Remarque 3.4.1. Fonction implicite de l'algorithme EM

L'algorithme EM est un système dynamique discret $\phi_{n+1} = T_{EM}(\phi_n)$, où la fonction T_{EM} est définie de manière implicite par $T_{EM}(\phi) = \arg \max_{\phi' \in \Phi} E_{\phi} [L_c(\mathbf{V}, \phi') | \mathbf{u}]$. Ainsi le comportement asymptotique de l'algorithme peut être déduit des propriétés de la fonction T_{EM} , ou de celles de son jacobien (s'il existe), noté $\partial_{\phi} T_{EM}$. Si la fonction T_{EM} est continue et que l'algorithme EM converge, ses limites possibles sont les points fixes stables de T_{EM} . Dans ce cas-là, la vitesse de convergence peut être déduite de l'étude des valeurs propres de ∂T_{EM} , qui s'exprime en fonction de l'information de Fisher des modèles complets et incomplets. Nous introduisons les matrices d'information observées incomplètes et complètes $I(\mathbf{u}, \phi) = -\nabla^2 L(\mathbf{u}, \phi)$ et

$I_c(\mathbf{v}, \phi) = -\nabla^2 L_c(\mathbf{v}, \phi)$, ainsi que l'espérance conditionnellement à $\mathbf{U} = \mathbf{u}$ de cette dernière, notée $I_m(\mathbf{u}, \phi) = -E_\phi [\nabla^2 L_c(\mathbf{V}, \phi) | \mathbf{u}]$ (I_m est souvent appelée la matrice d'information manquante). Si ϕ^* est un point fixe attracteur de T_{EM} , nous avons alors

$$\begin{aligned} \partial_\phi T_{EM}(\phi^*) &= (I_d - I_c(\mathbf{u}, \phi^*)^{-1} I(\mathbf{u}, \phi)) \\ &= I_c(\mathbf{u}, \phi^*)^{-1} I_m(\mathbf{v}, \phi^*) \end{aligned}$$

Cette dernière égalité s'interprète en disant que la vitesse de convergence de EM est égale à la proportion d'information manquante sur ϕ due à l'observation partielle de $\mathbf{V}$, (chapitre 3 [108]).

3.4.2 Application à l'estimation des mélanges finis

Dans le cas des mélanges finis, la vraisemblance complète s'écrit simplement, et EM donne des formules de mises à jour des paramètres communes aux MMCA-BI. En effet, leur log-vraisemblance complète s'écrit :

$$L_c(\phi) = \log(p(\mathbf{x}, A)) + \sum_{i=1}^N \sum_{k=1}^K \log(f(y_i, \theta_k)) \times 1_k(x_i) \quad (3.12)$$

La fonction Q à maximiser est :

$$Q(\phi, \phi_n) = E_{\phi_n} [\log(p(\mathbf{X}_N, A)) | \mathbf{y}] + \sum_{i=1}^N \sum_{k=1}^K P(X_i = k | \mathbf{y}, \phi_n) \log(f(y_i, \theta_k)) \quad (3.13)$$

Ainsi, la mise à jour $\phi_{n+1} = (A_{n+1}, (\theta_k^{(n+1)})_{1 \leq k \leq K})$ basée sur une maximisation globale s'obtient par deux maximisations sur les paramètres de la loi du processus a priori X et sur les lois des observations :

1. $A_{n+1} = \arg \max_A E_{\phi_n} [\log(p(\mathbf{X}_N, A)) | \mathbf{y}]$
2. pour $1 \leq k \leq K$, $\theta_k^{(n+1)} = \arg \max_{\theta_k} \sum_{i=1}^N P(X_i = k | \mathbf{y}, \phi_n) \log(f(y_i, \theta_k))$

Nous obtenons des procédures de ré-estimation pour les lois des observations qui sont indépendantes de la structure cachée choisie, puisque celle-ci n'intervient que par les probabilités marginales a posteriori $P(X_i = k | \mathbf{y}, \phi_n)$. Par la suite, nous noterons

$$\forall i, k, l, n, \pi_{i,k}^{(n)} = P(X_i = k | \mathbf{y}, \phi_n) \text{ et } p_i^{(n)}(k, l) = P(X_i = k, X_{i+1} = l | \mathbf{y}, \phi_n)$$

Nous obtenons alors une expression analytique des paramètres de la loi de $\mathbf{X}$ et la formule de ré-estimation de la probabilité de transition $A_{n+1} = (a_{ij}^{(n+1)})$ est

$$\forall k, l \leq K, \hat{a}_{kl}^{(n+1)} = \frac{\sum_{i=1}^{N-1} p_i^{(n)}(k, l)}{\sum_{i=1}^{N-1} \pi_{i,k}^{(n)}} \quad (3.14)$$

Dans le cas des chaînes de Markov cachées à espace d'état discret, les probabilités (3.4.2) sont faciles à obtenir par le biais des algorithmes avant-arrière. Pour des structures cachées plus complexes tels que les champs markoviens ou les réseaux bayésiens, le calcul des marges a posteriori est plus

délicat, voir impossible. Plusieurs méthodes ont été proposées pour éviter l'étape E ou pour avoir une approximation des probabilités a posteriori :

- par simulation selon la loi a posteriori de $\mathbf{X}_N$ conditionnellement à $\mathbf{y}$, algorithmes Monte Carlo EM de Wei et Tanner [154] et Stochastic Approximation EM de Lavielle, Delyon et Moulines [51] ;
- par approximation déterministe, algorithme “Iterative Conditional Mode” de Besag (ICM), [20] ;
- par algorithme de propagation des croyances généralisant les récursions avant-arrière, [70].

3.4.3 Algorithme *Stochastic EM*

L'algorithme EM est souvent considéré comme la méthode de référence pour l'estimation des modèles à données manquantes, ou ayant des structures particulièrement complexes. Il possède néanmoins des limitations qui ont stimulé la proposition de nombreux algorithmes itératifs d'estimation inspirés du “principe EM”. Les défauts connus de l'algorithme sont la sensibilité à la valeur initiale ϕ_0 , la faible vitesse de convergence, la difficulté du calcul de la loi a posteriori $P(\mathbf{V}|\mathbf{u}, \phi_n)$ ou du calcul de l'espérance, la difficulté de la maximisation de la phase M. Selon les modèles, ce sont les étapes E ou M, voire les deux, qui peuvent donner lieu à des difficultés. L'algorithme *Stochastic-EM* (SEM) de Celeux et Diebolt ([35]) pallie ces deux problèmes en proposant une procédure d'imputation en remplacement de l'étape E. Le schéma général de l'algorithme est une alternance entre simulation et maximisation de la log-vraisemblance “complétée”, jusqu'à obtention d'un comportement stationnaire de la suite des paramètres. Partant d'une valeur initiale ϕ_0 , nous construisons itérativement la suite aléatoire $(\phi_n)_{n \geq 1}$:

Espérance Stochastique $\mathbf{u}^{(n+1)} \sim p(\mathbf{u}|\mathbf{v}, \phi_n)$

Maximisation $\phi_{n+1} = \arg \max_{\phi} L_c(\mathbf{u}^{(n+1)}, \phi)$

Si nous appelons M la transformation qui associe à $\mathbf{u}^{(n+1)}$ le nouveau paramètre ϕ_{n+1} , nous pouvons réécrire SEM comme le système dynamique suivant :

$$\phi_{n+1} = T_{EM}(\phi_n) + V(\phi_n, \mathbf{u}^{(n+1)}) \quad (3.15)$$

où $V(\phi_n, \mathbf{u}^{(n+1)})$ est par définition égal à $M(\mathbf{u}^{(n+1)}) - T_{EM}(\phi_n)$. La suite aléatoire ainsi créée $(\phi_n)_{n \geq 1}$ est une chaîne de Markov homogène, qui peut s'interpréter comme une perturbation aléatoire de l'algorithme EM.

L'ergodicité et la convergence de (ϕ_n) vers une loi stationnaire $\pi_{\phi, SEM}$ ont été montrés lorsque les observations manquantes appartiennent à un compact ou lorsque le modèle est une famille exponentielle par Ip [93, 57] ; ou encore dans le cas indépendant (sans contraintes fortes sur la forme paramétrique, mais l'EMV du modèle complet doit exister et être asymptotiquement normal) par Nielsen [115]. La difficulté essentielle de l'algorithme SEM réside en la relation entre la moyenne de la loi stationnaire $\pi_{\phi, SEM}$ et l'EMV. Ce lien est en général complexe : Ip a montré dans le cas exponentiel que ces deux quantités diffèrent d'un ordre $1/N$, où N est la taille de l'échantillon. Lorsque N tend vers l'infini, Nielsen et Ip montrent tous les deux la convergence de la loi stationnaire vers une loi normale centrée sur l'EMV.

La loi stationnaire limite a pour moyenne un maximum de la vraisemblance, et les perturbations aléatoires dues à l'imputation permettent non seulement de sortir de certains "pièges", mais aussi d'avoir une image de la log-vraisemblance aux environs du maximum atteint, ce qui permet d'obtenir aussi une approximation de la variance de l'estimateur, [57]. D'un point de vue pratique, malgré une vitesse de convergence plus faible (accompagnée souvent d'un temps de calcul plus long en raison de la phase de simulation) l'avantage de SEM sur EM est d'être moins sensible à l'initialisation que EM, et de se comporter dans certains cas mieux que EM, [34].

Pour les CMCa-BI, l'algorithme SEM consiste en la simulation de $\tilde{\mathbf{X}}^{(n)} = (\tilde{X}_i^{(n)})_{1 \leq i \leq N}$ tirée selon la loi a posteriori $P(\mathbf{X}|\mathbf{y}, \phi_n)$, puis en le calcul au sein de chaque classe du maximum de vraisemblance dans le cas d'observations indépendantes :

$$\forall k, \theta_k^{(n+1)} = \arg \max_{\theta_k} \prod_{i|\tilde{x}_i^{(n)}=k} f(y_i, \theta_k)$$

$$\forall k, l, a_{kl}^{(n+1)} = \frac{\sum_{i=1}^{N-1} 1_{kl}(x_i, x_{i+1})}{\sum_{i=1}^N 1_{kl}(x_i)}$$

3.5 Fonction estimante et Estimation Conditionnelle Itérative

Des méthodes alternatives au maximum de vraisemblance ont été proposées pour l'estimation des CMCa, comme les moindres carrés (voir par exemple Mevel [112]), la maximisation de la vraisemblance marginale $\sum_{i=1}^N \log(p(y_i, \phi))$ (voir [103]) ou encore la "vraisemblance des données séparées" égale à $\sum_{i=1}^n \log(p_m((y_{m(i-1)+1}, \dots, y_{mi}), \phi))$ (où $p_m(\cdot, \phi)$ est la densité de $(Y_1, \dots, Y_m)$, voir [141]). Cependant, l'EMV est utilisé très majoritairement pour l'inférence des modèles à données manquantes grâce à la possibilité d'approcher l'EMV par EM⁸ ainsi que par l'existence de résultats théoriques sur sa convergence et son efficacité.

Nous proposons dans cette section une méthode d'estimation générale des modèles à données manquantes, basée sur la théorie des fonctions estimantes. Nous rappelons les définitions et propriétés fondamentales de ces dernières dans la section 3.5.1.1 (en nous inspirant de [79, 89]). L'objectif est alors de construire une famille de fonctions estimantes pour les modèles à données manquantes à partir des fonctions estimantes du modèle à données complètes. Le passage du cas complet au cas incomplet est réalisé par la projection des fonctions estimantes par l'opérateur d'espérance conditionnelle. Cette idée, latente dans l'algorithme EM⁹, a été généralisée par Heyde et Morton [90] pour proposer un nouvel algorithme récursif d'estimation de paramètres (que nous rappelons dans la section 3.5.3.3). Cependant, Heyde *et al.* n'étudie pas les propriétés des fonctions estimantes ainsi obtenues (hormis leur optimalité, que nous définissons par la suite), mais insistent sur les différences entre les algorithmes selon les fonctions estimantes et les projecteurs utilisés (sans s'attarder sur un en particulier).

En nous concentrant sur l'espérance conditionnelle, nous montrons que les fonctions projetées conservent certaines propriétés des fonctions sur données complètes, fournissant ainsi un cadre

⁸ou ses variantes, voir section 3.4.3 ou [108].

⁹comme le remarque Efron dans la discussion de l'article de Dempster *et al.*, la justification de EM repose sur l'identité de Fisher qui lie le score du modèle complet au score du modèle incomplet : $\nabla L(\mathbf{u}, \phi) = E[\nabla L_c(\mathbf{v}, \phi)|\mathbf{u}]$.

original et une ouverture possible à des études théoriques ultérieures pour une théorie générale de l'estimation des modèles à données incomplètes. Nous introduisons alors un algorithme de résolution des équations obtenues, que nous appelons Estimation Conditionnelle Itérative (ECI), parce qu'il permet de retrouver entre autres le principe ECI de Pieczynski [122]. Nous montrons que son analyse théorique peut être faite de manière similaire à l'algorithme EM (contrairement à l'algorithme général de Heyde et Morton), et qu'il permet de le retrouver, ainsi que plusieurs autres algorithmes d'estimation de modèles à données manquantes.

3.5.1 Fonction estimante et données manquantes

3.5.1.1 Rappels sur les fonctions estimantes

La théorie des fonctions estimantes permet de réunir dans le même formalisme de nombreuses méthodes d'estimation : moindres carrés, maximum de vraisemblance, moments, . . . Ces méthodes aboutissent à la recherche des racines d'une certaine fonction g qui sont, sous des conditions convenables, des estimateurs consistants des paramètres recherchés. L'approche classique consiste à montrer alors que ces estimateurs sont sans biais, ou optimaux (par exemple de variance minimale). La théorie des fonctions estimantes porte son intérêt sur la propriété de la fonction g elle-même, car les propriétés des estimateurs sont déduites de celles de g .

Nous définissons tout d'abord une fonction estimante et donnons quelques uns des concepts qui permettent de voir cette théorie comme une généralisation de la méthode du maximum de vraisemblance.

Définition 3.5.1. *Fonction estimante*

Soit $\mathcal{P} = \{p(\cdot, \phi), \phi \in \Phi\}$ un modèle paramétrique statistique indexé par $\Phi \subset \mathbb{R}^p$ et défini sur l'espace $\mathcal{V}$. La fonction $g : \mathcal{V} \times \Phi \longrightarrow \mathbb{R}^p$ est une fonction estimante si elle est non-biaisée, i.e.

$$\forall \phi \in \Phi, E_{\phi} [g(\mathbf{V}, \phi)] = 0 \quad (3.16)$$

Le vecteur aléatoire $g(\mathbf{V}, \phi)$ est parfois noté $g(\phi)$.

Parmi l'ensemble des fonctions estimantes, on s'intéresse plus particulièrement à celles qui sont carré intégrables. Nous en considérons alors un sous-ensemble¹⁰ que nous notons $\mathcal{H}_{\mathbf{V}}$. A partir d'une fonction estimante $g \in \mathcal{H}_{\mathbf{V}}$, nous pouvons déduire un estimateur de ϕ en résolvant, pour $\mathbf{v}$ donné, l'équation (si il existe une solution)

$$g(\mathbf{v}, \phi) = 0 \quad (3.17)$$

L'étude des propriétés asymptotiques de la racine $\hat{\phi}(\mathbf{v})$ obtenue en résolvant l'Eq. (3.17) se fait de manière similaire à celles de l'EMV et aboutit à la généralisation du vecteur score (le gradient de la log-vraisemblance) ainsi qu'à la définition d'un critère d'optimalité pour les fonctions estimantes. Afin d'étudier le comportement asymptotique de la fonction g (et de $\hat{\phi}(\mathbf{v})$), il est commode de la normaliser, étant donné que g et Ag (où A est une matrice inversible dépendant éventuellement de ϕ) possèdent les mêmes racines. Si g est une fonction estimante pour le modèle $\{p(\cdot, \phi), \phi \in \Phi\}$, nous définissons la fonction estimante normalisée

¹⁰cela peut-être l'ensemble des fonctions estimantes linéaires en $\mathbf{v}$

$$\forall \mathbf{v}, \phi \quad g^{(s)}(\mathbf{v}, \phi) = -E_\phi [\partial_\phi g(\mathbf{V}, \phi)]' E_\phi [g(\mathbf{V}, \phi)g(\mathbf{V}, \phi)']^{-1} g(\mathbf{v}, \phi)$$

où $\partial_\phi g$ représente le jacobien de g relativement à ϕ . Afin de comparer les fonctions estimantes entre elles, nous utilisons la matrice $\mathcal{E}(g) = E_\phi [\partial_\phi g(\phi)]' E_\phi [g(\phi)g(\phi)']^{-1} E_\phi [\partial_\phi g(\phi)]$, qui est aussi égal¹¹ à la variance de $g^{(s)}$, i.e.

$$\mathcal{E}(g) = E_\phi [g^{(s)}(\phi)g^{(s)}(\phi)']$$

Ces matrices s'interprètent comme des critères d'information, et nous pouvons les comparer entre elles au sens de l'ordre partielle des matrices semi-définies positives. $\mathcal{E}(g)$ généralise l'information de Fisher parce qu'elle la redonne lorsque nous utilisons le score, et qu'elle est reliée elle aussi à la variance asymptotique de l'estimateur de ϕ . En effet, si la taille de l'échantillon n tend vers l'infini, et que nous avons la normalité asymptotique de $g_n(\phi) = g_n(\mathbf{V}_n, \phi)$, i.e.

$$E_\phi [g_n(\phi)g_n(\phi)']^{-1} g_n(\phi) \longrightarrow_n N(0, I_p)$$

alors l'estimateur $\hat{\phi}_n(\mathbf{v})$ est lui aussi normal :

$$\mathcal{E}(g_n) (\hat{\phi}_n - \phi) \longrightarrow_n N(0, I_p)$$

Le critère $\mathcal{E}(g)$ permet de définir l'optimalité, à horizon fini, au sein d'une famille $\mathcal{H}_\mathbf{v}$ de fonctions estimantes :

Définition 3.5.2. Quasi-score

g^ est une fonction estimante optimale dans $\mathcal{H}_\mathbf{v}$ si*

$$\forall g \in \mathcal{H}_\mathbf{v}, \forall \phi \in \Phi, \mathcal{E}(g^*) \geq \mathcal{E}(g) \quad (3.18)$$

au sens de l'ordre des matrices semi-définies positives. La fonction g^ est alors appelée un quasi-score, et un estimateur déduit de g est appelé un estimateur de la quasi-vraisemblance.*

Lorsque le score existe, il est la fonction estimante optimale de la famille $\{g \in L^2 \mid E_\phi [g(\phi)] = 0\}$ (sans contrainte sur la forme des fonctions). Les quasi-scores vérifient la propriété remarquable suivante :

$$E_\phi [\partial_\phi g^{*(s)}(\phi)] = -E_\phi [g^{*(s)}(\phi)g^{*(s)}(\phi)'] \quad (3.19)$$

qui est donc une généralisation du fait que la variance du score soit égale à l'opposé de l'espérance du Hessien de la log-vraisemblance. Ainsi, comme $\mathcal{E}(g) = E_\phi [g^{(s)}(\phi)g^{(s)}(\phi)']$, l'inégalité (3.18) peut se voir comme une réinterprétation de l'inégalité de Cramer-Rao (et du théorème de Gauss-Markov pour les estimateurs linéaires) comme une borne sur la variance des fonctions estimantes, et non plus comme une borne sur les variances des estimateurs.

¹¹De manière générale, $\mathcal{E}(g) = \mathcal{E}(Ag)$ si A est une matrice inversible.

3.5.1.2 Construction de fonctions estimantes dans les modèles à données manquantes

Les fonctions estimantes fournissent des estimateurs ad hoc dans des modèles complexes, notamment lorsque la vraisemblance du modèle est compliquée voire inaccessible (entre autres dans les champs aléatoires ou dans les processus à temps continu) et d'obtenir des estimateurs consistants sous des conditions assez faibles. L'objectif de cette partie est d'introduire une méthode de construction de fonctions estimantes pour les modèles à données manquantes à partir des fonctions estimantes du modèle complet.

En effet, nous montrons que l'opération de projection des fonctions estimantes $g(\phi)$ du modèle complet $\{p_c(\cdot, \phi), \phi \in \Phi\}$ permet d'obtenir des fonctions estimantes pour le modèle incomplet $\{p(\cdot, \phi), \phi \in \Phi\}$ qui héritent alors des "bonnes propriétés" de $g(\phi)$.

Proposition 3.5.1. Projection des fonctions estimantes

Soit $\mathbf{V}$ une v.a. sur $(\mathcal{V}, \mathcal{B}(\mathcal{V}))$ et $\mathbf{U} = \varphi(\mathbf{V})$ v.a. sur $(\mathcal{U}, \mathcal{B}(\mathcal{U}))$, avec $\varphi : (\mathcal{V}, \mathcal{B}(\mathcal{V})) \longrightarrow (\mathcal{U}, \mathcal{B}(\mathcal{U}))$ fonction mesurable et déterministe. Si g est une fonction estimante carré intégrable et sans biais de $\{p_c(\cdot, \phi), \phi \in \Phi\}$ tel que

$$\begin{aligned} g : \mathcal{V} \times \Phi &\longrightarrow \mathbb{R}^p \\ (\mathbf{v}, \phi) &\longmapsto g(\mathbf{v}, \phi) \end{aligned}$$

alors la fonction G sur données incomplètes définie par

$$\begin{aligned} G : \mathcal{U} \times \Phi &\longrightarrow \mathbb{R}^p \\ (\mathbf{u}, \phi) &\longmapsto E_\phi[g(\mathbf{V}, \phi) | \mathbf{u}] = \int g(\mathbf{v}, \phi) p(\mathbf{v} | \mathbf{u}, \phi) \mu(d\mathbf{v}) \end{aligned}$$

est une fonction estimante sans biais, carré intégrable.

Démonstration. Nous notons $G(\phi) = G(\mathbf{U}, \phi)$ et $g(\phi) = g(\mathbf{V}, \phi)$. La fonction G est carré intégrable puisque c'est la projection de $g(\phi)$ appartenant à L^2 et G est sans biais parce que

$$\forall \phi, E_\phi[G(\phi)] = E_\phi[E_\phi[g(\phi) | \mathbf{U}]] = E_\phi[g(\phi)] = 0$$

□

Cette méthode de construction est particulièrement simple parce qu'il n'est pas nécessaire d'utiliser la projection sur un espace de fonctions estimantes prédéterminées $\mathcal{H}_{\mathbf{u}}$, comme le font Heyde et Morton. De plus, nous pouvons avoir une forme explicite de la fonction projetée dans de nombreux cas.

Le maximum de vraisemblance est l'estimateur obtenu lorsque la fonction estimante est le vecteur score. La proposition suivante montre que les estimateurs du maximum de vraisemblance complet et incomplet sont directement reliés.

Proposition 3.5.2. Identité de Fisher

Soit $s(\phi) = \nabla_\phi L_c(\mathbf{V}, \phi)$ le vecteur score complet, et $S(\phi) = \nabla_\phi L(\mathbf{U}, \phi)$ le vecteur score incomplet. Si nous pouvons intervertir intégration et dérivation, alors

$$S(\phi) = E_\phi[s(\phi) | \mathbf{U}] \tag{3.20}$$

Démonstration. Nous supposons que $p(\mathbf{v}|\mathbf{u}, \phi)$, la densité de $\mathbf{V}$ conditionnellement à $\mathbf{U}$, est telle que son support $H_{\mathbf{u}} = \{\mathbf{v} \in \mathcal{V} | \varphi(\mathbf{v}) = \mathbf{u}\}$ soit indépendant de ϕ . Nous avons alors :

$$\begin{aligned}
E_{\phi}[s(\phi)|\mathbf{u}] &= \int_{H_{\mathbf{u}}} \frac{\nabla_{\phi} p_c(\mathbf{v}, \phi)}{p_c(\mathbf{v}, \phi)} p(\mathbf{v}|\mathbf{u}, \phi) \nu(d\mathbf{v}) \\
&= \int_{H_{\mathbf{u}}} \frac{\nabla_{\phi} p(\mathbf{v}|\mathbf{u}, \phi) p(\mathbf{u}, \phi) + p(\mathbf{v}|\mathbf{u}, \phi) \nabla_{\phi} p(\mathbf{u}, \phi)}{p(\mathbf{v}|\mathbf{u}, \phi) p(\mathbf{u}, \phi)} p(\mathbf{v}|\mathbf{u}, \phi) \nu(d\mathbf{v}) \\
&= \int_{H_{\mathbf{u}}} \nabla_{\phi} p(\mathbf{v}|\mathbf{u}, \phi) \nu(d\mathbf{v}) + \int_{H_{\mathbf{u}}} \frac{\nabla_{\phi} p(\mathbf{u}, \phi)}{p(\mathbf{u}, \phi)} p(\mathbf{v}|\mathbf{u}, \phi) \nu(d\mathbf{v}) \\
&= \frac{\nabla_{\phi} p(\mathbf{u}, \phi)}{p(\mathbf{u}, \phi)}
\end{aligned}$$

□

Ainsi la projection de la fonction estimante optimale (à horizon fini) sur les données complètes est la fonction estimante optimale (à horizon fini) sur les données incomplètes : la proposition suivante montre que l'optimalité se transmet de manière générale par projection. Nous redémontrons cette proposition déjà montré dans [90], mais en utilisant la caractérisation de l'espérance conditionnelle. Nous notons $\Pi_{\mathbf{u}}(\mathcal{H}_{\mathbf{v}})$ l'ensemble des fonctions estimantes sur données incomplètes obtenues par la proposition (3.5.1).

Proposition 3.5.3. Projection des quasi-scores

Si $g^*(\phi)$ est un quasi-score dans la famille convexe $\mathcal{H}_{\mathbf{v}}$, alors $G^*(\phi) = E_{\phi}[g^*(\phi)|\mathbf{U}]$ est un quasi-score dans la famille $\Pi_{\mathbf{u}}(\mathcal{H}_{\mathbf{v}})$.

Démonstration. C'est une conséquence des propriétés de l'espérance conditionnelle, comme projecteur orthogonal dans l'espace de Hilbert des fonctions carré intégrables. En effet, il existe plusieurs caractérisations de l'optimalité d'une fonction estimante :

$$g^* \text{ optimal dans } \mathcal{H}_{\mathbf{v}} \text{ si et seulement si } g^* = \arg \min_{g \in \mathcal{H}_{\mathbf{v}}} E_{\phi} \left[(g - s(\phi))(g - s(\phi))' \right]$$

ce qui signifie que le quasi-score est la fonction "la plus corrélée" au score $s(\phi)$, lorsque ce dernier existe. Heyde a montré dans [90] que si Q est un quasi-score au sein de $\Pi_{\mathbf{u}}(\mathcal{H}_{\mathbf{v}})$, alors il doit vérifier

$$Q = \arg \min_{G \in \Pi_{\mathbf{u}}(\mathcal{H}_{\mathbf{v}})} E_{\phi} \left[(G - g^*)(G - g^*)' \right]$$

En particulier, nous avons aussi

$$Tr(E_{\phi} \left[(Q - g^*)(Q - g^*)' \right]) = \inf_{G \in \Pi_{\mathbf{u}}(\mathcal{H}_{\mathbf{v}})} Tr(E_{\phi} \left[(G - g^*)(G - g^*)' \right])$$

soit encore $E_{\phi} \left[\|Q - g^*\|^2 \right] = \inf_{G \in \Pi_{\mathbf{u}}(\mathcal{H}_{\mathbf{v}})} E_{\phi} \left[\|G - g^*\|^2 \right]$. Or cette borne inférieure est atteinte de manière unique par la projection orthogonale G^* (l'ensemble $\Pi_{\mathbf{u}}(\mathcal{H}_{\mathbf{v}})$ est convexe), donc nous avons $Q = G^*$.

□

Pour le calcul de la fonction normalisée $G^{(s)}$ ainsi que du critère d'information $\mathcal{E}(G)$, nous avons besoin des expressions de la variance de G et du jacobien $\partial_\phi G$. L'expression de la variance de G (notée $\text{cov}_\phi(G)$) se déduit de la formule de la variance conditionnelle :

$$\text{cov}_\phi(G(\phi)) = \text{cov}_\phi(g(\phi)) - E_\phi [\text{cov}_\phi(g(\mathbf{V}, \phi) | \mathbf{U})] \quad (3.21)$$

où $\mathbf{u} \mapsto \text{cov}_\phi(g(\mathbf{V}, \phi) | \mathbf{u})$ représente la covariance de $g(\phi)$ conditionnellement à $\mathbf{U}$ évaluée en $\mathbf{u}$.

De même, nous explicitons le lien entre le jacobien de la fonction estimante projetée et celui de celui de g . Il apparaît que cela correspond alors à une généralisation de la formule de Louis, pour n'importe quelle fonction estimante g , et plus seulement pour le score.

Proposition 3.5.4. Formule de Louis Généralisée

Si nous pouvons intervertir dérivation et intégration, nous avons

$$\partial_\phi G(\mathbf{u}, \phi) = E_\phi [\partial_\phi g(\mathbf{V}, \phi) | \mathbf{u}] + \text{cov}_\phi(g(\mathbf{V}, \phi), s(\mathbf{V}, \phi) | \mathbf{u}) \quad (3.22)$$

Démonstration. Nous calculons les dérivées de chaque composante de $G(\mathbf{u}, \phi) = (G_i(\mathbf{u}, \phi))_{1 \leq i \leq d}$, et notons $g(\mathbf{v}, \phi) = (g_i(\mathbf{v}, \phi))_{1 \leq i \leq d}$.

$$\begin{aligned} \nabla_\phi G_i(\mathbf{u}, \phi) &= \nabla_\phi \int_{H_{\mathbf{u}}} g_i(\mathbf{v}, \phi) p(\mathbf{v} | \mathbf{u}, \phi) \nu(d\mathbf{v}) \\ &= \int_{H_{\mathbf{u}}} \nabla_\phi g_i(\phi) p(\mathbf{v} | \mathbf{u}, \phi) \nu(d\mathbf{v}) + \int_{H_{\mathbf{u}}} g_i(\mathbf{v}, \phi) \nabla_\phi p(\mathbf{v} | \mathbf{u}, \phi) \nu(d\mathbf{v}) \end{aligned} \quad (3.23)$$

Or nous avons

$$\nabla_\phi p(\mathbf{v} | \mathbf{u}, \phi) = \frac{\nabla_\phi p(\mathbf{v}, \phi) p(\mathbf{u}, \phi) - p(\mathbf{v}, \phi) \nabla_\phi p(\mathbf{u}, \phi)}{p(\mathbf{u}, \phi)^2}$$

ce qui nous permet de réécrire le second terme de droite de (3.23), soit :

$$\begin{aligned} \int_{H_{\mathbf{u}}} g_i(\mathbf{v}, \phi) \nabla_\phi p(\mathbf{v} | \mathbf{u}, \phi) \nu(d\mathbf{v}) &= \int_{H_{\mathbf{u}}} g_i(\mathbf{v}, \phi) \frac{\nabla_\phi p(\mathbf{v}, \phi)}{p(\mathbf{u}, \phi)} \nu(d\mathbf{v}) \\ &\quad - \int_{H_{\mathbf{u}}} g_i(\mathbf{v}, \phi) \frac{p(\mathbf{v} | \mathbf{u}, \phi) \nabla_\phi p(\mathbf{u}, \phi)}{p(\mathbf{u}, \phi)} \nu(d\mathbf{v}) \\ &= \int_{H_{\mathbf{u}}} g_i(\mathbf{v}, \phi) \nabla_\phi \log p(\mathbf{v}, \phi) p(\mathbf{v} | \mathbf{u}, \phi) \nu(d\mathbf{v}) \\ &\quad - \frac{\nabla_\phi p(\mathbf{u}, \phi)}{p(\mathbf{u}, \phi)} \int_{H_{\mathbf{u}}} g_i(\mathbf{v}, \phi) p(\mathbf{v} | \mathbf{u}, \phi) \nu(d\mathbf{v}) \\ &= E_\phi [g_i(\mathbf{V}, \phi) \nabla_\phi \log p(\mathbf{V}, \phi) | \mathbf{u}] - G_i(\mathbf{u}, \phi) S(\mathbf{u}, \phi) \end{aligned}$$

Nous avons donc

$$\begin{aligned} \int_{H_{\mathbf{u}}} g(\mathbf{v}, \phi) \nabla_\phi p(\mathbf{v} | \mathbf{u}, \phi)' \nu(d\mathbf{v}) &= E_\phi [g(\mathbf{V}, \phi) s(\mathbf{V}, \phi)' | \mathbf{u}] - E_\phi [g(\mathbf{V}, \phi) | \mathbf{u}] E_\phi [s(\mathbf{V}, \phi)' | \mathbf{u}] \\ &= \text{cov}_\phi(g(\mathbf{V}, \phi), s(\mathbf{V}, \phi) | \mathbf{u}) \end{aligned}$$

□

Remarque 3.5.1. Nous retrouvons bien la décomposition de Louis (voir chapitre 4, [108]) lorsque nous utilisons comme fonction estimante sur les données complètes le score $g(\mathbf{v}, \phi) = s(\mathbf{v}, \phi)$ dans la formule (3.22) :

$$\partial_\phi S(\mathbf{u}, \phi) = E_\phi [\partial_\phi s(\mathbf{V}, \phi) | \mathbf{u}] + \text{cov}_\phi (s(\mathbf{V}, \phi), s(\mathbf{V}, \phi) | \mathbf{u})$$

Et en introduisant les informations de Fisher observées pour données complètes et incomplètes $I_c(\mathbf{v}, \phi) = -\partial_\phi s(\mathbf{v}, \phi)$ et $I(\mathbf{u}, \phi) = -\partial_\phi S(\mathbf{u}, \phi)$, la décomposition de Louis se réécrit :

$$I(\mathbf{u}, \phi) = E_\phi [I_c(\mathbf{V}, \phi) | \mathbf{u}] - \text{cov}_\phi (s(\mathbf{V}, \phi), s(\mathbf{V}, \phi) | \mathbf{u})$$

que l'on appelle aussi le principe d'information manquante, [108].

Rappelons l'expression de l'information de Fisher généralisée

$$\mathcal{E}(G) = E_\phi [\partial_\phi G(\phi)]' E_\phi [G(\phi)G(\phi)']^{-1} E_\phi [\partial_\phi G(\phi)]$$

et récapitulons l'expression du jacobien et de la variance :

$$\begin{aligned} E_\phi [\partial_\phi G(\mathbf{U}, \phi)] &= E_\phi [\partial_\phi g(\mathbf{V}, \phi)] + E_\phi [\text{cov}_\phi (g(\mathbf{V}, \phi), s(\mathbf{V}, \phi) | \mathbf{U})] \\ \text{cov}_\phi (G(\phi)) &= \text{cov}_\phi (g(\phi)) - E_\phi [\text{cov}_\phi (g(\mathbf{V}, \phi), s(\mathbf{V}, \phi) | \mathbf{U})] \end{aligned} \quad (3.24)$$

Si le calcul effectif de ces grandeurs pour un modèle précis est difficile, il fournit des indications (finalement assez intuitive) sur l'influence de l'observation partielle de $\mathbf{V}$ sur la qualité de l'estimation. La variance des estimateurs des paramètres augmente nécessairement lorsque nous perdons des observations, mais il apparaît que la variance asymptotique des paramètres est d'autant plus petite que la covariance de g conditionnellement à $\mathbf{U}$ est forte et que g est corrélée au score (complet). Les expressions rassemblées dans (3.24) nous indiquent par exemple que nous pouvons avoir intérêt à choisir une fonction estimante g_1 sur données complètes moins corrélée au score qu'une autre fonction g_2 , si la variance conditionnelle de g_1 est bien supérieure à celle de g_2 .

Nous venons de voir que nous conservons des propriétés intéressantes par projection, notamment l'optimalité. Nous nous posons maintenant la question de savoir sous quelles conditions la consistance des estimateurs se transmet elle aussi par projection. Pour cela, nous commençons tout d'abord par remarquer que la démonstration de la consistance des racines des fonctions estimantes suit une méthode similaire à celle de Cramer pour la consistance de l'EMV. Contrairement à la méthode de Wald, celle de Cramer fournit un résultat plus faible puisqu'elle ne montre que la consistance d'une suite de racines de la log-vraisemblance vers le vrai paramètre, et ne l'identifie pas à la suite des maxima globaux de la vraisemblance. Ainsi, lorsque nous trouvons plusieurs racines à des fonctions estimantes quelconques, se pose le problème du choix de la racine. Néanmoins, cette propriété de consistance peut être montrée sous des conditions assez faibles, comme le montre le théorème suivant.

Théorème 3.5.1. Critère de consistance ([89], p. 183)

Soit $(g_n(\phi))_{n \geq 1}$ une suite de fonctions estimantes continues $P_{\phi_0} - ps$ sur Φ , telle que $\exists \delta_0$, t.q. $\forall 0 < \delta < \delta_0$, $P_{\phi_0} - ps$, $\exists \epsilon > 0$ tel que

$$\limsup_{n \rightarrow \infty} \left(\sup_{\|\phi - \phi_0\| = \delta} (\phi - \phi_0)' g_n(\phi) \right) < -\epsilon \quad (3.25)$$

Alors il existe une suite d'estimateurs $(\hat{\phi}_n)_{n \geq 1}$, telle que $\hat{\phi}_n(\mathbf{V}) \rightarrow \phi_0 P_{\phi_0} - ps$ et $g_n(\hat{\phi}_n(\mathbf{V})) = 0$ pour n assez grand.

Remarque 3.5.2. Par un corollaire du théorème du point fixe de Brouwer, la condition (3.25) permet d'assurer l'existence d'une racine à l'équation $g_n(\phi) = 0$. En effet, si une application $g_n : \mathbb{R}^p \rightarrow \mathbb{R}^p$ continue vérifie

$$\forall \phi \setminus \|\phi - \phi_0\| = \delta, (\phi - \phi_0)' g_n(\phi) < 0 \quad (3.26)$$

alors il existe ϕ^* tel que $\|\phi^* - \phi_0\| \leq \delta$ et $g_n(\phi^*) = 0$. Géométriquement, la propriété (3.26) signifie que la fonction $g_n(\phi)$ définit un flux de vecteurs entrant dans la boule $\{\phi \mid \|\phi - \phi_0\| \leq \delta\}$. La continuité implique alors que ce champ de vecteur doit s'annuler au moins une fois dans la boule.

Le théorème du point fixe de Brouwer affirme que tout fonction continue de C dans C , où C est un convexe compact de $\mathbb{R}^p$, admet (au moins) un point fixe.

Nous avons vu que la projection des fonctions estimantes permet de déduire un certain nombre de propriétés des fonctions estimantes sur données complètes, pour lesquelles il est souvent plus facile de démontrer des résultats. Il est particulièrement intéressant de savoir s'il est possible de démontrer la consistance des fonctions projetées de la même façon. Nous donnons ci-dessous une heuristique pour transférer la consistance sous des hypothèses générales. En effet, la condition (3.25) peut être conservée par projection, sous des conditions que nous mettons en évidence ci-dessous.

Nous notons pour $\delta > 0$, $\mathcal{C}_\delta = \{\phi \mid \|\phi - \phi_0\| = \delta\}$ et $\mathcal{B}_\delta = \{\phi \mid \|\phi - \phi_0\| \leq \delta\}$. Nous faisons un développement de Taylor au premier ordre, en introduisant $\tilde{\phi}_n(\mathbf{v}) \in \mathcal{B}_\delta$:

$$\forall \phi \in \mathcal{C}_\delta, g_n(\mathbf{v}, \phi) = g_n(\mathbf{v}, \phi_0) + \partial_\phi g_n(\mathbf{v}, \tilde{\phi}_n(\mathbf{v})) (\phi - \phi_0)$$

Donc pour $G_n(\mathbf{u}, \phi)$ nous pouvons écrire

$$\begin{aligned} (\phi - \phi_0)' G_n(\mathbf{u}, \phi) &= \int (\phi - \phi_0)' g_n(\mathbf{v}, \phi_0) p_n(\mathbf{v} | \mathbf{u}, \phi) \mu(d\mathbf{v}) \\ &\quad + \int (\phi - \phi_0)' \partial_\phi g_n(\mathbf{v}, \tilde{\phi}_n(\mathbf{v})) (\phi - \phi_0) p_n(\mathbf{v} | \mathbf{u}, \phi) \mu(d\mathbf{v}) \end{aligned}$$

Si les fonctions $\mathbf{v} \mapsto g_n(\mathbf{v}, \phi_0) p_n(\mathbf{v} | \mathbf{u}, \phi)$ et $\mathbf{v} \mapsto (\phi - \phi_0)' \partial_\phi g_n(\mathbf{v}, \tilde{\phi}_n(\mathbf{v})) (\phi - \phi_0) p_n(\mathbf{v} | \mathbf{u}, \phi)$ sont

dominées uniformément pour tout $\phi \in \mathcal{C}_\delta$, alors nous pouvons affirmer que

$$\begin{aligned} \sup_{\phi \in \mathcal{C}_\delta} (\phi - \phi_0)' G_n(\mathbf{u}, \phi) &\leq \int \sup_{\phi \in \mathcal{C}_\delta} (\phi - \phi_0)' g_n(\mathbf{v}, \phi_0) p_n(\mathbf{v}|\mathbf{u}, \phi) \mu(d\mathbf{v}) \\ &\quad + \int \sup_{\phi \in \mathcal{C}_\delta} (\phi - \phi_0)' \partial_\phi g_n(\mathbf{v}, \tilde{\phi}_n(\mathbf{v})) (\phi - \phi_0) p_n(\mathbf{v}|\mathbf{u}, \phi) \mu(d\mathbf{v}) \end{aligned}$$

Nous réécrivons cette inégalité sous forme d'espérance conditionnelle par rapport au vrai paramètre ϕ_0

$$\begin{aligned} \sup_{\phi \in \mathcal{C}_\delta} (\phi - \phi_0)' G_n(\mathbf{u}, \phi) &\leq E_{\phi_0} \left[\sup_{\phi \in \mathcal{C}_\delta} (\phi - \phi_0)' g_n(\mathbf{V}, \phi_0) \frac{p_n(\mathbf{V}|\mathbf{u}, \phi)}{p_n(\mathbf{V}|\mathbf{u}, \phi_0)} | \mathbf{u} \right] \\ &\quad + E_{\phi_0} \left[\sup_{\phi \in \mathcal{C}_\delta} Q_n(\mathbf{V}, \phi) \frac{p_n(\mathbf{V}|\mathbf{u}, \phi)}{p_n(\mathbf{V}|\mathbf{u}, \phi_0)} | \mathbf{u} \right] \end{aligned} \quad (3.27)$$

avec $Q_n(\mathbf{V}, \phi) = (\phi - \phi_0)' \partial_\phi g_n(\mathbf{V}, \tilde{\phi}_n(\mathbf{V})) (\phi - \phi_0)$. Nous pouvons assurer l'existence d'un ϵ' tel que $\limsup \left(\sup_{\phi \in \mathcal{C}_\delta} (\phi - \phi_0)' G_n(\mathbf{u}, \phi) \right) < -\epsilon'$, si nous pouvons montrer que chaque terme de l'expression de droite de l'inégalité (3.27) est strictement négatif pour un δ assez petit.

Si nous supposons que la fonction estimante g_n vérifie la propriété (3.25), et que le jacobien $\partial_\phi g_n(\mathbf{V}, \phi)$ converge uniformément pour $\phi \in \mathcal{B}_\delta$ vers une matrice définie négative¹² $P_{\phi_0} - ps$, la majoration par $-\epsilon' < 0$ est assurée si $\liminf_n \frac{p_n(\mathbf{V}|\mathbf{u}, \phi)}{p_n(\mathbf{V}|\mathbf{u}, \phi_0)}$ reste bornée inférieurement par une fonction positive à support non-vide $P_{\phi_0} - ps$, pour tout $\phi \in \mathcal{C}_\delta$. Dans ce cas-là, les deux espérances ne peuvent pas être nulles et sont bien strictement négatives, et le théorème (3.5.1) s'applique.

3.5.2 Estimation Conditionnelle Itérative

La fonction estimante $(\mathbf{u}, \phi) \mapsto G(\mathbf{u}, \phi) = \int g(\mathbf{v}, \phi) p(\mathbf{v}|\mathbf{u}, \phi) \mu(d\mathbf{v})$ possède une expression qui la rend difficilement exploitable en pratique. Il est toujours possible de calculer point par point l'intégrale sur un maillage de l'espace des paramètres, puis d'en rechercher les zéros, mais cela n'est pas envisageable dès que l'espace des paramètres est grand et que les modèles deviennent complexes. Plusieurs méthodes numériques existent pour la recherche des racines d'une équation (Newton, Gradient, Score de Fisher, ...), mais elles supposent de savoir calculer la fonction G , ainsi que ses dérivées. Nous proposons ici une méthode pratique qui évite ces problèmes en tirant partie de la forme particulière de $G(\phi)$ et de l'existence d'un modèle complet. Le principe d'estimation conditionnelle itérative, similaire celui à de EM, consiste à dédoubler le paramètre ϕ , en introduisant la fonction $h(\phi_1, \phi_2)$ définie par

$$\forall \phi_1, \phi_2, h(\phi_1, \phi_2) = E_{\phi_1} [g(\mathbf{V}, \phi_2) | \mathbf{u}] \quad (3.28)$$

L'objectif est de construire, à partir de la fonction h et d'un paramètre initial ϕ_0 , une suite $(\phi_n)_{n \geq 1}$ obtenue en résolvant en ϕ_{n+1} l'équation

$$h(\phi_n, \phi_{n+1}) = 0 \quad (3.29)$$

¹²Si g_n est un quasi-score nous avons $E_{\phi_0} [\partial_\phi g_n(\mathbf{V}, \phi_0)] = -E_{\phi_0} [g_n(\phi_0) g_n(\phi_0)']$, et donc $\partial_\phi g_n(\mathbf{V}, \phi_0)$ converge dans L^1 vers une matrice définie négative.

Si la suite converge vers une valeur ϕ_∞ , alors la valeur limite vérifiera $E_{\phi_\infty} [g(\mathbf{V}, \phi_\infty) | \mathbf{u}] = 0$ et sera donc une racine de l'équation $G(\mathbf{u}, \phi) = 0$.

Nous donnons dans cette section des conditions permettant d'assurer l'existence et la convergence d'une telle suite. A ϕ_0 fixé, le théorème des fonctions implicites permet de déduire du comportement local de h l'existence d'une racine à l'équation $h(\phi_0, \phi) = 0$. Nous notons $\partial_i h$, $i = 1, 2$ la différentielle partielle par rapport à la i ème variable. Si h est de classe C^1 et si ϕ_0, ϕ'_0 sont deux points dans l'intérieur de Φ , alors l'inversibilité de la matrice $\partial_2 h(\phi_0, \phi'_0)$ permet d'affirmer qu'il existe une fonction $\mathcal{T}$, de classe C^1 définie sur un voisinage $\mathcal{V}(\phi_0)$ de ϕ_0 tel que

$$\forall \phi \in \mathcal{V}(\phi_0), h(\phi, \mathcal{T}(\phi)) = 0 \quad (3.30)$$

Notre objectif est donc de déterminer les points fixes de la fonction $\mathcal{T}$ qui sont racines de l'équation $G(\phi) = 0$. Nous supposons que la fonction $\mathcal{T}$ est telle que $\forall \phi \in \mathcal{V}(\phi_0), \mathcal{T}(\phi) \in \mathcal{V}(\phi_0)$, et que de plus $\mathcal{V}(\phi_0)$ est compact et convexe. Ceci permet d'une part de rendre possible la construction de la suite $(\phi_n)_{n \geq 1}$ et d'autre part d'assurer par l'existence d'au moins un point fixe par le théorème de Brouwer. Parmi les points fixes de $\mathcal{T}$, certains peuvent être attractifs ou répulsifs, et le comportement de la suite $(\phi_n)_{n \geq 1}$ dépend alors du jacobien de $\mathcal{T}$ en ces points. L'expression du jacobien au point initial ϕ_0 est donné par le théorème des fonctions implicites

$$\partial_\phi \mathcal{T}(\phi_0) = - (\partial_2 h|_{(\phi_0, \mathcal{T}(\phi_0))})^{-1} \partial_1 h|_{(\phi_0, \mathcal{T}(\phi_0))} \quad (3.31)$$

Si les modules de toutes les valeurs propres de $\partial_\phi \mathcal{T}(\phi_0)$ sont strictement inférieures à 1 (i.e. son rayon spectral noté $\|\partial_\phi \mathcal{T}(\phi_0)\|$ est inférieur à 1), nous pouvons affirmer que le rayon spectral du jacobien est inférieur à 1 sur un voisinage $\mathcal{V}'(\phi_0) \subset \mathcal{V}(\phi_0)$. Si $\mathcal{V}'(\phi_0)$ est à son tour compact, l'application $\mathcal{T}$ sera contractante parce que nous aurons $\|\mathcal{T}(\phi_1) - \mathcal{T}(\phi_0)\| \leq \left\{ \sup_{\phi \in \mathcal{V}'(\phi_0)} \|\partial_\phi \mathcal{T}\| \right\} \|\phi_1 - \phi_0\|$. Sous cette condition, $\mathcal{T}$ aura donc un unique point fixe (attracteur) $\tilde{\phi}$ sur $\mathcal{V}'(\phi_0)$ et vers lequel converge $(\phi_n)_{n \geq 1}$. Les expressions des jacobiens de h $\partial_1 h(\phi_1, \phi_2) = \partial_{\phi_1} (E_{\phi_1} [g(\mathbf{V}, \phi_2) | \mathbf{u}]) = \text{cov}_{\phi_1} (g(\mathbf{V}, \phi_2), s(\mathbf{V}, \phi_1) | \mathbf{u})$ et $\partial_2 h(\phi_1, \phi_2) = E_{\phi_1} [\partial_{\phi_2} g(\mathbf{V}, \phi_2) | \mathbf{u}]$ permettent donc d'affirmer que si pour tout $\phi \in \mathcal{V}'(\phi_0)$, la plus grande valeur propre (en module) de la matrice

$$\text{cov}_\phi (g(\mathbf{V}, \mathcal{T}(\phi)), s(\mathbf{V}, \phi) | \mathbf{u})^{-1} E_\phi [\partial_\phi g(\mathbf{V}, \mathcal{T}(\phi)) | \mathbf{u}] \quad (3.32)$$

est plus petite que 1, $(\phi_n)_{n \geq 1}$ converge vers un point de $\mathcal{V}'(\phi_0)$. Nous avons donc le résultat suivant :

Théorème 3.5.2. *Si la fonction $\mathcal{T}$ définie implicitement par l'équation (3.30) peut être restreinte à un voisinage compact de ϕ_0 , noté $\mathcal{C}_{\phi_0}$, tel que*

(i) $\mathcal{T}(\mathcal{C}_{\phi_0}) \subset \mathcal{C}_{\phi_0}$

(ii) $\sup_{\phi \in \mathcal{C}_{\phi_0}} \left\| \text{cov}_\phi (g(\mathbf{V}, \mathcal{T}(\phi)), s(\mathbf{V}, \phi) | \mathbf{u})^{-1} E_\phi [\partial_\phi g(\mathbf{V}, \mathcal{T}(\phi)) | \mathbf{u}] \right\| < 1$

alors $\mathcal{T}$ possède un seul point fixe $\tilde{\phi}$ dans $\mathcal{C}_{\phi_0}$ et toute suite $\phi_{n+1} = \mathcal{T}(\phi_n)$ converge vers $\tilde{\phi}$.

Les hypothèses de ce théorème sont difficiles à vérifier et ne constituent un critère pratique pour étudier le comportement asymptotique de la suite $(\phi_n)_{n \geq 1}$.

Si la suite $(\phi_n)_{n \geq 1}$ converge vers un point fixe $\tilde{\phi}$, le jacobien de $\mathcal{T}$ évalué au point $\tilde{\phi}$ permet de déterminer la vitesse de convergence asymptotique de la suite. Le jacobien au point fixe vaut

$\partial_\phi \mathcal{T}(\tilde{\phi}) = - \left(E_{\tilde{\phi}} \left[\partial_{\tilde{\phi}} g(\mathbf{V}, \tilde{\phi}) | \mathbf{u} \right] \right)^{-1} \text{cov}_{\tilde{\phi}} \left(g(\mathbf{V}, \tilde{\phi}), s(\mathbf{V}, \tilde{\phi}) | \mathbf{u} \right)$. La formule de Louis généralisée (3.22), donne l'expression de $\left(E_{\tilde{\phi}} \left[\partial_{\tilde{\phi}} g(\mathbf{V}, \tilde{\phi}) | \mathbf{u} \right] \right)$ en fonction de la fonction estimante projetée G , ce qui permet de réécrire le jacobien

$$\partial_\phi \mathcal{T}(\tilde{\phi}) = \left(Id_p - \text{cov}_{\tilde{\phi}} \left(g(\mathbf{V}, \tilde{\phi}), s(\mathbf{V}, \tilde{\phi}) | \mathbf{u} \right)^{-1} \partial_\phi G(\mathbf{u}, \tilde{\phi}) \right)^{-1} \quad (3.33)$$

où Id_p est la fonction identité sur $\mathbb{R}^p$. La vitesse de convergence est alors réglée par $\lambda(G)$, le module de la plus grande valeur propre de $\text{cov}_{\tilde{\phi}} \left(g(\mathbf{V}, \tilde{\phi}), s(\mathbf{V}, \tilde{\phi}) | \mathbf{u} \right)^{-1} \partial_\phi G(\mathbf{u}, \tilde{\phi})$. La vitesse sera élevée si la matrice $\text{cov}_{\tilde{\phi}} \left(g(\mathbf{V}, \tilde{\phi}), s(\mathbf{V}, \tilde{\phi}) | \mathbf{u} \right)$ est grande (au sens de l'ordre des matrices symétriques définies positives), c'est-à-dire si la fonction estimante complète est très corrélée au score (complet) conditionnellement aux observations. Nous retrouvons ici une des conditions pour que la variance asymptotique des estimateurs (égale à $\mathcal{E}(G)^{-1}$) soit faible.

En conclusion, si ϕ_0 est un paramètre quelconque tel que $E_{\phi_0} [\partial_\phi g(\mathbf{V}, \phi) | \mathbf{u}]$ soit une matrice inversible, et que les hypothèses du théorème 3.5.2 sont vérifiées alors la suite $(\phi_n)_{n \geq 1}$ converge vers une racine de l'équation $G(\mathbf{v}, \phi) = 0$. Sous ces conditions, nous pouvons définir l'algorithme suivant, appelé algorithme d'Estimation Conditionnelle Itérative (ECI) :

Algorithme 3.5.1. Algorithme ECI

Soit ϕ_0 une valeur initiale appartenant à Φ . La suite $(\phi_n)_{n \geq 1}$ est définie récursivement par

$$\forall n \geq 1, \phi_{n+1} \text{ est solution en } \phi \text{ de l'équation } E_{\phi_n} [g(\mathbf{V}, \phi) | \mathbf{u}] = 0$$

Remarque 3.5.3. ECI Stochastique

De même que pour EM, nous pouvons proposer une version stochastique d'ECI, en remplaçant l'espérance E_{ϕ_n} par la simulation des données manquantes selon $p(\mathbf{v} | \mathbf{u}, \phi_n)$ et la résolution de l'équation avec données complètes. De manière similaire à SEM, la suite des paramètres $(\phi_n)_{n \geq 0}$ ainsi construite est une chaîne de Markov, et le caractère aléatoire permet une meilleure exploration de l'espace des paramètres que par ECI déterministe.

L'algorithme EM est un algorithme ECI, appliqué à la fonction score des données complètes. En effet, pour un modèle à données incomplètes, la suite des paramètres $(\phi_n)_{n \geq 0}$ de l'algorithme EM est obtenue par la résolution successive du problème suivant (en rassemblant en une seule équation les étapes E et M)

$$(\nabla_\phi E_{\phi_n} [L_c(\mathbf{V}, \phi) | \mathbf{u}])|_{\phi=\phi_{n+1}} = 0$$

Si nous pouvons intervertir dérivation et intégration, une étape EM correspond donc à la résolution de l'équation

$$\begin{aligned} E_{\phi_n} [\nabla_\phi L_c(\mathbf{V}, \phi) | \mathbf{u}]|_{\phi=\phi_{n+1}} &= 0 \\ \iff E_{\phi_n} [s(\mathbf{V}, \phi) | \mathbf{u}]|_{\phi=\phi_{n+1}} &= 0 \end{aligned}$$

La convergence de la suite $(\phi_n)_{n \geq 0}$ construite par EM provient de la croissance de la fonction $(L(\mathbf{u}, \phi_n))$, (voir théorèmes de Wu, chapitre 3 [108]), contrairement à ECI pour qui cela repose sur

la propriété d'attraction d'un point fixe de l'application $\mathcal{T}$. Ainsi la convergence d'ECI ne peut se déduire en générale de la maximisation d'une certaine fonction, et la vérification de l'attractabilité se fait par l'examen de quantités similaires à celles qui sont étudiées pour la vitesse de convergence de l'algorithme EM. En pratique, cette condition est rarement vérifiable (d'autant qu'elle dépend clairement du point ϕ_0 choisie), et une méthode pragmatique est de relancer cet algorithme jusqu'à obtenir convergence, en utilisant différentes initialisations.

Malgré tout, la convergence d'ECI vers une racine de la fonction estimante n'est en aucun cas la garantie de l'obtention d'un bon estimateur de ϕ . Ainsi, le problème de l'influence de la valeur initiale est une limitation commune à EM. Ce problème peut être traité en lançant en parallèle plusieurs algorithmes ECI, ce qui donne lieu à un problème de sélection de racines. De manière générale, lorsque plusieurs racines sont trouvées à une fonction estimante, un test d'ajustement aux données permet de sélectionner la meilleure ([89], chapitre 13).

3.5.3 Quelques algorithmes ECI

L'algorithme ECI est une méthode de recherche de racines, et il est donc nécessaire de préciser la fonction estimante qu'on lui associe lors de l'estimation. Nous avons vu que lorsque nous prenons le score, nous obtenons l'algorithme EM (si nous pouvons intervertir intégration et dérivation). Dans les CMCA, la facilité de calcul de la vraisemblance et la faible complexité des modèles pour les lois d'émission (il s'agit souvent de modèles exponentiels) a rarement incité aux développements d'autres estimateurs non basés sur la vraisemblance (nous rappelons néanmoins la méthode des moindres carrés développé dans la thèse de Mevel [112] ou bien l'estimateur du maximum de la vraisemblance par bloc, proposée par Rydén [141]). C'est essentiellement dans les champs cachés que des méthodes alternatives ont été proposées, et nous donnons ci-dessous deux réinterprétations originales d'algorithmes classiques en termes de fonction estimante et d'algorithme ECI. L'utilisation (originale) d'ECI pour des chaînes sera exposée dans les chapitres 3 et 4, lorsqu'il s'agira d'estimer des copules.

3.5.3.1 Projection d'estimateur

Le "principe ECI" proposé par Pieczynski [122] part d'un estimateur $\mathbf{T}$ du modèle complet pour construire un estimateur du modèle incomplet en en prenant l'espérance conditionnelle. Si ϕ^* est le vrai paramètre, nous définissons

$$\tilde{\mathbf{T}}_{\phi^*}(\mathbf{U}) = E_{\phi^*} [\mathbf{T}(\mathbf{V}) | \mathbf{U}] \quad (3.34)$$

Comme $\tilde{\mathbf{T}}_{\phi^*}$ dépend du paramètre que l'on veut estimer, il n'est pas possible de le calculer : une procédure itérative est proposée alors, qui prend la forme

$$\forall n, \phi_{n+1} = E_{\phi_n} [\mathbf{T}(\mathbf{V}) | \mathbf{U}] \quad (3.35)$$

Nous sommes souvent amenés à remplacer, comme dans SEM, l'espérance conditionnelle par une phase d'imputation aléatoire des données manquantes par la loi a posteriori de $\mathbf{V}$ sachant $\mathbf{U}$ (lorsque $\phi = \phi_n$), et nous construisons alors une chaîne de Markov $(\phi_n)_{n \geq 1}$ explorant l'espace des paramètres. L'estimation retenue est alors la valeur moyenne lorsque la chaîne est stabilisée.

Le principe ECI permet de fournir des estimateurs satisfaisants, tant du point de vue de la qualité des résultats, que de la facilité de mise en oeuvre¹³ pour des modèles variées : en sonar [113], en analyse multispectrale [132], en modélisation de contour [55], en modélisation hiérarchique [97], ainsi que dans des contextes non-complètement probabilistes tel que la théorie de l'évidence [33] ou les chaînes de Markov cachées floues [17].

Nous pouvons réinterpréter le principe ECI en terme de projection de fonctions estimantes. En effet, dans le cas où l'estimateur $\mathbf{T}$ est un estimateur sans biais, nous pouvons lui associer la fonction estimante g

$$\forall \phi, g(\mathbf{V}, \phi) = \mathbf{T}(\mathbf{V}) - \phi$$

qui est carré intégrable si $\mathbf{T}$ est de variance finie. Si nous notons G la projection de g sur l'espace des fonctions estimantes en $\mathbf{U}$, nous avons :

$$\forall \phi, G(\mathbf{U}, \phi) = E_{\phi}[\mathbf{T}(\mathbf{V}) | \mathbf{U}] - \phi \quad (3.36)$$

En considérant directement la fonction estimante G plutôt que la statistique $\tilde{\mathbf{T}}_{\phi}$, nous contournons la difficulté liée à la dépendance de l'estimateur "naïf" $\tilde{\mathbf{T}}_{\phi}$ en le paramètre à estimer, car cette dépendance est déjà intégrée dans les fonctions estimantes g et G .

L'utilisation de l'algorithme ECI pour la résolution de $G(\mathbf{u}, \phi) = 0$ donne directement la formulation récursive (3.35). Pour cette fonction estimante, la résolution de l'équation est immédiate, et nous n'avons donc pas à vérifier l'existence de la suite des itérées. Cette difficulté laisse place au problème du calcul de l'espérance $E_{\phi}[\mathbf{T}(\mathbf{V}) | \mathbf{U}]$ qui nécessite souvent le recours à la simulation et l'utilisation d'un algorithme ECI stochastique. De plus, l'étude théorique du système dynamique stochastique (3.35) est difficile car nous n'avons pas une caractérisation claire de la limite $\phi_{\infty, N}(\mathbf{U})$ (si l'algorithme converge) à une taille d'échantillon N fixé (mise à part qu'elle soit égale à $E_{\phi_{\infty, N}(\mathbf{U})}[\mathbf{T}(\mathbf{V}) | \mathbf{U}]$), ni de son comportement lorsque N tend vers l'infini.

Récemment, un algorithme ECI (stochastique) a été utilisé pour l'estimation des champ de Markov couple pour lesquels la loi de $\mathbf{Y}$ conditionnellement à $\mathbf{X}$ est un champ de Markov Gaussien [14]. Afin d'obtenir des formules d'estimation simples, une estimation par moindres carrés du paramètre α caractérisant la dépendance du champ $\mathbf{X}$ (proposée par Derin et Elliot, voir chapitre 15 [156]) a été combinée à une estimation des paramètres du champ gaussien puis conditionnée par les observations. Les estimateurs obtenus permettent d'obtenir des procédures de segmentation non-supervisée qui peuvent améliorer dans certains cas celles obtenues par un algorithme de gradient stochastique (cherchant à maximiser la log-vraisemblance observée).

3.5.3.2 EM Gibbsien

L'algorithme EM Gibbsien est un algorithme d'estimation de champ markovien partiellement observé. Proposé par Chalmond [37, 159], cet algorithme cherche à maximiser la pseudo-vraisemblance au lieu de la vraisemblance. La pseudo-vraisemblance (introduite par Besag [19]) est définie par une technique de codage, c'est-à-dire par l'extraction d'un ensemble d'indices $L \subset S$

¹³Le principe ECI ne requiert pas "l'écriture de lignes de codes supplémentaires" par rapport à une procédure de segmentation non-supervisée, parce que nous avons uniquement besoin de savoir calculer et/ou échantillonner la loi a posteriori, et d'avoir un estimateur à données complètes (nécessaire dans les phases d'apprentissage des paramètres pour la segmentation supervisée).

maximal (au sens de l'inclusion) tel que les variables $(Z_s)_{s \in L}$ soient indépendantes entre elles conditionnellement aux variables restantes $(Z_s)_{s \notin L}$. La $(\log)L$ -pseudo-vraisemblance est définie par :

$$p_L(\phi) = \sum_{s \in L} \log p(z_s | (z_t)_{t \neq s}, \phi) \quad (3.37)$$

Son gradient permet de définir une fonction estimante car nous avons pour tout L

$$\begin{aligned} E \left[\nabla_\phi \sum_{s \in L} \log p(Z_s | (Z_t)_{t \neq s}, \phi) \right] &= E \left[E \left[\sum_{s \in L} \nabla_\phi \log p(Z_s | (Z_t)_{t \neq s}, \phi) \mid (Z_t)_{t \notin L} \right] \right] \\ &= E \left[\sum_{s \in L} E [\nabla_\phi \log p(Z_s | (Z_t)_{t \in V_s \cap L^c}, \phi) \mid (Z_t)_{t \in V_s \cap L^c}] \right] \\ &= 0 \end{aligned}$$

A tout codage L , nous pouvons associer le L -pseudo-score $s_L(\phi) = \nabla_\phi p_L(\phi)$ et nous pouvons améliorer la qualité de l'estimation en combinant plusieurs de ces fonctions estimantes (i.e. en les sommant, voir chapitre 6 [89]). La pseudo-vraisemblance est définie par

$$\tilde{L}(\phi) = \sum_{s \in S} \log p(z_s | (z_t)_{t \neq s}, \phi) \quad (3.38)$$

Son gradient $\nabla_\phi \tilde{L}(\phi)$ (appelé pseudo-score), qui est la somme des L -pseudo-scores, est encore une fonction estimante pour le champ markovien complet. Pour un champ partiellement observé, $\nabla_\phi \tilde{L}(\phi)$ est projeté sur l'espace des observations, et l'algorithme ECI est défini à partir de la fonction $h(\phi_1, \phi_2) = E_{\phi_1} [\nabla_\phi \tilde{L}(\phi) |_{\phi_2} | \mathbf{y}]$. L'algorithme obtenu ainsi est l'algorithme EM à la Gibbs. Younes dans [159] a montré dans le cas d'un champ markovien exponentiel que la projection du pseudo-score et l'algorithme ECI correspondant sont consistants *localement* :

- Il existe une racine à l'équation estimante qui converge vers le vrai paramètre ;
- l'EM gibbsien converge vers cette racine si la fonction h possède de "bonnes" propriétés.

3.5.3.3 Généralisation et algorithme "Projeter et Résoudre"

Pour conclure, nous rappelons les travaux de Heyde et Morton [90] qui permettent de considérer une classe plus large d'estimateurs (et par conséquent d'algorithmes d'estimation) mais qui se prête moins facilement à une étude théorique. Leur idée n'est plus de voir l'étape E de l'algorithme EM comme une espérance conditionnelle mais comme une étape de projection sur des espaces particuliers. Ceci les amène à définir l'algorithme Projeter et Résoudre (*Project and Solve*, noté P-S). Dans la méthode de projection que nous avons étudié, nous projetons une fonction estimante (complète) sur l'ensemble des fonctions estimantes incomplètes. Heyde *et al.* propose en toute généralité de projeter sur un espace contraint de fonctions estimantes incomplètes. Les auteurs considèrent $\mathcal{H}_\mathbf{v}$ une famille de fonctions estimantes (de moyenne nulle, et carré intégrable) utilisant les données complètes, et $\mathcal{H}_\mathbf{u}$ une famille de fonctions estimantes utilisant les données incomplètes (typiquement un sous espace vectoriel de $\mathcal{H}_\mathbf{v}$). A partir d'une fonction estimante $g \in \mathcal{H}_\mathbf{v}$, on construit une fonction estimante G appartenant à $\mathcal{H}_\mathbf{u}$ par un critère de moindre carré :

$$G = \arg \min_{H \in \mathcal{H}_{\mathbf{u}}} E_{\phi} [(H - g)(H - g)'] \quad (3.39)$$

où l'ordre utilisé pour définir la borne inférieure est celui sur les matrices définies positives. Le plus souvent, ceci se ramène à calculer la projection orthogonale de la fonction g sur l'espace $\mathcal{H}_{\mathbf{u}}$ (lorsque ce dernier est un sous-espace vectoriel de $\mathcal{H}_{\mathbf{v}}$ ou un convexe, par le théorème de projection dans les espaces de Hilbert). Dans de nombreuses situations intéressantes, comme par exemple le cas où $\mathcal{H}_{\mathbf{u}}$ est l'ensemble des fonctions estimantes linéaires en $\mathbf{u}$, la projection orthogonale sur $\mathcal{H}_{\mathbf{u}}$ n'est pas l'espérance conditionnelle (qui est le projecteur sur l'espace des fonctions de $\mathbf{u}$).

La définition de la fonction estimante $G(\mathbf{u}, \phi)$ par une projection la rend inexploitable, car les espérances intervenant dans le critère des moindres carrés (3.39) font intervenir le paramètre que l'on cherche. Heyde *et al.* suggèrent la méthode itérative suivante, partant d'une valeur initiale arbitraire ϕ_0 , dans laquelle nous notons $H(\phi | \phi_0, \mathbf{u})$ la fonction telle que

$$H(\phi | \phi_0, \mathbf{u}) = \arg \min_{H \in \mathcal{H}_{\mathbf{u}}} E_{\phi_0} [\|H(\mathbf{u}, \phi) - g(\mathbf{V}, \phi)\|^2]$$

Nous construisons itérativement ϕ_{n+1} en résolvant

$$H(\phi_{n+1} | \phi_n, \mathbf{u}) = 0 \quad (3.40)$$

Si la suite $(\phi_n)_{n \geq 0}$ converge vers une valeur $\tilde{\phi}$, alors elle vérifie $H(\tilde{\phi} | \tilde{\phi}, \mathbf{u}) = G(\mathbf{u}, \tilde{\phi})$, et par (3.40) est une racine de G . De plus, la proposition 3.5.3 est encore vraie dans ce contexte : si la fonction g est un quasi-score, alors il en est de même pour sa projection $G(\mathbf{u}, \phi)$ parmi l'ensemble $\mathcal{H}_{\mathbf{u}}$. Ainsi, l'ajout de contrainte sur le type des fonctions estimantes de $\mathcal{H}_{\mathbf{u}}$ (notamment la linéarité) permet de proposer des algorithmes d'estimation nouveau, et dont nous pouvons contrôler l'expression fonctionnelle ¹⁴.

Nous pensons que cette approche est une alternative très intéressante à l'EMV pour les modèles complexes, car elle permet de proposer de nombreuses solutions pragmatiques pour le calcul des paramètres : l'EMV perd de sa pertinence dans les CMCa car la multiplicité des racines de la log-vraisemblance ne permet plus d'identifier la racine à sélectionner comme celle étant un maximum global. Ainsi un résultat plus faible théoriquement qui ne garantit que la convergence d'une suite de racines, est en fait équivalent en pratique par cause de la difficulté à déterminer le maximum global. Cependant un aspect important à traiter ultérieurement est la vitesse de convergence des estimateurs ainsi obtenus.

¹⁴Heyde *et al.* donnent un exemple d'estimation dans un modèle à données manquantes où $\mathcal{H}_{\mathbf{u}}$ est l'ensemble des fonctions linéaires estimantes. La fonction estimante obtenue $G(\phi) = H(\phi, \phi, \mathbf{u})$ est alors différente de $E[g(\phi) | \mathbf{u}]$.

Chapitre 4

Modèles CMCa-BI multivariées

Nous traitons ici de l'estimation de modèles CMCa-BI multivariées pour différentes modélisations des lois d'émission. Selon le type des données à traiter et les hypothèses concernant les fluctuations aléatoires au sein de chaque classe, il est nécessaire de choisir un modèle pertinent pour décrire la physique du phénomène observé. Très communément, la loi normale est utilisée pour réaliser la classification de données ou la segmentation de signaux. Dans le contexte de la classification d'observations, il s'agit d'une hypothèse classique grâce à laquelle la notion de groupes est bien appréhendée par un profil moyen (la moyenne) et par une dispersion intrinsèque (la variance). En imagerie radar, la normalité (ou l'utilisation de lois déduites de la loi normale comme la loi du khi-2 pour la modélisation des intensités) provient de l'application d'un théorème central limite, en supposant que le nombre de réflecteurs au sein de chaque pixel est suffisamment grand pour faire cette approximation. Enfin, la possibilité de pouvoir mener complètement et exactement les calculs, constitue l'argument le plus fort en faveur de l'utilisation de la loi normale.

Il est cependant fréquent que les données traitées en image ou en signal s'écartent significativement de la loi normale, en raison de la complexité des phénomènes réels (nécessité de la prise en compte de queues de distribution épaisses [1] ou de dynamiques complexes [87]). De plus, des modèles théoriques de formation des signaux reçus justifient l'abandon de la normalité et incitent à l'utilisation d'autres modèles paramétriques pour la modélisation des intensités reçues en imagerie, des impulsions radar ou des données multicateurs.

Dans ce chapitre, nous présentons plusieurs modèles paramétriques pertinents pour la segmentation d'images, que nous pouvons rattacher à 3 types de modèles statistiques : les lois exponentielles, les lois elliptiques et les modèles spécifiés par copules. Après une description de ces modèles et de leur propriétés, nous donnons les méthodes d'estimation associées dans le cadre des CMCa-BI.

Il s'agit d'appliquer, selon le type des lois d'émission, les algorithmes de calcul d'estimateurs présentés dans le chapitre 2. Parmi ceux-ci, nous avons privilégié les estimateurs capables de fournir des algorithmes d'estimation simples à implémenter, et nécessitant peu ou pas de phases d'optimisation numérique multiparamètres (pouvant entraîner des temps de calcul importants ou/et des ressources machines importantes). Dans le cas des modèles exponentiels et elliptiques, le calcul (approché) de l'EMV par EM (ou SEM) permet d'avoir des calculs récursifs faciles à implémenter et nécessitant peu de calculs lourds. Nous utilisons aussi des modèles spécifiés par copules, ce

qui constitue la première utilisation des copules dans le contexte de la segmentation bayésienne. Pour ces nouveaux modèles de mélanges, nous proposons un estimateur ECI et utilisant la théorie que nous avons développé dans le chapitre 2. Nous obtenons alors un algorithme original et général d'estimation de mélanges de modèles multivariés. De plus, cet algorithme est modulaire : la modification de la forme paramétrique de la copule ou de la marginale d'une loi d'émission n'a pas d'impact sur les formules d'estimation de ses autres paramètres, ce qui permet d'avoir un programme d'estimation non-supervisée facilement modifiable.

Dans les 3 cas traités, nous considérons que nous avons un échantillon $\mathbf{y} \in \mathcal{Y}^N$ issues d'une CMCA-BI, dont les densités des lois d'émission $f(y, \theta_k)$ appartiennent toutes au même modèle paramétrique.

4.1 Mélange de lois exponentielles

4.1.1 Définition

Définition 4.1.1. Famille exponentielle régulière minimale (Barndorff-Nielsen [12])

Une famille exponentielle régulière sur $\mathbb{R}^d$ est une famille de probabilités $\mathcal{P}(\Theta)$ sur $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ dont les densités relativement à une mesure μ sont de la forme

$$\forall y \in \mathbb{R}^d, \forall \theta \in \Theta \subset \mathbb{R}^k, f(y, \theta) = \exp(\langle T(y), \theta \rangle - \varphi(\theta)) \quad (4.1)$$

où Θ est un ouvert de $\mathbb{R}^k$ et $T : \mathbb{R}^d \longrightarrow \mathbb{R}^k$ est la statistique privilégiée du modèle. Si T est de rang plein (i.e. les fonctions $(T_i)_{1 \leq i \leq k}$ sont linéairement indépendantes), la famille est dite minimale. La forme paramétrique donnée par (4.1) est appelée paramétrisation canonique.

Alternativement, une famille exponentielle peut être indexée par la paramétrisation naturelle, définie par $\eta = E_\theta[T(Y)]$. La relation de passage entre les paramètres θ et η est donnée par

$$\nabla_\theta \varphi(\theta) = \eta$$

L'EMV est l'estimateur privilégié des familles exponentielles en raison des propriétés d'existence et d'unicité du maximum de la log-vraisemblance sous des conditions faibles. Si nous estimons θ à partir d'un échantillon $\mathbf{y} = (y_1, \dots, y_N)$, alors si la famille est minimale et que $T(\mathbf{y})$ appartient à l'intérieur de $\nabla_\theta \varphi(\Theta)$, le maximum de la log-vraisemblance existe, est unique et est atteint par le point vérifiant l'équation $T(\mathbf{y}) = \eta$, [12, 100]. Pour déterminer l'EMV $\hat{\theta}_{MV}$ il s'agit d'inverser la relation $T(\mathbf{y}) = \nabla_\theta \varphi(\hat{\theta}_{MV})$, i.e.

$$\hat{\theta}_{MV} = (\nabla_\theta \varphi)^{-1}(T(\mathbf{y})) \quad (4.2)$$

De nombreuses lois usuelles sont des familles exponentielles, telles que les lois gamma ou gamma inverse, les lois normales, ou encore les lois de Von Mises-Fisher et de Wishart.

La log-vraisemblance d'un mélange de lois exponentielles se met sous une forme particulièrement facile à manipuler, et les algorithmes EM déduits ont généralement des expressions simples. Par conséquent, pour les modèles que nous présentons, nous calculons une approximation de l'EMV des mélanges de lois exponentielles par EM.

4.1.2 Estimation par EM de modèles exponentiels

Les étapes de l'algorithme EM prennent une forme particulièrement simple lorsque nous utilisons la paramétrisation canonique. Nous nous en servons alors pour déterminer les algorithmes de mise à jour des modèles exponentiels, paramétrés de manière classique.

Si nous avons K classes distinctes, indexées par un paramètre θ_k (et dont la statistique privilégiée commune est notée T), la fonction $Q(\phi, \phi_n)$ introduite dans la section 3.4.1 se décompose en $Q(\phi, \phi_n) = Q(A) + Q(\theta, \theta_n)$. Comme les mises à jour de A et θ sont effectuées indépendamment l'une de l'autre, nous donnons uniquement le calcul de θ_{n+1} :

$$Q(\theta, \theta_n) = \sum_{i,k=1}^{N,K} \pi_{i,k}^{(n)} (\langle T(y_i), \theta_k \rangle - \varphi(\theta_k))$$

L'étape M consiste donc en la résolution de l'équation habituelle pour les modèles exponentiels :

$$\forall k \leq K, \nabla_{\theta_k} \varphi(\theta_k) = \left(\overline{T(\mathbf{y})} \right)_{k,n} \quad (4.3)$$

où nous avons noté $\left(\overline{T(\mathbf{y})} \right)_{k,n} = \frac{\sum_{i=1}^N \pi_{i,k}^{(n)} T(y_i)}{\sum_{i=1}^N \pi_{i,k}^{(n)}}$. Nous donnons ci-dessous différents exemples de modèles exponentiels couramment utilisés.

4.1.2.1 Loi normale

Comme nous l'avons indiqué en introduction, la loi normale est souvent utilisée en segmentation (ou en classification, voir notamment [109, 69]) en raison de la possibilité de calculer aisément les quantités d'intérêt, notamment l'estimateur du maximum de vraisemblance. L'algorithme EM est alors couramment utilisé, et les étapes E et M s'obtiennent en considérant directement la paramétrisation classique¹ (m, Σ) (m et Σ sont respectivement la moyenne et la variance). Les formules de mises à jour sont :

$$\forall k \leq K, \begin{cases} m_k = \frac{\sum_{i=1}^N \pi_{i,k}^{(n)} y_i}{\sum_{i=1}^N \pi_{i,k}^{(n)}} \\ \Sigma_k = \frac{\sum_{i=1}^N \pi_{i,k}^{(n)} (y_i - m_k)(y_i - m_k)'}{\sum_{i=1}^N \pi_{i,k}^{(n)}} \end{cases} \quad (4.4)$$

4.1.2.2 Loi Gamma

Les lois gamma sont souvent utilisées pour modéliser l'intensité dans les images radar et permettent de généraliser la loi du khi-2. La loi gamma est spécifiée par deux paramètres a, b (respectivement de forme et d'échelle) et a une densité $f(y, (a, b)) = \frac{1}{\Gamma(a)b^a} y^{a-1} \exp(-\frac{y}{b}) 1_{\mathbb{R}^+}(y)$. Elle

¹Les algorithmes EM obtenus avec la paramétrisation $(m_k, \Sigma_k)_{1 \leq k \leq K}$ ou $\theta = (\theta_k)_{1 \leq k \leq K}$ sont équivalents grâce à l'invariance par reparamétrisation des étapes E et M

est notée $\gamma(a, b)$ et sa forme canonique est :

$$\begin{cases} T(y) = (-y, \log(y))' \\ \theta = (\frac{1}{b}, a - 1) \\ \varphi(\theta) = -(\theta_2 + 1) \log(\theta_1) + \log(\Gamma(\theta_2 + 1)) \end{cases}$$

et pour chaque itération n de EM nous devons résoudre le système :

$$\forall k, \begin{cases} \frac{\theta_{2,k}^{n+1} + 1}{\theta_{1,k}^{n+1}} = (\bar{\mathbf{y}})_{k,n} \\ -\log(\theta_{1,k}^{n+1}) + \psi(\theta_{2,k}^{n+1} + 1) = (\overline{\log \mathbf{y}})_{k,n} \end{cases}$$

où $\psi(x) = \frac{d \log \Gamma(x)}{dx}$ est la fonction digamma, dérivée logarithmique de la fonction Γ d'Euler. En réécrivant les équations en fonction des paramètres usuels a, b , nous obtenons les formules de mise à jour suivantes :

$$\forall k, \begin{cases} a_k^{n+1} = \xi^{-1}((\overline{\log \mathbf{y}})_{k,n} - \log((\bar{\mathbf{y}})_{k,n})) \\ b_k^{n+1} = \frac{(\bar{\mathbf{y}})_{k,n}}{a_k^{n+1}} \end{cases} \quad (4.5)$$

avec ξ fonction telle que $\xi(x) = \psi(x) - \log(x)$.

4.1.2.3 Loi de Von Mises - Fisher

Les lois de Von Mises-Fisher ([107]) sont une famille exponentielle sur la sphère unité $\mathbb{S}^{d-1} \subset \mathbb{R}^d$, dont la densité relativement à la mesure uniforme sur $\mathbb{S}^{d-1}$ vaut :

$$\forall y \in \mathbb{S}^{d-1}, f(y, (\xi, \kappa)) = \left(\frac{\kappa}{2}\right)^{\frac{d}{2}-1} \frac{1}{\Gamma(\frac{d}{2}) I_{d/2-1}(\kappa)} \exp(\kappa \xi' y)$$

où $\xi \in \mathbb{S}^{d-1}$ est appelé direction moyenne et $\kappa \geq 0$ est le paramètre de concentration. Cette loi est souvent utilisée pour la segmentation d'information directionnelle [109], mais aussi pour la classification de données textuelles et d'expression de gènes [7]. Nous verrons dans le chapitre 5, que nous pouvons représenter partiellement l'information Doppler ou polarimétrique par un point de la sphère $\mathbb{S}^{d-1}$. La paramétrisation canonique est :

$$\begin{cases} T(y) = y \\ \theta = \kappa \xi \\ \varphi(\theta) = (\frac{d}{2} - 1) \log(\frac{\kappa}{2}) - \log(\Gamma(\frac{d}{2})) - \log(I_{d/2-1}(\kappa)) \end{cases}$$

où $I_{d/2-1}$ représente la fonction de Bessel de première espèce [5]. Si nous notons $A_d(\kappa) = \frac{I_{d/2}(\kappa)}{I_{d/2-1}(\kappa)}$, les équations de vraisemblance pour chaque classe sont :

$$\forall k, \begin{cases} \hat{\xi}_k = \frac{(\bar{\mathbf{y}})_{k,n}}{\|(\bar{\mathbf{y}})_{k,n}\|} \\ \text{et} \\ \hat{\kappa}_k = A_d^{-1} \left(\|(\bar{\mathbf{y}})_{k,n}\| \right) \end{cases} \quad (4.6)$$

$\hat{\kappa}_k$ est déterminé par inversion de la fonction A_d . Plusieurs approximations de A_d ont été proposé afin d'alléger les calculs nécessaires à l'estimation du paramètre de concentration², voir [107] et [7].

Remarque 4.1.1. Loi de Von Mises-Fisher complexe

Nous pouvons étendre ce modèle exponentiel à la sphère unité complexe de $\mathbb{C}^d$, noté $\mathbb{CS}^{d-1} = \{z \in \mathbb{C}^d \mid \sum_{i=1}^d |z_i|^2 = 1\}$, en considérant le modèle tel que $f(y, (\xi, \kappa)) \propto \exp(\kappa \Re(\xi^* y))$ ($\Re$ désigne la partie réelle d'un vecteur complexe et $\Im$ sa partie imaginaire). Nous appelons ce modèle la loi Von Mises-Fisher complexe, et elle correspond alors à la loi de Von Mises réel sur la sphère $\mathbb{S}^{2d-1}$, de direction moyenne égale à $(\Re(\xi)' \Im(\xi)')'$.

4.2 Mélange de lois elliptiques

Nous présentons ici le modèle de vecteur aléatoire sphériquement invariant pour la modélisation des lois d'émission de CMCA-BI. Ce modèle a un double intérêt pour les applications :

- il s'agit d'une généralisation assez simple de la loi normale, et dont l'introduction dans le domaine de la classification peut être assimilée à la robustification (par rapport aux valeurs aberrantes) des procédures de segmentation non-supervisée (voir [109] et les références à l'intérieur) ;
- un raisonnement physique sur la formation des images et des signaux radar motivent leur utilisation dans les traitements de segmentation statistique, d'autant plus que celui-ci est validé sur données réelles [44].

Nous introduisons tout d'abord les modèles elliptiques, dont les vecteurs aléatoires sphériquement invariants (Spherically Invariant Random Vector, noté SIRV par la suite) sont un cas particulier, et nous donnons quelques unes de leurs propriétés. Nous nous intéressons alors au problème de l'estimation par maximum de vraisemblance de tels modèles, qui, à part dans le cas gaussien, conduisent à la résolution d'une équation normale non-linéaire.

La résolution de cette équation par un algorithme EM dans le cas de la loi de Student³ est un exemple classique de l'élégance de ce dernier. Après avoir donné la forme générale de l'algorithme EM pour les SIRV, nous rappelons son expression dans le cas d'une seule composante ([108, 111]) et d'un mélange ([109]). Nous calculons alors les expressions analytiques des algorithmes EM similaires pour la loi K multivariée. Les algorithmes d'estimation, obtenus indépendamment de l'article de Roberts et Furui [139], en sont une généralisation au cas multivarié réel et de mélange, et s'adaptent directement au cas complexe. L'algorithme EM est une alternative très intéressante aux méthodes classiques d'estimation pour estimer les paramètres d'échelle et de forme de la

²Ceci est une difficulté inhérente aux modèles paramétriques pour données directionnelles, pour lesquelles la détermination de la constante de normalisation est très difficile. L'estimation des paramètres, y compris dans les modèles exponentiels, posent des difficultés dues à la résolution d'équations non-linéaires complexes, faisant intervenir des fonctions spéciales, voir [107]. Ceci apparaît comme un frein à l'utilisation de ces modèles.

³que le nombre de degrés de liberté soit connu ou inconnu.

loi K (dans le cas scalaire), tant du point de vue calculatoire que de la qualité (voir [139] pour des expérimentations). Les techniques classiques utilisent la méthode des moments, l'optimisation numérique ou encore l'approximation asymptotique par une loi gamma de la loi K. Dans le cas multivarié (réel ou complexe), l'algorithme que nous donnons est la seule méthode proposée jusqu'à présent pour l'estimation conjointe de la moyenne, de la variance et du paramètre de forme.

4.2.1 Lois elliptiques et Vecteurs Aléatoires Sphériquement Invariants

Les lois elliptiques sont considérées comme une généralisation de la loi normale multivariée parce qu'elles sont, comme cette dernière, indexées par un paramètre de position et d'échelle, et que les courbes d'isoprobabilités de la densité sont de forme elliptique. Elles se distinguent de celle-ci par un troisième paramètre, fonctionnel ou de dimension finie (nous dirons euclidien par la suite), qui vient régler l'épaisseur des queues, c'est à dire la fréquence d'observations de valeurs élevées en module. Pour cette raison, ces modèles sont utilisés lorsque la loi normale est incapable d'expliquer l'occurrence d'un trop grand nombre de valeurs extrêmes, ou que les procédures d'estimation et/ou de décision ne sont pas robustes aux valeurs aberrantes ([77] et les références citées).

Pour une présentation théorique et détaillée des nombreuses propriétés des lois elliptiques, nous renvoyons à l'ouvrage de Gupta et Varga [82]. Nous donnons tout d'abord deux définitions (remarquablement) équivalentes des modèles elliptiques.

Définition 4.2.1. Loi elliptique

Le vecteur aléatoire $Y \in \mathbb{R}^d$ (et de support égale à $\mathbb{R}^d$) a une loi à contour elliptique, ou plus simplement elliptique s'il vérifie l'une des deux propriétés équivalentes suivantes :

1. Sa densité relativement à la mesure de Lebesgue s'écrit

$$\forall y \in \mathbb{R}^d, f(y, (\mathbf{m}, \Sigma, h)) = |\Sigma|^{-1/2} h\left(\|\Sigma^{-1/2}(y - \mathbf{m})\|^2\right) \quad (4.7)$$

avec h fonction continue telle que $\int_0^\infty h(s) s^{\frac{d}{2}-1} ds = \Gamma(\frac{d}{2}) \pi^{-\frac{d}{2}}$. Dans le cas où $\int_0^\infty h(s) s^{\frac{d}{2}} ds < \infty$, la moyenne et la covariance de Y existent et valent respectivement $\mathbf{m}$ et $\lambda \Sigma$, avec $\lambda > 0$.

2. Y admet la représentation stochastique suivante

$$Y = \mathbf{m} + \mathbf{R} \mathbf{A} U \quad (4.8)$$

avec R v.a.r positive et U v.a. uniformément répartie sur la sphère S^{d-1} . R et U sont des v.a. indépendantes, et $\mathbf{A}$ est une matrice non-singulière.

Nous avons la relation $\mathbf{A} \mathbf{A}^* = \Sigma$. $\mathbf{m}$ et Σ sont appelés paramètres euclidiens et h est appelé le paramètre fonctionnel.

Remarque 4.2.1. Si Y est un vecteur elliptique, alors le vecteur transformé $\tilde{Y} = \mathbf{H}Y + \mathbf{B}$ (où $\mathbf{H}$ est une matrice carrée de dimension d inversible et $\mathbf{B}$ est un vecteur de $\mathbb{R}^d$) est encore elliptique, mais de paramètres euclidiens égaux à $(\mathbf{H}\mathbf{m} + \mathbf{B}, \mathbf{H}\mathbf{A}\mathbf{H}')$. La loi de la v.a.r R (ou encore la fonction h) est inchangée.

Remarque 4.2.2. La seconde propriété est souvent exploitée pour construire un test d'ellipticité. En effet, la représentation stochastique (4.8) d'un vecteur elliptique centré Y s'obtient directement par normalisation, i.e. $R = \|Y\|$ et $U = \frac{Y}{\|Y\|}$. Le test d'ellipticité d'un échantillon $y_1, \dots, y_N$ consiste à tester la répartition uniforme sur la sphère des parties angulaires $\frac{y_i}{\|y_i\|}$, ainsi que l'indépendance de $\frac{y_i}{\|y_i\|}$ et $\|y_i\|$. Ce test a été utilisé dans [44] pour vérifier l'hypothèse d'ellipticité de mesures radar réelles.

Remarque 4.2.3. Pour identifier un modèle elliptique, nous avons besoin de contraintes liant $\mathbf{m}, \Sigma$ et h . En effet, nous n'avons pas en général

$$\forall y \in \mathbb{R}^d, f(y, (\mathbf{m}_1, \Sigma_1, h_1)) = f(y, (\mathbf{m}_2, \Sigma_2, h_2)) \implies \mathbf{m}_1 = \mathbf{m}_2, \Sigma_1 = \Sigma_2, h_1 = h_2$$

Ceci provient du fait que le facteur d'échelle est contenu à la fois dans la matrice Σ et dans la fonction h . Une contrainte classique pour obtenir un modèle identifiable est de supposer que $\text{Tr}(\Sigma) = d$.

Il est facile de construire des modèles elliptiques sur $\mathbb{R}^d$, puisqu'il suffit de choisir des fonctions h tels que la fonction $s \mapsto h(s)s^{\frac{d}{2}-1}$ soit intégrable. La seule difficulté consiste alors en le calcul de la constante de normalisation. Cependant pour les applications que nous traitons, nous nous intéressons à la sous-famille des SIRV, construite à partir de la loi normale.

Définition 4.2.2. Vecteur Aléatoire Sphériquement Invariant

Un vecteur aléatoire Y dans $\mathbb{R}^d$ est un vecteur aléatoire sphériquement invariant (Spherically Invariant Random Vector, noté par la suite SIRV) s'il admet la représentation stochastique suivante :

$$Y = U\epsilon \tag{4.9}$$

avec U, ϵ , deux variables aléatoires indépendantes, telles que U soit positive et $\epsilon \sim N(0, \Sigma)$, avec Σ matrice semi-définie positive.

Cette définition est celle utilisée en traitement du signal Radar [135] (voir aussi la bibliographie pour des références antérieures sur l'introduction de ce modèle), cependant dans la suite nous désignerons par SIRV les vecteurs aléatoires dont la représentation stochastique est $Y = \mathbf{m} + U\epsilon$, avec $\mathbf{m}$ vecteur de $\mathbb{R}^d$: les procédures d'estimation et de segmentation proposées ne sont pas modifiées par l'ajout d'un paramètre de position $\mathbf{m}$. Nous désignerons par SIRV centré, les modèles classiques utilisées en radar. Les données utilisées en radar sont complexes, et nous pouvons généraliser la définition des SIRV réelles au SIRV complexes, en considérant que ϵ est un vecteur gaussien complexe circulaire, voir annexe A. Afin de faciliter l'obtention des procédures d'estimation, nous utilisons la représentation stochastique suivante

$$Y = \mathbf{m} + U^{-1/2}\epsilon \tag{4.10}$$

où U a une densité g relativement à la mesure de Lebesgue. Dans le contexte Radar, où les SIRV ont été introduits pour modéliser les fouillis non-gaussiens [118, 135, 104, 117], le processus U (défini dans l'équation (4.9)) est appelé texture, et le processus gaussien ϵ est appelé "speckle" (chatoiement). Nous continuons à appeler texture le processus U intervenant dans (4.10).

4.2.2 Liens entre loi d'un SIRV et loi de la texture associée

Nous nous intéressons maintenant au lien entre la loi de U et la loi de Y , i.e. au lien entre les fonctions g et h de (4.7). La définition d'un SIRV nous permet de dire que $Y|U = u$ est une loi normale $N(\mathbf{m}, \frac{\Sigma}{u})$ et de déduire sa variance directement (par application de la formule de la variance conditionnelle)

$$V(Y) = E[U^{-1}] \times \Sigma \quad (4.11)$$

La texture U est une modulation de la puissance du signal rétrodiffusé. Nous pouvons aussi réécrire l'Eq. (4.7) en introduisant la forme quadratique $q(y) = \frac{1}{2}(y - \mathbf{m})' \Sigma^{-1}(y - \mathbf{m})$

$$f(y, (\mathbf{m}, \Sigma, g)) = \frac{|\Sigma|^{-1/2}}{(2\pi)^{d/2}} \int_0^\infty u^{\frac{d}{2}} e^{-uq(y)} g(u) du \quad (4.12)$$

La densité $f(\cdot | \mathbf{m}, \Sigma, g)$ est donc la transformée de Laplace de la fonction $g_d : u \mapsto \frac{|\Sigma|^{-1/2}}{(2\pi)^{d/2}} u^{\frac{d}{2}} g(u)$, évaluée au point $q(y)$. Nous pouvons, comme dans la remarque 4.2.3, imposer une contrainte sur la trace de la matrice Σ , pour assurer l'identifiabilité d'un SIRV lorsqu'il est indexé par le triplet $(\mathbf{m}, \Sigma, g)$. Cependant, pour l'estimation il est intéressant de mettre une contrainte sur la loi de U plutôt que sur la matrice Σ . La modélisation SIRV fait apparaître de manière particulièrement marquante le problème d'identifiabilité du modèle : la définition d'un SIRV par le produit de la texture U et du processus gaussien ϵ montre que la puissance du signal reçu Y peut être attribuée aussi bien à U qu'à ϵ :

$$\forall \lambda > 0, Y = \lambda^{1/2} \epsilon / (\lambda U)^{1/2} \quad (4.13)$$

Autrement dit, les paramètres (Σ, g) et $(\lambda \Sigma, \frac{1}{\lambda} g(\frac{\cdot}{\lambda}))$ donne la même densité $f(y, (\mathbf{m}, \Sigma, g))$. Par conséquent, nous imposons une contrainte sur le paramètre d'échelle de la loi de U . Un modèle SIRV sera noté $\mathfrak{E}(\mathbf{m}, \Sigma, \vartheta)$ si la loi de U est indexée par un paramètre $\vartheta \in \Xi$. Nous ne considérons par la suite que la loi de Student et la loi K (déduts de la loi gamma), qui sont les modèles les plus utilisés dans les applications qui nous intéressent et qui ont de plus l'avantage de posséder une expression analytique pour la densité (ce qui n'est généralement pas le cas des SIRV).

Nous rappelons tout d'abord la définition de la loi gamma inverse, ainsi que quelques propriétés de la loi gamma.

Définition 4.2.3. Loi Gamma et Gamma Inverse

Une variable aléatoire positive V suit une loi gamma inverse, notée $i\gamma(a, b)$, si V^{-1} suit une loi $\gamma(a, b)$. Sa densité relativement à la mesure de Lebesgue est alors

$$g(v) = \frac{1}{\Gamma(a)b^a} \frac{1}{v^{a+1}} e^{-1/bv} \mathbf{1}_{\mathbb{R}^+}(v)$$

Propriété 4.2.1. Moments des lois Gamma et Gamma Inverse

Si $U \sim \gamma(a, b)$, alors $\lambda U \sim \gamma(a, \lambda b)$. De plus, nous avons

$$\begin{cases} E[U] &= ab \\ E[\log U] &= \psi(a) + \log(b) \end{cases} \quad (4.14)$$

Si $V \sim i\gamma(a, b)$, alors $\lambda V \sim i\gamma(a, \frac{b}{\lambda})$. De plus, nous avons

$$\begin{cases} E[V] &= \frac{1}{(a-1)b} \\ E[\log V] &= -\psi(a) - \log(b) \end{cases} \quad (4.15)$$

Le paramètre b est le facteur d'échelle des lois gamma et gamma inverse à contrôler pour garantir l'identifiabilité du modèle SIRV déduit. Dans les deux cas, la contrainte que nous choisissons est $b = \frac{1}{a}$, ce qui se traduit par le fait que nous ne considérons que des lois gamma de moyenne égale à 1 et des lois gamma inverse de moyenne $\frac{a}{a-1}$. Si $U \sim \gamma(\frac{\nu}{2}, \frac{2}{\nu})$, alors Y suit une loi de Student (noté loi T) à ν degrés de liberté, notée $\mathbf{T}_{\mathbf{m}, \Sigma, \nu}$:

$$f(y, (\mathbf{m}, \Sigma, \nu)) = \frac{\Gamma(\frac{\nu+d}{2})|\Sigma|^{-1/2}}{(\pi\nu)^{d/2}\Gamma(\frac{\nu}{2})} \left(1 + \frac{2q(y)}{\nu}\right)^{-\frac{\nu+d}{2}} \quad (4.16)$$

Par l'Eq. (4.11) nous retrouvons que dès que $\nu > 2$, la variance d'une loi de Student existe et vaut

$$V_{T, \Sigma, \nu}(Z) = \frac{\nu}{\nu - 2} \times \Sigma$$

la moyenne étant $\mathbf{m}$.

Si $U \sim i\gamma(a, \frac{1}{a})$, alors Z a une loi de type K notée $\mathbf{K}_{\mathbf{m}, \Sigma, a}$:

$$f(y, (\mathbf{m}, \Sigma, a)) = \frac{2a^a |\Sigma|^{-1/2}}{(2\pi)^{d/2} \Gamma(a)} \left(\sqrt{\frac{q(y)}{a}} \right)^{a-\frac{d}{2}} K_{a-\frac{d}{2}} \left(2\sqrt{aq(y)} \right) \quad (4.17)$$

où K_α est la fonction de Bessel modifiée de première espèce, [5]. La moyenne de la loi K existe pour tout a , ainsi que sa variance qui vaut :

$$V_{K, \Sigma, a}(Y) = \Sigma$$

La loi normale apparaît comme un cas limite lorsque les paramètres $a, \nu \rightarrow \infty$. En pratique, la loi K tend très vite vers la loi normale (dès que $a \simeq 10$). Pour la loi T, l'approximation gaussienne est correcte pour $\nu \simeq 100$. Par contre, pour les faibles valeurs, les paramètres a et ν ont une influence opposée sur la forme de la densité :

1. La loi $\mathbf{T}_{\mathbf{m}, \Sigma, \nu}$ est une loi à queue épaisse, parce qu'il s'agit d'une loi de type puissance (à l'extrême, pour $\nu = 1$, Y suit une loi de Cauchy qui n'a pas de moyenne).
2. À l'opposé, le paramètre a est tel que $\mathbf{K}_{\mathbf{m}, \Sigma, a}$ tend vers une dirac en $\mathbf{m}$ lorsque a tend vers 0. Lorsque $a = 1$, la loi K est parfois appelée loi de Laplace généralisée. L'approximation asymptotique classique de la fonction K_α (voir [5])

$$\forall \alpha > 0, K_\alpha(x) \sim_{+\infty} \sqrt{\frac{\pi}{2x}} e^{-x} \quad (4.18)$$

permet de montrer que les queues décroissent toujours à une vitesse exponentielle (l'équivalent est $\|y\|^{a-\frac{d+1}{2}} e^{-2\sqrt{a}\|y\|}$), mais beaucoup moins rapide que la loi normale.

Pour cette raison, les lois T et K sont des distributions à queues épaisses (voir les figures 4.1 des marginales et les figures 4.2 pour des exemples de densités en dimension 2).

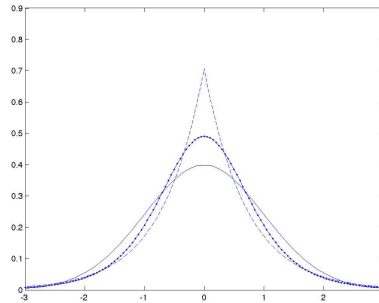


FIG. 4.1 – Marginales des lois elliptiques : loi K ('-'), loi T '*-', loi Normale (-)

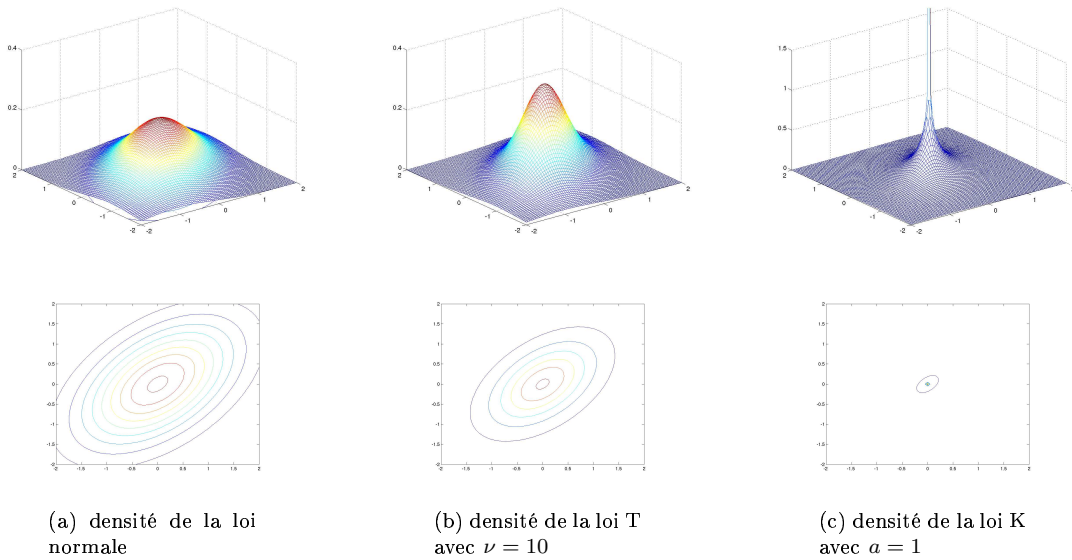


FIG. 4.2 – Graphes 2D et contours des densités elliptiques normal, T et K de paramètre $\mathbf{m} = 0$ et $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$.

La loi SIRV unidimensionnelle ne correspond pas à la définition classique de la loi K ([9, 49]), qui est initialement un modèle de densité sur $\mathbb{R}^+$ pour la modélisation des amplitudes non-Rayleigh. Une variable aléatoire suit une distribution de Rayleigh si sa densité est $f(r) = \frac{r}{\sigma^2} \exp\left(-\frac{r^2}{2\sigma^2}\right)$, ce qui est la loi de l'amplitude d'une loi gaussienne complexe circulaire. Il est équivalent de dire que la loi de l'intensité I (i.e. l'amplitude au carré) est une loi exponentielle de paramètre $\gamma(1, \sigma^2)$. La loi K est introduite en considérant que l'intensité I' effectivement reçue est une perturbation aléatoire multiplicative de l'intensité I . L'intensité reçue $I' = U^{-1/2}I$ a pour densité $f_{I'}(y) = \int_0^{+\infty} \frac{u}{\sigma^2} \exp\left(-u\frac{y}{\sigma^2}\right) g(u) du$. Si g suit une loi inverse gamma, alors nous obtenons la loi K, notée K_{a, σ^2} :

$$f_I(y, (a, \sigma^2)) = \frac{2a^a}{\sigma^2 \Gamma(a)} \left(\frac{y}{a\sigma^2}\right)^{(a-1)/2} K_{a-1}\left(2\frac{\sqrt{ay}}{\sigma}\right) 1_{\mathbb{R}^+}(y) \quad (4.19)$$

De par les propriétés d'indépendance entre partie angulaire et norme des SIRV, $f_{I'}$ est aussi la densité de la norme d'un SIRV de $\mathbb{R}^2$, de densité $\mathbf{K}_{0, \sigma^2 I_2, a}$.

Remarque 4.2.4. *En utilisant l'approximation asymptotique (4.18), nous trouvons que $f_I(y) \sim Cy^{a-1} e^{-2\frac{\sqrt{ay}}{\sigma}}$ ce qui vient justifier l'approximation de la loi des intensités par une loi gamma.*

Remarque 4.2.5. *Signalons l'intérêt de la densité $\mathbf{K}_{0, \sigma^2, a}$ définie sur $\mathbb{R}$ pour la modélisation des sorties de filtres d'analyses d'images (par exemple des coefficients d'ondelettes), pour la segmentation de fouillis, la reconnaissance de cibles, et plus généralement la reconnaissance de formes. La densité $\mathbf{K}_{0, \sigma^2, a}$ a été obtenue par U. Grenander et A. Srivastava sous des hypothèses assez générales comme étant la forme analytique des densités des contenus spectraux d'une image extraits par des filtres linéaires [81, 145].*

4.2.3 Estimation par Maximum de vraisemblance

4.2.3.1 Généralités

Si nous avons une loi elliptique $\mathfrak{E}(\mathbf{m}, \Sigma, \vartheta)$, nous sommes souvent intéressés par l'estimation des paramètres position-échelle $(\mathbf{m}, \Sigma)$, ou alors du paramètre complet $(\mathbf{m}, \Sigma, \vartheta)$. Nous nous intéressons ici à leur estimation par l'EMV. Dans le premier cas, si nous supposons que la fonction h est connue, la différentiation de la log-vraisemblance (dans le cas d'un échantillon i.i.d. $\mathbf{y} = (y_1, \dots, y_N)$) donne le système suivant :

$$\begin{cases} \mathbf{m} = \frac{\sum_{i=1}^N w(r_i) y_i}{\sum_{i=1}^N w(r_i)} \\ \Sigma = \frac{1}{N} \sum_{i=1}^N w(r_i) (y_i - \mathbf{m})(y_i - \mathbf{m})' \end{cases} \quad (4.20)$$

avec $r_i = \|\Sigma^{-1/2}(y_i - \mathbf{m})\|^2$ et $w(r) = -\frac{h'(r)}{h(r)}$.

Les estimateurs se mettant sous cette forme ont de “bonnes” propriétés de convergence et de robustesse (en termes de sensibilité à des valeurs aberrantes, [91]), si la fonction w décroît assez vite. En effet, la fonction $r \mapsto w(r)$ peut s'interpréter comme une fonction de poids qui pondère l'influence des observations dans l'estimation des paramètres. Cette fonction est décroissante (et souvent réécrite sous la forme $w(r) = \frac{\psi(r)}{r}$), de telle sorte que le poids de l'observation z_i soit faible si celle-ci est loin de la moyenne (relativement à la distance de Mahalanobis), et important si elle

en est proche. D'autres fonctions w que celles obtenues à partir de la vraisemblance peuvent être proposées pour l'estimation des paramètres $\mathbf{m}$ et Σ . Les estimateurs sont appelés M-estimateurs, et sous des conditions sur la fonction w , la consistance et la robustesse aux valeurs aberrantes est garantie (en sacrifiant l'efficacité). On peut par exemple citer la fonction (impair) de Huber définie par :

$$\psi(s) = \begin{cases} s, & |s| \leq a \\ \text{signe}(s) \times a, & |s| > a \end{cases}$$

où a est un paramètre à régler.

Cependant, la résolution analytique de (4.20) en $(\mathbf{m}, \Sigma)$ est impossible, hormis pour quelques fonctions particulières (par exemple la loi normale pour laquelle la fonction w est constante et égale à 1). Dans le contexte radar, une méthode itérative particulièrement intéressante a été développée pour les SIRV de moyenne nulle, appelée "estimateur du point fixe" (voir [77, 118]). En effet, de manière remarquable, l'EMV est le point fixe d'une fonction indépendante de h , pour laquelle il est possible de construire une suite de matrice convergeant vers ce point fixe. Outre sa facilité d'implémentation, cet indépendance de la forme paramétrique de $\mathfrak{E}(\mathbf{m}, \Sigma, \vartheta)$ le rend très intéressant pour son exploitation sur données réelles.

Nous nous intéressons par la suite à l'estimation du triplet $(\mathbf{m}, \Sigma, \vartheta)$, qui si elle oblige à spécifier la forme paramétrique de $\mathfrak{E}(\mathbf{m}, \Sigma, \vartheta)$, permet d'apporter une information utile sur la vitesse de décroissance des queues et d'avoir une caractérisation plus fine des données. La recherche de l'EMV aboutit à un système d'équations non-linéaires dans lesquelles sont liées les 3 paramètres. Nous présentons dans le paragraphe suivant une méthode itérative de maximisation de la log-vraisemblance des SIRV basée sur l'algorithme EM (déjà proposé pour l'estimation des paramètres $(\mathbf{m}, \Sigma)$ et $(\mathbf{m}, \Sigma, \vartheta)$, voir [140, 108, 111]).

4.2.3.2 Recherche de maxima de la vraisemblance par EM

Supposons que nous avons un échantillon i.i.d $\mathbf{y} = (y_1, \dots, y_N) \sim \mathfrak{E}(\mathbf{m}, \Sigma, \vartheta)$. Si nous observons U et Y , la log-vraisemblance (complète) s'écrit :

$$L_c(\mathbf{m}, \Sigma, \vartheta) = -\frac{N}{2} \log((2\pi)^d |\Sigma|) + \sum_{i=1}^N \left\{ \frac{d}{2} \log(u_i) - u_i q(y_i) + \log(g(u_i, \vartheta)) \right\} \quad (4.21)$$

Sans préciser la densité g , nous écrivons le schéma général de l'algorithme EM pour une loi $\mathcal{E}(\mathbf{m}, \Sigma, \vartheta)$. Partant d'une initialisation $\phi_0 = (\mathbf{m}^{(0)}, \Sigma^{(0)}, \vartheta^{(0)})$, la suite des paramètres $\phi_n = (\mathbf{m}^{(n)}, \Sigma^{(n)}, \vartheta^{(n)})_{n \geq 1}$ est obtenue par la résolution de 2 problèmes d'optimisation indépendants :

$$\begin{cases} \min_{\mathbf{m}, \Sigma} & \frac{N}{2} \log(|\Sigma|) + \sum_{i=1}^N w_i^{(n)} q(y_i) \\ \text{et} \\ \max_{\vartheta} & \sum_{i=1}^N E_{\phi_n} [\log(g(U_i, \vartheta)) | \mathbf{y}] \end{cases} \quad (4.22)$$

où $w_i^{(n)} = E_{\phi_n} [U_i | \mathbf{y}]$. Nous pouvons donner l'expression analytique des réestimations de $\mathbf{m}$ et Σ :

$$\begin{cases} \mathbf{m}^{(n+1)} &= \frac{\sum_{i=1}^N w_i^{(n)} y_i}{\sum_{i=1}^N w_i^{(n)}} \\ \text{et} \\ \Sigma^{(n+1)} &= \frac{1}{N} \sum_{i=1}^N w_i^{(n)} (y_i - \mathbf{m}^{(n+1)})(y_i - \mathbf{m}^{(n+1)})' \end{cases} \quad (4.23)$$

Si nous pouvons dériver par rapport à θ (et intervertir dérivation et intégration), nous devons résoudre en toute généralité l'équation :

$$\sum_{i=1}^N E_{\phi_n} [\nabla_{\vartheta} \log(g(U_i, \vartheta)) | \mathbf{y}] = 0 \quad (4.24)$$

Le plus souvent, cette équation est relativement facile à résoudre : on peut se ramener à une équation du genre $a(\vartheta) = a^{(k)}$, où $a^{(k)}$ est une constante calculée à partir des données, et a est une fonction croissante.

L'algorithme EM est donc un algorithme itératif de résolution des équations normales (4.20) d'un modèle elliptique, qui consiste en une mise à jour successive des paramètres selon les directions $\mathbf{m}$, Σ puis ϑ , au lieu d'être une méthode de recherche globale. Si nous n'estimons par exemple que la moyenne et la variance (en supposant que ϑ est connu), nous avons une méthode de type "Iteratively Reweighted Least Square", c'est à dire une correction successive des calculs de la moyenne et de la variance, par un poids venant pénaliser les observations atypiques, et surpondérer les observations "proches" de la moyenne.

La difficulté de cet algorithme réside dans l'étape E, pour laquelle il est nécessaire de déterminer la loi a posteriori U_i conditionnellement à Y_i (les observations sont toutes indépendantes), ou tout du moins de connaître les moments a posteriori de U et $\nabla_{\vartheta} \log(g(U, \vartheta))$. La loi de U a posteriori est

$$p(u | y) \propto u^{d/2} e^{-up(y)} g(u) \quad (4.25)$$

Nous avons alors deux possibilités qui consistent soit à déterminer explicitement cette loi a posteriori et d'en calculer analytiquement les moments d'intérêt, soit de procéder à une approche par simulation, conduisant ainsi à un algorithme de type Monte Carlo EM (i.e. l'étape E est remplacée par une simulation et une estimation empirique de la fonction Q à maximiser). Il s'agit d'un problème classique en analyse bayésienne, et l'existence de formule analytique pour la densité a priori est assurée en utilisant des lois conjuguées. Ceci est notamment le cas de la loi de Student, pour laquelle la loi de U est la loi conjuguée naturelle de la loi normale (voir Robert, [137]). Dans ce cas-là, nous avons

$$p(u | y) \propto u^{\frac{d}{2} + a - 1} \exp\left(-u\left(q(y) + \frac{1}{b}\right)\right) \quad (4.26)$$

Le modèle de Student se prête donc particulièrement bien à une estimation par EM, mais aussi à une simulation basée sur l'échantillonneur de Gibbs, car la loi a posteriori est connue et facile à simuler. Nous donnons ci-dessous les lois a posteriori et les formules exactes de réestimation dans les cas de la loi T et K.

4.2.3.3 Loi a posteriori et algorithme EM pour la loi de Student

Pour une loi de Student à ν degrés de libertés, nous pouvons directement expliciter les poids $w_i^{(n)}$ et l'estimation de ν , puisque la loi de U_i a posteriori est une gamma $\Gamma(\frac{\nu+d}{2}, \frac{2}{2q(y_i)+\nu})$, dont la moyenne (voir remarque 4.2.1) est

$$\begin{cases} E_\phi[U_i | \mathbf{y}] &= \frac{d+\nu^{(n)}}{2q(y_i)+\nu^{(n)}} \\ E_\phi[\log(U_i) | \mathbf{y}] &= \psi(\frac{d+\nu}{2}) - \log(\frac{2q(y_i)+\nu}{2}) \end{cases} \quad (4.27)$$

L'équation (4.24) prend la forme :

$$-\psi(\frac{\nu}{2}) + \log(\frac{\nu}{2}) + 1 + \frac{1}{N} \sum_{i=1}^N \log(w_i^{(n)}) - w_i^{(n)} + \psi(\frac{\nu_i^{(n)}+d}{2}) - \log(\frac{\nu_i^{(n)}+d}{2}) = 0 \quad (4.28)$$

Nous notons $\varphi(x) = \psi(x) - \log(x)$ et $\tilde{\varphi}^{(n)} = 1 + \frac{1}{N} \sum_{i=1}^N \log(w_i^{(n)}) - w_i^{(n)} + \psi(\frac{\nu_i^{(n)}+d}{2}) - \log(\frac{\nu_i^{(n)}+d}{2})$. La mise à jour du paramètre ν s'écrit

$$\varphi(\frac{\nu}{2}) = \tilde{\varphi}^{(n)} \quad (4.29)$$

Si $\tilde{\varphi}^{(n)} > 0$, il existe toujours une unique solution parce que la fonction $x \mapsto \varphi(x)$ est strictement croissante sur $\mathbb{R}^{+*}$. En conclusion, nous obtenons, l'algorithme EM pour une loi $\mathbf{T}_{\mathbf{m}, \Sigma, \nu}$:

1. Initialisation : $\mathbf{m}^{(0)} = \frac{1}{N} \sum_{i=1}^N y_i$, $\Sigma^{(0)} = \frac{1}{N} \sum_{i=1}^N (y_i - \mathbf{m}^{(0)}) (y_i - \mathbf{m}^{(0)})'$, $\nu^{(0)} = 100$ (estimation dans le cas gaussien)
2. Pour tout n , calculer $q^{(n)}(y_i) = \frac{1}{2} (y_i - \mathbf{m}^{(n)})' \Sigma^{(n)-1} (y_i - \mathbf{m}^{(n)})$.
Calculer les poids $w_i^{(n)} = \frac{d+\nu^{(n)}}{2q^{(n)}(y_i)+\nu^{(n)}}$.
Calculer la moyenne et variance des observations $(y_i)_{1 \leq i \leq N}$ en pondérant avec $\mathbf{m}^{(n+1)}$ et $\Sigma^{(n+1)}$ par les poids $w_i^{(n)}$.
Calculer $\tilde{\varphi}^{(n)}$ et $\nu^{(n+1)} = 2\varphi^{-1}(\tilde{\varphi}^{(n)})$.

4.2.3.4 Loi a posteriori et algorithme EM pour la loi K

Malgré des expressions plus complexes dans la loi a posteriori, il est toujours possible de calculer les moments a posteriori nécessaire dans les équations (4.23) et (4.24). En effet, nous pouvons intégrer l'expression (4.25), ce qui donne l'expression suivante pour la densité a posteriori de la texture :

$$p(u | \mathbf{y}) = \frac{u^{(\frac{d}{2}-a)-1} \exp\left(-\left(uq(y) + \frac{a}{u}\right)\right)}{2\left(\sqrt{\frac{q(y)}{a}}\right)^{a-\frac{d}{2}} K_{a-\frac{d}{2}}(2\sqrt{aq(y)})} \quad (4.30)$$

Nous pouvons alors calculer tous les moments a posteriori de U qui interviennent dans l'expression de Q :

$$\left\{ \begin{array}{lcl} E_\phi[U|\mathbf{y}] & = & \frac{K_{\frac{d}{2}-a+1}(2\sqrt{aq(y)})}{\sqrt{\frac{q(y)}{a}} K_{\frac{d}{2}-a}(2\sqrt{aq(y)})} \\ E_\phi[\frac{1}{U}|\mathbf{y}] & = & \frac{\sqrt{\frac{q(y)}{a}} K_{\frac{d}{2}-a-1}(2\sqrt{aq(y)})}{K_{\frac{d}{2}-a}(2\sqrt{aq(y)})} \\ E_\phi[\log(U)|\mathbf{y}] & = & H(\sqrt{\frac{q(y)}{a}}, a) \end{array} \right. \quad (4.31)$$

avec $H(x, a) = -\log(x) + \partial_\alpha \log K_\alpha(2ax)|_{\alpha=\frac{d}{2}-a}$.

La mise à jour du paramètre de queue a aboutit à la résolution de l'équation

$$\varphi(a) = \varphi^{(n)} \quad (4.32)$$

avec $\varphi^{(n)} = 1 - \left(\frac{1}{N} \sum_{i=1}^N E_{\phi_n}[\frac{1}{U_i}|\mathbf{y}] + E_{\phi_n}[\log(U_i)|\mathbf{y}] \right)$. Pour résumer, l'algorithme permettant d'estimer une loi $\mathbf{K}_{\mathbf{m}, \Sigma, a}$ est :

1. Initialisation : $\mathbf{m}^{(0)} = \frac{1}{N} \sum_{i=1}^N y_i$, $\Sigma^{(0)} = \frac{1}{N} \sum_{i=1}^N (y_i - \mathbf{m}^{(0)}) (y_i - \mathbf{m}^{(0)})'$, $a^{(0)} = 10$ (estimation dans le cas gaussien)
2. Pour n , calculer $q^{(n)}(y_i) = \frac{1}{2} (y_i - \mathbf{m}^{(n)})' \Sigma^{(n)-1} (y_i - \mathbf{m}^{(n)})$.
Calculer les poids $w_i^{(n)} = \frac{K_{\frac{d}{2}-a(n)+1}(2\sqrt{a^{(n)}q^{(n)}(y_i)})}{\sqrt{\frac{q^{(n)}(y_i)}{a^{(n)}}} K_{\frac{d}{2}-a(n)}(2\sqrt{a^{(n)}q^{(n)}(y_i)})}$.
Calculer la moyenne et variance $\mathbf{m}^{(n+1)}$ et $\Sigma^{(n+1)}$ des observations $\mathbf{y} = (y_i)_{1 \leq i \leq N}$ pondérées par les poids $w_i^{(n)}$.
Calculer $\varphi^{(n)}$ selon l'Eq. (4.32) et calculer $a^{(n+1)} = \varphi^{-1}(\varphi^{(n)})$.

4.2.4 Estimation de mélanges de lois elliptiques

Nous proposons dans cette section un algorithme d'estimation par EMV d'une CMCa-BI dont les lois d'émission appartiennent au modèle SIRV. Nous exploitons la double structure cachée pour un mélange de K modèles elliptiques : le processus $\mathbf{U}$ qui vient perturber la normalité des observations, et le processus $\mathbf{X}$ qui vient modifier les paramètres selon une dynamique markovienne. Les lois d'émission $f(y, \theta_k) \in \mathcal{E}(\mathbf{m}_k, \Sigma_k, \vartheta_k)$ appartiennent toutes au même modèle paramétrique. Le processus $\mathbf{U}$ de texture est alors tel que (U_i, Y_i) soient indépendants conditionnellement à $\mathbf{X}$ et $P(\mathbf{U}, \mathbf{Y} | \mathbf{X}) = \prod_{i=1}^N P(Y_i | U_i, X_i) P(U_i | X_i)$. La loi conditionnelle de Y_n conditionnellement à U_n, X_n est une loi normale $N(\mathbf{m}_{X_n}, U_{X_n} \Sigma_{X_n})$. La vraisemblance complète du modèle de mélange est donc une légère modification de la log-vraisemblance (4.21) par l'ajout du processus $\mathbf{X}$. Finalement, la maximisation de la fonction $Q(\phi, \phi_n)$ donne les formules de réestimation suivante :

$$\left\{ \begin{array}{l} \mathbf{m}_k^{(n+1)} = \frac{\sum_{i=1}^N w_{i,k}^{(n)} \pi_{i,k}^{(n)} y_i}{\sum_{i=1}^N \pi_{i,k}^{(n)} w_i^{(n)}} \\ \text{et} \\ \Sigma_k^{(n+1)} = \frac{1}{N} \sum_{i=1}^N \pi_{i,k}^{(n)} w_{i,k}^{(n)} (y_i - \mathbf{m}_k^{(n+1)}) (y_i - \mathbf{m}_k^{(n+1)})' \end{array} \right. \quad (4.33)$$

Comme dans les équations (4.33), les moments a posteriori $E_\phi[U_i | \mathbf{y}, X_i = k]$, $E_\phi[1/U_i | \mathbf{y}, X_i = k]$ et $E_\phi[\log(U_i) | \mathbf{y}, U_i = k]$ sont multipliés par les probabilités a posteriori $(\pi_{i,k}^{(n)})_{i,k}$. Les paramètres

ν_k d'un mélange de lois T sont obtenus par la résolution des équations suivantes :

$$\forall k \leq K, \varphi\left(\frac{\nu_k^{(n+1)}}{2}\right) = \log\left(\frac{\nu_k^{(n)} + M}{2}\right) - \psi\left(\frac{\nu_k^{(n)} + M}{2}\right) - 1 - \frac{1}{\sum_{i=1}^N \pi_{i,k}^{(n)}} \sum_{i=1}^N \pi_{i,k}^{(n)} \left\{ \log(w_{i,k}^{(n)}) - w_{i,k}^{(n)} \right\} \quad (4.34)$$

De même, les paramètres a_k d'un mélange de lois K sont solutions des équations :

$$\forall k \leq K, \varphi(a_k^{(n+1)}) = \frac{1}{\sum_{i=1}^N \pi_{i,k}^{(n)}} \sum_{i=1}^N \pi_{i,k}^{(n)} E\left[\frac{1}{U_i} + \log(U_i) | y_i, X_i = k\right] - 1 \quad (4.35)$$

L'algorithme EM proposé tire partie de la compatibilité des structures cachées $\mathbf{U}$ et $\mathbf{X}$: chaque type de paramètres des lois d'émission de chaque classe est mis à jour séparément, rendant ainsi l'implémentation de ces algorithmes et donc l'utilisation des procédures de segmentation non-supervisée envisageable pour des traitements rapides. Il est donc possible théoriquement de proposer le même genre d'approche pour le calcul de l'EMV des autres lois elliptiques. Dans les cas où il n'est pas possible de déduire la forme analytique de la loi a posteriori de la texture, il reste envisageable d'utiliser des méthodes de simulation pour calculer l'étape E avec un algorithme MCEM.

Remarque 4.2.6. *Les algorithmes EM proposés dans section se généralisent directement aux SIRV complexes (voir annexe A).*

4.2.5 Expérimentations

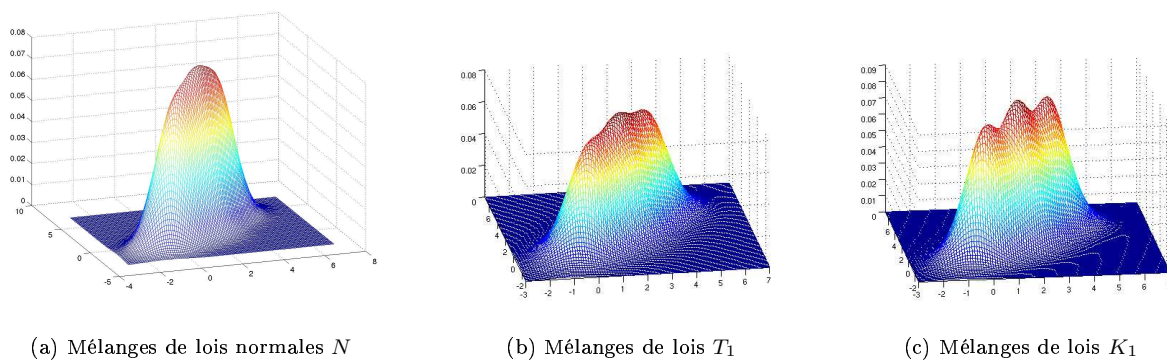
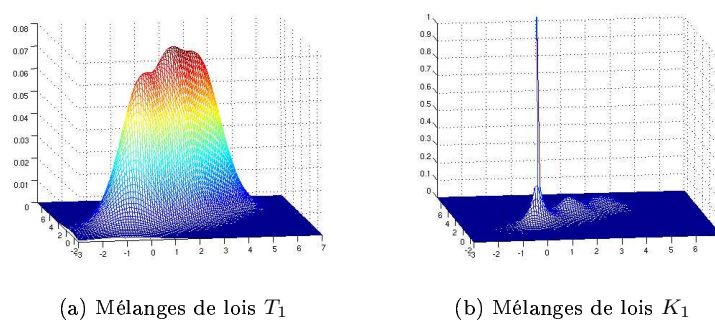
4.2.5.1 SIRV réels

Nous mettons en oeuvre dans cette section les algorithmes d'estimation de mélanges de loi de Student et de loi K dans le cadre de la segmentation non-supervisée (par MPM) de chaînes de Markov cachées CMCA-BI. Nous comparons les différents modèles SIRV : loi normale, loi T et loi K afin de mettre en évidence l'apport du paramètre de queue pour l'estimation et la segmentation.

Nous traitons le cas d'une CMCA-BI stationnaire à 3 classes. Le processus caché est plutôt "stable", avec une matrice de transition $A = (a_{ij})_{1 \leq i, j \leq 3}$, $a_{ii} = 0.8$ et $a_{ij} = 0.1$ quand $i \neq j$. Chaque classe est définie par les paramètres de moyennes et variances suivants : $\mu_1 = (0, 0)'$, $\mu_2 = (1.5, 1.5)'$, $\mu_3 = (3, 3)'$, $\Sigma_1 = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 1 \end{bmatrix}$, $\Sigma_2 = \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix}$ et $\Sigma_3 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$. Nous testons alors 5 scénarii différents :

1. Un mélange de 3 lois normales (mélange $\mathcal{N}$)
2. Un mélange de 3 lois T, de paramètres de queue identiques $\nu_1 = \nu_2 = \nu_3 = 10$ (mélange T_1)
3. Un mélange de 3 lois T, de paramètres de queue $\nu_1 = 5$, $\nu_2 = 10$ and $\nu_3 = 15$ (mélange T_2)
4. Un mélange de 3 lois K, de paramètres de queue identiques $a_1 = a_2 = a_3 = 5$ (mélange K_1)
5. Un mélange de 3 lois K, de paramètres de queues $a_1 = 1$, $a_2 = 2$, $a_3 = 4$ (mélange K_2)

Les densités des 5 mélanges obtenus sont tracées dans les fig. (4.3) et (4.4). Nous comparons les procédures de segmentation non-supervisée pour ces 5 mélanges en utilisant une CMCA-BI dont les lois d'émission sont soit des lois normales, des lois de Student ou des lois K. Les modèles correspondant sont notés respectivement modèles N, T et K. L'échantillon que nous segmentons est de taille $N = 1000$.

FIG. 4.3 – Mélanges de SIRV ayant même paramètre de queue (texture) N, K_1, T_1 FIG. 4.4 – Mélanges de SIRV ayant des paramètres de queue distincts K_2, T_2

mélanges	Taux d'erreur
$\mathcal{N}$	12,4
T_1 et T_2	14,5
K_1 et K_2	10,9

TAB. 4.1 – Taux d'erreur (%) en segmentation supervisée

Les taux d'erreur théorique (en contexte supervisé) sont estimés par Monte Carlo et sont donnés dans le tableau (4.1). Malgré des paramètres de queue différents, les mélanges de lois K (K_1 et K_2) et de lois T (T_1 et T_2) possèdent des taux d'erreur théoriques très proches, que nous considérons comme égaux.

Les mélanges de Student sont les plus difficiles à restaurer (en raison de la décroissance en puissance des queues, ce qui augmente la taille des zones de confusion), alors que les mélanges de loi K sont les plus faciles à segmenter (ce qui est bien illustré par les 3 pics séparés du mélange).

L'estimation des paramètres est effectuée par l'algorithme EM de la section 4.2.4, pour un nombre de 100 itérations. L'initialisation est réalisée en estimant les paramètres des lois d'émission à partir des classes construites par un algorithme au K plus proches voisins. Les résultats de la segmentation non-supervisée sont rassemblés dans le tableau (4.13), et ont été obtenu par Monte Carlo (sur 500 simulations).

Lors de la restauration non-supervisée, nous pouvons dissocier le cas où les queues sont égales de celui où elles sont distinctes. En effet, pour les mélanges $\mathcal{N}$, T_1 , K_1 , les 3 modèles donnent des performances similaires, avec une dégradation relativement faible des taux d'erreur par rapport au cas supervisé. La restauration la plus difficile est celle du mélange T_1 , pour lequel le modèle K donne des performances équivalentes (légèrement meilleure) que le vrai modèle.

Pour les mélanges T_2 , K_2 , il apparaît que l'estimation est affecté par les changements de "texture", ce qui dégrade les performances de segmentation. Il est au contraire notable que les modèles T et K deviennent d'une part équivalent dans un contexte hétérogène (pour les textures), et d'autre part plus performant dans le cas du mélange K_2 (nous pouvons considérer que les performances des modèles K et N pour les mélanges T_1 et T_2 sont équivalentes).

Ceci signifie que le paramètre de queue est une caractéristique des classes qui permet de mieux discriminer les classes. En effet, le modèle K et T réussissent à capturer cette différence de texture, même si l'algorithme EM ne permet pas de déterminer exactement les valeurs des paramètres de queue :

- Modèle T_2 : $\hat{\nu}_1 = 5,1$; $\hat{\nu}_2 = 14,2$; $\hat{\nu}_3 = 15$ (modèle T) et $\hat{a}_1 = 2$; $\hat{a}_2 = 4,4$; $\hat{a}_3 = 5$ (modèle K)
- Modèle K_2 : $\hat{\nu}_1 = 2,9$; $\hat{\nu}_2 = 6,4$; $\hat{\nu}_3 = 11$ (modèle T) et $\hat{a}_1 = 1$; $\hat{a}_2 = 4,4$; $\hat{a}_3 = 4,8$ (modèle K)

Les paramètres de moyenne et de variance sont estimés par contre sans biais.

4.2.5.2 SIRV complexes

Les SIRV utilisés en radar sont de moyenne nulle, ce qui est une difficulté supplémentaire pour la segmentation, car la valeur moyenne est un paramètre très discriminant entre les classes. Nous

Mélange Estimé	$\mathcal{N}$	T_1	K_1	T_2	K_2
N	13,4	16	13,1	16,9	13
T	13,4	15,6	13,8	15,7	11,7
K	13,5	15,5	13,1	15,7	11,7

TAB. 4.2 – Taux d’erreur (%) en segmentation non-supervisée pour les mélanges N , T_1 , T_2 , K_1 , K_2

mélanges	Taux d’erreur
CN	23,5
CT_1	24,9
CK_1	24,8
CT_2	25
CK_2	23,5

TAB. 4.3 – Taux d’erreur (%) en segmentation supervisée

montrons néanmoins que les SIRV ne possédant que des variances et des paramètres de queue distincts peuvent toujours être séparés, et que l’estimation se comporte suffisamment bien pour permettre des procédures de segmentation non-supervisée acceptables. Nous comparons pour les SIRV complexes (centrées), les performances des modèles de mélange complexe CN, CT et CK (respectivement pour loi normale complexe, loi T complexe et loi K complexe). Comme dans le cas réel, nous testons 5 scénarii, selon le type de la loi d’émission et l’égalité des paramètres de la texture.

Le processus caché a 3 états et a pour matrice de transition $A = (a_{ij})_{1 \leq i, j \leq 3}$, $a_{ii} = 0.8$ et $a_{ij} = 0.1$ quand $i \neq j$. Les classes sont de moyenne nulle et ont pour variance : $\Sigma_1 = \begin{bmatrix} 1 & 0.2 + 0.8i \\ 0.2 - 0.8i & 1 \end{bmatrix}$, $\Sigma_2 = \begin{bmatrix} 1 & -0.4 + 0.1i \\ -0.4 + 0.1i & 1 \end{bmatrix}$ et $\Sigma_3 = \begin{bmatrix} 1 & 0.5 + 0.5i \\ 0.5 - 0.5i & 1 \end{bmatrix}$. Nous testons alors 5 scénarii différents :

1. Un mélange de 3 lois normales (mélange CN)
2. Un mélange de 3 lois T, de paramètres de queue identiques $\nu_1 = \nu_2 = \nu_3 = 10$ (mélange CT_1)
3. Un mélange de 3 lois T, de paramètres de queue $\nu_1 = 5$, $\nu_2 = 10$ and $\nu_3 = 15$ (mélange CT_2)
4. Un mélange de 3 lois K, de paramètres de queue identiques $a_1 = a_2 = a_3 = 5$ (mélange CK_1)
5. Un mélange de 3 lois K, de paramètres de queue $a_1 = 1$, $a_2 = 2$, $a_3 = 4$ (mélange CK_2)

Nous donnons dans le tableau (4.3) les taux d’erreur en segmentation supervisée de ces 5 mélanges complexes. Les résultats de la segmentation non-supervisée sont rassemblées dans le tableau (4.4). Les 5 modèles testés ont des covariances complexes, et possèdent les mêmes paramètres de queue que dans le cas réel. Nous pouvons constater que les taux d’erreur sont beaucoup plus importants que dans le cas où la moyenne est non nulle, malgré des variances bien distinctes, et il est clair que la possibilité d’une restauration convenable est essentiellement due à la markovianité du processus des états (et à la forte probabilité de rester dans le même état). Les taux d’erreur sont cependant tous à peu près équivalents, et de l’ordre de 25%.

Des résultats de simulation, il est clair que le modèle N est très peu robuste à la présence de textures (et donc à la présence de queues de distribution épaisses), et ceci d'autant plus que les classes ont des textures différentes (nous constatons une très forte dégradation entre CK_1 et CK_2). Enfin, nous pouvons constater que les performances des modèles K et T sont équivalentes lorsque les textures sont identiques, mais se différencient légèrement lorsque les textures sont distinctes. D'ailleurs contrairement au cas où les moyennes étaient non nulles, l'hétérogénéité des textures dégrade la procédure de segmentation non-supervisée (particulièrement dans le cas où le mélange de lois K). Cependant l'algorithme EM en donne (en moyenne) une estimation acceptable :

- Modèle CT_2 : $\hat{\nu}_1 = 5,9$; $\hat{\nu}_2 = 10,4$; $\hat{\nu}_3 = 13,1$ (modèle T) et $\hat{a}_1 = 2,5$; $\hat{a}_2 = 4,4$; $\hat{a}_3 = 5,4$ (modèle K)
- Modèle CK_2 : $\hat{\nu}_1 = 2,2$; $\hat{\nu}_2 = 5,9$; $\hat{\nu}_3 = 9,7$ (modèle T) et $\hat{a}_1 = 1$; $\hat{a}_2 = 2,1$; $\hat{a}_3 = 3,4$ (modèle K)

Mélange Estimé	CN	CT ₁	CK ₁	CT ₂	CK ₂
CN	27	40,3	31,2	42,6	41,9
CT	27,5	29	28	29,3	31,9
CK	27,8	29,5	28	31,9	28,1

TAB. 4.4 – Taux d'erreur (%) en segmentation non-supervisée pour des mélanges CN , CT_1 , CT_2 , CK_1 , CK_2

Comme précédemment, les estimations de la matrice de covariance sont convenables pour les modèles K et T, alors que les estimations données par le modèle N en sont des dilatations.

4.2.5.3 Conclusion

Nous avons testé les algorithmes EM d'estimation de mélanges de SIRV dans 10 scénarii différents. Nous pouvons conclure que la loi normale (ou le mélange N) est à éviter lorsque nous avons des mélanges de SIRV, alors que les lois K et T ont des performances similaires dans la plupart des cas. Ainsi, comme l'algorithme EM pour l'estimation d'un mélange de loi T est moins coûteux en termes de calcul (contrairement à la loi K qui nécessite l'évaluation de la fonction spéciale K_α), il peut être préférable en pratique d'utiliser un modèle T (ou CT) pour la segmentation de lois K.

Nous pouvons affirmer aussi que les paramètres de texture sont des grandeurs identifiables et peuvent permettre utilement de caractériser des classes, même si leur estimation est plus sensible que celles des paramètres de moyenne ou de variance. Nous remarquons que l'estimation est de meilleure qualité dans le cas de mélange de SIRV centrées que dans le cas de SIRV décentrées. En effet, dans ce dernier cas, l'estimation des paramètres de queues est polluée car les observations issues des autres classes sont considérées comme des atypiques. Ainsi dans le contexte des applications radar, le mélange de SIRV peut servir à caractériser efficacement les zones de texture correspondantes à des matrices de variance et des textures différentes.

4.3 Les copules

Nous introduisons dans cette section les copules pour la modélisation des lois d'émission dans le cadre de la segmentation d'observations multivariées par CMCA-BI. Les copules sont utilisées

pour la première fois dans le contexte de la segmentation bayésienne, et sont aussi un outil de modélisation multivariée nouveau en traitement d'image et du signal. L'intérêt de ces modèles est de permettre d'une part d'introduire de nouvelles formes de dépendance entre les différentes composantes des observations, mais aussi de contrôler les lois unidimensionnelles des lois multivariées (pour lesquelles nous avons vu qu'il est possible d'avoir des a priori quant à leur forme, par exemple pour des intensités, les densités marginales sont $K_{a,\sigma^2}, \mathbf{K}_{0,\sigma^2,a}, \gamma, \dots$). Nous obtenons donc des modèles statistiques qui respectent mieux certaines connaissances a priori, et qui sont plus flexibles en permettant des formes variées pour les densités (par opposition au SIRV par exemple).

Nous donnons un bref aperçu des propriétés et des familles paramétriques de copules, ainsi que des méthodes d'estimation utilisées dans le cadre d'une seule copule. Nous décrivons alors l'estimation des CMCa-BI par projection d'une fonction estimante, ce qui permet d'éviter des maximisations pénibles et pénalisantes pour les applications. Nous montrons l'intérêt des copules en construisant aisément des lois gammas multivariées utiles pour la segmentation d'image multicomposantes. Nous comparons alors les performances des deux types de modèles multivariés vu dans ce chapitre : les SIRV et les copules, sur des données simulées et réelles.

4.3.1 Définition et propriétés

Pour une introduction à la théorie des copules, nous renvoyons au livre de Nelsen ([114]) qui présente les copules dans un cadre très général, ou à celui de Joe ([94]) qui traite de nombreux aspects de la modélisation statistique multivariée avec les copules.

Définition 4.3.1. Copule

Une copule de dimension d est une fonction telle que :

1. $C : [0, 1]^d \longrightarrow [0, +\infty[$
2. Si $\exists i \leq d$, tq $u_i = 0$, alors $C(u_1, \dots, u_d) = 0$
3. $\forall i \leq d$, $C(1, \dots, u_i, \dots, 1) = u_i$
4. $\forall \mathbf{u}, \mathbf{v} \in [0, 1]^d$ tq $\forall i, u_i \leq v_i$, $\Delta_{\mathbf{u}}^{\mathbf{v}} C \geq 0$

avec $\Delta_{\mathbf{u}}^{\mathbf{v}} C = \Delta_{u_d}^{v_d} \Delta_{u_{d-1}}^{v_{d-1}} \dots \Delta_{u_2}^{v_2} \Delta_{u_1}^{v_1} C(\mathbf{x})$. L'opérateur $(\Delta_{u_k}^{v_k})$ est défini par

$$\forall \mathbf{x}, \Delta_{u_k}^{v_k} C(\mathbf{x}) = C(x_1, \dots, v_k, \dots, x_d) - C(x_1, \dots, u_k, \dots, x_d)$$

$\Delta_{\mathbf{u}}^{\mathbf{v}} C$ est appelé le C – volume de l'hypercube défini par les coordonnées $(\mathbf{u}, \mathbf{v})$. Dans le cas d'une copule bivariée, la condition 4 de positivité du 2 – volume s'écrit :

$$C(v_1, v_2) - C(v_1, u_2) - C(u_1, v_2) + C(u_1, u_2) \geq 0$$

Les propriétés qui découlent directement de cette définition sont qu'une copule est une fonction croissante, continue, presque partout dérivable, et que c'est en fait une fonction de répartition sur $[0, 1]^d$, dont les marges sont uniformes sur le segment $[0, 1]$. Le résultat fondamental de la théorie des copules est le théorème suivant :

Théorème 4.3.1. Théorème de Sklar

Soit F une fonction de répartition sur $\mathbb{R}^d$ dont les distributions marginales ont pour fonctions de répartitions (f.d.r) $F_1, \dots, F_d$. Alors il existe une copule C telle que

$$\forall \mathbf{x} \in \mathbb{R}^d, F(x_1, \dots, x_d) = C(F_1(x_1), \dots, F_d(x_d)) \quad (4.36)$$

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

Le théorème de Sklar donne la forme explicite de la classe de Fréchet $\mathcal{F}(F_1, \dots, F_d)$, où celle-ci désigne l'ensemble de distributions sur $\mathbb{R}^d$ ayant pour marges les distributions $F_1, \dots, F_d$.

Le théorème de Sklar nous dit que $\mathcal{F}(F_1, \dots, F_d)$ est décrit par l'ensemble des copules, c'est-à-dire des distributions sur $[0, 1]^d$ à marges uniformes. L'unicité de la décomposition de Sklar lorsque les F_i sont continues nous permet alors d'en avoir une indexation (injective). Il est donc possible de construire (resp. d'estimer) une loi multivariée de façon univoque en spécifiant (resp. en estimant) ses distributions marginales (appelées aussi marges) et sa copule.

Le théorème de Sklar prend la forme suivante lorsque la f.d.r est dérivable : si $X_1, \dots, X_d$ sont des v.a.r ayant une densité f_i relativement à λ (mesure de Lebesgue) et de densité jointe f par rapport à $\lambda^{\otimes d}$, alors les densités marginales et jointe sont liées de la manière suivante :

$$\forall \mathbf{x}, f(\mathbf{x}) = \prod_{i=1}^d f_i(x_i) \times c(F_1(x_1), \dots, F_d(x_d)) \quad (4.37)$$

où la fonction c est une densité sur $[0, 1]^d$ dont les marges sont uniformes, telle que $c = \partial_{1\dots d} C$, C étant la copule définie par l'eq. (4.36). c est appelée la densité de la copule. Les densités conditionnelles se calculent directement à partir de l'eq. (4.37) :

$$\text{Si } f(x_1, x_2) = f_1(x_1)f_2(x_2)c(F_1(x_1), F_2(x_2)) \text{ alors } f(x_1 | x_2) = f_1(x_1)c(F_1(x_1), F_2(x_2))$$

De la même façon, nous pouvons calculer directement la loi d'un groupe de variables, conditionnellement à un autre sous-groupe de variables si l'on connaît toutes les marges et la copule globale C . Soit $\mathbf{X} = (X_i)_{i \in [1..d]}$ une variable aléatoire dans $\mathbb{R}^d$ de copule C , avec des densités marginales f_i . Si I est inclus dans $[1..d]$, tel que $|I| = k$, alors la copule C_I du vecteur $\mathbf{X}_I = (X_i)_{i \in I}$ est :

$$C_I(u_{i_1}, \dots, u_{i_k}) = C(1, \dots, u_{i_1}, 1, \dots, u_{i_k}, 1, \dots, 1) \quad (4.38)$$

C_I est dite sous-copule de C . Nous notons c_I la densité associée. La densité de $\mathbf{X}$ conditionnellement à $\mathbf{X}_I$ est :

$$\forall \mathbf{x} \in \mathbb{R}^d, f(\mathbf{x} | \mathbf{x}_I) = \prod_{i \notin I} f_i(x_i) \times \frac{c(F_1(x_1), \dots, F_d(x_d))}{c_I(F_{i_1}(x_{i_1}), \dots, F_{i_k}(x_{i_k}))} \quad (4.39)$$

Remarque 4.3.1. Classes de Fréchet

Les classes de Fréchet sont complètement identifiées dans le cas où toutes les marges sont univariées, mais le problème de leur détermination reste en toute généralité très complexe et nécessite d'être examiné au cas par cas. Le problème de Fréchet sous sa forme la plus générale est : soit $F_1, \dots, F_n$ des fonctions de répartition sur $\mathbb{R}^{d_i}$, $d_i \geq 1$, ayant éventuellement des marges communes. Quelle est l'ensemble des probabilités sur $\mathbb{R}^m$, $m \leq d_1 + \dots + d_n$ ayant pour marges les distributions $F_1, \dots, F_n$?

S'il n'y a pour l'instant pas de réponse générale, il y a deux types de réponses partielles au pro-

blème : la détermination de contraintes de compatibilité entre les probabilités F_i pour que la classe de Fréchet soit non vide, et la construction de familles de lois possédant les marges désirées. Joe traite plusieurs cas particuliers de problèmes de Fréchet, selon le type de lois pour $F_1, \dots, F_n$, voir chapitres 3, 4 dans [94].

Remarque 4.3.2. Le cas de l'indépendance est évidemment retrouvé par ce que l'on appelle la copule produit : $C^\perp(\mathbf{u}) = \prod_{i=1}^d u_i$, et nous avons alors $c^\perp(\mathbf{u}) = 1$.

L'intérêt de la théorie des copules pour la modélisation, et plus généralement pour les statistiques est multiple :

1. Nous pouvons construire potentiellement une multitude de nouveaux modèles multivariés sur $\mathbb{R}^d$ ([94, 67]), et proposer une réelle alternative pratique à la loi normale, ou à ces prolongements (comme les modèles elliptiques, section 4.2.1).
2. Les modèles multivariés peuvent être construits de manière à valoriser l'information acquise par l'étude univariée de chacune des marges et en proposant ensuite un modèle de dépendance crédible entre chacune des composantes.
3. Nous pouvons proposer ou revisiter de nouvelles mesures de dépendance entre variables aléatoires : fonctionnelle (la copule elle-même), les mesures d'association de Kendall, Spearman, information mutuelle, ... (voir [114, 146])
4. Construction de nouvelles procédures d'estimation de lois multivariées : paramétriques, semi-paramétriques, non-paramétriques [144]
5. Construction de nouveaux tests d'indépendance [47]
6. Construction de nouveaux modèles de séries temporelles stationnaires non-linéaires, [94, 41, 46]

Nous donnons tout d'abord quelques exemples de modèles paramétriques de copules couramment utilisés. A ce sujet, nous signalons deux méthodes différentes de construction de copules : une méthode de construction directe, le plus souvent de copules bivariées, qui aboutit essentiellement aux familles de copules archimédiennes. L'autre consiste à déterminer à l'aide du théorème de Sklar la copule sous-jacente à des modèles multivariés usuels, dont les représentants les plus nombreux sont les copules elliptiques.

Définition 4.3.2. Copule Archimédienne

Soit φ une fonction vérifiant les propriétés suivantes :

$$\begin{cases} \varphi : [0, 1] \longrightarrow [0, +\infty] \\ \varphi(1) = 0, \varphi(0) = +\infty \\ \varphi \text{ continue, strictement décroissante} \end{cases}$$

Une copule archimédienne est alors définie par $C(u, v) = \varphi^{-1}(\varphi(u) + \varphi(v))$. φ est appelé le générateur de la copule. Si φ est deux fois dérivable et que $\varphi' > 0$, la densité de la copule est $c(u, v) = (\varphi^{-1})^{(2)}(\varphi(u) + \varphi(v)) \times \varphi'(u)\varphi'(v)$.

Nous renvoyons aux livres de Nelsen et Joe pour une liste assez complète des copules archimédiennes (le tableau (4.5) rassemble les plus couramment utilisées).

Copules	$t \mapsto \varphi_\theta(t)$	θ	C_θ
Clayton	$\frac{1}{\theta}(t^{-\theta} - 1)$	$[-1, \infty) - \{0\}$	$C_\theta(u, v) = \max((u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}, 0)$
Gumbel	$(-\log t)^\theta$	$[1, \infty)$	$C_\theta(u, v) = \exp\left(-\left[(-\log u)^\theta + (-\log v)^\theta\right]^{1/\theta}\right)$
Frank	$(-\log \frac{e^{-\theta t} - 1}{e^{-\theta} - 1})$	$\mathbb{R} - \{0\}$	$C_\theta(u, v) = -\frac{1}{\theta} \log\left(1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1}\right)$

TAB. 4.5 – Quelques exemples de copules archimédiennes

Les expressions des copules archimédiennes peuvent être relativement simples, cependant les densités correspondantes (que nous aurons à utiliser dans les applications de segmentation et classification) ont des expressions compliquées et peu maniables. Un reproche souvent fait à ces modèles est l'indexation par un paramètre monodimensionnel, ce qui constitue souvent une limitation à l'adéquation aux données. Il s'agit aussi souvent de copules bivariées, mais elles sont extensibles à la dimension d . Sous certaines conditions sur le générateur φ , la fonction à d arguments définie par

$$\forall \mathbf{u} \in [0, 1]^d, C(\mathbf{u}) = \varphi_\theta^{-1}\left(\sum_{i=1}^d \varphi_\theta(u_i)\right) \quad (4.40)$$

est une copule.

Les copules elliptiques sont une autre famille très riche de copules multivariées, à laquelle font partie les copules gaussiennes et Student données par les définitions suivantes.

$$\text{Copule Gaussienne : } c_\rho(u_1, \dots, u_d) = |\rho|^{-1/2} \exp\left(-\frac{1}{2}\zeta'(\rho^{-1} - I_d)\zeta\right) \quad (4.41)$$

où $\rho = (r_{ij})_{1 \leq i, j \leq M}$ est une matrice de corrélation, et $\zeta = (\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d))$, Φ étant la f.d.r d'une gaussienne centrée réduite.

$$\text{Copule Student : } c_\rho(u_1, \dots, u_d) = |\rho|^{-1/2} \frac{\Gamma(\frac{\nu+d}{2})\Gamma(\frac{\nu}{2})^{d-1}}{\Gamma(\frac{\nu+1}{2})^d} \frac{\left(1 + \frac{1}{\nu}\zeta'\rho^{-1}\zeta\right)^{-\frac{\nu+d}{2}}}{\prod_{k=1}^d \left(1 + \frac{\zeta_k^2}{\nu}\right)^{-\frac{\nu+1}{2}}} \quad (4.42)$$

où $\rho = (r_{ij})_{1 \leq i, j \leq M}$ est une matrice de corrélation, et $\zeta = (T_\nu^{-1}(u_1), \dots, T_\nu^{-1}(u_d))$, T_ν étant la fdr d'une loi de Student centrée réduite.

Ces deux copules sont les copules des lois elliptiques correspondantes, et dont l'expression a été obtenue en appliquant le théorème de Sklar (car nous connaissons la loi des marges et la loi jointe). Ainsi la copule gaussienne est la seule à redonner une densité gaussienne lorsque l'on lui applique des marges de gaussienne univariée. De même, la copule de Student est la seule copule qui redonne une loi de Student lorsque les marges sont de type Student (et que tous les paramètres de queue sont égaux).

Nous utiliserons essentiellement ces 2 modèles par la suite, parce que nous avons une expression explicite, aisément manipulable et calculable de la densité (que nous utilisons constamment dans les modèles MMCA). Malheureusement, ces copules font presque figure d'exception car les copules elliptiques ont rarement d'expression aussi facile. Fan *et al.* ([67]) ont utilisé la relation bijective entre lois des marges et lois jointe pour étudier de manière générale les copules elliptiques et leur expression rend prohibitive leur utilisation. Par exemple, nous donnons l'expression explicite de la

copule de la loi K

$$\forall \mathbf{u}, c(\mathbf{u}) = \frac{\Gamma(a)^{d-1}}{(2a^a)^{d-1}} \sqrt{2a}^{\frac{(d-1)(2a+1)}{2}} \frac{\sqrt{(\zeta \rho^{-1} \zeta')^{a-d/2}} K_{a-d/2}(\sqrt{2a} \zeta \rho^{-1} \zeta')}{\prod_i |\zeta_i| \prod_i K_{a-d/2}(\sqrt{2a} |\zeta_i|)} \quad (4.43)$$

où $\zeta = (\mathcal{K}_a^{-1}(u_1), \dots, \mathcal{K}_a^{-1}(u_d))$, avec $\mathcal{K}_a$ fonction de répartition d'une loi K centrée réduite, de paramètre a . L'évaluation de cette densité implique donc des calculs complexes (entre autres l'intégration numérique pour calculer les f.d.r.), et interdit son utilisation de manière intensive.

Mesures d'association La copule représente de manière exhaustive la dépendance entre des variables aléatoires. L'information qu'elle délivre est malheureusement trop complexe pour pouvoir être interprétée en tant que lien de dépendance entre variables. Pour cette raison, des mesures d'association numérique sont utilisées pour avoir un résumé interprétable de la copule. De manière remarquable, un certain nombre de mesures classiques d'association entre couple de variables aléatoires réelles (X, Y) sont fonction uniquement de la copule C_{XY} [114]. Les mesures les plus courantes sont le tau de Kendall τ et le Rho de Spearman ρ , dont les définitions sont les suivantes.

Si nous avons deux copies indépendantes $(X_i, Y_i)_{i=1,2}$ du couple (X, Y) , nous avons

$$\tau = P((X_1 - X_2)(Y_1 - Y_2) > 0) - P((X_1 - X_2)(Y_1 - Y_2) < 0) \quad (4.44)$$

Le tau de Kendall τ est la différence entre la probabilité d'avoir une paire concordante et une paire discordante. Nelsen a montré que

$$\tau_{XY} = 4 \int_{[0,1]^2} C_{XY}(x, y) dC_{XY}(x, y) - 1$$

Le Rho de Spearman est le coefficient de corrélation (au sens usuel) des rangs à un facteur près, i.e.

$$\tilde{\rho}_{XY} = E[(F_Y(Y) - E[F_Y(Y)])(F_X(X) - E[F_X(X)])] \quad (4.45)$$

Nous avons alors

$$\tilde{\rho}_{XY} = 12 \int_{[0,1]^2} C(u, v) du dv - 3$$

Pour l'étude de la dépendance entre deux variables aléatoires, ces mesures sont donc préférables au coefficient de corrélation de Pearson entre X, Y qui dépend de la copule mais aussi des marginales. Cette dépendance apparaît nettement au travers de l'identité de Hoeffding :

Proposition 4.3.1. Identité de Hoeffding

Si X et Y sont deux variables aléatoires réelles intégrables, telles que $E[|XY|] < \infty$, et si nous notons F_X, F_Y, F_{XY} les fonctions de répartition respectives de X, Y et du couple, nous avons alors

$$\text{cov}(X, Y) = \int_{\mathbb{R}^2} (F_{XY}(x, y) - F_X(x)F_Y(y)) dx dy$$

Par conséquent si X et Y sont deux variables centrées réduites, leur coefficient de corrélation ρ (de Pearson) est égale à

$$\rho = E[XY] = \int_{\mathbb{R}^2} (C_{XY}(F_X(x), F_Y(y)) - F_X(x)F_Y(y)) \, dx dy$$

et dépend des marginales.

Lorsque nous utilisons une copule paramétrique, il est par conséquent intéressant d'en exprimer les mesures τ et $\tilde{\rho}$ en fonction de ces paramètres, afin de retrouver la même interprétation intuitive du coefficients de corrélation dans les modèles classiques. Sur ce point, les copules elliptiques peuvent être mal interprété car elles sont paramétrées par une matrice de corrélation (de Pearson) ρ . Comme nous l'avons vu, cette matrice ne peut pas être égale (en toute généralité) à la matrice de covariance Σ d'un vecteur l'ayant pour copule, car Σ dépend des marges. En fait pour les copules elliptiques, la matrice $\rho = (\rho_{ij})$ est reliée au tau de Kendall (τ_{ij}) :

$$\forall i \neq j, \rho_{ij} = \sin\left(\frac{\pi}{2}\tau_{ij}\right) \quad (4.46)$$

De plus, la nullité du paramètre ρ ne correspond pas à la situation d'indépendance entre les variables (X_i) , mis à part pour la copule gaussienne. Ceci se traduit par le fait que la famille des copules Student ne contient pas la copule d'indépendance $c^\perp$, contrairement au cas de la copule normale pour laquelle $\rho = Id_d \implies c_\rho = c^\perp$.

4.3.2 Inférence d'une copule

Avant l'introduction de l'estimation de mélanges de copules, nous rappelons différentes méthodes utilisées pour l'inférence paramétrique d'un modèle spécifié à l'aide d'une copule. Nous souhaitons estimer les paramètres $(\theta = (\theta_i)_{1 \leq i \leq d}, \eta)$ d'une densité multivariée, dont les densités marginales appartiennent aux modèles $\{f_{\theta_i}, \theta_i \in \Theta_i\}$ et dont la copule appartient au modèle $\{c_\eta, \eta \in \Upsilon\}$.

Nous nous concentrons sur le maximum de vraisemblance, car cette méthode d'estimation est souvent considérée comme la méthode de référence. Nous verrons de plus que le problème d'optimisation que l'on doit résoudre peut être contourné par l'usage des fonctions estimantes. Si $\mathbf{y} = (y_1, \dots, y_N)$ où $y_i = (y_i^j)_{1 \leq j \leq d}$ est un échantillon i.i.d, nous devons maximiser la fonctionnelle $\mathcal{L}(\theta_j, \eta)$

$$\mathcal{L}(\theta, \eta) = \sum_{i=1}^N \left\{ \sum_{j=1}^d \log f_{\theta_j}(y_i^j) + \log c_\eta(F_{\theta_1}(y_i^1), \dots, F_{\theta_d}(y_i^d)) \right\} \quad (4.47)$$

Il n'y a généralement pas de solution analytique à ce problème de maximisation, et la résolution numérique en est complexe, car tous les paramètres sont liés par la copule. En effet, le vecteur score $s(\theta, \eta) = \left(\nabla_{\theta_1} \mathcal{L}' \dots \nabla_{\theta_d} \mathcal{L}' \nabla_\eta \mathcal{L}' \right)'$ donnent les équations normales suivantes :

$$\begin{cases} \forall j \leq d, \sum_{i=1}^N \nabla_{\theta_j} \log f_{\theta_j}(y_i^j) + \partial_j \log c_\eta(F_{\theta_1}(y_i^1), \dots, F_{\theta_d}(y_i^d)) \times \nabla_{\theta_j} \log f_{\theta_j}(y_i^j) = 0 \\ \text{et} \\ \nabla_\eta \mathcal{L} = \sum_{i=1}^N \nabla_\eta \log c_\eta(F_{\theta_1}(y_i^1), \dots, F_{\theta_d}(y_i^d)) = 0 \end{cases} \quad (4.48)$$

∂_j est la dérivée partielle relativement à la j -ième variable. La difficulté à calculer l'EMV nécessite de trouver des méthodes alternatives “efficaces” afin d'éviter de tomber sur de “mauvais” maxima locaux, aboutissant à de mauvais estimateurs. Très rapidement avec l'estimation des premiers modèles multivariés spécifiés à l'aide de copule, des techniques ont été proposées pour contourner ce problème, et se ramener à des calculs beaucoup plus simples, parfois même explicites : c'est le cas de la méthode *Inference For Margins* (IFM, [144, 94]).

L'idée est de découper la recherche du maximum global (4.47) en 2 étapes, en mettant à profit le fait que la log-vraisemblance $\mathcal{L}(\theta, \eta)$ soit la somme de deux termes : la somme des log-vraisemblances des marges (notées $\mathcal{L}_j$ pour $1 \leq j \leq d$) et un terme qui ne dépend que de la copule $\nabla_\eta \mathcal{L}$.

Nous procédons alors en deux temps : on détermine les EMV $\hat{\theta}_j$ des paramètres des marges comme dans le cas indépendant, puis ils sont utilisés pour calculer $\nabla_\eta \mathcal{L}$, qui est maximisé à son tour en η .

$$\begin{cases} \hat{\theta}_j = \arg \max_{\theta_j} \sum_{i=1}^N \log f_{\theta_j}(y_i^j) \\ \hat{\eta} = \arg \max_{\eta \in E} \sum_{i=1}^N \log c_\eta(F_{\hat{\theta}_1}(y_i^1), \dots, F_{\hat{\theta}_d}(y_i^d)) \end{cases} \quad (4.49)$$

Cette simplification algorithmique correspond à l'utilisation d'une fonction estimante différente du score $s(\theta, \eta)$. La maximisation en deux temps revient à résoudre le système

$$\begin{cases} \forall j \leq d, \nabla_{\theta_j} \mathcal{L}_j = 0 \\ \nabla_\eta \mathcal{L} = 0 \end{cases} \quad (4.50)$$

qui est légèrement différent de celui obtenu pour l'EMV. Le système (4.50) correspond à la fonction estimante $g_{\text{IFM}}(\theta, \eta) = \left(\nabla_{\theta_1} \mathcal{L}'_1 \dots \nabla_{\theta_d} \mathcal{L}'_d \nabla_\eta \mathcal{L}' \right)'$. Ceci permet d'affirmer sous des conditions similaires à celles nécessitant la consistance de l'EMV que les estimateurs $\left((\hat{\theta}_{i,\text{IFM}})_{1 \leq i \leq d}, \hat{\eta}_{\text{IFM}} \right)$ issus de ce programme de maximisation sont consistants et asymptotiquement normaux. La perte par rapport à l'estimateur du maximum de vraisemblance se situe au niveau de l'efficacité de l'estimateur : la variance asymptotique $\mathcal{E}(g_{\text{IFM}})^{-1}$ (voir [94, 95]) est supérieure à la variance de l'EMV. Néanmoins d'un point de vue pratique, l'estimateur IFM fournit le plus souvent une estimation de meilleure qualité qu'une maximisation globale. L'estimateur IFM affiche même une variance inférieure que “l'EMV calculé”, ainsi qu'un biais moindre dans le cas d'erreur de spécifications de la loi des marges (la comparaison par Monte-Carlo des 2 estimateurs a été effectuée dans différents contextes par Joe et Xu [94] et Fermanian et Scaillet [68]).

Remarque 4.3.3. *Sur les aspects de robustesse de l'estimation de la copule relativement aux erreurs sur les marges, nous signalons qu'un estimateur semi-paramétrique appelé omnibus, permet d'éviter la propagation de cette erreur à la copule, voir [74]. Au lieu d'utiliser les fonctions de répartition $F_{\hat{\theta}_j}(y_i^j)$ pour déterminer l'EMV de la copule, on utilise $\hat{F}_j$ l'estimateur empirique de la f.d.r. L'estimation de la copule n'est donc pas perturbée par une mauvaise estimation et/ou spécification des marges (voir [68] pour une comparaison entre optimisation globale, IFM et omnibus). Cet estimateur est notamment employé pour l'estimation de modèle semi-paramétrique de séries temporelles stationnaires, dont seule la copule est définie par un modèle paramétrique [41].*

4.3.3 Inférence d'un mélange de copules

Pour l'inférence des CMCA-BI, nous décomposons les lois d'émission $f(y, \theta_k)$ en terme copules et marges, de telle sorte que nous ayons :

$$f(y, \theta_k) = \prod_{i=1}^d f_{\theta_{i,k}}(y^i) \times c_{\eta_k}(F_{\theta_{1,k}}(y^1), \dots, F_{\theta_{d,k}}(y^d)) \quad (4.51)$$

les marges appartiennent au modèle $\{f_{\theta_{i,k}}, \theta_{i,k} \in \Theta_{i,k}\}$ (de f.d.r. $F_{\theta_{i,k}}$) avec $1 \leq i \leq d$, $1 \leq k \leq K$, et les copules appartiennent au modèle $\{c_{\eta_k}, \eta_k \in \Upsilon_k\}$. Nous notons $\theta_k = ((\theta_{i,k})_{1 \leq i \leq d}, \eta_k)$, et le paramètre ϕ de la CMCA-BI est $\phi = (A, (\theta_k)_k) = (A, (\theta_{i,k})_{i,k}, (\eta_k)_k)$.

4.3.3.1 Identifiabilité des lois

Pour assurer l'identifiabilité des mélanges de la forme (4.51), nous supposons que les marges et les copules sont identifiables, et que les f.d.r $F_{\theta_{i,k}}$ sont continues :

$$\forall i \in [1..d], \sum_{k=1}^K \pi_k f_{\theta_{i,k}}(y_i) = \sum_{k=1}^K \pi'_k f_{\theta'_{i,k}}(y_i) \mu - \text{ps} \implies \forall k \leq K, \pi_k = \pi'_k \text{ et } \theta_{i,k} = \theta'_{i,k} \quad (4.52)$$

$$\forall \mathbf{u} \in [0, 1]^d, \sum_{k=1}^K \pi_k c_{\eta_k}(\mathbf{u}) = \sum_{k=1}^K \pi'_k c_{\eta'_k}(\mathbf{u}) \mu - \text{ps} \implies \forall k \leq K, \pi_k = \pi'_k \text{ et } \eta_k = \eta'_k \quad (4.53)$$

Sous cette condition, le modèle $f(y, \theta_k)$ est identifiable. En effet, si nous avons :

$$\sum_{k=1}^K \pi_k f(y, \theta_k) = \sum_{k=1}^K \pi'_k f(y, \theta'_k)$$

l'égalité est alors vraie composante par composante, et l'hypothèse d'identifiabilité des marginales permet d'affirmer que nous avons $\forall k, \forall i, \theta_{i,k} = \theta'_{i,k}$ et $\pi_k = \pi'_k$. Ceci implique en particulier que nous avons $\forall i, k, \forall y_i, F_{\theta_{i,k}}(y_i) = F_{\theta'_{i,k}}(y_i)$. Nous pouvons donc nous ramener à l'hypercube $[0, 1]^d$, et l'identifiabilité des mélanges de copules permet de conclure.

Enfin l'identifiabilité d'un mélange de copules peut souvent se déduire de l'identifiabilité de modèles multivariés classiques. Soit $\{C_\eta, \eta \in \Upsilon\}$ une famille de copules et si nous avons

$$\forall \mathbf{u} \in [0, 1]^d, \sum_{k=1}^K \pi_k C_{\eta_k}(\mathbf{u}) = \sum_{k=1}^K \pi'_k C_{\eta'_k}(\mathbf{u})$$

Alors pour tout $H_1, \dots, H_d$ f.d.r , nous avons

$$\forall \mathbf{y}, \sum_{k=1}^K \pi_k C_{\eta_k}(H_1(y_1), \dots, H_d(y_d)) = \sum_{k=1}^K \pi'_k C_{\eta'_k}(H_1(y_1), \dots, H_d(y_d))$$

Il suffit alors de choisir les fonctions H_i de manière à ce que le problème ait déjà été résolu pour la famille multivariée $\{C_\eta(H_1, \dots, H_d), \eta \in E\}$ (les f.d.r. H_i peuvent aussi dépendre d'un paramètre θ). Ainsi pour les copules normales et Student ayant même nombre de degrés de libertés, il suffit

de prendre les f.d.r Φ et T_ν pour obtenir l'identifiabilité.

4.3.3.2 Estimation par projection des équations IFM

La difficulté intrinsèque de la détermination d'un estimateur d'une CMCa-BI, et en particulier de l'EMV, est accentuée par la spécification des lois d'émission à l'aide des copules. Une estimation classique d'un mélange de copules par algorithme EM aboutit à la maximisation de la fonction Q suivante

$$\begin{aligned} Q(\phi, \phi_n) = & E[\log(p(\mathbf{X}, A)) | \mathbf{y}, \phi_n] + \sum_{i,j,k=1}^{N,d,K} \pi_{i,k}^{(n)} \log(f_{\theta_{j,k}}(y_i^j)) \\ & + \sum_{i,k=1}^{N,K} \pi_{i,k}^{(n)} \log(c_{\eta_k}(F_{\theta_{1,k}}(y_i^1), \dots, F_{\theta_{d,k}}(y_i^d))) \end{aligned}$$

où $\phi_n = (A_n, (\theta_{i,k}^{(n)})_{i,k}, (\eta_k^{(n)})_k)$ est l'estimation courante du paramètre ϕ . La vraisemblance complète projetée a la même forme que l'équation (4.47) et pose par conséquent les mêmes difficultés au moment de la phase de maximisation.

De manière similaire au cas de l'estimation d'une seule copule, nous pouvons contourner cette difficulté en utilisant la méthode IFM et l'estimation par projection que nous avons développé dans la section 3.5. Nous introduisons la fonction estimante sur données complètes g_{IFM}

$$g_{\text{IFM}} : ([1..K] \times \mathcal{Y})^N \times \mathcal{S}^K \times \prod_{i,j=1}^{K,d} \Theta_{i,j} \times \prod_{k=1}^K E_k \longrightarrow \mathbb{R}^{K(K-1) + \sum_{i,j=1}^{K,d} |\Theta_{i,j}| + \sum_{k=1}^K |E_k|}$$

Nous notons L_c la log-vraisemblance complète du modèle et $L_{c,j} = \sum_{i,k=1}^{N,K} 1_k(x_i) \log(f_{\theta_{j,k}}(y_i^j))$ la log-vraisemblance complète du processus marginal $(\mathbf{X}, \mathbf{Y}^j) = (X_n, Y_n^j)_{1 \leq n \leq N}$. La fonction estimante IFM est

$$g_{\text{IFM}}((\mathbf{x}, \mathbf{y}), \phi) = \left(\nabla_A L'_c \left(\nabla_{\theta_{k,j}} L'_{c,j} \right)_{1 \leq k \leq K} \left(\nabla_{\eta_k} L'_c \right)_{1 \leq k \leq K} \right)' \quad (4.54)$$

Notre objectif est maintenant de déterminer les racines de la fonction estimante projetée

$$G_{\text{IFM}}(\mathbf{y}, \phi) = E_\phi[g_{\text{IFM}}((\mathbf{x}, \mathbf{y}), \phi) | \mathbf{y}]$$

Pour cela, nous utilisons l'algorithme ECI correspondant (voir section 3.5.2). Partant d'une valeur initiale ϕ_0 , cet algorithme, que nous appellerons IFM-ECI, se met sous la forme suivante

$$\forall j \in [1..d], \forall k \in [1..K], \begin{cases} \theta_{j,k}^{(n+1)} \text{ tel que } \sum_i^N \pi_{i,k}^{(n)} \nabla_{\theta_k} \log f_{\theta_{j,k}}(y_i^j) = 0 \\ \text{et} \\ \eta_k^{(n+1)} \text{ tel que } \sum_{i=1}^N \pi_{i,k}^{(n)} \nabla_{\eta_k} \log c_{\eta_k}(F_{\theta_{1,k}^{(n+1)}}(y_i^1), \dots, F_{\theta_{d,k}^{(n+1)}}(y_i^d)) = 0 \end{cases} \quad (4.55)$$

et peut s'interpréter comme une maximisation en deux temps de la fonctionnelle $Q(\phi, \phi_n)$.

Remarque 4.3.4. Algorithme GEM et ECI

Malgré la similitude entre les fonctions IFM et le score, la convergence de l'algorithme IFM-ECI ne peut se voir comme un simple corollaire de la convergence des algorithmes GEM. En pratique, les itérations (4.55) aboutissent souvent à une croissance de la log-vraisemblance observée, mais la décomposition de la variation $Q(\phi_{n+1}, \phi_n) - Q(\phi_n, \phi_n)$ ne permet pas de conclure en général quant à la croissance de la vraisemblance. La variation est la somme de 3 termes, dont un n'est que partiellement contrôlé. Pour alléger les formules, nous noterons pour tout paramètre $\forall i, k, c_{\eta_k}(F_{\theta_{1,k}}(y_i^1), \dots, F_{\theta_{d,k}}(y_i^d)) = c_{\eta_k}(F_{\theta_{j,k}})$.

$$\begin{aligned} Q(\phi_{n+1}, \phi_n) - Q(\phi_n, \phi_n) &= E[\log(p(\mathbf{X}, A_{n+1})) | \mathbf{y}, \phi_n] - E[\log(p(\mathbf{X}, A_n)) | \mathbf{y}, \phi_n] \\ &\quad + \sum_{i,j,k=1}^{N,d,K} \pi_{i,k}^{(n)} \log(f(y_i^j, \theta_{j,k}^{(n+1)})) - \sum_{i,j,k=1}^{N,d,K} \pi_{i,k}^{(n)} \log(f(y_i^j, \theta_{j,k}^{(n)})) \\ &\quad + \sum_{i,k=1}^N \pi_{i,k}^{(n)} \log(c_{\eta_k^{(n+1)}}(F_{\theta_{j,k}^{(n+1)}})) - \sum_{i=1}^N \pi_{i,k}^{(n)} \log(c_{\eta_k^{(n)}}(F_{\theta_{j,k}^{(n)}})) \end{aligned}$$

Le troisième terme de la somme (noté $\Delta L_C(\eta)$) se décompose en :

$$\begin{aligned} \Delta L_C(\eta) &= \sum_{i,k=1}^N \pi_{i,k}^{(n)} \log(c_{\eta_k^{(n+1)}}(F_{\theta_{j,k}^{(n+1)}})) - \sum_{i=1}^N \pi_{i,k}^{(n)} \log(c_{\eta_k^{(n)}}(F_{\theta_{j,k}^{(n)}})) \\ &= \left\{ \sum_{i=1}^N \pi_{i,k}^{(n)} \log(c_{\eta_k^{(n+1)}}(F_{\theta_{j,k}^{(n+1)}})) - \sum_{i=1}^N \pi_{i,k}^{(n)} \log(c_{\eta_k^{(n)}}(F_{\theta_{j,k}^{(n+1)}})) \right\} \\ &\quad + \left\{ \sum_{i=1}^N \pi_{i,k}^{(n)} \log(c_{\eta_k^{(n)}}(F_{\theta_{j,k}^{(n+1)}})) - \sum_{i=1}^N \pi_{i,k}^{(n)} \log(c_{\eta_k^{(n)}}(F_{\theta_{j,k}^{(n)}})) \right\} \end{aligned}$$

Le premier terme est positif (parce que nous maximisons la vraisemblance de la copule une fois que les paramètres des marges ont été mis à jour). Par contre, nous ne contrôlons pas le sens des variations du second terme. Lorsque nous constatons la croissance de la log-vraisemblance observée, cela signifie que les éventuelles discordances entre les marges estimées à l'instant $n+1$ par rapport à la copule estimée à l'instant n sont compensées par l'accroissement de vraisemblance dû à la mise à jour des paramètres par IFM.

4.3.4 Expérimentations

Les expérimentations que nous menons dans cette partie nous permettent d'une part de mettre en évidence la richesse et la flexibilité de la modélisation par copules, et d'autre part d'illustrer l'algorithme ECI et sa capacité à produire des algorithmes d'estimation simples et originaux. Nous comparons tout d'abord les performances des algorithmes IFM-ECI et EM dans le cas de mélanges elliptiques, et nous introduisons de nouveaux modèles multivariés intéressants pour la segmentation d'images et de signaux, et plus généralement pour la classification de données multidimensionnelles.

4.3.4.1 Comparaison des algorithmes IFM et EM

L'algorithme IFM-ECI (ou sa version stochastique) pour l'estimation des mélanges de copules peut être utilisé pour estimer n'importe quel mélange de lois multivariées. En effet, en vertu du théorème de Sklar, toute loi d'émission peut être décomposée en un modèle copule et marginale, pour lequel nous pouvons utiliser IFM-ECI en reparamétrisant le modèle. Ceci peut être en particulier appliqué aux modèles SIRV pour lesquels cette reparamétrisation est obtenue facilement (nous n'examinons que le cas de mélange de lois normales et de loi de Student pour lesquelles les copules ont des expressions simples et facilement calculables). Nous proposons donc une nouvelle famille d'estimateur de mélanges multivariés dont les algorithmes d'estimation ne sont pas des algorithmes GEM (voir remarque 4.3.4), puisqu'il n'accroisse pas nécessairement la vraisemblance.

Nous reprenons le mélange $\mathcal{N}$ de lois normales étudié dans la section 4.2.5.1 pour un échantillon de taille $N = 500$. Nous estimons les paramètres par EM, ou par IFM-ECI. Nous faisons 500 itérations de l'algorithme EM, et 300 dans le cas de l'algorithme IFM-ECI. Sur un total de 500 simulations, nous avons obtenu les paramètres moyens, que nous avons rassemblé dans les tableaux 4.6, 4.7 et 4.8. Nous pouvons constater que les algorithmes EM et IFM-ECI fournissent en moyenne des estimations identiques des paramètres, à horizon finie. L'algorithme IFM-ECI donne une bonne estimation des paramètres des marges pour les 3 classes, alors qu'il a tendance à sous-estimer les coefficients de corrélation (de manière un peu plus marqué que l'algorithme EM). Ces derniers ont tendance à être mieux estimés par l'algorithme EM.

	EM	IFM-ECI
Classe 1	(-0,01 -0,01)	(-0,00 -0,00)
Classe 2	(1,5 1,5)	(1,47 1,47)
Classe 3	(3,01 3,05)	(3,00 3,00)

TAB. 4.6 – Moyennes estimées

	EM	IFM-ECI
Classe 1	(0,96 0,97)	(0,98 0,96)
Classe 2	(0,96 0,98)	(0,96 0,98)
Classe 3	(0,96 0,96)	(0,99 0,99)

TAB. 4.7 – Variances estimées

	EM	IFM-ECI
Classe 1	0,39	0,38
Classe 2	0,19	0,18
Classe 3	0,5	0,49

TAB. 4.8 – Coefficients de corrélation estimés

L'algorithme EM donne des estimations équivalentes à IFM-ECI pour le modèle $\mathcal{N}$, et il semble préférable dans cette situation d'utiliser EM, en raison de sa rapidité. Les écart-types des

estimateurs sont cependant similaires, et nous obtenons des taux d'erreur (en segmentation non-supervisée) semblables : l'algorithme EM donne une taux erreur légèrement meilleur de 14,2% contre 14,4% pour l'IFM-ECI. Ceci nous permet de constater qu'un meilleur estimateur des paramètres ne donnent pas nécessairement des taux d'erreur non-supervisés significativement supérieurs à ceux d'un estimateur théoriquement moins bon.

Nous menons la même comparaison pour le mélange de loi de Student T_1 , pour lequel tous les paramètres de queue sont égaux à 10. Nous utilisons alors une modification de l'algorithme IFM-ECI, dans lequel le nombre de degrés de libertés des copules n'est pas estimé et est fixé à 20 pour toutes les classes. Seuls les paramètres de queue des marginales sont donc réestimés : chaque loi d'émission à des densités marginales qui sont des lois T contraintes à avoir le même paramètre de queue. Le mélange estimé par IFM-ECI ne correspond donc pas exactement au modèle T_1 , mais malgré cela, nous obtenons des résultats d'estimation meilleurs pour IFM-ECI que pour EM pour l'estimation de ν_1, ν_2, ν_3 . Ces derniers sont surestimées par IFM-ECI, mais beaucoup moins que par EM (cf. tableau (4.12)). Par contre, l'estimation des termes diagonaux de Σ_2 est moins bonne par IFM-ECI (tableau (4.10)).

	EM	IFM-ECI
Classe 1	(0 -0,01)	(0 0)
Classe 2	(1,49 1,51)	(1,5 1,52)
Classe 3	3 2,99)	(3 2,99)

TAB. 4.9 – Moyennes estimées

	EM	IFM-ECI
Classe 1	(1 0,96)	(0,99 0,96)
Classe 2	(0,97 0,99)	(0,92 0,94)
Classe 3	(1 1,02)	(0,97 0,99)

TAB. 4.10 – Entrées diagonales de la matrice Σ (du modèle $T_{m,\Sigma,\nu}$)

	EM	IFM-ECI
Classe 1	0,38	0,38
Classe 2	0,17	0,12
Classe 3	0,5	0,5

TAB. 4.11 – Coefficients de corrélation estimées

Les écart-types des estimateurs sont similaires pour EM et IFM-ECI, sauf pour les paramètres de queue. Pour ceux-ci, l'écart-type est entre 20 et 30 pour l'algorithme EM (pour les 3 classes), et entre 7 et 10 pour IFM-ECI (pour les classes). Enfin, nous obtenons des taux d'erreur en segmentation non-supervisée égaux à 17,1% dans les deux cas pour les deux estimateurs.

Remarque 4.3.5. *Lorsque nous utilisons le vrai nombre de degrés de libertés pour la copule dans la procédure ECI-IFM, les résultats d'estimation sont équivalents pour les paramètres de moyenne*

	EM	IFM-ECI
Classe 1	18,4	13,85
Classe 2	23,9	13,6
Classe 3	18,3	13,6

TAB. 4.12 – Paramètres de queue estimés, ν

et de variance, mais sont moins bons pour les paramètres de queue : nous obtenons en moyenne $\hat{\nu}_1 = 15,8$, $\hat{\nu}_2 = 17,2$, $\hat{\nu}_3 = 14,3$. Cependant, le taux d'erreur reste sensiblement le même à 16,9%.

Dans cette étude, nous avons montré que les algorithmes EM et IFM-ECI avait des comportements similaires pour la segmentation non-supervisée (en termes de taux d'erreur), et pour l'estimation lorsque les modèles estimés sont les mêmes. De manière générale, les variances des estimateurs sont similaires hormis pour les paramètres de queue, pour lequel il y a un réel avantage à utiliser IFM-ECI.

Cependant, il est difficile de séparer le comportement théorique des estimateurs EMV et “fonction IFM projetée” de celui des algorithmes itératifs que nous avons utilisé pour les calculer. Le fait que nous constatons que les estimateurs soient proches en moyenne des vrais paramètres permet d'affirmer que les deux algorithmes ont convergé vers une bonne racine de l'équation. Ceci permet de valider empiriquement la méthode IFM-ECI, dont nous avons vu dans la section 3.5.2 qu'elle était fondée sur des propriétés locales de la fonction estimante difficile à vérifier (possibilité d'itérer la fonction $\mathcal{T}$, et propriété de contractivité sur un voisinage compact). Cet aspect est directement relié au problème de l'initialisation de l'algorithme itératif. Nous avons utilisé pour IFM-ECI le paramètre ϕ_0 calculés à partir d'une segmentation des observations basées sur une CMCa-BI avec des lois d'émission de type Student. Nous obtenons donc une première segmentation assez proche de la vraie de la segmentation, ce qui nous permet de démarrer l'algorithme dans un voisinage proche du vrai paramètre (d'où la convergence systématique d'IFM-ECI vers une bonne racine).

4.3.4.2 Lois elliptiques généralisées

Nous présentons dans cette section les lois elliptiques généralisées, initialement introduites par Fang, Fang et Kotz ([67]), et qui sont une généralisation des modèles SIRV. En effet, pour les SIRV, la connaissance de la loi marginale est équivalente à la connaissance de la loi du vecteur (ce lien a été explicité par Rangaswamy dans [135]) : soit $Y = (Y_1, \dots, Y_d)$ un SIRV centré et $t \mapsto \psi(t)$ la fonction caractéristique de la loi de Y_k (et dépend de la fonction h définie dans (4.7)), alors la densité de la loi jointe est :

$$f(y, (0, \Sigma, h)) = \frac{|\Sigma|^{-1/2}}{(2\pi)^{d/2}} q(y)^{\frac{2-d}{4}} \int_0^\infty \psi(t) t^{\frac{d}{2}} J_{\frac{d-2}{2}}(t\sqrt{q(y)}) dt \quad (4.56)$$

où J_α est la fonction de Bessel d'indice α [5]. Nous rappelons que $q(y) = \frac{1}{2}y' \Sigma^{-1}y$.

Cette propriété (4.56) apparaît plutôt comme un inconvénient pour la modélisation multivariée (par exemple pour les images multicateurs, ou pour la classification de données de manière générale) car elle implique que la forme d'une marginale “commande” la manière dont sont liées les différentes composantes d'un vecteur. Autrement dit dans un modèle SIRV, le choix d'une marginale implique celui de la forme de la copule (et de toutes les autres marges). Il est possible

d'alléger cette contrainte en choisissant la copule indépendamment des marges, ce qui amène, en particulier, à la définition des lois T et K généralisées (ou asymétriques), en utilisant des marges T ou K.

Définition 4.3.3. Loi T généralisée

Un vecteur aléatoire $\mathbf{Y}$ de $\mathbb{R}^d$ suit une loi de Student généralisée si $\forall i \in [1..d]$, $Y_i \sim \mathbf{T}_{m_i, \sigma_{ii}^2, \nu_i}$ et que sa copule est une copule de Student de paramètre ρ et de degrés ν_c .

Si nous avons $\forall i, \nu_i = \nu_c$, $\mathbf{Y}$ suit une loi de Student $\mathbf{T}_{\mathbf{m}, \Sigma, \nu_c}$, avec $\mathbf{m} = (m_1, \dots, m_d)'$, et $\Sigma = (\rho_{ij}\sigma_i\sigma_j)_{i,j}$. La famille des lois normales est aussi incluse dans les lois de Student généralisée, en prenant $\nu_i = \nu_c = \infty$. La même généralisation peut être proposée pour la loi K.

Définition 4.3.4. Loi K généralisée

Un vecteur aléatoire $\mathbf{Y}$ de $\mathbb{R}^d$ suit une loi K généralisée si $\forall i \in [1..d]$, $Y_i \sim \mathbf{K}_{m_i, \sigma_{ii}^2, a_i}$ et que sa copule est une copule K de paramètre ρ et de degrés a_c .

Bien sûr, de manière très générale, nous pouvons utiliser n'importe quelle type de copule. Ces modèles ont été proposé par Fang *et al.* dans une étude générale sur l'expression des copules elliptiques. L'intérêt de cette généralisation est que les lois ainsi construites restent proches des modèles à contour elliptiques (les contours obtenus sont des ellipses déformées par les paramètres ν_i), et peuvent fournir un moyen original et relativement simple de tester l'ellipticité de la loi (qui revient à tester l'égalité des paramètres de queues des marginales et de la copule). C'est ce modèle que nous avons utilisé dans la section précédente pour la comparaison d'IFM-ECI et EM dans les mélanges de Student.

Cependant, Fang *et al.* ne propose pas ces modèles elliptiques généralisés dans le cadre de mélange. Cette étude constitue donc la première application de ces modèles à la segmentation. De plus, nous proposons une méthode d'estimation de ce mélange, en utilisant les fonctions estimantes IFM-ECI.

Nous comparons ci-dessous les performances des méthodes d'estimation par EM et par IFM-ECI dans le cadre de la segmentation non-supervisée d'une CMCA-BI. Dans le cadre de l'estimation basée sur la décomposition copule-marginales, nous ne faisons pas d'estimation sous contrainte d'égalité des paramètres de queues des marginales avec celui de la copule (comme dans la section 4.3.4.1). Le paramètre de queue de la copule reste fixe (supposé connu) pour l'estimation IFM-ECI. Ceci évite une phase de maximisation supplémentaire pour la détermination de ν_c (et pour laquelle il n'existe pas de solution analytique).

Modèle Utilisé	N	T	K
T	13,4	15,7	11,7
K	13,5	15,7	11,7
CopN-T	13,5	16,7	13,1
CopT-T	13,6	16,1	13,3

TAB. 4.13 – Taux d'erreur en mode non-supervisée pour des mélanges de SIRV et SIRV généralisées

Nous avons des différences plus nettes entre EM et IFM-ECI dans cette situation que dans le paragraphe précédent. Nous pouvons observer que les performances sont dégradées en segmentation

non-supervisée en utilisant des SIRV généralisées, par rapport au modèle SIRV classique. D'une part, ceci provient de l'absence de contraintes d'égalité sur les paramètres de queues, ce qui fait que nous avons plus de paramètres à estimer, et augmente la variance des estimateurs et les taux d'erreur. D'autre part, nous avons échantillon de plus grande taille ($N = 1000$), ce qui facilite la recherche des maxima de la log-vraisemblance de celle-ci car la fonction est plus lisse.

4.3.4.3 Lois gamma multivariées et images multicapteurs

Les lois gamma sont considérées comme un modèle acceptable pour modéliser les intensités des images ou des signaux. Cependant, dans le cas des signaux multicapteurs, l'utilisation de modèles multivariés qui puissent incorporer cette information est souvent compromise par la complexité des lois en jeu. En effet, les modèles gamma multivariés les plus souvent utilisés sont définis à partir de leur fonction caractéristique, pour lesquelles il n'est pas possible d'avoir une expression analytique de la densité (voir par exemple [39] pour une application des lois gamma multivariées et une illustration des problèmes d'estimation qu'elles engendrent). Ceci rend ces modèles difficilement envisageables dans des applications de segmentation (supervisée ou non-supervisée) pour lesquelles la densité est évaluée en permanence. Ainsi dans les applications de segmentation de signaux et d'images, l'incorporation de telles contraintes est difficilement envisageable, et la loi normale est alors le choix par défaut.

Dans l'approche que nous avons développée, nous pouvons facilement proposer des lois multivariées à marges gamma (spécifiée par copule) qui se prêtent aux calculs nécessaires de la segmentation bayésienne. Nous donnons une application originale (dans le contexte de la segmentation de données multicapteurs) des copules à la segmentation d'une image multispectrale dans [26]. Nous considérons une image CASI (*Compact Airborne Spectrographic Imager*⁴), dont seulement 4 bandes spectrales ont été sélectionnées pour la segmentation après réduction de données. Nous observons alors dans chacune des bandes l'intensité reçue. Nous proposons par conséquent pour les lois d'émission une description copule-marginales, en forçant les marges à appartenir à la famille gamma. Nous comparons alors les résultats de segmentation obtenues en utilisant différentes modélisations des lois multivariées :

1. les lois elliptiques, qui correspondent à une recherche de classes "sans a priori", uniquement basée sur les notions de moyenne et de dispersion (symétrique autour de la moyenne).
2. la méthode de décorrélation basée sur une analyse en composante indépendantes (ACI) et le modèle de Pearson : nous supposons qu'il existe une transformation linéaire des données par laquelle les composantes deviennent indépendantes et dont les lois appartiennent au système de Pearson $\mathcal{F}$. Cette méthode permet de généraliser aux données multidimensionnelles la recherche de la loi d'émission dans une grande famille de lois paramétriques.
3. les composantes sont supposées indépendantes, mais appartiennent toutes au modèle gamma. Ceci correspond au cas d'un modèle dont les marginales sont bonnes, mais avec une structure de dépendance incorrecte.
4. les marginales sont de type gamma et la dépendance est modélisée par une copule gaussienne.

⁴Le CASI est un spectroradiomètre imageur permettant d'enregistrer des images sur un support magnétique dans un large spectre s'étalant du visible (environ 400nm) au proche infrarouge (environ 900nm). Il utilise une matrice composée de détecteurs CCD disposés perpendiculairement au sens de déplacement de l'avion. Le CASI fournit des enregistrements en mode spatial et spectral.

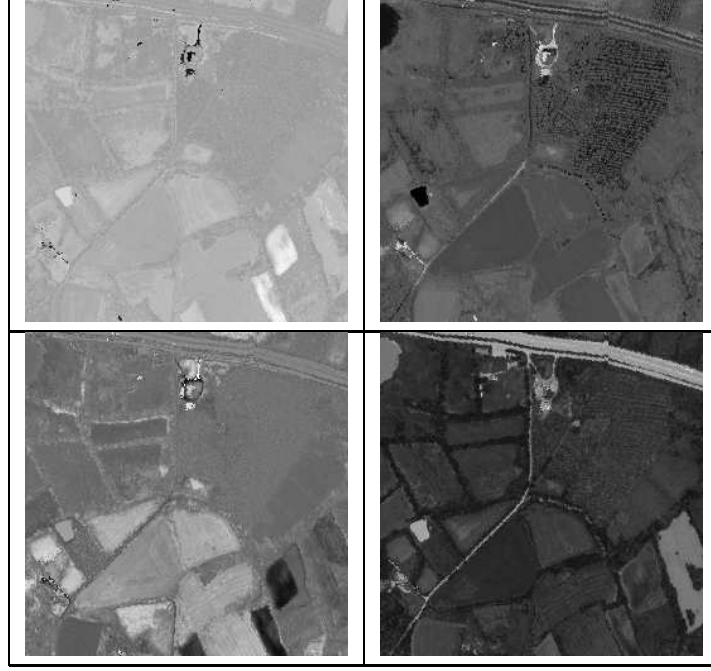


FIG. 4.5 – Image Casi - 4 bandes spectrales

Les résultats obtenus permettent de mettre en avant l'importance d'une spécification correcte des lois marginales, et de la prise en compte de la dépendance entre les canaux. Il apparaît que le modèle copule gaussienne et marges gamma possède les qualités des modèles mélange de "gamma indépendant" et mélange gaussien pour fournir une meilleure segmentation.

De même, nous avons réalisé la segmentation d'une image satellite bidimensionnelle (deux capteurs) et nous comparons les résultats obtenus avec plusieurs modèles classiques de lois bivariées. Les deux images ont été obtenus par le satellite JERS1⁵, avec une résolution pour chaque pixel de 25×25 mètres carrés. Elle représente une partie de la Rondonie, qui est une région de l'Amazonie dans laquelle les cultures sont faites sur brûlis : les arbres sont coupés et brûlés et la terre est transformée en pâturage ou en terre cultivée. Nous avons finalement 4 classes : brûlis, pâturages, forêts et culture.

Nous essayons alors de retrouver automatiquement ces 4 classes à partir des images (4.7) en utilisant des CMCA-BI ayant des lois d'émission multivariées de type "gamma indépendante", ACI + système de Pearson, loi K multivariée et gamma bivariée (avec une copule de Student à 5 degrés de liberté).

Remarque 4.3.6. *Les lois elliptiques (loi normale, T ou K) donnent des résultats en segmentation tout à fait similaires. Nous ne présentons que les résultats obtenus pour la loi K , plus adaptée à la modélisation des images (cf. remarque 4.2.5).*

⁵Il s'agit d'un satellite japonais d'observation de la terre dédié particulièrement à l'étude du sol, de l'agriculture, des pêcheries et des forêts. Il est équipé d'une radar à synthèse d'ouverture à émission en bande L (de 1 à 2 GHz, soit une longueur d'onde entre 15 et 30 cm).

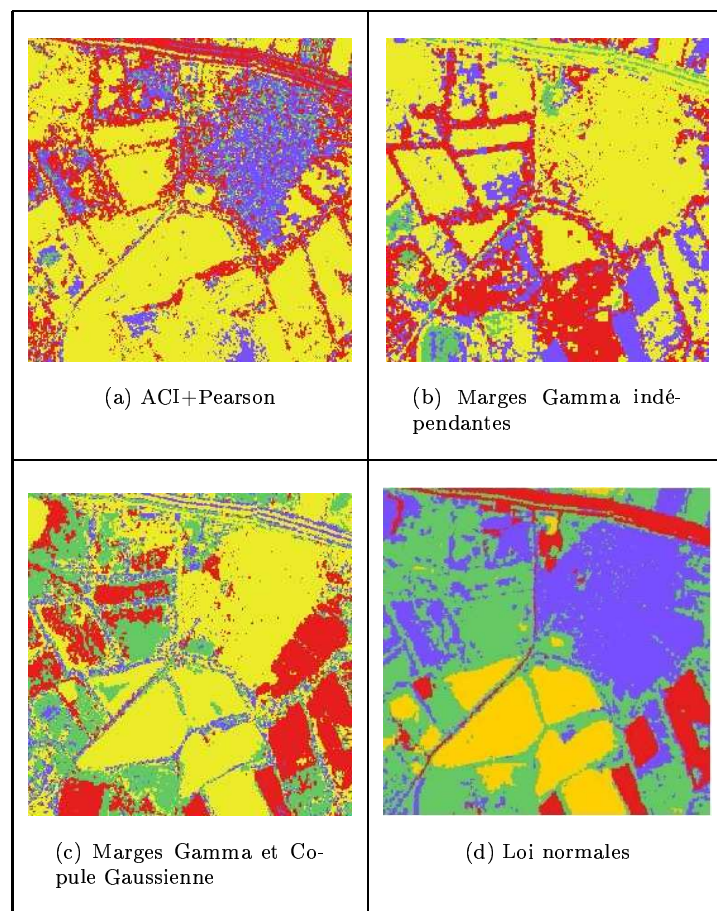


FIG. 4.6 – Images segmentées

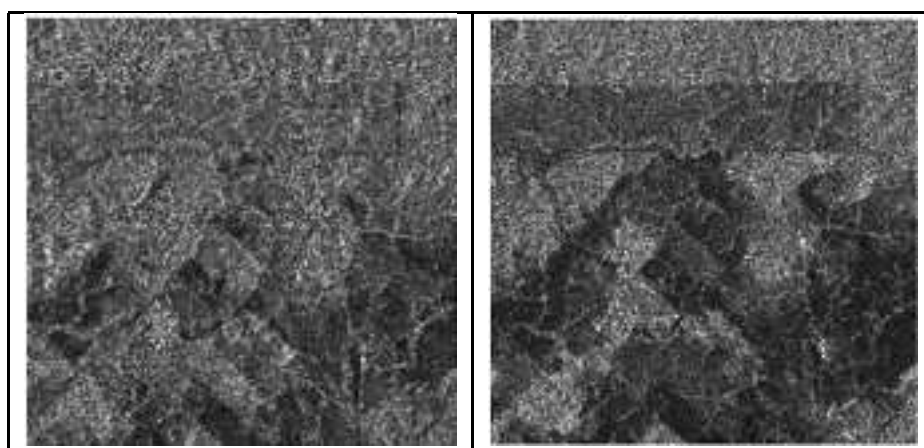


FIG. 4.7 – Images observées de Rondonie.



FIG. 4.8 – Vérité terrain de Rondonie. Blanc : forêt, Gris clair : pâturage récent, Gris foncé : brûlis, Noir : cultures

Les segmentations obtenues ne donnent pas de cartes avec des zones bien délimitées, en raison de l'interpénétration des zones de cultures, forêts et brûlis. Il apparaît clairement que le modèle gamma bivarié (avec copule) permet de séparer assez nettement la forêt des zones modifiées par l'homme, contrairement aux autres lois d'émission. Il y a cependant une confusion importante entre pâturage récent (celle-ci est cependant bien retrouvée par le modèle gamma avec copule), culture et brûlis (pour ce dernier, cela est du en partie à son faible effectif relatif). La segmentation obtenue avec le modèle "marginales gamma indépendantes" indique que les bonnes performances pour la séparation de la forêt des autres zones sont dues en partie à la modélisation par des marges gamma. L'utilisation de la corrélation des canaux permet d'obtenir des régions plus homogènes. Les modèles avec des lois d'émission de type K ou des lois obtenues par ACI et système de Pearson ne permettent pas de séparer les différentes zones, malgré la prise en compte de la corrélation.

En conclusion, sur ces deux images réelles multivariées, il est apparaît que l'utilisation de la loi gamma multivariée permet d'obtenir des images segmentées de meilleur qualité que celles obtenues par des modèles multivariées plus classiques. Cela nous permet de penser qu'une prise en compte de contraintes simples sur les marginales (densités sur $\mathbb{R}^+$ et distribution asymétrique) peut améliorer la modélisation statistique des données réelles manipulées en traitement d'images.

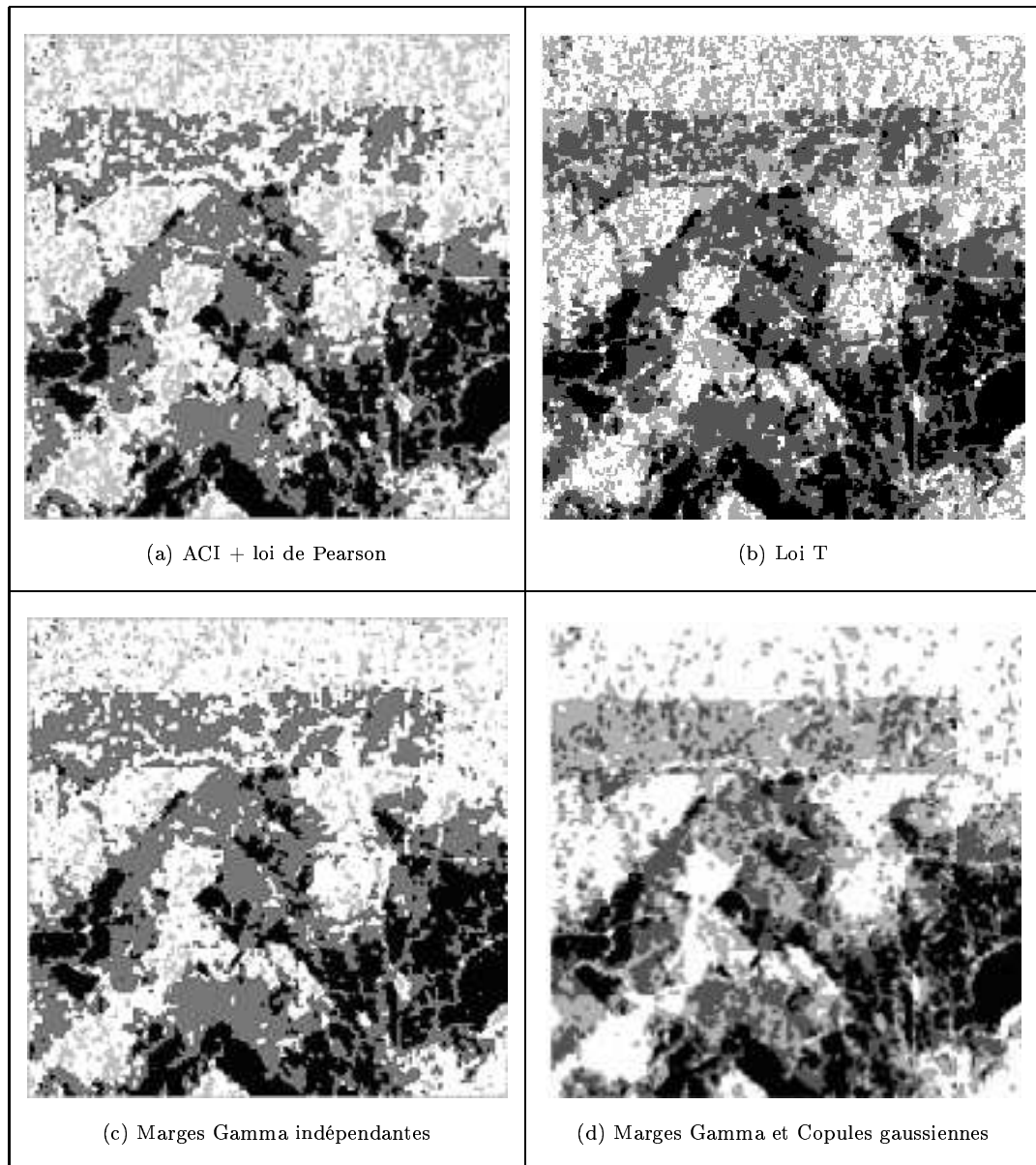


FIG. 4.9 – Segmentation pour différentes lois d'émission

Chapitre 5

Chaînes de Markov Couples et Copules

Ce chapitre reprend et développe l'article [28]. Nous étudions l'utilisation des chaînes de Markov couples pour la segmentation. Pour cela, nous procédons à une reformulation originale du problème de la classification dans laquelle les copules se révèlent pertinentes pour la description d'un cadre général de la classification des données issus d'un processus stationnaire. Le nouveau modèle que nous proposons alors (et étudions particulièrement) est une chaîne de Markov cachée et nous discutons des différences entre les modèles que nous proposons et les modèles CMCa existants dans la littérature. C'est aussi l'occasion de préciser l'apport des CMCa par rapport aux CMCa-BI du point de vue de la description de la structure de dépendance des données observées. Pour cela, nous donnons la structure de la fonction d'autocovariance des CMCa, et nous proposons une analyse de ces modèles basée sur les coefficients de mélangeance (utilisés classiquement pour l'étude des processus stationnaires). Nous montrons ensuite qu'il est possible d'estimer ces nouveaux modèles par la méthode de projection de fonctions estimantes développées dans la section 3.5. De même, nous montrons qu'il est possible de calculer une approximation de l'estimateur par un algorithme ECI. Nous obtenons alors des procédures d'estimation simples, évitant par exemple les optimisations complexes impliquées par une maximisation de la log-vraisemblance. Finalement, nous montrons que la méthode d'estimation développée pour les CMCa s'étend directement aux CMCo générales.

5.1 Segmentation des processus stationnaires

5.1.1 Processus stationnaires et copules

Soit $\mathbf{Y}$ un processus stationnaire et $\mathbf{y} = (y_1, \dots, y_N)$ une réalisation que nous voulons segmenter. De nombreuses méthodes existent pour la classification de données et peuvent être, pour une grande part, réunies dans la discipline de Reconnaissance de Formes (*Pattern Recognition*) qui recouvrent entre autres l'analyse discriminante linéaire, les algorithmes de "K plus proches voisins", les arbres de classification, les réseaux de neurones (voir [56] pour un panorama des techniques de reconnaissance de formes) ou encore les machines à support de vecteur (*Support Vector*

Machine [149]). Ces méthodes supposent le plus souvent que les observations $(y_1, \dots, y_N)$ constituent un échantillon i.i.d, et elles cherchent alors à déterminer les caractéristiques des observations qui soient les plus discriminantes. Elles se prêtent généralement assez mal à la classification d'observations dépendantes. La classification basée sur des modèles probabilistes génératifs (parfois désignée par *Maximum Likelihood Classification*) adopte le paradigme bayésien en considérant que les observations sont la réalisation d'un mélange de K lois, telle que

$$\forall y_n \in \mathcal{Y}, p(y_n) = \sum_{k=1}^K \pi_k p(y_n | k) \quad (5.1)$$

La grande flexibilité de l'approche bayésienne (via la modélisation hiérarchique) permet alors d'intégrer des connaissances complexes sur les liens entre les observations, typiquement par le biais du processus latent $\mathbf{X}$ associé au mélange de lois ($\forall k \leq K, \pi_k = P(X_n = k)$). Classifier les observations revient à estimer le processus $\mathbf{X}$ qui a le plus souvent une interprétation physique, de telle sorte que nous ayons deux connaissances a priori :

- nous connaissons partiellement le lien entre $\mathbf{X}$ et $\mathbf{Y}$ (i.e. la densité $p(\mathbf{y} | \mathbf{x})$) via la connaissance de la densité $p(y_n | x_n)$, que nous savons être égale à $f_{x_n}(y_n) \cdot \{f_k(y), k \in [1..K]\}$ est un ensemble de lois a priori que nous devons le plus souvent estimer, et qui synthétise une connaissance a priori des phénomènes physiques en jeu ;
- nous connaissons la structure (de dépendance) du processus $\mathbf{X}$.

De manière générale, si nous avons invariance des phénomènes observés par translation des indices n , nous pouvons dire que le processus $\mathbf{Y}$ est un mélange stationnaire. L'objectif de cette section est de faire le lien entre la loi (connue) de Y_n et la loi jointe de $\mathbf{Y}_N = (Y_1, \dots, Y_N)$ lorsque cette dernière est obtenue après spécification de la loi conditionnelle $p(\mathbf{y} | \mathbf{x})$. Si nous notons $\mathbf{x}_{-n} = (x_1, \dots, x_{n-1}, x_{n+1}, \dots, x_N) \in \mathcal{X}^{N-1}$, nous avons

$$p(y_n | x_n) = \sum_{\mathbf{x}_{-n}} p(y_n | \mathbf{x}) p(\mathbf{x}_{-n} | x_n) \quad (5.2)$$

Il est donc nécessaire que les modèles proposés pour la densité conditionnelle $p(\mathbf{y} | \mathbf{x})$ aient des marges qui soient compatibles avec les lois d'émission f_k selon l'eq. (5.2). Les CMCa-BI sont des modèles spécifiant les interactions entre les observations (la densité $p(\mathbf{y} | \mathbf{x})$), et qui sont en accord avec les densités $p(y_n | x_n)$ (et $p(y_n)$) que l'on se donne a priori. Mais en toute généralité, la compatibilité entre $p(\mathbf{y} | \mathbf{x})$ et $p(y_n | x_n)$ est difficile à réaliser. Par conséquent, nous faisons l'hypothèse que le lien entre $\mathbf{X}_N$ et $\mathbf{Y}_N$ est tel que

$$\forall n \leq N, p(y_n | \mathbf{x}) = p(y_n | x_n) \quad (5.3)$$

Dans ce cas, l'eq. (5.2) est systématiquement réalisée, et la densité conditionnelle $p(\mathbf{y} | \mathbf{x})$ a pour marges les densités $\{p(y_n | x_n)\}_{1 \leq n \leq N}$. Lorsque $\mathcal{Y} = \mathbb{R}$, les copules (voir section 4.3) nous permettent de décrire l'ensemble des densités $p(\mathbf{y} | \mathbf{x})$ vérifiant cette condition : si $F(\mathbf{y} | \mathbf{x})$ et $F(y_n | x_n)$ désignent (resp.) les fonctions de répartition de $p(\mathbf{y} | \mathbf{x})$ et $p(y_n | x_n)$, nous avons

$$F(\mathbf{y} | \mathbf{x}) = C_{\mathbf{x}}(F(y_1 | x_1), \dots, F(y_N | x_N)) \quad (5.4)$$

$C_{\mathbf{x}}$ est une copule, dépendante du processus caché, qui est unique si toutes les f.d.r $F(y_n | x_n)$ sont continues. De plus, si nous notons $\bar{F}(y) = \sum_{k=1}^K \pi_k F(y | x_k)$ et $\bar{C}$ la copule de la loi jointe de $\mathbf{Y}_N$, nous avons alors l'égalité

$$\bar{C}(\bar{F}(y_1), \dots, \bar{F}(y_N)) = \sum_{\mathbf{x}} p(\mathbf{x}) C_{\mathbf{x}}(F(y_1 | x_1), \dots, F(y_N | x_N))$$

Si les f.d.r de toutes les classes sont continues, l'identification de la copule $\bar{C}$ est possible grâce au théorème de Sklar :

$$\forall u_1, \dots, u_N, \bar{C}(u_1, \dots, u_N) = \sum_{\mathbf{x}} p(\mathbf{x}) C_{\mathbf{x}}\left(F\left(\bar{F}^{-1}(u_1) | x_1\right), \dots, F\left(\bar{F}^{-1}(u_N) | x_N\right)\right)$$

La copule $\bar{C}$ représente toute la structure de dépendance du processus $\mathbf{Y}_N$. Il apparaît donc que dans un mélange stationnaire, la dépendance modélisée entre les observations Y_n dépend non seulement de la dépendance conditionnelle introduite par $p(\mathbf{y} | \mathbf{x})$ (et de $p(\mathbf{x})$) mais aussi des lois marginales de Y_n choisies.

En plus de vérifier la propriété (5.3), nous supposons que le mélange stationnaire $\mathbf{Y}_N = (Y_n)_{n \geq 1}$ est telle que $\mathbf{Z}_N = (\mathbf{X}_N, \mathbf{Y}_N)$ soit aussi une chaîne de Markov. Ceci implique alors que $\mathbf{Y}_N$ conditionnellement à $\mathbf{X}_N$ soit une chaîne de Markov, i.e. la densité de la copule $C_{\mathbf{x}}$ (définie dans l'équation (5.4)) peut se factoriser

$$c_{\mathbf{x}}(u_1, \dots, u_N) = \prod_{i=2}^N c_{x_{i-1}, x_i}(u_{i-1}, u_i) \quad (5.5)$$

avec c_{x_{i-1}, x_i} la copule de la $p(y_{i-1}, y_i | x_{i-1}, x_i)$.

Ceci revient alors à relâcher l'hypothèse d'indépendance conditionnelle dans le modèle CMCa-BI. Une CMCa stationnaire réelle sera complètement décrite par la matrice de transition de la chaîne cachée $\mathbf{X}_N$, ses densités marginales $p(y_1 | x_1)$ et les copules c_{x_1, x_2} . Dans le cas vectoriel, cette description est plus difficile puisque la décomposition de Sklar qui permet la décomposition (5.4) ne se généralise pas aux vecteurs aléatoires : l'ensemble des CMCa stationnaires à valeurs dans $\mathbb{R}^d$ est identifiable à l'ensemble des densités $p(y_1, y_2 | x_1, x_2)$ sur $\mathbb{R}^{2d}$ ayant pour marges $p(y_i | x_i)$, $i = 1, 2$, mais il n'est pas identifiable à l'ensemble de copules. En effet, le théorème d'impossibilité démontré par Genest *et al.* dans [31] montre que la seule copule bivariable $\mathbf{C} : [0, 1] \times [0, 1] \rightarrow \mathbb{R}$ telle que pour toutes f.d.r F_1 et F_2 sur $\mathbb{R}^d$, la fonction $F = \mathbf{C}(F_1, F_2)$ définie sur $\mathbb{R}^{2d}$ par

$$F\left((y_1^i)_{1 \leq i \leq d}, (y_2^j)_{1 \leq j \leq d}\right) = \mathbf{C}\left(F_1\left((y_1^i)_{1 \leq i \leq d}\right), F_2\left((y_2^j)_{1 \leq j \leq d}\right)\right)$$

soit une f.d.r. sur $\mathbb{R}^{2d}$, est la copule d'indépendance $C^\perp(u, v) = uv$. Par conséquent, à matrice de transition et marges $(f_k)_{1 \leq k \leq K}$ fixées, l'ensemble des CMCa correspondant peut-être décrit partiellement selon les marges choisies. Nous en donnerons dans la suite plusieurs exemples.

5.1.2 Le modèle CMCa stationnaire avec copules

Nous décrivons dans cette section le modèle paramétrique que nous avons utilisé pour les CMCa stationnaires réelles. Il est particulièrement maniable pour les applications parce qu'il est

finalment très similaire au modèle CMCa-BI. Nous donnons aussi quelques cas particuliers de CMCa vectorielles, non étudiées par la suite, en mettant en évidence certaines des difficultés qui leur sont inhérentes.

5.1.2.1 Observations réelles

Le modèle paramétrique CMCa que nous utilisons par la suite est décrit par la matrice de transition $A \in \mathcal{S}^K$, dont la loi stationnaire π est supposée unique. Nous supposons que les lois d'émission ont une densité (par rapport à la mesure de Lebesgue) appartenant à un même modèle paramétrique $f(y, \theta_k)$, $\theta_k \in \Theta$ (avec des fonctions de répartition $F(y, \theta_k)$ continues et $\theta = (\theta_k)_{1 \leq k \leq K}$) : il s'agira le plus souvent de lois normale, gamma ou weibull. Les copules bivariées modélisant la dépendance conditionnelle ont une densité $c(u, v; \eta_{kl})$, $\eta_{kl} \in \Upsilon$ (nous notons $\eta = (\eta_{kl})_{1 \leq k, l \leq K}$).

Conditionnellement au processus caché, le processus des observations est une chaîne de markov non-homogène dont le noyau de transition admet la densité suivante (relativement à la mesure de Lebesgue)

$$p(y_n | y_{n-1}, x_n, x_{n-1}) = f(y_n, \theta_{x_n}) c_{\eta_{x_{n-1}, x_n}}(F(y_{n-1}, \theta_{x_{n-1}}), F(y_n, \theta_{x_n})) \quad (5.6)$$

Le noyau de transition du processus complet $\mathbf{Z}$ a pour densité

$$\forall z_1, z_2, p(z_2 | z_1) = a_{x_1 x_2} \times f(y_2, \theta_{x_2}) c_{\eta_{x_1, x_2}}(F(y_1, \theta_{x_1}), F(y_2, \theta_{x_2})) \quad (5.7)$$

La loi du processus $\mathbf{Y}$ est donc indexée par $\phi = (A, \theta, \eta) \in \mathcal{S}^K \times \Theta^K \times \Upsilon^{K^2}$, et a pour densité

$$p(\mathbf{y}, \phi) = \sum_{x_1, \dots, x_N} \pi_{x_1} a_{x_1 x_2} \dots a_{x_{N-1} x_N} \prod_{i=1}^N f(y_i, \theta_{x_i}) \prod_{i=2}^N c_{\eta_{x_{i-1}, x_i}}(F(y_{i-1}, \theta_{x_{i-1}}), F(y_i, \theta_{x_i})) \quad (5.8)$$

Le modèle CMCa-BI décrit par $\{A, f(y, \theta_k), \theta_k \in \Theta\}$ est un sous-modèle du modèle CMCa lorsque les familles de copules $c(u, v; \eta_{kl})$, $\eta_{kl} \in \Upsilon$ contiennent la copule $c^\perp$. C'est le cas de la copule gaussienne pour laquelle nous avons $c_0 = c^\perp$, ou de certaines copules archimédiennes, par exemple la famille de Gumbel-Hougaard d'expression $C_\eta(u, v) = \exp\left(-((- \log u)^\eta + (- \log v)^\eta)^{1/\eta}\right)$, pour laquelle $C_1 = C^\perp$, ou encore la famille de Gumbel $C(u, v; \eta) = uv \exp(-\eta \log u \log v)$ pour laquelle $C_0 = C^\perp$. Nous notons $\eta^\perp$ un tel paramètre lorsqu'il existe¹, et $\eta^\perp = (\eta_{kl}^\perp)$ lorsque cela est possible pour toutes les classes. Si $\eta^\perp \in \Upsilon$ alors le modèle CMCa-BI est emboîté dans le modèle CMCa-BI.

Nous abordons ici l'identifiabilité du modèle CMCa. La loi du processus $\mathbf{Y}$ est identifiable si la loi de (Y_1, Y_2) est identifiable. Par conséquent, il suffit de supposer l'identifiabilité du modèle paramétrique pour les densités $p(y_1, y_2 | x_1, x_2)$, i.e. l'identifiabilité du mélange des lois bivariées de f.d.r $C_{\eta_{kl}}(F(y, \theta_k), F(y, \theta_l))$. Or nous savons que si les mélanges finis des lois d'émission $f(y, \theta)$, $\theta \in \Theta$ et des copules $c_\eta(u, v)$, $\eta \in \Upsilon$ sont identifiables, alors les lois bivariées sont identifiables d'après la partie 4.3.3.1. Nous utilisons dans la suite les copules gaussiennes qui sont faciles à implémenter, et suffisamment riches pour mettre en évidence l'intérêt de l'apport de la dépendance conditionnelle.

¹ $\eta^\perp$ n'existe pas toujours, comme pour les copules elliptiques autre que la copule gaussienne. Dans ce cas-là, le modèle CMCa-BI n'est pas emboîté dans le modèle CMCa.

5.1.2.2 Observations vectorielles

Comme nous n'avons plus le théorème de Sklar qui nous permet de contrôler les marges souhaitées, le cas des CMCa vectorielles ne peut plus être traitée de manière générale. Nous donnons alors quelques exemples de familles paramétriques de CMCa stationnaires vectorielles.

Le cas gaussien Une construction simple de CMCa est celui où les densités $p(y_1|x_1)$ sont gaussiennes. En effet, si nous supposons que le couple (Y_1, Y_2) conditionnellement à (X_1, X_2) est un vecteur gaussien, nous pouvons garantir la contrainte sur les marges par un simple contrôle des moyennes et variances. Nous notons les moyennes et variances des lois d'émission $(\mathbf{m}_k)_{1 \leq k \leq K} \in \mathbb{R}^d$ et $(\Sigma_k)_{1 \leq k \leq K} \in \mathcal{SPD}(d)$, de telle sorte que le vecteur gaussien (Y_1, Y_2) conditionnellement à $(X_1 = k, X_2 = l)$ ait pour moyenne et variance

$$\mathbf{m}_{kl} = \begin{bmatrix} \mathbf{m}_k \\ \mathbf{m}_l \end{bmatrix} \quad \mathbf{S}_{kl} = \begin{bmatrix} \Sigma_k & \Sigma_{kl} \\ \Sigma'_{kl} & \Sigma_l \end{bmatrix} \quad (5.9)$$

Σ_{kl} est une matrice $d \times d$ et représente exactement la covariance entre les observations conditionnellement aux états k, l . Ainsi, l'ensemble des CMCa à marges et à couples gaussiens conditionnellement aux états est décrit par l'ensemble des matrices Σ_{kl} telles que $\mathbf{S}_{kl}$ soient définies positives. Nous donnons ci-dessous une condition nécessaire et suffisante pour que cela soit le cas, conséquence de la décomposition de Choleski de la matrice $\mathbf{S}_{kl}$.

Proposition 5.1.1. *Soit $\rho = \begin{bmatrix} \rho_1 & r \\ r' & \rho_2 \end{bmatrix}$ une matrice de taille $2d \times 2d$, telles $(\rho_i)_{i=1,2}$ soient définies positives. La matrice ρ est définie positive si et seulement si $\rho_2 - r' \rho_1 r > 0$.*

Nous pouvons construire un modèle similaire lorsque nous supposons que toutes les densités $p(y_1|x_1)$ sont des lois elliptiques $\mathfrak{E}(\mathbf{m}_k, \Sigma_k, \vartheta)$ (dans le cas des lois T et K, elles ont le même paramètre de queue). Si (Y_1, Y_2) conditionnellement à $(X_1 = k, X_2 = l)$ est lui aussi un SIRV de loi $\mathfrak{E}(\mathbf{m}_{kl}, \Sigma_{kl}, \vartheta)$ (égaux aux paramètres de (5.9)), nous retrouvons les marges $p(y_1|x_1)$.

Le cas général Il est possible d'adapter la modélisation par copule aux CMCa vectorielles, mais la description obtenue est beaucoup plus lourde que dans le cas scalaire (ou que dans le cas gaussien précédent). En effet, si $p(y_1|x_1)$ a pour densités marginales $f_{x_1}^1(y_1^1), \dots, f_{x_1}^d(y_1^d)$, de f.d.r correspondantes $F_{x_1}^1(y_1^1), \dots, F_{x_1}^d(y_1^d)$, la fonction de répartition de (Y_1, Y_2) conditionnellement à (X_1, X_2) s'écrit

$$F(y_1^1, \dots, y_1^d, y_2^1, \dots, y_2^d | x_1, x_2) = C_{x_1, x_2}(F_{x_1}^1(y_1^1), \dots, F_{x_1}^d(y_1^d), F_{x_2}^1(y_2^1), \dots, F_{x_2}^d(y_2^d)) \quad (5.10)$$

La copule C_{x_1, x_2} vérifie la propriétés suivantes sur les sous-copules

$$\begin{aligned} C_{[1..d], x_1, x_2} &= C_{x_1} \\ C_{[d+1..2d], x_1, x_2} &= C_{x_2} \end{aligned}$$

où C_{x_1} et C_{x_2} sont les copules obtenues par la décomposition de Sklar de $p(y_1|x_1)$ et $p(y_2|x_2)$.

Le contrôle des sous-copules d'une copule est difficile, parce que cela revient à résoudre le problème des classes de Fréchet dans le cas général (voir le chapitre 2 de [94] pour des réponses

partielles à ce problème). Il est plus facile de partir de copules multivariées connues et d'en déduire les sous-copules afin d'utiliser ces dernières pour modéliser la structure de dépendance des $p(y_1 | x_1)$. Ceci est possible notamment pour les copules archimédiennes et elliptiques. En effet, si $\forall \mathbf{u} \in [0, 1]^{2d}$, $C(\mathbf{u}) = \varphi_\theta^{-1} \left(\sum_{i=1}^{2d} \varphi_\theta(u_i) \right)$, les sous-copules sont encore archimédiennes et égales à $C_{[1..d]}(u_1, \dots, u_d) = \varphi_\theta^{-1} \left(\sum_{i=1}^d \varphi_\theta(u_i) \right)$ et $C_{[d+1..2d]}(u_{d+1}, \dots, u_{2d}) = \varphi_\theta^{-1} \left(\sum_{i=d+1}^{2d} \varphi_\theta(u_i) \right)$. Cependant l'utilisation des copules archimédiennes pour la modélisation est limitée car elle suppose que la structure de dépendance soit la même entre deux composantes du même vecteur ou encore à deux instants différents, ce qui est une hypothèse assez lourde. Dans le contexte de la segmentation, cette contrainte implique que seules les lois marginales des observations peuvent changer avec les classes.

Les copules elliptiques sont de ce point de vue plus intéressantes, car il est possible de contrôler les sous-copules, tout en ayant des liens différents entre les composantes du vecteur (Y_1, Y_2) .

Proposition 5.1.2. *Soit une copule elliptique C de taille M de matrice de corrélation ρ . Si $A \subset [1..M]$, la sous-copule (de C) C_A est une copule elliptique du même type que C , paramétrée par la matrice de corrélation $\rho_A = (\rho_{ij})_{i,j \in A}$.*

En particulier, si $M = 2d$ et que $\rho = \begin{bmatrix} \rho_1 & r \\ r' & \rho_2 \end{bmatrix}$ est la décomposition par bloc de taille $(d \times d)$ de la matrice ρ , les sous-copules $C_{[1..d]}$ (resp. $C_{[d+1..2d]}$) sont des copules elliptiques de paramètre ρ_1 (resp. ρ_2).

Démonstration. Il suffit d'exploiter le fait que la copule d'un vecteur $(U_1, \dots, U_M)$ est invariante par transformation continue et croissante composante par composante. Ainsi, pour déterminer une sous-copule, il suffit de trouver une "bonne" transformation $g = (g_1, \dots, g_M)$ telle que les lois de $(V_1, \dots, V_M) = (g_1(U_1), \dots, g_M(U_M))$ et de $(V_i)_{i \in A}$ soient faciles à identifier. Pour les copules elliptiques, cette transformation est $g = (\Phi^{-1}, \dots, \Phi^{-1})$, où Φ est la f.d.r. inverse de la marginale (centrée réduite) de la loi elliptique correspondant à la copule. Nous obtenons alors un vecteur elliptique dont les lois des sous-vecteurs appartiennent à la même famille, et pour lesquelles l'identification des copules est immédiate. □

Si nous supposons que la loi de chaque classe k est décrite par une copule gaussienne (resp. Student) de dimension d , ayant une matrice de corrélation ρ_k , alors une famille de copules $2d$ dimensionnelles convenant pour la modélisation CMCa est la famille de copule gaussienne (resp. Student) indexée par les matrices de la forme $\rho_{kl} = \begin{bmatrix} \rho_k & r_{kl} \\ r'_{kl} & \rho_l \end{bmatrix}$, avec r_{kl} telles que ρ_{kl} soit une matrice définie positive.

L'ensemble des CMCa dont la loi du couple a une copule elliptique est donc décrite par l'ensemble des matrices $(\rho_k)_{1 \leq k \leq K}$ et des matrices $(r_{kl})_{1 \leq k, l \leq K}$ vérifiant la condition de positivité donnée par la proposition 5.1.1.

5.1.3 Liens avec les processus autorégressifs à changements de régimes markovien

Nous avons déjà évoqué dans la section 3.2.3 les processus autorégressifs à changements de régimes markovien (PARCM). Nous précisons ici les différences essentielles entre ces modèles et le modèle de mélange stationnaire introduit dans la section précédente. Un PARCM est défini par deux processus $\mathbf{X} = (X_n)_{n \geq 1}$ et $\mathbf{Y} = (Y_n)_{n \geq -s+1}$ (s entier positif), tel que $\mathbf{X}$ soit une chaîne de Markov possédant une densité de transition $q(x_n, x_{n+1})$, et tel que $\mathbf{Y}$ conditionnellement à $\mathbf{X}$ soit une chaîne de Markov de mémoire s . Le noyau de transition de Y_n conditionnellement à $\bar{Y}_{n-1} = (Y_{n-1}, \dots, Y_{n-s})$ (pour $n \geq 1$) ne dépend que de X_n (parmi les $(X_k)_{1 \leq k \leq n}$). Un exemple de PARCM, justifiant l'appellation *autorégressive* est basée sur la décomposition “espace d'états” suivante

$$\begin{cases} X_{n+1} & \sim q(x_n, \cdot) \\ Y_{n+1} & = f(\bar{Y}_n, X_n, \epsilon_{n+1}) \end{cases} \quad (5.11)$$

avec $(\epsilon_n)_{n \geq 2}$ bruit blanc indépendant du processus $\mathbf{X}$ tel que ϵ_{n+1} soit indépendant de $(Y_1, \dots, Y_n)$, et f fonction mesurable.

Cette modélisation est employée dans de nombreux domaines pour décrire des changements abrupts (lorsque l'espace $\mathcal{X}$ est fini), ou des dynamiques complexes entre autres en automatique [61], en analyse d'image [116, 36], ou encore en météorologie [2]. Parmi les plus usuelles, nous signalons le processus AR linéaire d'ordre 1 (scalaire ou vectoriel), avec $\mathcal{X}$ fini :

$$\begin{cases} X_{n+1} & \sim q(x_n, \cdot) \\ Y_{n+1} & = A_{X_n} Y_n + B_{X_n} + \epsilon_{n+1} \end{cases} \quad (5.12)$$

avec $\forall k, A_k \in \mathbb{M}_d$ et $B_k \in \mathbb{R}^d$.

Le processus joint $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$ est donc une chaîne de Markov, et nous pouvons considérer qu'il s'agit d'un cas particulier de chaîne de Markov Couple. La différence entre un mélange stationnaire et un PARCM réside dans la description du modèle et sa paramétrisation : le modèle PARCM est un modèle dynamique, décrit par ses transitions, par opposition au mélange stationnaire qui est un modèle “statique” au sens où l'on spécifie la loi stationnaire de Y_n , et ensuite la dépendance. La conséquence en est que la densité de transition du processus $\mathbf{Y}$ conditionnellement à $\mathbf{X}$ dépend de *deux* états successifs X_n et X_{n+1} , y compris lorsque la copule ne dépend pas des états, voir eq. (5.6). De même, lorsque nous avons une CMCa stationnaire gaussienne, se mettant sous une forme similaire à (5.12), la densité de transition dépend elle aussi de X_n et X_{n+1} , par les relations (5.9). Ceci empêche l'application direct des résultats récents concernant l'EMV, démontrés pour les PARCM (voir section 3.2.3), et nécessitent donc des développements ultérieurs pour les mélanges stationnaires avec copules (non traités dans cette thèse).

A l'inverse, les CMCa stationnaires contournent deux problèmes auxquelles doivent faire face la modélisation PARCM :

- La nécessité de démontrer la stabilité du système étudié, c'est-à-dire l'existence (et l'unicité) d'un processus stationnaire au sens strict, vérifiant les équations (5.11). Nous pouvons trouver

dans [59, 2] plusieurs conditions.

- La méconnaissance de la vraisemblance due à la méconnaissance de la loi stationnaire de Y_n , ce qui nécessite de développer l'inférence des paramètres à partir de la vraisemblance conditionnelle $p(\mathbf{y} | y_0, \dots, y_{-r+1})$, comme cela est fait dans [59, 2].

5.2 Copule et dépendance temporelle

Le modèle CMCa-BI est un modèle de processus stationnaire tel que la dépendance soit uniquement la conséquence de du processus latent. Comme le remarque McKay [3] pour la construction d'un test d'adéquation des modèles CMCa-BI, la qualité de la représentation d'un processus stationnaire $\mathbf{Y}$ par un modèle CMCa-BI ne se résume pas seulement à la mesure de l'adéquation du mélange $\sum_{k=1}^K \pi_k f(y_n, \theta_k)$ avec la vraie loi stationnaire $p(y_n)$, mais aussi à la qualité de l'estimation des densités d'ordre supérieure $p(y_n, y_{n+1})$, $p(y_{n+1}, y_{n+2}, y_{n+3})$, ... Le modèle CMCa avec copule est un modèle plus complet que CMCa-BI parce qu'il permet de mieux modéliser les densités $p(y_n, y_{n+1})$, ce qui donne une chance de mieux décrire aussi les densités $p(\mathbf{y}_{n:n+m})$, où $m > 0$. Notre objectif est de montrer d'une part l'apport des CMCa pour la modélisation des densités d'ordre supérieure, et d'autre part de montrer leurs capacités (ou leurs limites) de modélisation de la dépendance. Cependant, les densités $p(\mathbf{y}_{n:n+m})$ ne sont pas facilement appréhendables dès que m devient grand, et nous nous contentons alors d'examiner l'autocovariance des modèles CMCa. Cette dernière devient elle aussi rapidement complexe, et nous nous intéressons aux propriétés de mélangeance des CMCa, afin de déduire le comportement asymptotique de la fonction d'autocovariance.

5.2.1 Dépendance dans les CMCa-BI et CMCa

Les fonctions d'autocovariance des processus CMCa-BI et CMCa se déduisent directement des hypothèses faites sur les modèles. Nous donnons aussi l'expression asymptotique de l'autocovariance des CMCa-BI.

Propriété 5.2.1. Covariance des CMCa-BI

Soit $(\mathbf{X}, \mathbf{Y})$ une CMCa-BI stationnaire, telle que la loi stationnaire de $\mathbf{X}$ soit $\pi = (\pi_1, \dots, \pi_K)'$. Nous notons pour tout n , $p_{kl}^{(n)} = P(X_1 = k, X_n = l)$ et nous supposons que $\forall k, E[|Y_1| | X_1] < \infty$, avec $E[Y_1 | X_1 = k] = m_k$. L'autocovariance de $\mathbf{Y}$ existe et vaut

$$\forall n, \text{cov}(Y_1, Y_n) = \sum_{k,l} \left(p_{kl}^{(n)} - \pi_k \pi_l \right) m_k m_l \quad (5.13)$$

Si la matrice de transition est irréductible et apériodique, nous pouvons affirmer

$$\exists 0 < \lambda < 1, m \text{ entier positif}, \text{cov}(Y_1, Y_n) \sim_{\infty} C n^{m-1} \lambda^n \quad (5.14)$$

Démonstration. Nous avons

$$\begin{aligned} E[|Y_1 Y_n|] &= E[E[|Y_1 Y_n| | X_1, X_n]] \\ &= E[E[|Y_1| | X_1] E[|Y_n| | X_n]] < \infty \end{aligned}$$

Par la même décomposition, nous avons :

$$\begin{aligned} E[Y_1 Y_n] - E[Y_1] E[Y_n] &= \sum_{k,l} p_{kl}^{(n)} m_k m_l - \left(\sum_k \pi_k m_k \right)^2 \\ &= \sum_{k,l} \left(p_{kl}^{(n)} - \pi_k \pi_l \right) m_k m_l \end{aligned}$$

Le théorème de Perron-Frobenius [23] permet d'avoir une description précise de la vitesse de décroissance de la covariance. En effet, si la matrice de transition A est irréductible et apériodique, nous avons

$$A^n = 1\pi' + O(n^{m-1}\lambda^n)$$

avec λ la valeur absolue de la seconde plus grande valeur propre de A (de module strictement plus petit que 1) et m sa multiplicité algébrique². Par conséquent,

$$\forall k, l, p_{kl}^{(n)} = \pi_k \pi_l + O(n^{m-1}\lambda^n)$$

et donc $\text{cov}(Y_1, Y_n) = \left(\sum_{k,l} m_k m_l \right) \times O(n^{m-1}\lambda^n)$.

□

Le modèle CMCa-BI donne donc peu de moyens de reproduire des corrélations complexes qui peuvent exister dans les observations que nous souhaitons segmenter. Un inconvénient majeur des modèles CMCa-BI apparaît en particulier lorsque les lois d'émission sont de moyennes nulles : l'éq. (5.13) implique que la fonction d'autocovariance est nulle³. En particulier, la modélisation par CMCa-BI et SIRV centrées des signaux radar impliquent que les observations soient décorréelées, or c'est une hypothèse mise en défaut sur les données réelles [152, 151]. La modélisation par CMCa permet de pallier ce défaut comme le montre l'expression de la covariance.

Propriété 5.2.2. Covariance des CMCa

Soit $(\mathbf{X}, \mathbf{Y})$ une CMCa stationnaire, telle que la loi stationnaire de $\mathbf{X}$ soit $\pi = (\pi_1, \dots, \pi_K)'$. Nous supposons que $\forall k, E[|Y_1| | X_1 = k] < \infty$, avec $E[Y_1 | X_1 = k] = m_k$, et les copules c_{kl} sont continues. Alors la covariance existe, et nous avons

$$\forall n \geq 2, \text{cov}(Y_1, Y_n) = \sum_{k,l} p_{kl}^{(n)} \text{cov}_{kl}(Y_1, Y_n) + \sum_{k,l} \left(p_{kl}^{(n)} - \pi_k \pi_l \right) m_k m_l \quad (5.15)$$

où $\text{cov}_{kl}(Y_1, Y_n)$ est la covariance de Y_1, Y_n conditionnellement à $X_1 = k, X_n = l$.

Démonstration. Comme nous avons $\forall k, E[|Y_1| | X_1 = k] < \infty$, et que les copules c_{kl} sont continues donc bornées sur $[0, 1]^2$, ceci assure que $E[|Y_1 Y_n|] < \infty$. La covariance se décompose alors en deux termes

²i.e. le nombre de fois que λ est racine du polynôme caractéristique de A .

³La formule (5.13) est encore valide pour des observations vectorielles.

$$\begin{aligned}
\text{cov}(Y_1, Y_n) &= E[Y_1 Y_n] - E[Y_1] E[Y_n] \\
&= \sum_{k,l} p_{kl}^{(n)} E[Y_1 Y_n | X_1 = k, X_n = l] - \sum_{k,l} \pi_k \pi_l m_k m_l
\end{aligned}$$

Nous réécrivons cette dernière expression sous la forme de l'eq. (5.15) en faisant apparaître le terme $\sum_{kl} p_{kl}^{(n)} m_k m_l$.

□

La covariance d'une CMCa est la somme de la covariance créée par le processus caché (égale à la variance dans le cas CMCa-BI), et de la covariance conditionnelle des observations. Cette deuxième contribution est d'expression complexe et il n'est pas possible d'en déduire une expression analytique, ni même une expression asymptotique de la covariance, comme dans les CMCa-BI. Nous apporterons une réponse partielle à cette question dans la section suivante.

5.2.2 Quelques remarques sur la dépendance des processus stationnaires

Nous introduisons tout d'abord les différents concepts de mélangeance utilisés en statistique pour étudier les propriétés des processus stationnaires. Les coefficients de mélangeance sont définis de manière "abstraite" entre deux tribus. Soit $(\Omega, \mathcal{A}, P)$ un espace probabilisé et $\mathcal{F}$ et $\mathcal{G}$ deux sous-tribus de $\mathcal{A}$. Nous définissons alors les coefficients de α -mélange, β -mélange et ϕ -mélange par

$$\begin{aligned}
\alpha(\mathcal{F}, \mathcal{G}) &= \sup_{A \in \mathcal{F}, B \in \mathcal{G}} |P(A \cap B) - P(A)P(B)| \\
\beta(\mathcal{F}, \mathcal{G}) &= E \left[\sup_{A \in \mathcal{F}} |P(A) - P(A|\mathcal{G})| \right] \\
\phi(\mathcal{F}, \mathcal{G}) &= \sup_{A \in \mathcal{F}, B \in \mathcal{G}/P(B)>0} |P(A) - P(A|B)|
\end{aligned} \tag{5.16}$$

Ces 3 coefficients de mélangeance sont liés par l'inégalité suivante

$$2\alpha \leq \beta \leq \phi \tag{5.17}$$

Si nous notons de manière générique $c(\mathcal{F}, \mathcal{G})$ les coefficients de mélangeance entre deux tribus définies par les équations (5.16), nous pouvons définir les coefficients de mélangeance d'un processus stationnaire $\mathbf{Y} = (Y_n)_{n \in \mathbb{Z}}$ par

$$\forall k \geq 1, c_{k, \mathbf{Y}} = c(\sigma((Y_n)_{n \leq 0}), \sigma((Y_n)_{n \geq k}))$$

Nous pouvons donc associer à $\mathbf{Y}$ les suites de coefficients de mélangeance $(\alpha_n)_{n \geq 1}$, $(\beta_n)_{n \geq 1}$, $(\phi_n)_{n \geq 1}$ en considérant les tribus des "événements passé et futur". Le processus $\mathbf{Y}$ est alors appelé α -mélangeant (ou fortement mélangeant), β -mélangeant ou ϕ -mélangeant si la suite des coefficients correspondants tend vers 0 : $\lim_{k \rightarrow \infty} c_{k, \mathbf{Y}} = 0$. Par les inégalités de l'eq. (5.17), la ϕ -mélangeance implique la β -mélangeance, qui implique la mélangeance forte. Ces coefficients, selon leur vitesse de décroissance vers 0, permettent d'identifier les cas pour lesquels nous avons toujours une loi forte des grands nombres, un théorème limite central, ... [62].

Ces coefficients sont importants parce qu'ils permettent de contrôler, entre autres, le comportement asymptotique des autocovariances d'un processus, grâce aux inégalités suivantes.

Lemme 5.2.1. Inégalité de Davydov

Soient U, V deux variables aléatoires réelles, tel que $U \in L^p(P)$, $V \in L^q(P)$. Soient $p, q, r > 0$, tels que $\frac{1}{p} + \frac{1}{q} + \frac{1}{r} = 1$ et $\alpha = \alpha(\sigma(U), \sigma(V))$. Nous avons alors

$$\text{cov}(U, V) \leq 2r (2\alpha)^{1/r} \|U\|_p \|V\|_q \quad (5.18)$$

Lemme 5.2.2. Inégalité d'Ibragimov

Soient U, V deux variables aléatoires réelles, tel que $U \in L^p(P)$, $V \in L^q(P)$. Soient $p, q > 0$, tels que $\frac{1}{p} + \frac{1}{q} = 1$, et $\phi = \phi(\sigma(U), \sigma(V))$ nous avons alors

$$\text{cov}(U, V) \leq 2\phi^{1/p} \|U\|_p \|V\|_q \quad (5.19)$$

Nous renvoyons à [62] pour une présentation des propriétés de dépendance dans les processus stationnaires, et aux résultats asymptotiques qui leur sont associés. Nous redémontrons ci-dessous la propriété "d'hérédité" de la mélangeance, qui montre qu'une fonction déterministe ne peut pas augmenter la dépendance entre deux tribus.

Théorème 5.2.1. Soient U, V deux variables aléatoires réelles définies sur un espace probabilisé, $f, g : (\mathbb{R}, \mathcal{B}(\mathbb{R})) \rightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ deux fonctions mesurables et $U' = f(U)$, $V' = g(V)$. Nous avons alors :

(i)

$$c(\sigma(U'), \sigma(V')) \leq c(\sigma(U), \sigma(V)) \quad (5.20)$$

(ii) Les coefficients de mélangeance ne dépendent que de la copule C_{UV} du couple (U, V)

(iii) Soit $\mathbf{Z}$ est un processus stationnaire de coefficients de mélangeance $(c_{k, \mathbf{Z}})_{k \geq 1}$ et le processus $\mathbf{Y} = (f(Z_n))_{n \in \mathbb{Z}}$ où f est une application mesurable. Alors nous avons

$$\forall k \geq 1, c_{k, \mathbf{Y}} \leq c_{k, \mathbf{Z}}$$

Démonstration. Ces résultats sont une conséquence directe de la propriété de base des tribus : si f est une application mesurable, alors $\sigma(f(U)) \subset \sigma(U)$. L'inégalité (5.20) découle alors de la définition des coefficients de mélangeance. Par exemple, pour le coefficient α , si nous notons $\sigma(U) = \mathcal{F}$, $\sigma(U') = \mathcal{F}'$ et $\sigma(V) = \mathcal{G}$, $\sigma(V') = \mathcal{G}'$, il est clair que

$$\sup_{A \in \mathcal{F}', B \in \mathcal{G}'} |P(A \cap B) - P(A)P(B)| \leq \sup_{A \in \mathcal{F}, B \in \mathcal{G}} |P(A \cap B) - P(A)P(B)|$$

Dans le cas où les applications f et g sont bijectives, ayant les inégalités dans les deux sens, nous avons donc égalité des coefficients de mélangeance $c(\sigma(U'), \sigma(V')) = c(\sigma(U), \sigma(V))$. En particulier, si nous prenons pour f et g les fonctions de répartition de U et V (notées respectivement F_U et F_V), nous en déduisons que les coefficients de mélangeance sont égaux à ceux de $F_U(U)$, $F_V(V)$ et donc ne dépendent que de la copule de (U, V) .

L'extension de l'inégalité (5.20) aux processus repose sur le même argument que dans le cas bivarié.

□

Ce théorème permet de déduire la structure de dépendance des chaînes de Markov partiellement observées de celle du processus complet. Si $\mathbf{Y}$ est une chaîne de Markov partiellement observée, il existe $\mathbf{Z}$ chaîne de Markov et une fonction mesurable f telles que

$$\forall n, Y_n = f(Z_n)$$

Dans le cas des chaînes couples et triplets, la fonction f est l'application coordonnée sur l'espace $\mathcal{Y}$. Sous des conditions assez générales (stationnarité et Harris-réurrence, voir annexe C), les chaînes de Markov sont ergodiques et β -mélangeantes. Sous ces conditions sur $\mathbf{Z}$, les chaînes de Markov partiellement observées (et en particulier les CMCa) sont elles aussi ergodiques et β -mélangeantes. De plus, nous connaissons un majorant de la vitesse de décroissance, ce qui nous permet de connaître grâce aux inégalités (5.17) et (5.18) la vitesse de décroissance de la fonction d'autocovariance. Par exemple, si $\mathbf{Z}$ est géométriquement ergodique (avec une décroissance exponentielle du coefficient β_n , ce qui est une propriété vérifiée par les modèles PARCM sous les hypothèses de Douc et al., [59]), nous pouvons affirmer la covariance décroît exponentiellement dès que $Y_n \in L^{2+\epsilon}$ avec $\epsilon > 0$. Si la chaîne de Markov est uniformément géométriquement ergodique (ce qui correspond à la décroissance exponentielle des ϕ_n), nous avons encore décroissance exponentielle de la covariance, avec des contraintes moins fortes sur l'existence de moments grâce à l'inégalité d'Ibragimov. Nous tirons de ces remarques deux conclusions :

- La vitesse de décroissance de la dépendance des CMCa (et notamment de la covariance) peut être obtenue par la connaissance des propriétés du processus markovien complet, plus facile à étudier [62].
- La modélisation par CMCa des processus de mélanges stationnaires semble plus adaptée aux processus dont la dépendance décroît rapidement. En effet, comme nous ne modélisons directement que les densités $p(y_n, y_{n+1})$, les densités d'ordre supérieure $p(y_n, y_{n+2}), \dots, p(y_n, y_{n+p})$ ne sont que des conséquences de l'hypothèse “processus markovien partiellement observé”. Un modèle CMCa tel que sa dépendance soit à décroissance rapide a donc des densités $p(y_n, y_{n+p})$ qui convergent vers rapidement vers $p(y_n)p(y_{n+p})$, ce qui permet de ne faire appel qu'à des hypothèses déjà formulées sur le processus (i.e. la loi de la marginale), et donc limite les risques de mauvaises adéquations aux données. Ceci peut consister en un test de cohérence du modèle par rapport aux données. Si nous voulons modéliser et segmenter des processus à dépendance longue, il semble préférable d'en faire l'hypothèse explicitement, par exemple par l'emploi de chaînes couple ou triplet partiellement markoviennes (voir [130] pour la segmentation des processus à mémoire longue).

5.3 Algorithme d'estimation

Nous abordons dans cette section l'estimation d'une CMCa scalaire, selon le modèle décrit dans la section 5.1.2.1. Nous rappelons brièvement l'application de la méthode “Inference For Margins” pour l'estimation d'une chaîne de Markov stationnaire $\mathbf{Y} = (Y_n)_{n \geq 1}$ spécifiée par une copule. Comme dans la section 4.3, nous en déduisons un estimateur ECI pour CMCa et pour lesquelles nous pouvons avoir des formules explicites.

5.3.1 Estimation des chaînes de Markov avec copules

Soit une chaîne de Markov $\mathbf{Y} = (Y_n)_{n \geq 1}$ réelle et stationnaire, telle que sa loi stationnaire et la copule de (Y_1, Y_2) admettent respectivement des densités $f_\theta, \theta \in \Theta$ (de f.d.r F_θ) et $c_\eta, \eta \in \Upsilon$. Si $\mathbf{y} = (y_1, \dots, y_N)$ est un échantillon, sa log-vraisemblance s'écrit :

$$\mathcal{L}(\theta, \eta) = \sum_{i=1}^N \log(f_\theta(y_i)) + \sum_{i=2}^N \log(c_\eta(F_\theta(y_{i-1}), F_\theta(y_i))) \quad (5.21)$$

Le calcul de l'EMV nécessite la résolution des équations normales suivantes

$$\begin{cases} \sum_{i=1}^N \nabla_\theta \log f_\theta(y_i) + \sum_{i=1}^N \nabla_\theta \log(c_\eta(F_\theta(y_{i-1}), F_\theta(y_i))) &= 0 \\ \sum_{i=2}^N \nabla_\eta \log(c_\eta(F_\theta(y_{i-1}), F_\theta(y_i))) &= 0 \end{cases} \quad (5.22)$$

Comme nous l'avons déjà vu dans la section 4.3, les équations normales ne permettent pas d'obtenir d'expression analytique pour le maximum de vraisemblance, et des procédures d'optimisation numérique sont nécessaires pour résoudre ce système.

Les méthodes d'estimation en deux étapes IFM et omnibus (voir section 4.3.2) présentées dans le cas multivariée s'appliquent encore (voir [41] pour une application d'omnibus à l'estimation semi-paramétrique d'une chaîne de Markov). La méthode IFM consiste en la résolution du système suivant

$$\begin{cases} \sum_{i=1}^N \nabla_\theta \log f_\theta(y_i) &= 0 \\ \sum_{i=2}^N \nabla_\eta \log(c_\eta(F_\theta(y_{i-1}), F_\theta(y_i))) &= 0 \end{cases} \quad (5.23)$$

pour laquelle nous obtenons plus facilement une solution analytique, ou pour lequel les optimisations numériques sont plus simples et performantes. Ce système correspond alors à la fonction estimante $G_{IFM}(\theta, \eta) = \begin{pmatrix} \nabla_\theta \mathcal{L}'_m & \nabla_\eta \mathcal{L}' \end{pmatrix}$, où nous avons noté $\mathcal{L}_m$ pour la log-vraisemblance de la marginale.

5.3.2 Estimation des CMCa avec copules

A partir des développements sur la projection des fonctions estimantes (section 3.5), nous proposons une nouvelle procédure d'estimation des paramètres d'un modèle CMCa. La log-vraisemblance complète du modèle (décrit dans la partie 5.1.2.1) s'écrit

$$L_c(\phi) = \log p(\mathbf{x}, A) + \sum_{i,k=1}^{N,K} 1_k(x_i) \log f(y_i, \theta_k) + \sum_{i,k,l=1}^{N,K,K} 1_{k,l}(x_i, x_{i+1}) \log c_{\eta_{kl}}(F(y_i, \theta_k), F(y_{i+1}, \theta_l))$$

Nous introduisons alors la log-vraisemblance marginale du processus $(Z_n)_{1 \leq n \leq N}$ définie par

$$\begin{aligned}
L_{c,m} &= \sum_{i=1}^N \log(p(z_i, \phi)) \\
&= \sum_{i,k} 1_k(x_i) \log(\pi_k f(y_i, \theta_k))
\end{aligned}$$

et la fonction estimante

$$g_{\text{IFM}}((\mathbf{x}, \mathbf{y}), \phi) = \left(\nabla_A L'_c \quad \left(\nabla_{\theta_k} L'_{c,m} \right)_{1 \leq k \leq K} \quad \left(\nabla_{\eta_{k,l}} L'_c \right)_{1 \leq k, l \leq K} \right)' \quad (5.24)$$

Nous proposons d'estimer les paramètres de la CMCa en résolvant l'équation définie par la projection de la fonction estimante g_{IFM} , i.e.

$$G_{\text{IFM}}(\mathbf{y}, \phi) = E_\phi [g_{\text{IFM}}((\mathbf{x}, \mathbf{y}), \phi) | \mathbf{y}] \quad (5.25)$$

Nous utilisons l'algorithme ECI pour la recherche des racines de l'équation $G_{\text{IFM}}(\mathbf{y}, \phi) = 0$, ce qui donne les formules de ré-estimations implicites suivantes pour les paramètres des lois d'émission et des copules

$$\text{pour } 1 \leq k, l \leq K, \quad \begin{cases} \theta_k^{(n+1)} \text{ tel que } \sum_{i=1}^N \pi_{i,k}^{(n)} \nabla_{\theta_k} \log f(y_i, \theta_k) = 0 \\ \eta_{kl}^{(n+1)} \text{ tel que } \sum_{i=1}^N p_i^{(n)}(k, l) \nabla_{\eta_{kl}} \log c_{\eta_{kl}}(F(y_i, \theta_k^{(n+1)}), F(y_{i+1}, \theta_l^{(n+1)})) = 0 \end{cases} \quad (5.26)$$

La version stochastique de ECI-IFM consiste en la résolution avec données complétées de l'équation (5.24), ce qui donne les mêmes formules de mises à jour que l'eq. (5.23) mais pour chaque classe k et chaque couple de classes (k, l) . Nous pouvons finalement réinterpréter la procédure IFM-ECI de la manière suivante :

1. Calculer les paramètres A_{n+1} et θ_{n+1} comme pour une CMCa-BI
2. Calculer les mises à jour des f.d.r des lois d'émission
3. Calculer η_{n+1} comme étant l'EMV des copules, basé sur les données mises à jour $\left(F(y_i, \theta_k^{(n+1)}) \right)_{1 \leq i, k \leq N, K}$

5.3.3 Expérimentations

Nous illustrons dans cette section l'influence de la dépendance conditionnelle pour les CMCa dans les phases d'estimation et de segmentation des mélanges stationnaires. Ceci nous permet de mettre en évidence l'intérêt de la prise en compte de la dépendance entre les observations pour

- l'estimation du mélange $\sum_{k=1}^K \pi_k f(y_i, \theta_k)$,
- l'estimation de l'autocovariance du processus observé $\mathbf{Y}$,
- le taux d'erreur en segmentation non-supervisée.

Nous illustrons sur des données radar réelles, le comportement des algorithmes de segmentation par CMCa-BI et CMCa avec copules.

5.3.3.1 Estimation et modélisation de la dépendance

Nous montrons tout d'abord à l'aide de quelques exemples que les CMCa-BI reproduisent mal la loi du processus observé $\mathbf{Y}$, lorsque nous avons une dépendance conditionnelle. Nous traitons pour cela 3 exemples : nous simulons dans le premier cas une CMCa à deux classes dont les lois d'émission sont des lois gamma, et telles que la dépendance conditionnelle soit portée par une copule gaussienne. Dans le deuxième et le troisième cas, il s'agit de mélange stationnaire dont nous ne connaissons pas la loi, mais dont nous sommes sûrs qu'il ne s'agit ni d'une CMCa-BI ni d'une CMCa : il s'agit d'un champ de Markov caché, dont les observations sont corrélées conditionnellement aux classes.

Pour mettre en évidence la qualité de l'estimation la loi du processus, nous utilisons les fonctions d'autocovariance et d'autocorrélation, mais aussi les coefficients de Kendall et de Spearman (voir 4.3.1) comme indicateur de la dépendance de $\mathbf{Y}$. L'avantage de ces deux dernières mesures de dépendance est de ne dépendre que de la copule, contrairement à l'autocorrélation. Leur examen nous permet de séparer la qualité de l'estimation du mélange de la qualité de l'estimation de la dépendance.

L'**exemple 1** est une CMCa à deux classes telles que les lois d'émission soient des lois gamma $\gamma(1, 1)$ et $\gamma(3, 2)$. La matrice de transition de la chaîne de Markov $\mathbf{X}$ est $A = \begin{bmatrix} 0,8 & 0,2 \\ 0,2 & 0,8 \end{bmatrix}$. Les dépendances conditionnelles sont représentées par des copules gaussiennes pour toutes les classes indexées par le coefficient de corrélation ρ_{kl} , $k, l = 1, 2$. Nous supposons que $\rho_{11} = \rho_{22} = 0,8$ et que $\rho_{12} = \rho_{21} = 0$. Cette dernière hypothèse signifie que les observations sont indépendantes conditionnellement aux états lorsque les états sont différents (ce qui correspond aux zones frontalières).

Nous simulons une chaîne de longueur $N = 1000$, et nous la restaurons de manière non-supervisée par une CMCa-BI et une CMCa, en utilisant un algorithme IFM-ECI (équivalent à l'algorithme EM dans le cas CMCa-BI). Nous déduisons des paramètres estimés les fonctions de dépendance théoriques pour les modèles CMCa et CMCa-BI⁴.

Il apparaît dans les quatre graphes de la figure 5.1 que les données issues du modèle CMCa sont bien estimées dans le modèle CMCa : précisément le tau de Kendall et le rho de Spearman sont très bien restitués sur les 9 premiers décalages. Par contre, le modèle CMCa-BI ne permet pas de restituer convenablement la dépendance pour les 3 premiers décalages correspondant aux lois $p(y_1, y_2)$, $p(y_1, y_3)$ et $p(y_1, y_4)$ (en ce qui concerne le rho de Spearman). Celle-ci décroît très rapidement avant de trouver la courbe asymptotique (à décroissance exponentielle) des données réelles (de même que la courbe théorique du processus CMCa étudié).

Les deuxième et troisième exemples auxquelles nous nous intéressons sont des mélanges stationnaires qui ne sont pas des processus CMCa. Il s'agit de champ de Markov à bruit corrélé : nous générons tout d'abord $\epsilon = (\epsilon_s)_{s \in S}$ une champ de markov gaussien (de type moyennes mobiles) de telle sorte que nous puissions connaître la loi marginale (loi normale centrée réduite). Par application de la fonction de répartition ϕ , nous obtenons alors une réalisation $\mathbf{u} = (u_s)_{s \in S}$ d'une copule gaussienne de type markovien. Nous pouvons générer alors un mélange stationnaire

⁴Dans le cas CMCa, la formule (5.15) n'est pas facilement calculable et la fonction théorique est obtenue par simulation et calcul empirique sur une chaîne de longueur 10000.

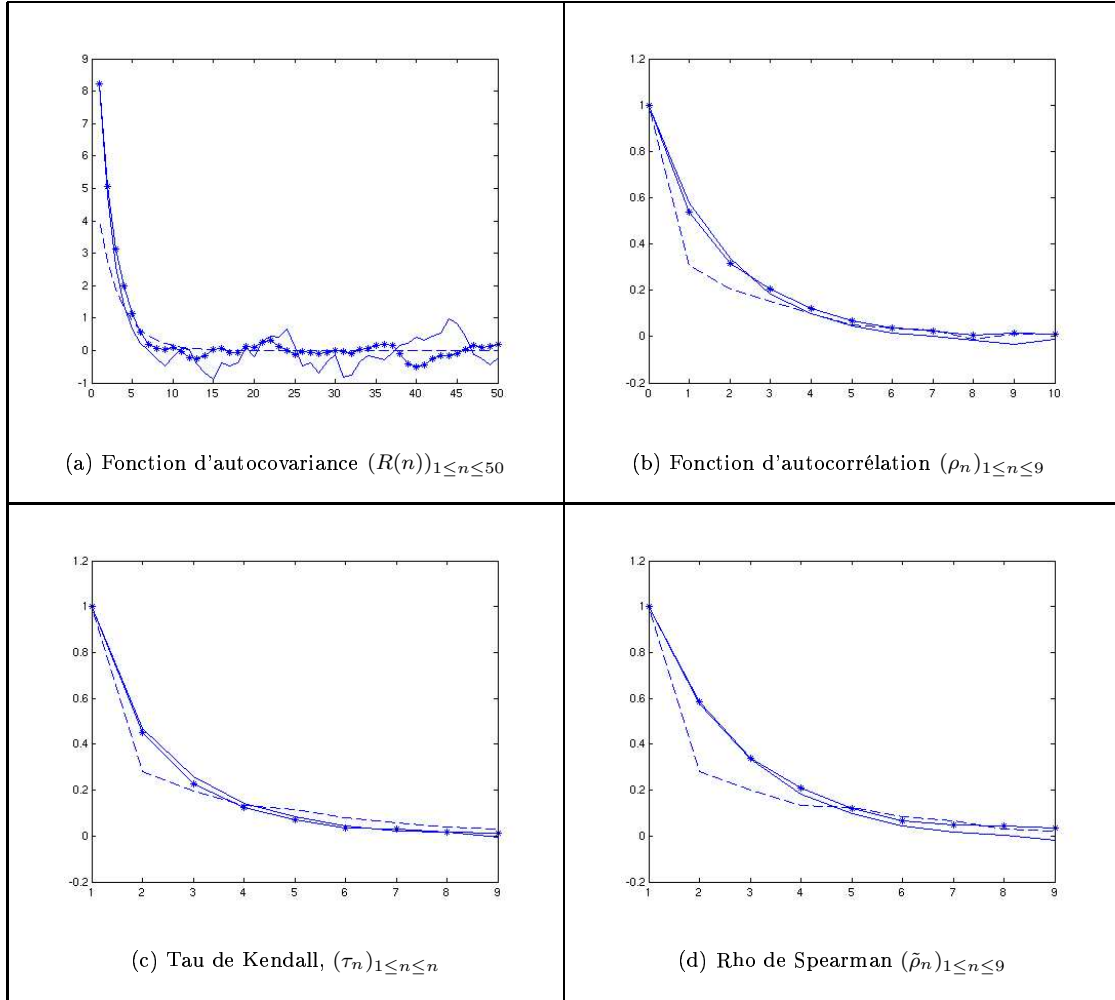


FIG. 5.1 – Estimation de la dépendance de la réalisation d'une CMCa à 2 classes et marges gamma (**Exemple 1**)

— : dépendance empirique (estimée sur l'échantillon), — — : dépendance théorique par CMCa-BI, * — * : dépendance théorique par CMCa

en appliquant la fonction de répartition inverse $F_{\theta_k}^{-1}$, où l'étiquette k est déterminée par un mesure de Gibbs à K classes, aux 4 plus proches voisins. La densité de ce champ est proportionnelle à $\exp\left(\alpha \sum_{\langle s,t \rangle} x_s x_t\right)$, où $\sum_{\langle s,t \rangle}$ désigne la somme sur l'ensemble des sites qui sont voisins. Le paramètre α est la force de la dépendance entre deux sites, et vaut 2 dans les simulations.

Le champ $\mathbf{y} = (y_s)_{s \in S}$ ainsi obtenu est alors transformé en chaîne par un parcours de Peano : la chaîne obtenue n'est pas markovienne (elle est de plus à dépendance longue portée de par la forme du parcours) mais nous connaissons sa loi stationnaire qui est le mélange $\sum_{k=1}^K \pi_k F_{\theta_k}(y)$. Nous considérons alors deux cas, pour lequel nous avons :

Exemple 2 2 classes avec des lois gamma $\gamma(1,1)$ et $\gamma(3,2)$ (voir figure 5.2)

Exemple 3 3 classes avec des lois normales $N(0;0,5)$, $N(0;1)$, $N(0;2)$ (voir figure 5.4)

Ainsi ni le modèle CMCa-BI, ni le modèle CMCa ne sont adaptés pour décrire ces données (en

raison des hypothèses de markovianité du processus $\mathbf{X}$ mais aussi du processus joint $\mathbf{Z}$).

Lorsque nous traçons les fonctions de dépendance estimées théoriquement et les fonctions de dépendance empiriques (voir figures 5.3 et 5.5), nous constatons que la dépendance longue des observations est mal reproduite par les CMCa (qui ont ici une décroissance de type exponentielle, voir section 5.2) qui la sous-estime.

Cependant dans les 2 champs traités, il apparaît que le modèle CMCa avec copules réussit à estimer correctement l'autocorrélation jusqu'à l'ordre 2 (voire 3), mais aussi les taux de Kendall et Rho de Spearman de $p(y_1, y_2), p(y_1, y_3)$ et dans une moindre mesure celles de $p(y_1, y_4)$. Il apparaît alors que le modèle CMCa estimé induit une décroissance exponentielle de la dépendance, ce qui donne un écart important avec les données pour les lois d'ordre supérieure $p(y_1, y_{1+p})$ avec $p > 4$. Le modèle CMCa-BI reproduit mal cette dépendance, qui décroît très fortement dès le premier décalage, et reste à un niveau très bas.

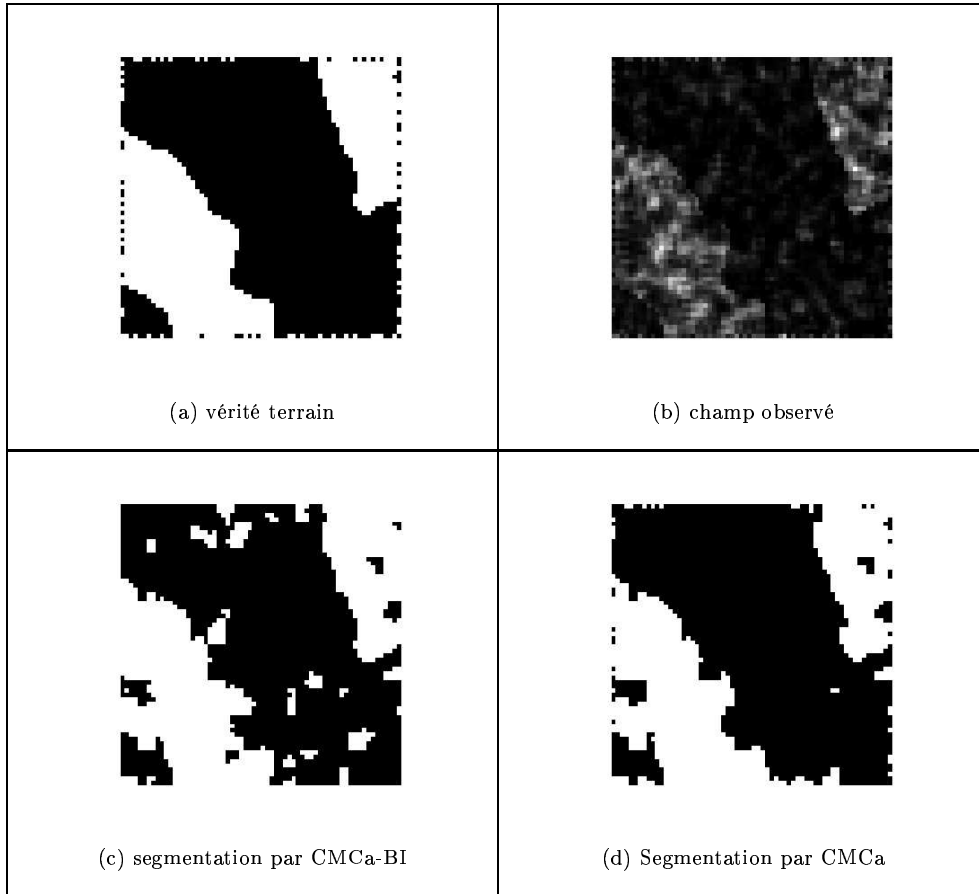


FIG. 5.2 – Champ couple à 2 classes et marges gamma (**Exemple 2**)

L'étude du champ à 3 classes permet de montrer que l'estimation des marges est biaisée par l'estimation de la dépendance. En effet, les 3 classes sont de moyennes nulles, ce qui implique qu'une estimation non-biaisée des paramètres par CMCa-BI donne une fonction d'autocovariance nulle par la formule (5.13). Or celles-ci sont strictement positives, et l'examen des paramètres

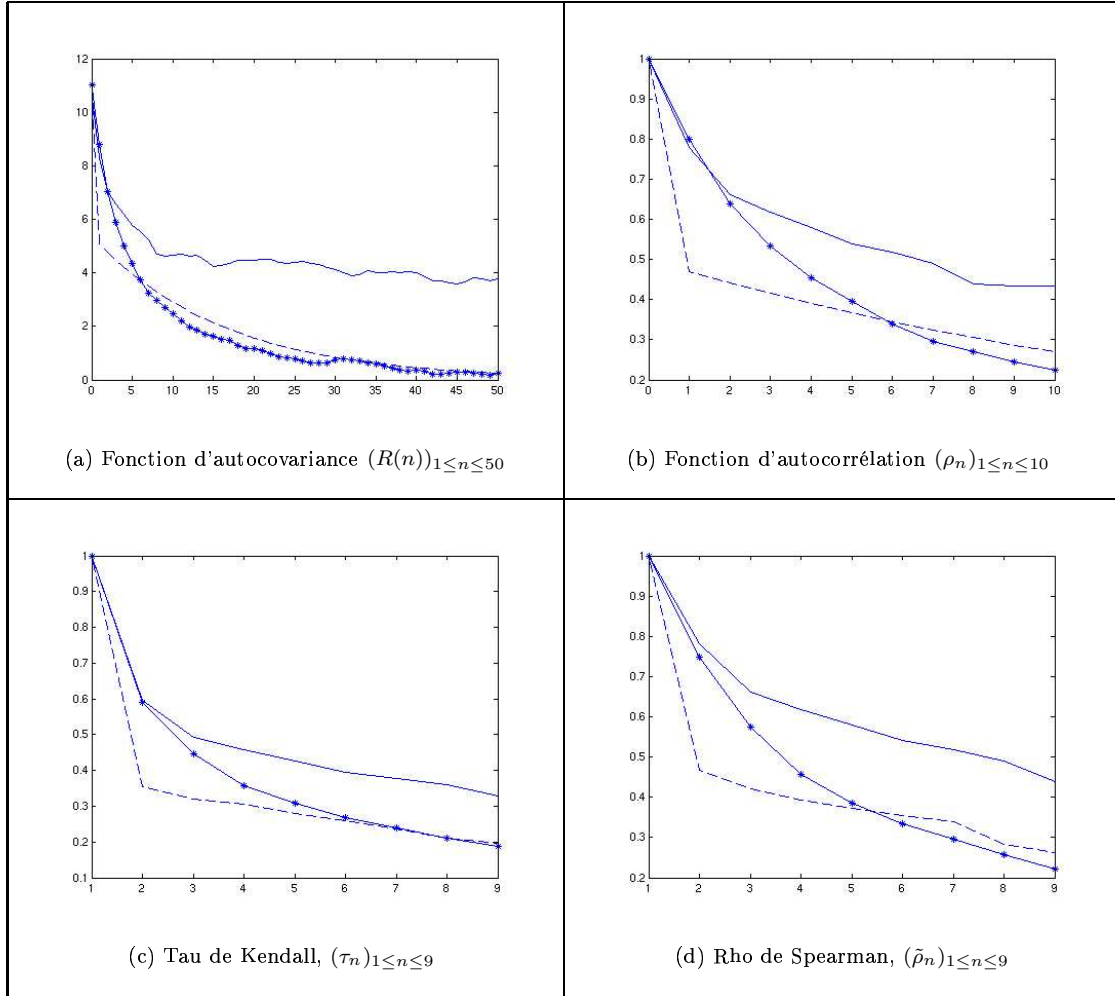


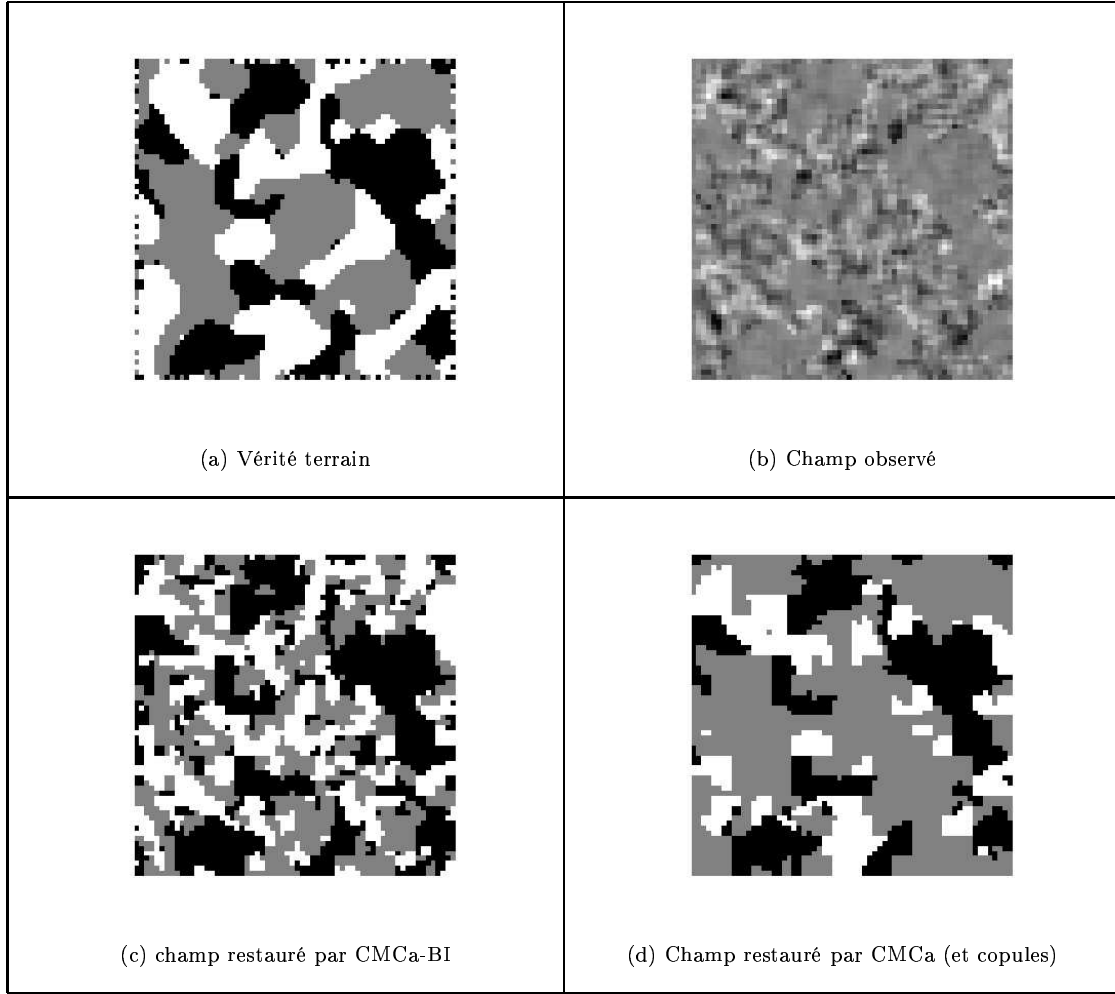
FIG. 5.3 – Estimation de la dépendance d'un mélange stationnaire à 2 classes et marges gamma (**Exemple 2**)

— : dépendance empirique (estimée sur l'échantillon), — — : dépendance théorique par CMCa-BI, * — * : dépendance théorique par CMCa

estimés pour les lois d'émission (cf. tableau 5.1), montre que l'estimation des moyennes est biaisée afin de reproduire la structure de dépendance des données. Nous remarquons que le modèle CMCa avec copule estime convenablement les paramètres de la loi d'émission, et que le modèle reproduit une partie de la corrélation des données en estimant des paramètres de copules non-nulles (correspondant au second terme de droite de l'eq. (5.15)), égaux à $\rho_{11} = 0,6$, $\rho_{22} = 0,57$ et $\rho_{33} = 0,69$. Les autres paramètres des copules sont forcés à être égaux à 0, pour éviter une très forte variance (et des instabilités numériques) dues au faible nombre d'observations (y_n, y_{n+1}) pour estimer la copule correspondant aux états $(X_n = k, X_{n+1} = l)$ lorsque $k \neq l$.

5.3.3.2 Estimation des paramètres

Nous reprenons ici les résultats de l'article [28], dans lequel nous nous sommes intéressés à l'estimation des paramètres et à la segmentation non-supervisée par CMCa, selon la force de la

FIG. 5.4 – Champ couple à 3 classes et marges gaussiennes centrées (**Exemple 3**)

dépendance (conditionnellement aux états). Nous prenons comme exemple une CMCa à deux classes et lois d'émission gamma, dont la matrice de transition $A = \begin{bmatrix} 0,8 & 0,2 \\ 0,2 & 0,8 \end{bmatrix}$. La première classe suit une loi de type $\gamma(1; 0, 5)$ et la seconde classe est une loi de type $\gamma(2; 1)$. La dépendance conditionnelle, modélisée par une copule gaussienne, est telle que le coefficient de corrélation soit nul pour $(X_1 = k, X_2 = l)$ lorsque $k \neq l$. Lorsque les classes sont identiques, le paramètre de la copule ρ peut varier entre 0 et 0,8. Les résultats des procédures d'estimation sont obtenues par simulation Monte Carlo, pour des séries de longueur $N = 1000$ et a été répété 100 fois.

Il apparaît que la dépendance a tendance à augmenter le taux d'erreur, mais que celui-ci se stabilise rapidement (pour $\rho > 0,5$). Les procédures de segmentation supervisée donnent des résultats proches tant que la dépendance n'est pas très forte (voir tableau 5.2) mais il est notable que la négligence de la dépendance conditionnelle dans les CMCa a tendance à dégrader les performances. Cependant cet écart important entre les 2 modèles est moindre lorsque nous comparons les taux d'erreur en mode non-supervisée. Dans ce contexte, il apparaît clairement que l'estimation des paramètres entraîne des erreurs importante qui ne permette plus d'avoir une telle différence entre

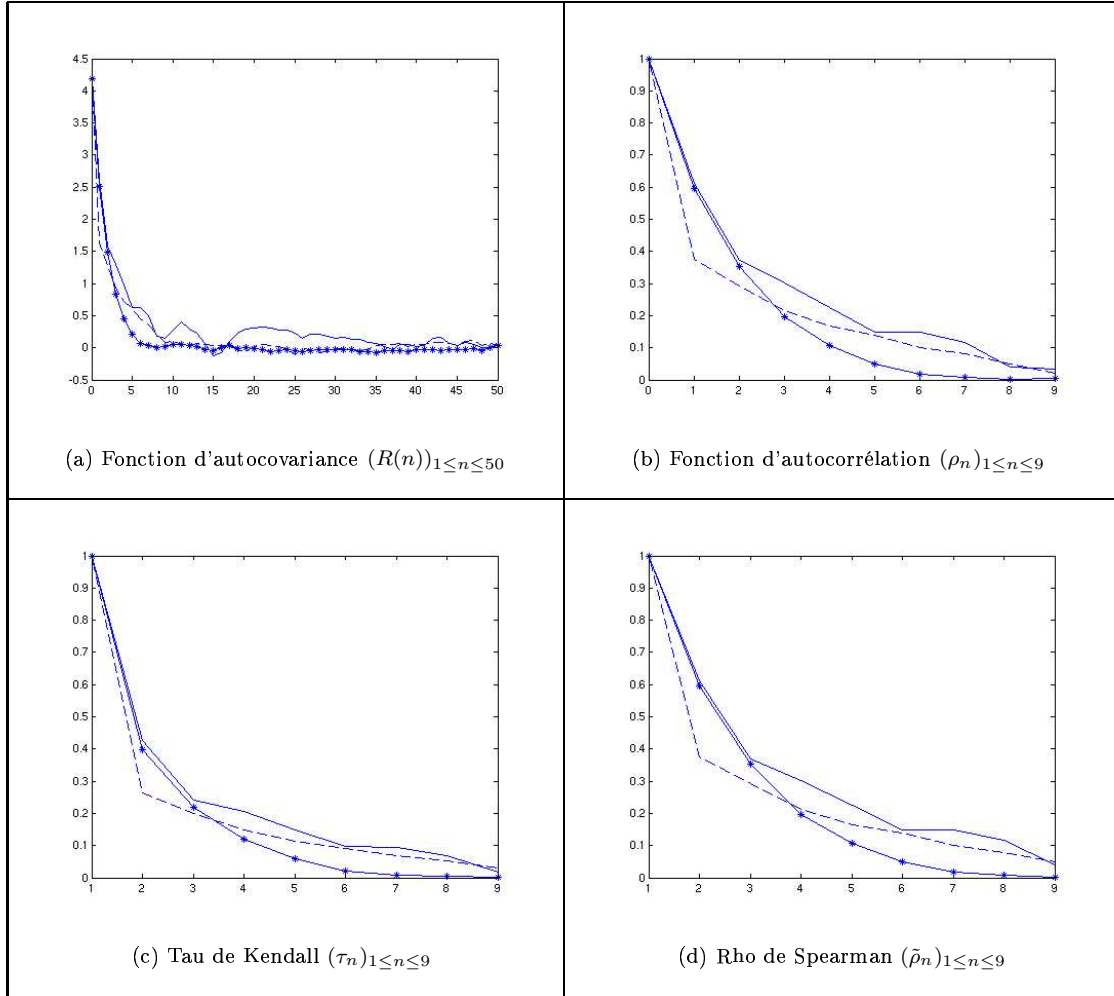


FIG. 5.5 – Estimation de la dépendance d'un mélange stationnaire de type champ couple à 3 classes et marges gaussiennes (**Exemple 3**)

— : dépendance empirique (estimée sur l'échantillon), — — : dépendance théorique par CMCa-BI, * — * : dépendance théorique par CMCa

les modèles CMCa et CMCa-BI. Tout d'abord en raison d'un plus grand nombre de paramètres pour les CMCa que pour les CMCa-BI, nous avons une plus grande variance des estimateurs (voir tableau 5.4) ce qui donne des performances moins bonnes pour la segmentation non-supervisée. Cependant lorsque la dépendance est forte ($\rho > 0,5$), le vrai modèle CMCa devient meilleur que le modèle CMCa-BI.

Si nous comparons les racines carrées des écarts quadratiques moyens des estimateurs, incorporant ainsi le biais de l'estimateur, nous constatons que ceux-ci sont beaucoup plus importants pour les CMCa-BI que pour les CMCa dès que $\rho = 0,4$ (pour les paramètres de la classe 2), voir tableau 5.5. Comme nous l'avons déjà observé dans la section précédente au sujet du mélange de 3 lois normales centrées (cf. tableau 5.1), nous avons une estimation biaisée des paramètres de la loi stationnaire lorsque nous avons une dépendance conditionnelle.

Classe	CMCa-BI		CMCa	
	$\hat{m}$	$\hat{\sigma}^2$	$\hat{m}$	$\hat{\sigma}^2$
1	0	0,56	-0,01	0,5
2	-1,73	1,83	-0,05	1,92
3	1,77	1,83	0,25	3,25

TAB. 5.1 – Paramètres estimés des marges gaussiennes (**exemple 3**)

ρ	CMCa-BI	CMCa
0	13,4	
0,2	15,8	15,6
0,4	17,7	16,8
0,6	19,5	17,5
0,8	22	17,5

TAB. 5.2 – Taux d'erreur supervisée en fonction de la corrélation

Nous nous intéressons maintenant à un mélange stationnaire obtenu à partir d'un champ de Markov stationnaire obtenue de manière similaire à l'**exemple 2**, que nous appelons **exemple 4**. Nous ajoutons par rapport à l'**exemple 2** une troisième classe dont la loi d'émission est $\gamma(2, 1)$ (tout en conservant le même paramètre α pour le champ markovien des classes). Dans ce cas-là, nous obtenons alors les résultats du tableau 5.6. Le modèle CMCa permet d'estimer correctement les paramètres des marges, alors que le modèle CMCa-BI fournit une estimation biaisée des paramètres d'émission. Pour la segmentation non-supervisée, nous obtenons alors un taux d'erreur de 15,4 % dans le cas CMCa, contre 22,9% dans le cas CMCa-BI, voir figure 5.6.

5.3.3.3 Segmentation de données réelles

Nous segmentons des données réelles issus d'un radar à impulsions Doppler émettant en bande X (i.e. un domaine de fréquence de 8-12 GHz, soit une longueur d'onde de l'ordre de 5 cm, voir section 6.1). Le signal est constitué de 256 cases distances (observations), et chaque observation $y_n = (y_n^1, \dots, y_n^{16})$ est un vecteur complexe de dimension 16 (i.e. la rafale contient 16 récurrences). Nous faisons une cartographie de la radiale en exploitant uniquement l'information de réflectivité et nous calculons l'amplitude moyenne reçue en chaque case $A_n = \frac{1}{16} \sum_{k=1}^{16} |y_n^k|$ pour une émission et une réception horizontale en polarimétrie. Le signal reçu est présenté dans la figure (5.7). Nous connaissons la vérité terrain pour cette rafale, et celle-ci est caractérisée par la variabilité physique du sol (dont nous donnons la composition dans le tableau (5.7)), mais aussi par les conditions météorologiques. Nous avons sur cette rafale deux fouillis de pluie (avec des vitesses radiale entre -5 et -12 m/s), qui sont étendus et possèdent une forte réflectivité. La position de ces deux fouillis est indiquée sur la figure (5.7), et correspond aux observations 1 à 131 et 241 à 256 (la localisation des fouillis a été effectuée après filtrage Doppler et obtention de la carte distance-Doppler de la rafale par une transformée de Fourier discrète sur 128 récurrences, voir le chapitre 5 et la figure 6.5). La pluie marque fortement le profil des réflectivités de la radiale, et nous avons essentiellement trois niveaux de réponse :

forte pluie + relief ou infrastructure

ρ	CMCa-BI	CMCa
0	14,4	14,8
0,2	16,3	17
0,4	18,7	18,9
0,6	20,4	20,4
0,8	23,6	21,8

TAB. 5.3 – Taux d’erreur non-supervisée en fonction de la corrélation

Corrélation		CMCa-BI		CMCa	
	classe	$\sigma_{\hat{a}}$	$\sigma_{\hat{b}}$	$\sigma_{\hat{a}}$	$\sigma_{\hat{b}}$
0	1	0,06	0,06	0,06	0,09
	2	0,23	0,07	0,23	0,08
0,2	1	0,08	0,08	0,08	0,11
	2	0,27	0,09	0,037	0,12
0,4	1	0,08	0,07	0,09	0,12
	2	0,38	0,09	0,47	0,14
0,6	1	0,1	0,07	0,12	0,17
	2	0,47	0,09	0,53	0,17
0,8	1	0,14	0,1	0,14	0,3
	2	0,76	0,11	0,67	0,26

TAB. 5.4 – Ecart-type des estimateurs

moyenne pluie + habitat dispersé ou verdure ou relief

faible temps sec + verdure

Pour la segmentation de cette rafale, nous utilisons un modèle CMCa et un modèle CMCa-BI, avec des lois d’émission γ . En raison du faible nombre d’observations et de transitions, le modèle CMCa que nous estimons est contraint à être tel que $\forall k \neq l, c(u, v; \rho_{kl}) = c^\perp$: dans le cas des copules gaussiennes, ceci se traduit par la condition $\forall k \neq l, \rho_{kl} = 0$.

Par minimisation du critère BIC (voir section 3.3.2), nous déterminons le nombre de classes pour les modèles CMCa et CMCa-BI (en faisant varier de 2 à 5 classes) : nous sélectionnons un modèle CMCa à 3 classes et un modèle CMCa-BI à 4 classes, voir tableau (5.8). Comme nous l’avons indiqué dans la section 3.3.2, le critère BIC⁵ pénalise plus la complexité que le critère AIC et a tendance à sélectionner des modèles plus parcimonieux (le critère AIC sélectionne un modèle à 5 classes pour CMCa et CMCa-BI). Nous pouvons constater que le terme de pénalisation croît suffisamment rapidement avec les données pour avoir un critère convexe à minimiser (ce qui n’est pas le cas de AIC qui oscille pour le modèle CMCa). De plus, les critères AIC et BIC font préférer le modèle CMCa au modèle CMCa-BI, ce qui constitue un élément supplémentaire pour une meilleure adéquation aux données par la prise en compte d’une dépendance conditionnelle (au moyen de

⁵Nous avons vu que le critère BIC, et de manière générale les critères de vraisemblance pénalisée sous certaines conditions, permettent de sélectionner le bon nombre de classes dans un modèle CMCa si celui-ci est β -mélangeant (voir remarque 3.3.1). Nous n’utilisons pas ici l’EMV dans le calcul de la vraisemblance pénalisée, mais l’estimateur de la projection de la fonction estimante IFM, en considérant qu’ils sont assez proches l’un de l’autre pour que cette approximation suffise à sélectionner le nombre de classes.

Corrélation		CMCa-BI		CMCa	
	classe	MSE($\hat{a}$)	MSE($\hat{b}$)	MSE($\hat{a}$)	MSE($\hat{b}$)
0	1	0,045	0,045	0,0046	0,009
	2	0,051	0,006	0,054	0,007
0,2	1	0,0075	0,007	0,0075	0,013
	2	0,14	0,012	0,15	0,014
0,4	1	0,0134	0,008	0,083	0,015
	2	0,37	0,026	0,25	0,02
0,6	1	0,025	0,014	0,014	0,033
	2	0,64	0,04	0,32	0,03
0,8	1	0,05	0,026	0,02	0,1
	2	1,66	0,073	0,46	0,07

TAB. 5.5 – Racine carrée des écarts quadratiques moyens

Classe	CMCa-BI		CMCa	
	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$
1	1,1	0,35	0,98	0,48
2	2,6	0,77	1,99	0,97
3	4,45	1,46	3	1,9

TAB. 5.6 – Paramètres estimées pour les marges gamma dans le champ 3

la copule gaussienne). Ceci est confirmé par la comparaison des lois stationnaires théoriques des CMCa et CMCa-BI avec la loi stationnaire des données, voir figure (5.8) ainsi que par la structure de dépendance estimée voir figure (5.9). La densité $p(y_1, y_2)$ est mieux estimée dans le modèle CMCa que CMCa-BI. L'examen de la dépendance du processus (fonction d'autocorrélation, tau de Kendall et rho de Spearman) montre que le modèle CMCa-BI décrit mieux la dépendance d'ordre 2 (i.e. la loi de Y_1, Y_3), mais nous avons au-delà des estimations proches l'une de l'autre pour les modèles CMCa-BI et CMCa (excepté pour le taux de Kendall). Les estimations des paramètres des copules sont $\hat{\rho}_{11} = 0,9$, $\hat{\rho}_{22} = 0,7$ et $\hat{\rho}_{33} = 0,64$, ce qui montrent un écart fort par rapport au modèle CMCa-BI.

Pour finir les cartes obtenues pour les deux rafales sont rassemblées dans le tableau (5.9). Les classes 1 du modèle CMCa et CMCa sont identiques et correspondent à une réflectivité faible et à la zone sans pluie. La classe 3 du modèle CMCa et la classe 4 du modèle CMCa-BI sont

Cases	Type
1-57	aérodrome
58-74	carrière
77	route
80-104	vignes-cultures
104	route
104-153	village-habitat dispersé
154-200	vignes-cultures
201-256	reliefs-hauteurs

TAB. 5.7 – Composition physique de la rafale

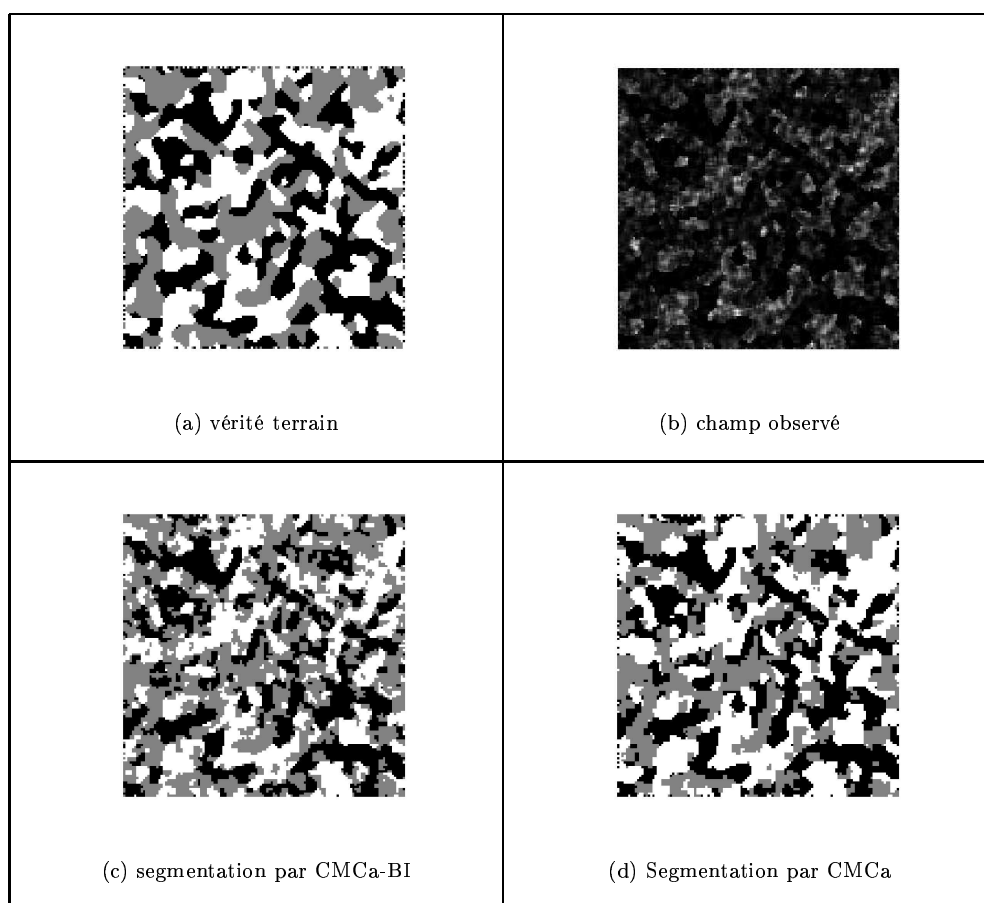


FIG. 5.6 – Champ couple à 3 classes et marges gamma

presque identiques (le modèle CMCa donnent des zones un peu plus étendues) et correspondent au zone de très forte réflectivité en raison de la pluie et d'obstacle très réfléchissant. La différence essentielle entre les deux modèles réside dans la segmentation de la classe des valeurs moyennes, correspondant à la classe 2 du modèle CMCa et qui a été séparée en 2 classes par le modèle CMCa-BI. Le village est séparé entre les classes 2 et 3 dans le modèle CMCa-BI alors qu'il ne l'est pas dans le modèle CMCa. Il en est de même pour la zone de hauteur en fin de radiale (correspondant à la zone 201-256).

En conclusion, le modèle CMCa apparaît comme un modèle pertinent pour décrire le processus des amplitudes sur les données traitées. Il permet d'une part de modéliser un environnement

critères	AIC		BIC	
	CMCa-BI	CMCa	CMCa-BI	CMCa
nombre de classes				
2	615,2	289	636,5	317,5
3	220,7	77,2	263,3	130,4
4	167,1	83,2	238	168,3
5	150,4	71,5	256,7	195,6

TAB. 5.8 – Critère d'information pour la sélection de modèles

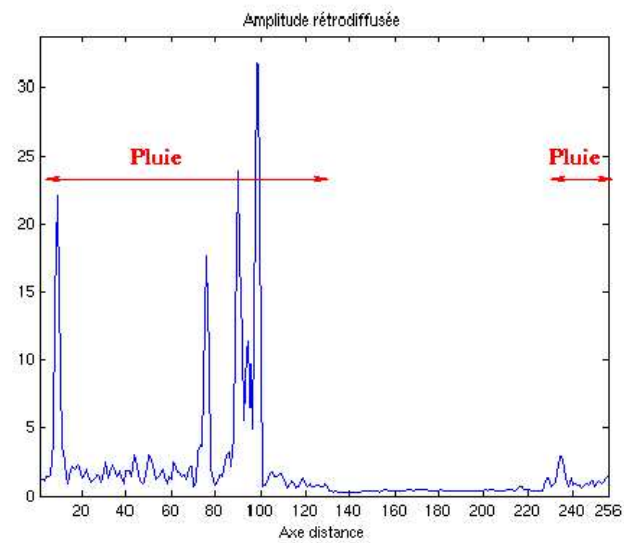


FIG. 5.7 – Amplitudes reçues sur le long de la rafale, les numéros des cases-distance sont en abscisse

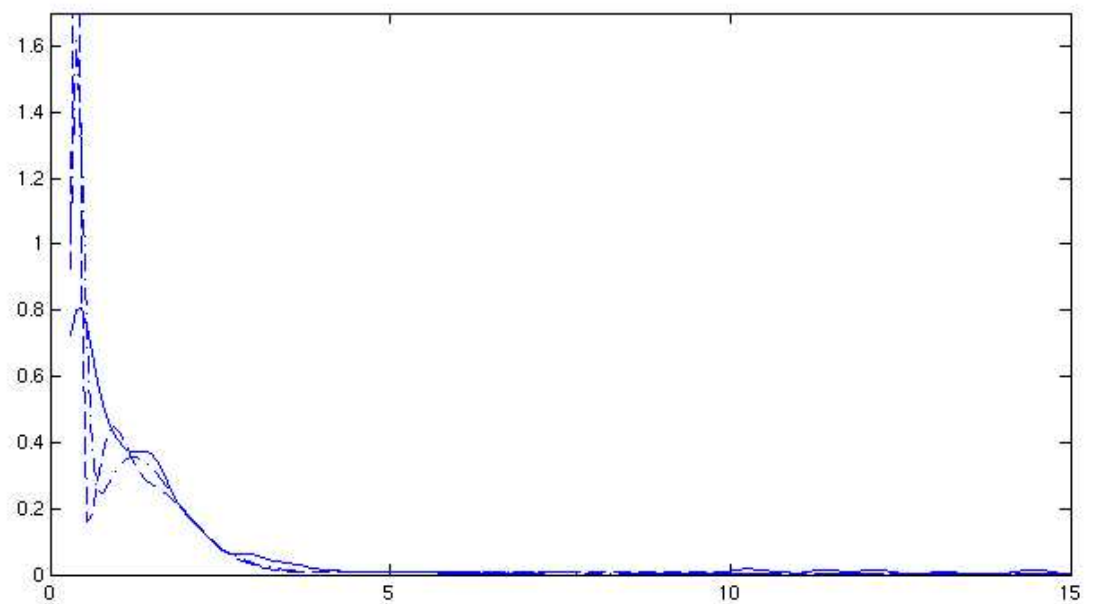


FIG. 5.8 – Densité stationnaire estimée $p(y_n, \phi)$.
 — : estimation de la densité par la méthode des noyaux (noyau gaussien et fenêtre choisie par plug-in), —· : densité théorique par CMCa (mélange de 3 densités gamma), — — : densité théorique par CMCa-BI (mélange de 4 densités gamma)

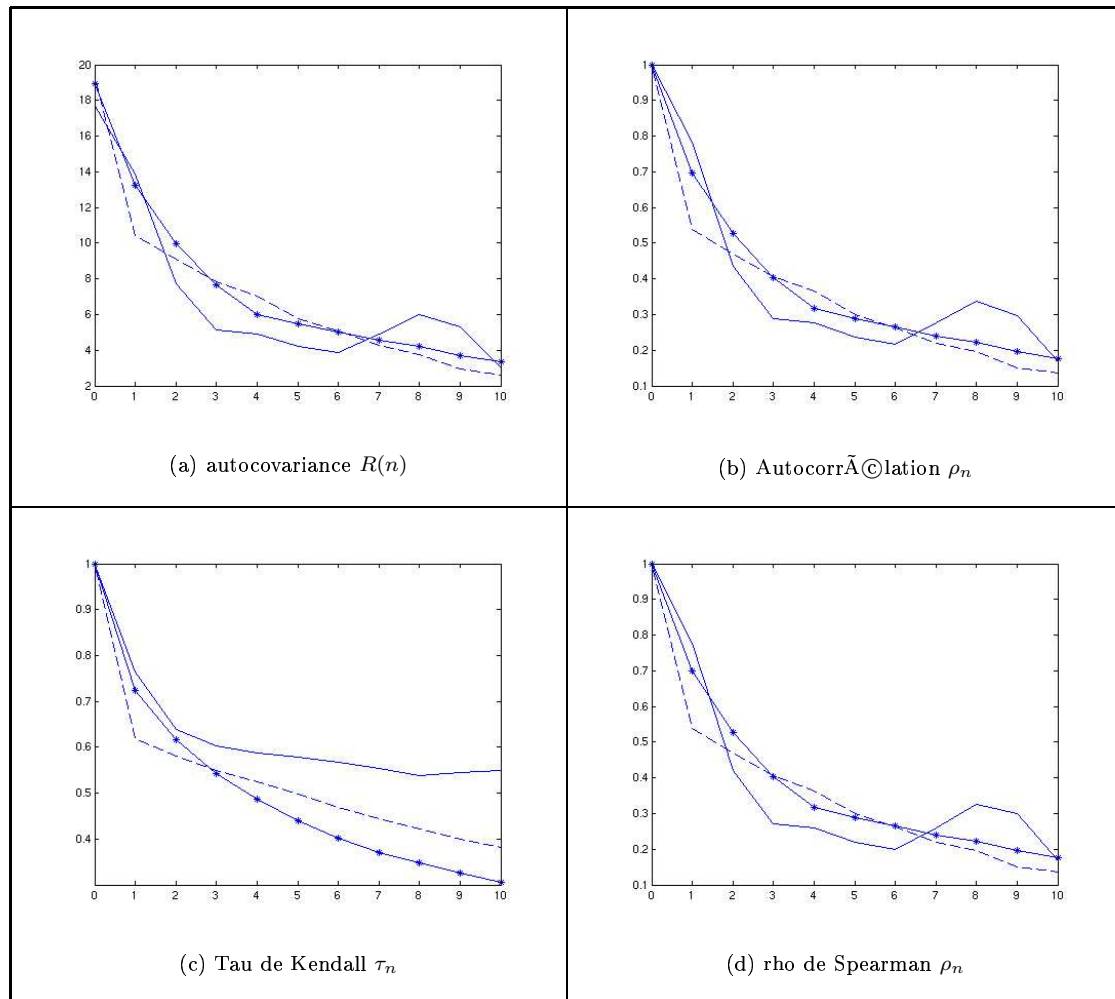


FIG. 5.9 – Structure de d pendance estim e

— : d pendance empirique (estim e sur l' chantillon), -- : d pendance th orique par CMCa-BI,
 * - * : d pendance th orique par CMCa

Localisation	CMCa-BI	CMCa
Classe 1	131-214, 218-226	131-214, 218-226
Classe 2	111-130, 215-217 227-232, 238-254	1-6, 13-69 78-87, 101-130 215-217, 227-256
Classe 3	1-6, 12-72 78-87, 101-110 233-237, 255-256	7-12, 70-77 88-100
Classe 4	7-11, 73-77 88-100	

TAB. 5.9 – Cartes obtenues par segmentation bay sienne

hétérogène, composé de zones aux propriétés de réflectivité très différentes, et de tenir compte de la corrélation de l'intensité en distance au sein des zones homogènes. La forte valeur des paramètres des copules gaussiennes et la bonne estimation des propriétés de dépendance de $p(y_1, y_2)$ sont de fortes indications en faveur de l'existence de cette dépendance, et pour le rejet du modèle CMCa-BI. Ce modèle de corrélation de l'intensité est très proche d'un modèle déjà été introduit par Watts dans [152, 153] pour la modélisation de la texture d'un signal radar (voir section 4.2.1). Watts *et al.* suppose que la texture est un processus ayant des marges gamma et une dépendance markovienne modélisée par une copule gaussienne. Nous avons deux innovations par rapport à ces travaux : nous introduisons un processus markovien caché $\mathbf{X}$ modélisant l'hétérogénéité de l'environnement, et nous estimons les paramètres du modèle ainsi que le nombre de classes (Watts *et al.* l'utilise seulement pour la génération de données).

5.4 Chaînes de Markov couples

L'utilisation des CMCo pour la segmentation des signaux impliquent une modification de l'interprétation de la classification. En effet, nous savons que si $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$ est une chaîne de Markov stationnaire, sa loi est décrite par les densités $p(x_1, x_2)$ et $p(y_1, y_2 | x_1, x_2)$. En toute généralité, nous avons $p(y_1 | x_1, x_2) \neq p(y_1 | x_1)$ et $p(y_2 | x_1, x_2) \neq p(y_2 | x_2)$, ce qui implique que les lois d'émission classiques $p(y_1 | x_1)$ et $p(y_2 | x_2)$ ne constituent plus les "briques élémentaires" du modèle de classification, mais sont à leur tour un mélange de lois :

$$\begin{aligned} \forall x_1, p(y_1 | x_1) &= \sum_{x'_2} p(x'_2 | x_1) p(y_1 | x_1, x'_2) \\ \forall x_2, p(y_2 | x_2) &= \sum_{x'_1} p(x'_1 | x_2) p(y_2 | x'_1, x_2) \end{aligned} \quad (5.27)$$

et nous pouvons dire finalement que la densité $p(y_n)$ est un mélange de mélange de lois. Ainsi, l'approche du modélisateur doit évoluer car il doit formuler sa connaissance de la réalité physique en décomposant la densité $p(y_1 | x_1)$ comme le mélange de plusieurs phénomènes. Il s'agit de traduire les différences de comportement entre zones homogènes (cas où $x_1 = x_2$) et différentes zones frontières (cas où $x_1 \neq x_2$). Ce raffinement peut être handicapant car il nécessite d'avoir des informations souvent inaccessibles sur les phénomènes considérés. Alternativement, nous pouvons voir cette approche comme une modélisation non-paramétrique des lois $p(y_1 | x_1)$. Mais contrairement à un modèle CMCa-BI pour lequel les lois d'émission seraient représentées par un mélange de lois, i.e. $p(y_1 | x_1) = \sum_{k=1}^M \lambda_k(x_1) f_k(y_1, x_1)$, les poids du mélange considéré dans les CMCo dépendent non seulement de l'état courant, mais aussi de l'état précédent.

Pour les CMCo, les modèles proposés seront construits à partir des densités $p(y_1, y_2 | x_1, x_2)$ et $p(x_1, x_2)$, cependant ces densités ne peuvent être choisies de manière indépendantes l'une de l'autre car nous devons respecter des contraintes, dues à la stationnarité du processus. Nous explicitons alors le choix de contraintes que nous avons fait, et nous donnons alors la forme générale des modèles paramétriques utilisés. Une méthode d'estimation, toujours basée sur les fonctions estimantes est proposée, et nous traitons quelques exemples.

5.4.1 Le modèle CMCo stationnaire

Dans la mesure où le processus $\mathbf{Z}$ est markovien, sa stationnarité est équivalente à l'égalité des densités $p(y_1, x_1)$ et $p(y_2, x_2)$. Ceci implique que les densités $p(x_1, x_2)$ et $p(y_1, y_2 | x_1, x_2)$ doivent vérifier

$$\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, \sum_{x_2} p(x, x_2) p_g(y | x, x_2) = \sum_{x_1} p(x_1, x) p_d(y | x_1, x) \quad (5.28)$$

Si nous choisissons librement la loi $p(x_1, x_2)$, nous devons choisir des modèles pour $p(y_1, y_2 | x_1, x_2)$ telles qu'ils soit possible de contrôler les marges $\{p_g(y | x_1, x_2)\}$ et $\{p_d(y | x_1, x_2)\}$, et d'assurer de plus que ces marges vérifient la contrainte (5.28).

Pour le premier point, il est possible d'utiliser les copules pour représenter la densité jointe $p(y_1, y_2 | x_1, x_2)$ (dont nous notons la densité de copule c_{x_1, x_2}) si nous nous restreignons au cas des observations réelles. Pour le deuxième point, nous supposons que nous avons

$$\forall y \in \mathcal{Y}, \forall k, l, \begin{cases} p_g(y | k, l) &= p_d(y | l, k) \\ p(k, l) &= p(l, k) \end{cases} \quad (5.29)$$

Sous ces hypothèses de symétrie, la contrainte (5.28) est donc toujours vérifiée. Nous supposons alors que les marginales $p_g(y | k, l)$ appartiennent aux modèles $\{f(y, \theta_{kl}), \theta_{kl} \in \Theta\}$ et que les c_{kl} appartiennent aux modèles $\{c_{\eta_{kl}}(\cdot, \cdot), \eta_{kl} \in \Upsilon\}$. Nous considérons donc des CMCo stationnaires paramétrées par une matrice de probabilités jointes symétriques $P = (p_{kl}) \in \mathbb{R}^{K \times K}$, les paramètres des marges gauches $\theta = (\theta_{kl}) \in \Theta^{K \times K}$ et les paramètres des copules $\eta = (\eta_{kl}) \in \Upsilon^{K \times K}$. Nous notons $\phi = (P, \theta, \eta) \in \Phi$. La loi de Z_1 s'écrit donc

$$p(z_1, \phi) = \sum_{x_2'} p_{x_1 x_2'} f(y_1, \theta_{x_1 x_2'}) \quad (5.30)$$

et la loi jointe de (Z_1, Z_2) a pour expression

$$p(z_1, z_2, \phi) = p_{x_1 x_2} f(y_1, \theta_{x_1 x_2}) f(y_2, \theta_{x_2 x_1}) c_{\eta_{x_1 x_2}}(F(y_1, \theta_{x_1 x_2}), F(y_2, \theta_{x_2 x_1})) \quad (5.31)$$

Remarque 5.4.1. *Nous utiliserons pour les marginales des modèles paramétriques simples tels que des lois gaussiennes ou gamma (avec Θ de petites dimensions), mais qui permettent de reproduire des lois complexes pour les lois d'émission $p(y_1 | x_1)$.*

5.4.2 Estimation des modèles couples

5.4.2.1 Estimation avec données complètes

De manière similaire aux CMCa, nous proposons une fonction estimante pour estimer les paramètres du modèle complet. Afin d'obtenir des formules d'estimation simple, nous construisons la fonction estimante de sorte que les estimations des paramètres P , θ et η soient découplées : ainsi

nous n'avons pas à envisager des méthodes numériques de recherche de racines dans des espaces de grande dimension. Nous notons $L_c(\phi) = \log p(\mathbf{z}, \phi)$ pour la log-vraisemblance du processus complet (définie à partir de (5.31)). La fonction $\tilde{G}$ que nous utilisons est

$$\begin{aligned} \tilde{G} : \mathcal{Z}^N \times \Phi &\longrightarrow \phi \\ (\mathbf{z}, \phi) &\longmapsto (G_1(\mathbf{z}, P), G_2(\mathbf{z}, P, \theta), G_3(\mathbf{z}, P, \theta, \eta)) \end{aligned}$$

où nous avons

$$\begin{cases} G_1(\mathbf{z}, P) &= P - \hat{P} \\ G_2(\mathbf{z}, P, \theta) &= \nabla_{\theta} L_{c, m_g} \\ G_3(\mathbf{z}, P, \theta, \eta) &= \nabla_{\eta} L_c \end{cases} \quad (5.32)$$

$\hat{P}$ est l'estimateur empirique symétrisé de la matrice de probabilités jointes

$$\forall k, l, \hat{P}_{kl} = \frac{1}{2(N-1)} \sum_{i=1}^{N-1} (1_{kl}(x_i, x_{i+1}) + 1_{lk}(x_i, x_{i+1})) \quad (5.33)$$

L_{c, m_g} est la log-vraisemblance de la marginale gauche (m_g) de la loi jointe (Y_1, Y_2, X_1, X_2) , c'est à dire de la loi de (Y_1, X_1, X_2) . Nous avons donc $L_{c, m_g} = \sum_{i,k,l=1}^{N,K,K} 1_{k,l}(x_i, x_{i+1}) \log(p_{kl} f(y_i, \theta_{kl}))$, et la fonction estimante correspondante est alors

$$G_2(\mathbf{z}, P, \theta) = \left(\sum_{i=1}^N 1_{k,l}(x_i, x_{i+1}) \nabla_{\theta_{kl}} \log f(y_i, \theta_{kl}) \right)_{1 \leq k, l \leq K} \quad (5.34)$$

Enfin la fonction estimante G_3 vaut

$$G_3(\mathbf{z}, P, \theta, \eta) = \left(\sum_{i=1}^N 1_{k,l}(x_i, x_{i+1}) \nabla_{\eta_{kl}} \log c_{\eta_{kl}}(F(y_i, \theta_{kl}), F(y_{i+1}, \theta_{lk})) \right)_{1 \leq k, l \leq K} \quad (5.35)$$

La fonction $\tilde{G}$ est une modification de la fonction estimante IFM, qui permet d'avoir une estimation séquentielle des paramètres P , Θ puis η . Cependant par définition de G_1, G_2, G_3 , nous avons toujours $E_{\phi} [\tilde{G}(\mathbf{Z}, P, \theta, \eta)] = 0$.

Remarque 5.4.2. La résolution de l'équation $\tilde{G}(\phi) = 0$ peut se voir comme un "pivot de Gauss" pour résoudre le système $\nabla_{\phi} L_c(\phi) = 0$ correspondant au maximum de vraisemblance.

5.4.2.2 Estimation avec données incomplètes

Lorsque nous n'observons que le processus $\mathbf{Y}$, nous utilisons alors la fonction estimante $\tilde{g}$ obtenue par projection de $\tilde{G}(\phi)$

$$g(\mathbf{y}, \phi) = E_{\phi} [\tilde{G}(\mathbf{Z}, \phi) | \mathbf{y}]$$

Pour déterminer une racine de l'équation, nous utilisons l'algorithme ECI, ce qui donne les formules suivantes de mises à jour des paramètres sont

$$\forall k, l, \begin{cases} p_{kl}^{(n+1)} = \frac{1}{2(N-1)} \sum_{i=1}^{N-1} \left(p_i^{(n)}(k, l) + p_i^{(n)}(l, k) \right) \\ \theta_{kl}^{(n+1)} \text{ tel que } \sum_{i=1}^N p_i^{(n)}(k, l) \nabla_{\theta_{kl}} f(y_i, \theta_{kl}) = 0 \\ \eta_{kl}^{(n+1)} \text{ tel que } \sum_{i=1}^N p_i^{(n)}(k, l) \nabla_{\eta_{kl}} c_{\eta_{kl}}(F(y_i, \theta_{kl}^{(n+1)}), F(y_{i+1}, \theta_{lk}^{(n+1)})) = 0 \end{cases} \quad (5.36)$$

Nous pouvons aussi proposer une version stochastique de l'algorithme ECI en simulant selon la loi a posteriori le processus $\mathbf{X}$, et en résolvant directement le système $\tilde{G}((\mathbf{x}, \mathbf{y}), \phi) = 0$. Dans ce cas là, l'intérêt pratique de cet algorithme apparaît clairement : outre la simplicité des calculs impliqués pour obtenir le système à résoudre (nous n'avons pas besoin de calculer les dérivées $\nabla_{\phi} p(y_1, y_2 | x_1, x_2, \phi)$), nous avons aussi une propriété de modularité de l'algorithme. Il est possible d'utiliser les formules (et donc les programmes déjà écrits) du maximum de vraisemblance dans le cas d'une seule classe. De plus le changement de modèle paramétrique pour une marge ou une copule ne vient pas modifier le système à résoudre pour les autres paramètres.

5.4.3 Expérimentations

Nous montrons un exemple de segmentation de chaînes couples, et comparons les performances de segmentation entre modèle CMCa et modèle CMCo. Nous simulons une chaîne CMCo à deux états, tels que $p_d(y | 1, 1)$ soit une loi $\gamma(3; 3)$, $p_d(y | 1, 2)$ soit une loi $\gamma(2; 2, 5)$. Les distributions marginales $p_d(y | 2, 1)$ et $p_d(y | 2, 2)$ sont respectivement des lois de Weibull⁶ $W(1, 2)$ et $W(1, 1)$. Nous supposons de plus que la matrice de probabilités jointes du processus caché est $P = \begin{bmatrix} 0,4 & 0,15 \\ 0,15 & 0,4 \end{bmatrix}$, et que les copules sont gaussiennes. Nous faisons varier comme précédemment l'intensité de la dépendance en modifiant les paramètres des copules. Nous comparons alors sur deux configurations les performances en segmentation non-supervisée des CMCo et des CMCa ($N = 500$) :

cas 1 $\rho_{11} = \rho_{22} = 0,1$ et $\rho_{12} = \rho_{21} = 0,5$

cas 2 $\rho_{11} = \rho_{22} = 0,5$ et $\rho_{12} = \rho_{21} = 0,8$

Il est délicat de comparer les modèles CMCa et CMCo parce que nous ne pouvons pas comparer la loi de Y_n dans l'un et l'autre modèle. Cependant, nous proposons pour le modèle CMCa une loi γ pour la classe 1, et une loi de Weibull pour la classe 2. Pour évaluer les différences entre ces modèles dans le cas non-supervisée, nous avons calculé les taux d'erreur dans le cas supervisé des 3 modèles CMCo, CMCa et CMCa-BI. Pour ces deux derniers modèles nous estimons les modèles CMCa approchés en utilisant la vérité terrain et des estimateurs à données complètes. Ceci nous permet d'avoir une estimation de la capacité à décrire le processus CMCo simulé. Nous avons alors des performances équivalentes pour les modèles CMCa et CMCa-BI, même si la dépendance du processus dans la configuration 2 est mieux capturée par le modèle CMCa.

Les taux d'erreur se dégradent fortement pour les modèles CMCa et CMCa-BI en mode non-supervisée mais il reste cependant proches. Le taux d'erreur diminue avec l'augmentation de la

⁶Une variable aléatoire positive suit une loi de weibull, notée $W(a, b)$ si sa densité de probabilité relativement à la mesure de Lebesgue est $g(y; a, b) = \frac{b}{a^b} y^{b-1} e^{-(\frac{y}{a})^b} 1_{\mathbb{R}^+}(y)$.

	CMCo		CMCa		CMCa-BI	
Cas	supervisée	non-supervisée	estimée avec apprentissage	non-supervisée	estimée avec apprentissage	non-supervisée
1	5,3	6,4	7,9	12,5	7,7	12,3
2	2,5	3,4	7	11	7,7	11,9

TAB. 5.10 – Taux d’erreur non-supervisée des chaînes couples

FIG. 5.10 – Champ de Markov couple à 2 classes et marges gamma (**exemple 2**) segmentée par une chaîne de Markov couple à 2 classes et marges gamma

dépendance, mais il reste bien supérieur au taux d’erreur non-supervisée de la vraie CMCo.

Dans cette situation, nous pouvons conclure que la différence majeure entre CMCo et CMCa-BI tient plus à la modélisation plus riche des densités marginales $p(y_n, \phi)$ qu’à l’introduction de la dépendance conditionnelle : même si le modèle CMCa la prend en compte, il n’est pas capable de reproduire la loi du processus CMCo significativement mieux que le processus CMCa-BI.

Nous segmentons avec une CMCo l’image traitée dans l’exemple 2 qui est la réalisation d’une champ de Markov Couple, dont les marges sont de type γ (voir section 5.3.3.1). Dans ce cas là, nous obtenons un taux d’erreur de 4% (contre 4,2% pour le modèle CMCa, et 9,9% pour le modèle CMCa-BI), et l’image segmentée est représentée dans la figure (5.10).

Paramètres estimés ($a\ b$)	(1 1)	(1 2)	(2 1)	(2 2)
CMCo	(0,95 1,1)	(1,5 0,4)	(4 1,5)	(2,2 2,5)
CMCa	(0,99 1)		(2,89 2,1)	
CMCa-BI	(1 0,9)		(2,6 2,1)	

TAB. 5.11 – Estimation des marges

Coefficients de corrélation	(1 1)	(1 2)	(2 1)	(2 2)
CMCo	0,7	0,5	0,8	0,65
CMCa	0,67	0,43	0,37	0,67

TAB. 5.12 – Estimation des copules gaussiennes

Probabilités	(1 1)	(1 2)	(2 1)	(2 2)
CMCo	0,61	0,01	0,01	0,35
CMCa	0,63	0,02	0,02	0,33
CMCa-BI	0,6	0,015	0,015	0,37

TAB. 5.13 – Estimation des probabilités jointes de (X_1, X_2)

Le modèle CMCa (et dans une moindre mesure le modèle CMCa-BI) donnent une bonne approximation des lois marginales. Le modèle CMCo sépare quant à lui chaque marge de $p(y_1, y_2)$ sous la forme d'un mélange de deux gamma : les composantes $p_d(y|1, 1)$ et $p_d(y|2, 2)$ ont des coefficients proches des vrais valeurs, mais ceux-ci tiennent compte du fait que la loi est modifiée lorsque nous avons des zones frontières et de la présence des densités $p_d(y|1, 2)$ et $p_d(y|2, 1)$. Les estimations des structures de dépendance (probabilités jointes de l'état caché et copules) sont tout à fait similaires (exceptées pour les paramètres de la copule de la densité $p(y_1, y_2|x_1 = 1, x_2 = 2)$). L'utilisation des densités $p_d(y|x_1, x_2)$ dans cette modélisation CMCo permet de prendre en compte la non-markovianité du processus caché dans cet exemple. Malgré cette amélioration du modèle par rapport aux CMCa, nous n'avons pas d'amélioration net des résultats de la segmentation.

Comme le montre [53], les modèles CMCo permettent d'avoir des performances en segmentation supérieure aux CMCa-BI, ce qui provient d'une meilleure modélisation de la dépendance, mais aussi de la densité $p(y_n)$ par un mélange. Le modèle CMCa apparaît alors comme un compromis entre ces deux modèles qui devient proche de l'un ou l'autre en fonction de la présence ou non de dépendance conditionnelle, et de la bonne adéquation de la loi marginale. Nous avons deux difficultés auxquelles nous devons faire face dans le cas CMCo général, qui sont des conséquences de la complexité des modèles, ce qui incite à l'utilisation de modèles CMCa utilisant de "bonnes" lois d'émission. Nous abordons cet aspect dans la section suivante.

5.4.4 Discussion sur la complexité des modèles

Les modèles CMCo deviennent rapidement très complexes, ce qui se traduit d'une part par des difficultés d'interprétation du modèle (et des paramètres estimés), et d'autre part par des difficultés d'estimation. En effet, les paramètres d'effet croisé θ_{kl} et η_{kl} avec $k \neq l$ sont souvent très difficiles à estimer en raison du type de données qui nous intéressent : lorsque nous segmentons une image ou un signal à l'aide de modèles markoviens, l'information spatiale que nous exploitons est basée sur l'a priori de l'existence de zones homogènes. Cela signifie que nous avons des taux de transitions faibles entre états différents (ce que l'on peut constater souvent par l'obtention de matrice de transition à diagonale "largement" dominante), ce qui implique que l'évènement $\{(X_n, X_{n+1}) = (k, l)\}$ avec $k \neq l$ a une probabilité très faible d'arriver. Par conséquent, nous avons peu d'observations pour estimer θ_{kl} et η_{kl} , et ceux-ci ont une très forte variance : il est même possible d'observer dans des versions stochastiques de l'algorithme ECI une disparition complète d'une transition (k, l) est de

se retrouver donc dans l'impossibilité de pouvoir estimer les paramètres correspondants. De plus, l'estimation séquentielle de IFM-ECI peut amplifier ce phénomène parce qu'une erreur importante sur l'estimation des paramètres des marges θ_{kl} aura une influence sur celle des copules η_{kl} , qui seront nécessairement mal estimées. La recherche d'un estimateur est rendue délicate par une faible taille d'échantillon ou un grand nombre de classes. Les CMCo entraînent une explosion du nombre de paramètres, ce qui est une autre cause de l'augmentation de la variance des estimateurs. Nous désignons $|\Theta|$ pour le nombre de paramètres estimés pour chaque densité modélisée ($p(y_1|x_1)$ pour les CMCa et $p(y_1|x_1, x_2)$ pour les CMCo) et $|\Upsilon|$ pour le nombre de paramètres estimés pour les copules. Le nombre global de paramètres estimés dans un modèle CMCa est $|\Phi_{\text{CMCa}}| = K \times (K - 1) + K \times |\Theta| + K^2 \times |\Upsilon|$ (et en particulier $|\Phi_{\text{CMCa-BI}}| = K \times (K - 1) + K \times |\Theta|$), alors que celui d'un modèle CMCo est $|\Phi_{\text{CMCo}}| = (K^2 - 1) + K^2 \times |\Theta| + K^2 \times |\Upsilon|$. La complexité du modèle varie donc en le carré du nombre de classes pour les modèles CMCo et CMCa, ce qui montre que l'estimation de modèles avec de nombreuses classes nécessite énormément de données. Si nous recherchons 5 classes avec les hypothèses de l'exemple précédent, nous avons $|\Theta| = 2$ et $|\Upsilon| = 1$ ce qui nous donne $|\Phi_{\text{CMCa-BI}}| = 30$, $|\Phi_{\text{CMCa}}| = 55$ et $|\Phi_{\text{CMCo}}| = 99$: cette augmentation du nombre de paramètres pose bien évidemment le problème de l'interprétation du modèle obtenu. Pour lutter contre cette augmentation de la complexité, il est possible de rajouter des contraintes telles $\theta_{kl} = \theta_k$ et $\rho_{kl} = 0$ si $k \neq l$. La première permet de passer du modèle CMCo au modèle CMCa, la deuxième permet de limiter (dans le cas CMCo) ou d'éliminer (dans le cas CMCa) les problèmes d'estimation dus aux faibles nombres de changement d'états. Sous ces deux contraintes, le modèle CMCa a une complexité égale à $K \times (K - 1) + K \times (|\Theta| + |\Upsilon|)$, qui reste du même ordre de grandeur que celle du modèle CMCa-BI. Ce modèle CMCa contraint est particulièrement intéressant en raison de sa parcimonie et de sa capacité à capturer des caractéristiques essentielles du processus observé.

Chapitre 6

Segmentation Doppler et polarimétrique de l'environnement Radar

Un radar (actif) est un système qui émet une onde électromagnétique de forme connue et qui reçoit les échos renvoyés par les différents obstacles ou cibles du paysage. Comme son nom l'indique, le radar (*Radio Detection And Ranging*) a parmi ses objectifs la détection des cibles et l'estimation de certaines de leur caractéristiques, comme leur position et leur vitesse. Les mesures effectuées par ce système sont perturbées aléatoirement par des parasites d'origines diverses :

- le bruit des récepteurs ;
- les parasites dus à l'environnement (brouilleurs, réflexions sur le paysage générant un fouillis) ;
- les distorsions dues à la propagation ;
- la rotation d'antenne (lorsque celle-ci est en mouvement).

L'objectif des traitements effectués sur les signaux reçus est alors de diminuer, voire d'éliminer ces perturbations afin de maximiser le rapport signal à bruit et les probabilités de détection, ainsi que la précision des estimateurs des caractéristiques des cibles. Les progrès techniques réalisés depuis une cinquantaine d'années ont permis le développement de radars capables de fournir une image de l'environnement, qui exploitent précisément le bruit issu du paysage (i.e. le fouillis) et donne de multiples informations sur le sol, la végétation, l'atmosphère, la mer,...

Ainsi aux radars classiques de surface tels que les radars de champ de bataille, de défense aérienne, côtier, météo, naval, se sont ajoutés les radars aéroportés ou satellite dits à synthèse d'ouverture (*Synthetic Aperture Radar*, SAR) qui permettent d'obtenir des images de l'environnement terrestre, exploitant les nombreux avantages du signal radar [106] :

- sa richesse physique : intensité rétrodiffusée, effet Doppler, polarimétrie, interférométrie ;
- son insensibilité aux conditions atmosphériques ;
- ses capacités de discrimination des matériaux.

Par des techniques de traitement d'image, ces images permettent de fournir des informations exploitables dans les applications militaires, ou des domaines tels que la météorologie, l'océanographie, l'agriculture, ou l'exploration pétrolière.

Ces techniques, entre autres la segmentation bayésienne, ont été rarement appliquées pour les radars classiques de surface qui donnent des images unidimensionnelles. Nous utilisons dans cette partie les algorithmes de segmentation développés dans les chapitres précédents pour proposer une méthodologie de cartographie de l'environnement de radars de surveillance exploitant les informations Doppler ou polarimétrique. Cependant, la méthodologie proposée pour la cartographie de l'environnement Doppler n'est nullement spécifique au radar, et peut être étendue aux traitements des signaux porteurs de cette information comme ceux obtenues par échographie Doppler (émission d'ultrasons), lidar (*Light Detection and Ranging*, laser pulsé) ou sonar (*Sound Navigation and Ranging*, émission d'ultrasons). Les procédures tiennent compte du format particulier des signaux radar qui sont des images unidimensionnelles de signaux multidimensionnels.

Nous rappelons tout d'abord les principes physiques du radar et décrivons brièvement la chaîne de traitement radar, et les opérations nécessaires à l'obtention des données sur lesquelles sont appliquées les traitements de segmentation. Nous expliquons alors le problème de la détection en radar, et nous faisons le lien avec la segmentation et la détection qui constitue une application potentielle des travaux de cette thèse.

Nous développons alors une méthodologie pour la segmentation de l'environnement Doppler, basée sur une modélisation autorégressive du signal radar. Nous présentons un nouveau modèle CMCa pour la segmentation d'une rafale, et donnons quelques exemples d'applications.

De même que pour l'information Doppler, nous rappelons les propriétés polarimétriques des ondes électromagnétiques, et nous donnons différentes représentations de l'information polarimétrique en retenant la représentation sur la sphère de Poincaré. Nous utilisons alors un modèle CMCa-BI pour la segmentation des données radar et la réalisation d'une cartographie polarimétrique de l'environnement.

L'essentiel de notre travail s'est porté sur le choix d'une "bonne" représentation et séparation des informations Doppler et polarimétriques, et sur leur modélisation probabiliste dans le cadre de la segmentation bayésienne par CMCa. Nous proposons alors des méthodes originales de segmentation Doppler ou polarimétrique, faisant usage des lois de Von Mises - Fisher [107].

6.1 Le signal radar

Nous décrivons dans cette section les principes physiques du radar et les moyens par lesquels les informations de localisation en distance et en vitesse d'une cible sont obtenues. Nous donnons aussi le fonctionnement de la chaîne de traitement radar, le principe de la détection et finalement nous formulons le problème de la cartographie radar en termes de problème de segmentation bayésienne.

6.1.1 Principes et objectifs du radar

Un radar émet une onde électromagnétique, dans une direction (correspondant à un azimut et une élévation donnée, voir figure (6.1)), que l'on représente habituellement sous la forme complexe suivante

$$s_e(t) = u(t)e^{2\pi i f_0 t} \quad (6.1)$$

où $u(t)$ est l'enveloppe complexe du signal émis, et f_0 est la fréquence de l'onde porteuse. Le signal physiquement émis par le radar est la partie réelle de $s_e(t)$, soit $\Re(u(t)e^{2\pi j f_0 t})$. Une cible située à une distance D du radar (fixe) et possédant une vitesse radiale (par rapport au radar) égale à v , réfléchit l'onde électromagnétique de telle sorte que le signal reçu par le radar soit la partie réelle de

$$s_r(t) = Au(t - \frac{2D}{c})e^{2\pi j f_0 t} e^{i\frac{2v}{c}t} \quad (6.2)$$

A est un coefficient d'atténuation complexe (comportant un terme de phase), et c représente la vitesse de la lumière. Le problème de base du radar est donc de mesurer le temps de retour et la fréquence d'un signal afin d'obtenir le retard ΔT et le décalage en fréquence Δf (du à l'effet Doppler) par rapport à l'onde émise, grâce auxquels nous pouvons déduire position et vitesse de la cible :

$$\begin{cases} D &= \frac{c\Delta T}{2} \\ v &= \frac{\lambda\Delta f}{2} \end{cases}$$

$\lambda = \frac{c}{f_0}$ est la longueur d'onde de l'onde émise.

Le signal émis est atténué par les phénomènes de propagation et de réflexion, et le lien entre puissance émise P_e et puissance reçue P_r dépend des propriétés du radar, de l'onde émise, de la cible et de l'environnement :

$$P_r = P_e \frac{G^2 \lambda^2 \sigma_0}{(4\pi)^3 D^4 a} \quad (6.3)$$

G est le gain (en émission et en réception) de l'antenne du radar et a représente les pertes liées à l'absorption des milieux traversés. σ_0 est la Surface Equivalente Radar (SER) de la cible et représente ses propriétés de réflectivité électromagnétique, qui dépendent de ses dimensions, de son orientation ainsi que de ses propriétés diélectriques (et notamment de la longueur d'onde émise λ).

Le radar permet d'acquérir les informations de position, de réflectivité électromagnétique et de vitesse le long d'une portée (appelée radiale ou encore axe distance), et dans tout l'environnement volumétrique en pointant différents sites (élévations) et gisements (azimuts). L'information polarimétrique, faisant intervenir la structure vectorielle du champ électromagnétique, est décrite dans la section 6.3.

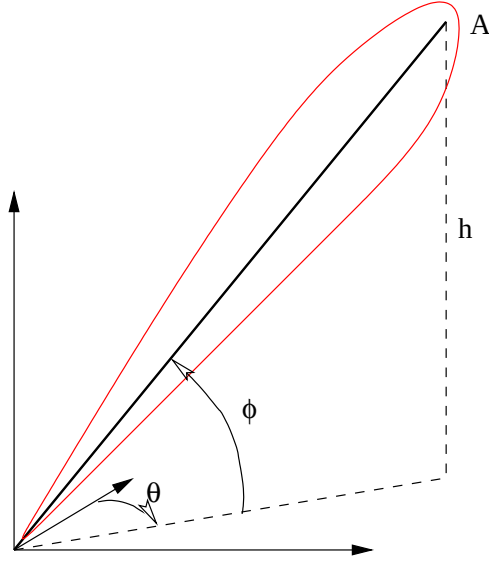


FIG. 6.1 – Emission d’une onde électromagnétique à l’élévation ϕ et l’azimut θ (h est la hauteur de la cible A). Le cône rouge représente l’énergie envoyée dans la direction de la cible A .

6.1.2 Chaîne de traitement d’un radar

Nous décrivons ici la chaîne de traitement typique d’un radar Doppler à impulsions. L’onde émise consiste en un train d’impulsions électromagnétiques qui forme une rafale, que l’on représente habituellement par la figure (6.2). Ainsi, avec l’enveloppe complexe $u(t)$, les grandeurs suivantes :

- la durée des impulsions τ ,
- la durée de la récurrence T_r ,
- la durée cohérente (ou de la rafale) T_{coh} ,
- le nombre d’impulsions n_i ,
- la longueur d’onde λ ,

définissent la forme de l’onde.

Ainsi l’impulsion est d’abord synthétisée, puis elle est “montée” (à l’aide de filtres analogiques) jusqu’à la fréquence f_0 du radar : c’est le mélange Fréquence Intermédiaire (mélange FI). L’onde électromagnétique est émise et réfléchiée par une cible (écho).

Le radar reçoit durant sa phase d’écoute l’onde réfléchiée et la fait redescendre en fréquence par une série de filtres analogiques. Le signal reçu est alors échantillonné pour pouvoir construire une image de la radiale, constituée de N cases-distance. La case-distance numéro n correspond alors à un échantillonnage du signal s_r aux instants $\tau + (n - 1)\Delta\tau + kT_r$ (où nous avons $\Delta\tau \leq \tau$) et pour k variant de 1 à n_i . Cet échantillonnage est effectué lors de la conversion analogique-numérique en respectant le théorème de Shannon, en exploitant le fait que la largeur de bande du signal reçu est égale à celle du signal émis, noté B par la suite. Par conséquent, la résolution (en mètre) des cases distances est alors égale à $c\frac{\Delta\tau}{2}$. Lors de la démodulation amplitude-phase, le signal échantillonné est transformé en complexe : à chaque case-distance n (et pour une rafale) est donc associé un vecteur complexe $s_n = (s_{n,1}, \dots, s_{n,n_i})$. Ces mesures $(s_n)_{1 \leq n \leq N}$, appelées données IQ

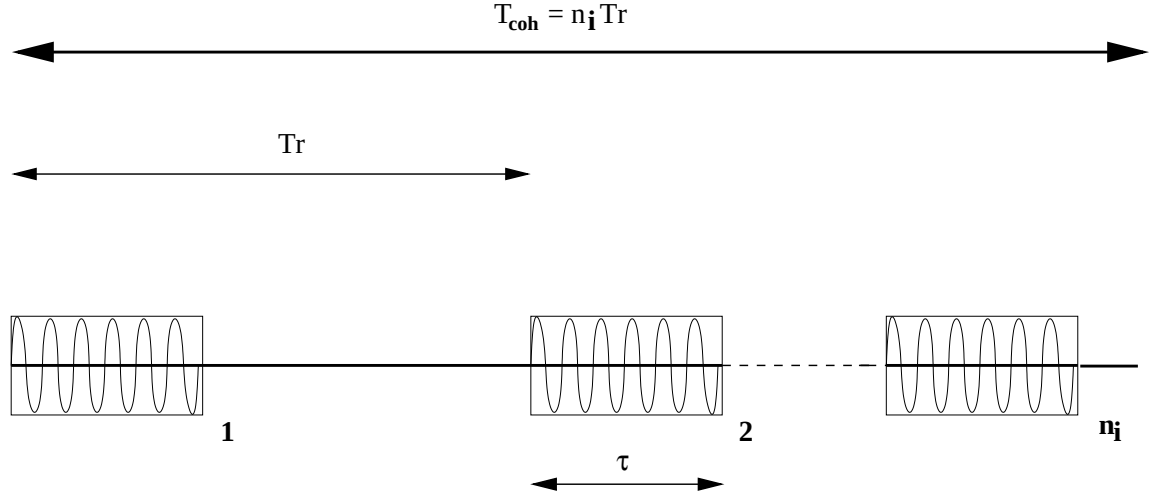


FIG. 6.2 – Rafale émise par un radar Doppler à impulsions

(*In phase-Quadrature*), sont alors utilisées pour faire la localisation en distance (via la compression d'impulsion) et en vitesse (via filtrage Doppler) (voir le diagramme de la figure (6.3)).

Après l'écoute d'une rafale, nous récupérons un tableau de données complexes structuré selon un axe distance et un axe récurrence. Afin d'augmenter le rapport signal à bruit qui est dégradé par l'addition du bruit thermique du radar (souvent considéré comme un bruit blanc gaussien), un filtrage adapté est réalisé, i.e. le signal reçu est corrélé au signal émis. En effet, le signal effectivement traité est $s_r(t) = As(t - \Delta T)e^{2i\pi\Delta f t} + b(t)$ et nous parlons de signal détecté (ou d'image détectée pour les radar SAR) pour le signal filtré $y(t) = \int s_r(u)s_e^*(u - t)du$. Outre l'amélioration du rapport signal à bruit, ce filtrage adapté (appelé aussi compression d'impulsion dans le cas d'une impulsion modulée en fréquence) permet d'améliorer la résolution du radar (i.e. le pouvoir discriminant en distance entre deux cibles). Pour améliorer l'efficacité la compression d'impulsion, le signal émis est un *chirp* quadratique, c'est-à-dire un signal presque sinusoïdal, centré sur la fréquence porteuse f_0 et modulé en fréquence (voir figure (6.4)) :

$$\forall t \in \left[-\frac{\tau}{2}, \frac{\tau}{2}\right], s_e(t) = S_0 e^{2i\pi(f_0 t + \frac{K_f t^2}{2})} \quad (6.4)$$

La compression d'impulsion est suivie d'un filtrage doppler qui permet de séparer les échos selon leur vitesse et de supprimer, par exemple, les échos fixes qui peuvent être plus forts que les cibles mouvantes d'intérêt (avions, ...). Un banc de filtres (ou une Transformée de Fourier Discrète, TFD) est utilisé, où le filtre k est adapté à la vitesse radiale v_k , de telle sorte que nous obtenons une image distance-Doppler de la rafale (voir figure (6.5)).

Remarque 6.1.1. *Les radars à impulsions sont soumis au problème de la résolution d'ambiguïté en distance causé par le train régulier d'impulsions : la distance D n'est connue que modulo une distance ambiguë notée $D_{amb} = \frac{cT_r}{2}$. De même, l'estimation de la vitesse radiale des cibles n'est possible que modulo la vitesse ambiguë $v_{amb} = \frac{\lambda}{2T_r}$ (en raison de la connaissance modulo 2π de la phase entre les impulsions régulièrement espacées). Le produit $D_{amb} \times v_{amb} = \frac{\lambda}{4}$ étant constant, il*

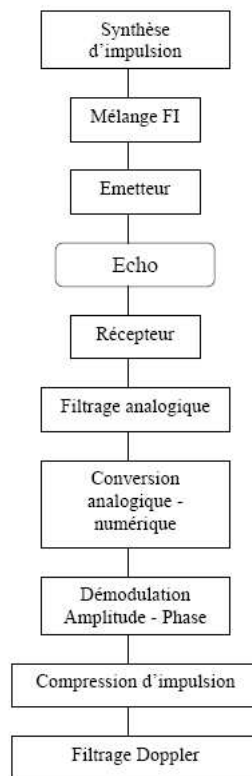


FIG. 6.3 – Chaîne de traitement radar

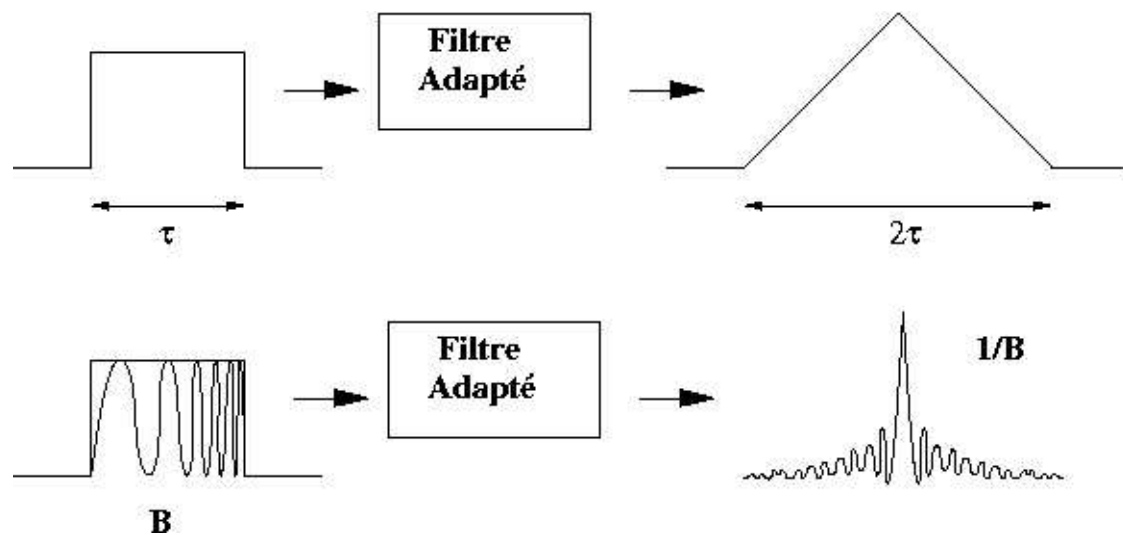


FIG. 6.4 – Compression d'impulsion de signaux non-modulés en fréquence et modulés en fréquence. En haut : à gauche, impulsion rectangulaire (une seule fréquence) émise de durée τ , et à droite le signal détecté (signal triangulaire de durée double). En bas : à gauche, chirp (modulation quadratique de la fréquence de largeur de bande B) émis, et à droite le signal détecté (sinus cardinal dont la largeur du pic à 3dB vaut $1/B$).

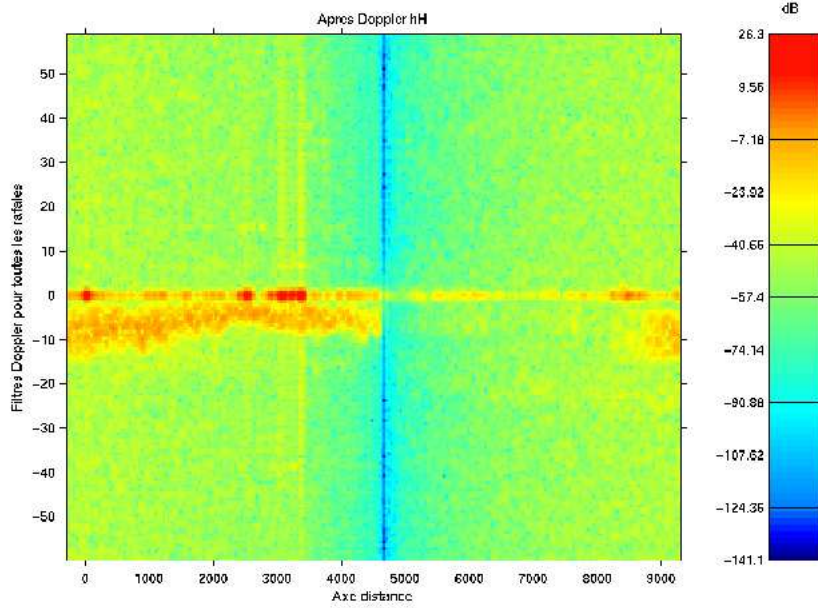


FIG. 6.5 – Exemple de carte distance-Doppler (représentant $20 \times \log_{10}(S(f))$) obtenue par un radar en bande X (bande de fréquence 8-12 GHz, soit une longueur d'onde λ de l'ordre de 3 cm), en polarisation horizontale (émission et réception). Nous avons 256 cases distances en abscisse (correspondant à une distance d'environ 9200 mètres) et 128 filtres Doppler en ordonnée (FFT réalisées sur 128 récurrences) échantillonnant l'intervalle des fréquences $[-\frac{1}{2}, \frac{1}{2}]$.

y a donc un compromis entre l'ambigüité en distance et en vitesse. Malgré tout, ces incertitudes peuvent être levées en réémettant plusieurs rafales à des durées de récurrence T_r ou à des fréquences λ différentes.

La carte distance-Doppler de la figure (6.5) fait apparaître clairement sur la première partie de la radiale un fouillis Doppler possédant une vitesse négative. Cela correspond à un nuage de pluie qui se rapproche du radar (de même en fin de portée). Nous pouvons aussi constater la présence de 3 réponses supérieures à 10 dB au Doppler nul, i.e. des pics de la densité spectrale, correspondant à des échos fixes (en l'occurrence des bâtiments massifs).

Chaque pixel de cette carte contient un complexe $\tilde{y}_{n,k}$ associé à une case distance n et au filtre k , de sorte qu'une cible est détectée si ce complexe est tel que $|\tilde{y}_{n,k}|^2 > \text{seuil}$, où le seuil dépend de n . Ce critère de détection correspond au test du rapport de vraisemblance (sous hypothèse de bruit $b(t)$ gaussien) :

$$\begin{cases} H_0 : & s_r(t) = b(t) \\ H_1 : & s_r(t) = As(t - \Delta T)e^{2i\pi\Delta f t} + b(t) \end{cases} \quad (6.5)$$

La probabilité de détection d'une cible, sous l'hypothèse H_1 est appelée la probabilité de détection P_d . Le risque de première espèce de ce test est appelé la probabilité de fausse alarme (P_{fa}) et doit

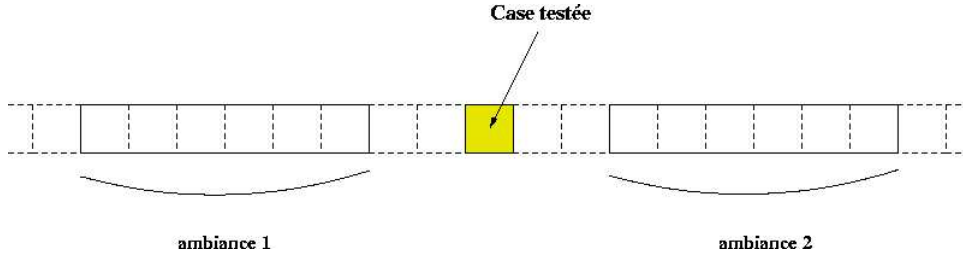


FIG. 6.6 – Détection par TFAC. Estimation de l’ambiance moyenne autour de la case testée (les cases directement adjacentes ne sont pas considérées par robustesse aux cibles étendues ou proches). Si les ambiances estimées sont les moyennes des cases droites et gauches, le TFAC utilise (par exemple) le maximum des deux ambiances estimées.

être impérativement fixé à une valeur très faible¹, le plus souvent 10^{-6} . Comme l’environnement radar est aléatoire, le seuil doit pouvoir s’adapter de manière à maintenir la P_{fa} à un niveau de 10^{-6} . Les traitements qui permettent ce contrôle de la fausse alarme sont appelés TFAC (pour “Taux de Fausse Alarme Constant”). Ces méthodes de régulation de la fausse alarme reposent essentiellement sur l’estimation du niveau moyen de “l’ambiance”, et la comparaison de la valeur courante à ce niveau moyen. Le bruit aléatoire qui constitue l’ambiance dans laquelle la cible est plongée est appelée le fouillis. Selon le type de fouillis (fouillis de sol, de mer, de pluie), les méthodes d’estimation de l’ambiance peuvent varier, mais gardent comme principe général la “moyennisation” sur les cases situées avant et après la case testée, voir figure (6.6).

Malgré l’utilisation de procédures statistiques pour améliorer la robustesse de cette estimation relativement aux valeurs atypiques (comme les statistiques de rang), l’estimation de l’ambiance peut être détériorée par un fouillis inhomogène (et augmenter ainsi la P_{fa} ou diminuer la P_d), avec la présence de changements abrupts. La segmentation bayésienne peut limiter ce problème en proposant une carte des différents fouillis et une localisation de ces changements. Deux pistes, non explorées dans la suite, peuvent être proposées : la carte des fouillis est utilisée pour déterminer, à l’aide d’un critère de vote, le type de TFAC à utiliser (garantissant la meilleure probabilité de fausse alarme possible, tout en maximisant la probabilité de fausse alarme [40]). Une autre optique n’est plus de considérer la carte elle-même, mais le modèle statistique ayant permis de la construire. Le modèle estimé (par exemple une CMCa) est une modélisation du mélange de fouillis rencontré sur la radiale, ainsi les cibles ponctuelles ayant des réflectivités particulièrement fortes ou avec des profils vitesse significativement différents de leur environnement peuvent être considérées comme des points atypiques de ce modèle. Par conséquent, une procédure de détection des atypiques d’un mélange, similaire à celles proposées par McLachlan (chapitre 7, dans [109]) peut être adaptée. Par exemple, dans le cas d’un mélange de lois SIRV, cela revient à utiliser un critère d’atypicité des observations y_i par rapport à la valeur moyenne définie par $q_k(y_i) = \frac{1}{2}(y_i - m_k)^* \Sigma_k^{-1}(y_i - m_k)$ et à comparer la valeur $\sum_{k=1}^K \pi_{i,k} q_k(y_i)$ (ou $\min_k \pi_{i,k} q_k(y_i)$) à un seuil. Si celle-ci est grande, nous considérons alors que l’observation y_i est un point atypique du modèle de fouillis, et qu’il s’agit donc d’une cible.

¹En effet, les conséquences d’une fausse alarme peuvent être très grave dans le domaine militaire. De plus, cela permet d’éviter de saturer le radar par des requêtes de confirmation de détection, ou par la création de fausses pistes pour le suivi de cibles.

6.1.3 Segmentation des fouillis

La méthodologie de segmentation de fouillis que nous proposons se situe en amont de l'analyse Doppler et de la phase de détection des cibles. Nous utilisons le signal reçu après compression d'impulsion, noté $(y_n)_{1 \leq n \leq N}$, et nous le segmentons selon 3 types d'information :

1. Réflectivité
2. Doppler
3. Polarimétrie

Dans les parties suivantes, nous proposons une méthodologie pour extraire et séparer ces 3 informations, de manière à fournir des cartes différentes et complémentaires de l'environnement radar. L'image que nous traitons n'est donc pas l'image distance-Doppler habituelle, mais est une séquence d'observations appartenant à des espaces de représentation adéquat. La modélisation statistique utilisée décrit donc le mélange des fouillis (Doppler ou polarimétrique) observé le long d'une radiale, et intègre l'information de dépendance spatiale. Nous utilisons alors un modèle CMCA, et la carte recherchée est l'estimateur MPM du processus caché.

Par la suite, nous décrivons les segmentations exploitant uniquement les informations Doppler ou polarimétrique. Nous ne traitons donc pas la segmentation de l'environnement radar basée sur la réflectivité, qui peut être réalisée directement avec les modèles CMCA présentés dans le chapitre 5, voir l'exemple 5.3.3.3.

6.2 Segmentation Doppler

L'information Doppler correspond à la densité spectrale que l'on peut associer à chaque train d'impulsion. Ainsi, elle est classiquement utilisée pour séparer une cible du fouillis de sol, possédant une forte réflectivité mais une fréquence Doppler nulle, ou du bruit thermique, souvent assimilable à un bruit blanc, qui a une densité spectrale constante. Entre ces deux comportements extrêmes, les fouillis peuvent avoir des spectres aux formes variées, présentant des pics à différentes vitesses. Ceci est le cas des fouillis de mer, des fouillis volumiques en présence d'hydrométéores (pluie, nuage de grêles), ou de turbulences atmosphériques.

L'objectif de la segmentation Doppler décrite ici est d'obtenir une classification des fouillis selon le profil de leur densité spectrale, i.e. sa forme (typiquement le nombre de pics et leur étalement). Les radars Doppler à impulsion permettent de réaliser une analyse Doppler haute-résolution, qui constitue une alternative à l'analyse Doppler classique par banc de filtre ou TFD (cette amélioration de la représentation de l'information Doppler a déjà été développée pour des données échographiques Doppler [11]). Cette approche est basée sur une modélisation autorégressive de chaque échantillon y_n , ce qui revient à ne considérer qu'une sous-famille de spectres possibles pour le fouillis. La segmentation Doppler proposée consiste alors en la classification, par un modèle CMCA, des paramètres indexant les spectres autorégressifs .

6.2.1 Information Doppler et modélisation autorégressive

La modélisation autorégressive des signaux radar est désignée dans le domaine radar par le terme d'analyse haute-résolution. Elle permet d'avoir une expression paramétrique et analytique de

la densité spectrale (que nous pouvons connaître alors à n'importe quelle résolution, contrairement au spectre obtenu par TFD) en chaque case distance n , correspondant aux données $(y_{n,1}, \dots, y_{n,n_i})$. Nous supposons que $(y_{n,1}, \dots, y_{n,n_i})$ est la réalisation d'un processus autorégressif $\mathbf{Y} = (Y_n)_{n \in \mathbb{Z}}$ d'ordre p , i.e. que $\mathbf{Y}$ vérifie

$$\forall n, Y_n = \sum_{k=1}^p a_k Y_{n-k} + \epsilon_n \quad (6.6)$$

avec (ϵ_n) bruit blanc centré, de variance σ_b^2 . Les coefficients complexes $(a_k)_{1 \leq k \leq p}$ sont appelés coefficients autorégressifs. S'ils sont tels que le polynôme $P(z) = 1 - \sum_{k=1}^p a_k z^{-k}$ ait toutes ses racines à l'intérieur du cercle unité (ce que nous supposons par la suite), il s'agit alors du filtre à minimum de phase et le bruit blanc $(\epsilon_n)_{n \in \mathbb{Z}}$ est le processus d'innovation de $\mathbf{Y}$. Dans ce cas, la suite des autocovariances $(R(n) = E[X_0 X_n^*])_{n \in \mathbb{Z}}$ (ou des autocorrélations $(\rho_n = \frac{R(n)}{R(0)})_{n \in \mathbb{Z}}$) est entièrement déterminée par les coefficients autorégressifs. La densité spectrale de puissance $S(f) = \sum_{n \in \mathbb{Z}} R(n) e^{2i\pi n f}$, définie sur $\mathcal{I} =]-\frac{1}{2}, \frac{1}{2}]$, est égale à

$$S(f) = \frac{\sigma_b^2}{|1 - \sum_{k=1}^p a_k e^{-2i\pi k f}|^2} \quad (6.7)$$

La modélisation autorégressive s'avère être une bonne description du fouillis de mer, parce qu'elle permet de restituer assez fidèlement sa dynamique au sein d'une même rafale [87]. De manière générale, les paramètres du modèle (6.6) sont considérés comme un bon résumé de la dynamique des fouillis Doppler, et possède l'avantage de découpler cette dynamique (représentée par le vecteur $(a_1 \dots a_p)'$) de la propriété de réflectivité de l'environnement (qui est portée par σ_b^2).

Cependant, plusieurs paramétrisations d'un modèle autorégressif sont possibles et nous utilisons parmi celles-ci les coefficients de corrélation partielle $(\mu_k)_{1 \leq k \leq p}$ (que nous appellerons indifféremment coefficients de réflexion) dont nous rappelons la définition dans l'annexe B, ainsi que les relations de passage entre les différents paramétrisations d'un processus AR : coefficients d'autocovariance, coefficients AR et coefficients de réflexion (CR). Les coefficients de réflexion possèdent trois avantages :

- calcul direct et rapide des CR $(\mu_k)_{1 \leq k \leq p}$ (et de la puissance du bruit σ_b^2) par l'algorithme de Burg à partir d'un échantillon $(y_1, \dots, y_{n_i})$ avec $n_i \geq p$;
- existence d'une version régularisée de l'algorithme de Burg qui permet l'estimation des CR sur de faibles tailles d'échantillon (une dizaine d'observations), en rajoutant un critère de douceur spectrale. Les coefficients μ_k sont "forcés" à décroître lorsque k augmente, ce qui permet de limiter la présence de pics parasites, voir annexe B.
- la seule contrainte que doit vérifier la suite des CR est

$$\mu_p = 0 \implies \forall n \geq p, \mu_n = 0 \quad (6.8)$$

Le plus souvent, en raison de la rotation d'antenne (qui empêche de pointer longtemps dans la même direction), ou de la non-stationnarité des phénomènes mesurés (lorsque l'antenne est fixe) nous avons une dizaine de récurrences dans une rafale. Ainsi il est important d'avoir un algorithme d'estimation des CR qui soit rapide et fiable. L'algorithme de Burg nous permet d'extraire très rapidement l'information Doppler des données IQ, et d'obtenir une image de l'environnement dans "l'espace Doppler". Le troisième point implique que les CR constituent un système de coordon-

nées indépendantes pour décrire l'espace des processus stationnaires à σ_b^2 fixé. En effet, pour tout processus stationnaire au second ordre, nous pouvons associer la suite des coefficients de corrélation partielle $(\mu_n)_{n \geq 1}$ dont la connaissance est équivalente à la connaissance des autocorrélations $(\rho_n)_{n \geq 1}$. Chaque coefficient μ_n ou ρ_n appartient à $\mathcal{D} = \{z \in \mathbb{C} \mid |z| \leq 1\}$, mais toute suite de paramètres de $\mathcal{D}^\infty$ ne correspond pas à une suite d'autocorrélations $(\rho_n)_{n \geq 1}$, en raison de la contrainte de positivité. Ceci n'est pas le cas des corrélations partielles : toute suite de CR $(\mu_n)_{n \geq 1} \in \mathcal{D}^\infty$ est acceptable, pour peu que nous respectons la contrainte (6.8). De plus, le fait de tronquer une suite de CR à l'ordre p , revient à considérer une approximation par un processus autorégressif d'ordre p . Cette faible contrainte sur les CR permet donc de proposer facilement des lois paramétriques sur l'espace $\mathcal{D}^p$, et donc de proposer des procédures de segmentation bayésienne par CMCa.

6.2.2 Segmentation bayésienne et cartographie Doppler

Nous appelons cartographie Doppler la segmentation de l'environnement radar en utilisant uniquement les CR calculées à partir de chaque échantillon $(y_1, \dots, y_{n_i})$ disponible dans les cases-distances. Ceci permet d'identifier des zones distinctes indépendamment de leur propriétés de réflectivité électromagnétique, parce que celle-ci est portée après analyse haute-résolution par le scalaire σ_b^2 . Notre objectif est donc de rassembler les vecteurs des “coordonnées spectrales” $\mu = (\mu_1, \dots, \mu_p) \in \mathcal{D}^p$ en utilisant une notion de similarité entre vecteurs μ , ce que nous faisons dans le cadre de la segmentation bayésienne en choisissant une famille paramétrique de lois d'émission. Pour cela, nous utilisons une décomposition polaire du vecteur μ :

$$\mu = \|\mu\| \times \bar{\mu} \quad (6.9)$$

où nous avons $\|\mu\| = \sqrt{\sum_{i=1}^p |\mu_i|^2}$ et $\bar{\mu} = \frac{\mu}{\|\mu\|}$, que nous appelons respectivement la norme et la direction de μ .

Remarque 6.2.1. *La norme de μ est reliée à la complexité du processus. En effet, si $(Y_n)_{n \in \mathbb{Z}}$ est un processus stationnaire au second ordre quelconque (non nécessairement autorégressif), nous avons*

$$\forall p \geq 1, \sigma_p^2 = \prod_{k=1}^p (1 - |\mu_k|^2) \times \sigma_0^2 \quad (6.10)$$

où $\sigma_0^2 = E[Y_n Y_n^*]$ est la variance de $\mathbf{Y}$ et σ_p^2 est la variance de l'erreur de prédiction linéaire de Y_n par $Y_{n-1}, \dots, Y_{n-p}$ (cf. annexe B.1, équation (B.5)). La suite des puissances σ_p^2 est décroissante et minorée par 0 : elle converge² donc, d'où l'on déduit que le produit infini des modules des coefficients de réflexion $\prod_k (1 - |\mu_k|^2)$ converge. Si pour tout k , $|\mu_k| \neq 1$, ceci est équivalent à la convergence de la série $\sum_k |\mu_k|^2$ (vers une limite notée $\|\mu\|_\infty^2$), et nous savons que dans ce cas-là, la limite du produit infini vérifie les inégalités suivantes³ :

$$1 - \|\mu\|_\infty^2 \leq \prod_{k=1}^\infty (1 - |\mu_k|^2) \leq \exp(-\|\mu\|_\infty^2) \quad (6.11)$$

La borne inférieure n'est pas très informative puisque le produit appartient à l'intervalle $[0, 1]$ et que

²la valeur de cette limite est donnée par la formule de Kolmogorov [24] : $\sigma_\infty^2 = \exp\left(\int_{-1/2}^{1/2} \log(S(f)) df\right)$.

³Nous utilisons l'identité $1 - (a + b) \leq (1 - a)(1 - b) \leq e^{-(a+b)}$.

dans beaucoup de situations, nous avons $\|\mu\|_\infty^2 > 1$, mais cela permet de montrer que la variance de l'innovation σ_∞^2 est contrôlée par la norme $\|\mu\|_\infty^2$.

Nous considérons alors que les classes que nous recherchons sont caractérisées par des valeurs moyennes autour desquelles les observations sont réparties symétriquement. Dans le cas où les observations sont des points de $\mathbb{R}^p$, la similarité entre deux observations (ou une observation et la valeur moyenne) est mesurée en terme de distance euclidienne, et la traduction probabiliste classique de cet a priori est l'utilisation de lois d'émission gaussienne. Cependant, l'hypothèse gaussienne n'est pas satisfaisante pour les coefficients de réflexion parce qu'ils sont contraints à appartenir au compact $\mathcal{D}^p$ et que la non-linéarité des transformations nécessaires à leur calcul (et le faible nombre des récurrences utilisées) ne permet pas d'avoir une approximation de cette loi.

L'utilisation de lois sur des espaces non-euclidiens pose rapidement des problèmes complexes pour la définition de la distance entre les observations et le profil moyen (par exemple lorsque $\mathcal{Y}$ possède une structure géométrique comme une variété [119]), et aussi pour le calcul effectif des densités (notamment les constantes de normalisation [107]). Celles-ci font intervenir des fonctions spéciales ce qui implique le plus souvent l'absence de formule analytique pour l'estimation des paramètres de ces lois. Pour contourner ces problèmes, nous utilisons une approche utilisée en classification consistant en l'utilisation partielle de l'information contenue dans les observations. Il s'agit d'exploiter uniquement l'information de direction portée par le vecteur μ , et de considérer que deux vecteurs μ_1 et μ_2 sont similaires si le produit scalaire des grandeurs normalisées $\Re(\langle \bar{\mu}_1, \bar{\mu}_2 \rangle)$ est grand (lorsque les vecteurs sont complexes) ou le produit scalaire $\langle \bar{\mu}_1, \bar{\mu}_2 \rangle$ (lorsque les vecteurs sont réels). Cette méthode a déjà été appliquée à des données textuelles et d'expression de gènes dont la classification est difficile parce qu'elles appartiennent à des espaces de grandes dimensions [8]. Dans notre cas, cela revient à considérer que l'information portée par la norme $\|\mu\|$ est beaucoup moins discriminante et utile que la direction $\bar{\mu}$ parce que tous les CR sont de norme inférieure à 1. Cela revient alors à utiliser comme indice de similarité le cosinus de l'angle formé par les deux vecteurs μ_1 et μ_2 . Malgré tout, nous avons vu précédemment que $\|\mu\|$ contenait une partie de l'information sur la complexité du processus, qui peut améliorer les segmentations obtenues avec les directions. Nous proposons alors deux méthodes de segmentation bayésienne utilisant soit uniquement $\bar{\mu}$, soit $\bar{\mu}$ et $\|\mu\|$.

6.2.3 Le modèle statistique paramétrique utilisé

Le modèle classique pour la classification de données directionnelles consiste en l'utilisation de lois de Von Mises - Fisher pour la modélisation des lois d'émission. Ici, dans le traitement du signal radar nous introduisons la dépendance spatiale par une modélisation CMCa-BI (McLachlan dans [109] ou Banerjee *et al.* dans [8] traitent le cas i.i.d), et le modèle que nous utilisons pour la loi d'émission est une loi de Von Mises-Fisher complexe.

Lorsque nous voulons exploiter à la fois $\|\mu\|$ et $\bar{\mu}$, nous supposons qu'il s'agit de deux grandeurs indépendantes, telles que la loi d'émission de l'observation $(\|\mu\|, \bar{\mu})$ s'écrive $p(\|\mu\|, \bar{\mu}) = p_1(\|\mu\|)p_2(\bar{\mu})$, où p_2 est une loi de Von Mises-Fisher complexe et p_1 est une loi gamma. Dans ce cas, le paramètre θ_k de la loi de la classe k est $\theta_k = (\theta_k^1, \theta_k^2) = (a_k, b_k, \kappa_k, \xi_k)$ où a_k et b_k sont les paramètres de la loi gamma, et κ_k et ξ_k sont ceux de la loi de Von Mises-Fisher.

Cependant, comme nous l'avons montré dans l'exemple de la section 5.3.3.3 pour l'intensité,

les signaux radar sont corrélés en distance, y compris dans des milieux homogènes. Nous rappelons que cette dépendance existe physiquement (entre autres dans le fouillis mer, voir les travaux de Watts [152, 153]), mais est aussi créé par la compression d'impulsion (qui réalise une convolution sur une fenêtre spatiale, voir section 6.1.2). Pour en tenir compte, nous supposons que nous pouvons la modéliser en utilisant un modèle CMCa. De même que pour les vecteurs aléatoires (théorème d'impossibilité de Genest *et al.* pour la généralisation des copules aux lois bivariées sur $\mathbb{R}^p \times \mathbb{R}^q$), il est difficile d'expliciter des lois sur $(\mathbb{R}^+ \times \mathbb{CS}^{p-1}) \times (\mathbb{R}^+ \times \mathbb{CS}^{p-1})$ dont les marges appartiennent à un modèle prédéfini. Afin d'introduire une éventuelle dépendance entre deux observations consécutives $(\|\mu_1\|, \overline{\mu}_1)$ et $(\|\mu_2\|, \overline{\mu}_2)$, nous introduisons la loi bivariée suivante :

$$f(\|\mu_1\|, \theta_1^1) f(\|\mu_2\|, \theta_2^1) c(F(\|\mu_1\|, \theta_1), F(\|\mu_2\|, \theta_2)) \times f(\overline{\mu}_1, \theta_1^2) f(\overline{\mu}_2, \theta_2^2) \quad (6.12)$$

où c est une copule (gaussienne dans les applications que nous traitons). Avec cette loi jointe, deux vecteurs de CR sont liés par leur norme : si la copule utilisée induit une dépendance positive (i.e. les deux variables ont tendance à varier dans le même sens, comme pour une copule gaussienne avec un coefficient de corrélation positif) nous aurons tendance à avoir des vecteurs voisins de même longueur (et des spectres possédant la même richesse spectrale). Finalement, pour la segmentation Doppler, nous pouvons utiliser les informations $(\|\mu_n\|, \overline{\mu}_n)$ et nous avons deux densités de transition du processus complet $Z_n = (X_n, \mu_n)$ est :

$$\begin{cases} \text{(directionnel)} & p(z_{n+1} | z_n) = p(x_{n+1} | x_n) p(\overline{\mu}_{n+1} | x_{n+1}) \\ \text{(complet)} & p(z_{n+1} | z_n) = p(x_{n+1} | x_n) p(\|\mu_{n+1}\| | \|\mu_n\|, x_n, x_{n+1}) p(\overline{\mu}_{n+1} | x_{n+1}) \end{cases} \quad (6.13)$$

Nous donnons plusieurs exemples des segmentations obtenues sur données simulées, ainsi que sur des données réelles.

6.2.4 Exemples

Nous montrons que le modèle paramétrique introduit dans la section précédente pour des raisons essentiellement techniques (contrôle des lois marginales, possibilité d'écrire facilement la densité des lois, existence de procédures d'estimation simples) est assez générale pour décrire correctement les fluctuations des observations simulées ou réelles. Dans les applications, nous utilisons 5 coefficients de réflexion pour décrire et segmenter l'environnement : nous travaillons donc avec des point de $\mathcal{D}^5$, soit des points de $\mathbb{R}^+ \times \mathbb{CS}^5$.

6.2.4.1 Données simulées

Pour réaliser la segmentation, nous calculons un résumé des données - le vecteur μ des coefficients de réflexion - dont la qualité dépend du nombre d'observations (et de leur loi). Si le nombre de récurrences disponibles en chaque case-distance augmente, les fluctuations aléatoires des coefficients de réflexion seront essentiellement dues aux fluctuations de l'environnement. Si les propriétés Doppler fluctuent peu au sein de chaque classe, nous aurons alors des nuages de points très concentrées. Nous illustrons cette situation (voir figure (6.7)) lorsque nous avons 128 récurrences pour une rafale de 500 cases-distance. Nous avons 3 classes correspondant à des processus autorégressifs distincts, de même variance σ_0^2 mais avec des coefficients de réflexion différents

$\mu_1 = \begin{pmatrix} 0,1 - 0,9i \\ 0,4i \\ 0 \end{pmatrix}$, $\mu_2 = \begin{pmatrix} 0,3 + 0,8i \\ 0,5 \\ 0,1 + 0,2i \end{pmatrix}$ et $\mu_3 = \begin{pmatrix} 0,1 - 0,9i \\ 0 \\ 0 \end{pmatrix}$. Les données sont simulées en faisant l'hypothèse que nous avons une CMCa-BI telle que le processus caché ait la matrice de transition $A = \begin{pmatrix} 0,9 & 0,05 & 0,05 \\ 0,05 & 0,9 & 0,05 \\ 0,1 & 0,1 & 0,8 \end{pmatrix}$. Les normes des vecteurs sont égales à $\|\mu_1\| = 0,98$, $\|\mu_2\| = 1,03$ et $\|\mu_3\| = 0,87$, et le profil des normes estimées le long de la radiale est donné par la figure (6.8).

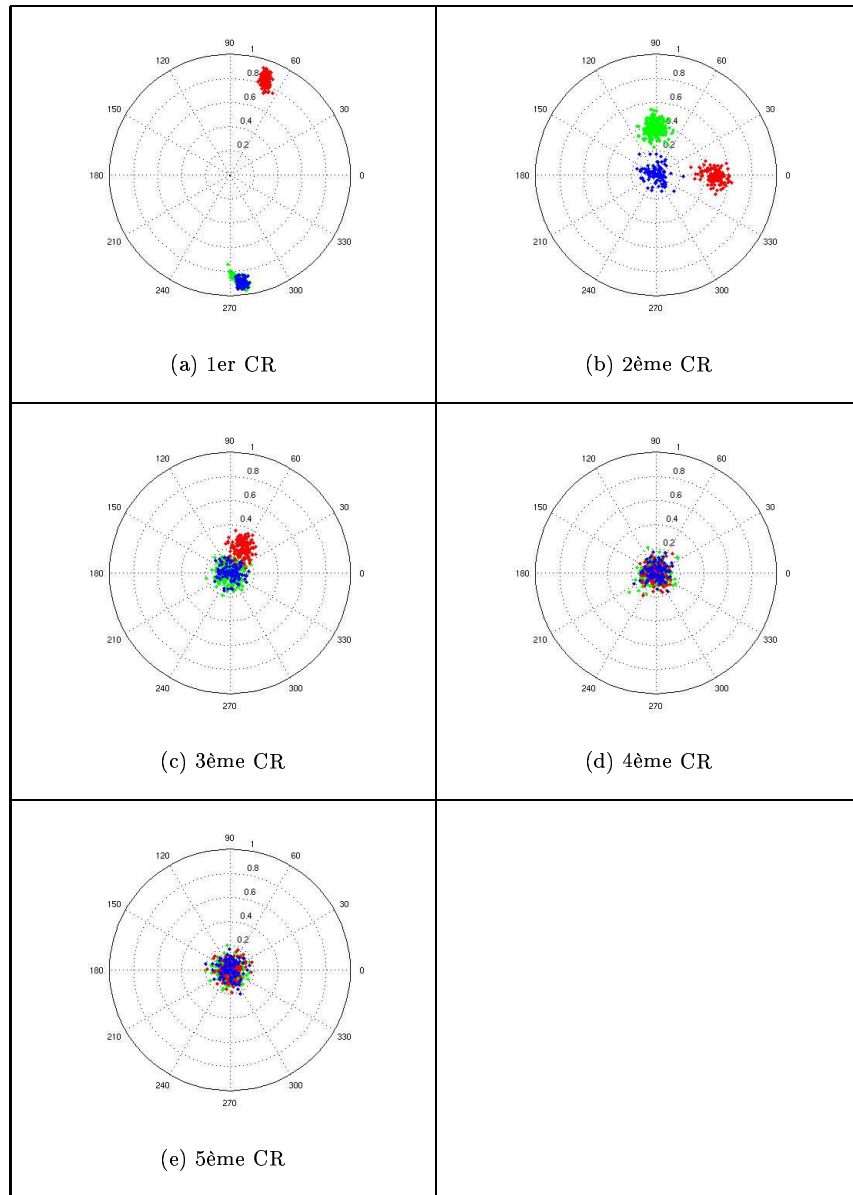


FIG. 6.7 – Nuage des coefficients de réflexion ($p = 5$).

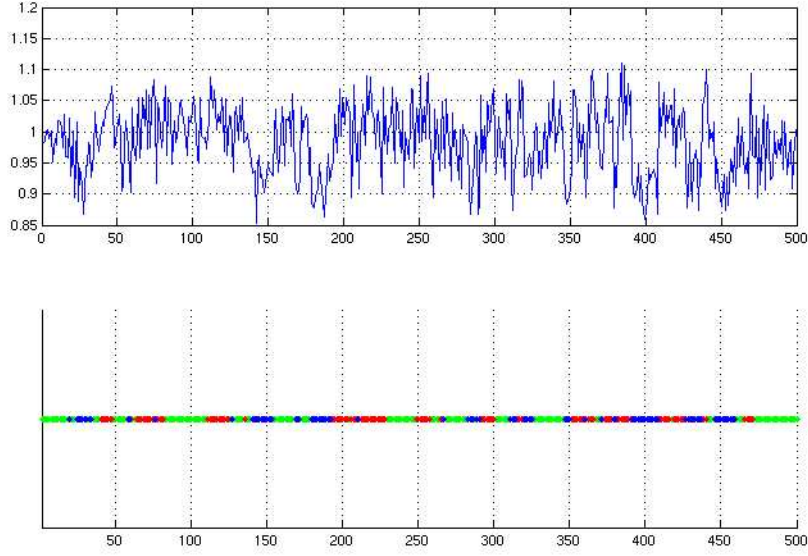


FIG. 6.8 – Normes des vecteurs ($\|\mu\|$) calculés avec 128 récurrences, et répartition des classes le long de la radiale.

Dans cette configuration, nous obtenons alors un taux d’erreur en segmentation non-supervisée égal à 0% lorsque nous utilisons le modèle CMCa-BI en exploitant les informations $(\|\mu\|, \bar{\mu})$ (nous avons imposé $c = c^\perp$ pour la copule gaussienne). Si nous utilisons une CMCa-BI pour la segmentation des directions $\bar{\mu}$ nous avons encore un taux d’erreur de 0% : la direction est donc une statistique totalement discriminante dans cet exemple.

Si nous avons seulement 16 récurrences par case-distance, la dispersion des coefficients de réflexion au sein de chaque classe est beaucoup plus grande. Nous montrons ci dessous les résultats des segmentations obtenues à partir de deux types des coefficients de réflexion calculées par un algorithme de Burg classique ou régularisé (voir section B.3 en annexe). Ceci permet d’évaluer l’influence de la dispersion des CR (due à la surestimation des CR d’ordre élevé lorsque nous avons peu d’observations pour les estimer). De plus, nous segmentons les données en exploitant soit les directions $\bar{\mu}$, soit les directions et les moyennes $\bar{\mu}$ et $\|\mu\|$. Les résultats de segmentation sont rassemblés dans le tableau (6.1) et ont été obtenues sur 500 itérations Monte Carlo.

	régularisé	non régularisé
$\bar{\mu}$	7,4 (2,3)	4,5 (1,7)
$\bar{\mu}$ et $\ \mu\ $	7,4 (2,3)	4,4 (1,7)

TAB. 6.1 – Taux d’erreur de segmentation de processus AR en % (et écart-type entre parenthèses), à partir de 16 récurrences.

La première remarque est que l’ajout de l’information de la norme $\|\mu\|$ n’améliore pas notre capacité de discrimination pour la segmentation. Ceci est explicable par le faible écart entre les normes $\|\mu_1\|$, $\|\mu_2\|$ et $\|\mu_3\|$ et par le caractère très “bruitée” de cette information avec seulement 16 récurrences. La deuxième remarque est que l’utilisation des CR régularisés a tendance à augmenter

le taux d'erreur des segmentations (ainsi que son écart-type). Ceci est une conséquence de la régularisation qui tend à faire converger les CR vers 0, ce qui diminue donc les différences entre les classes.

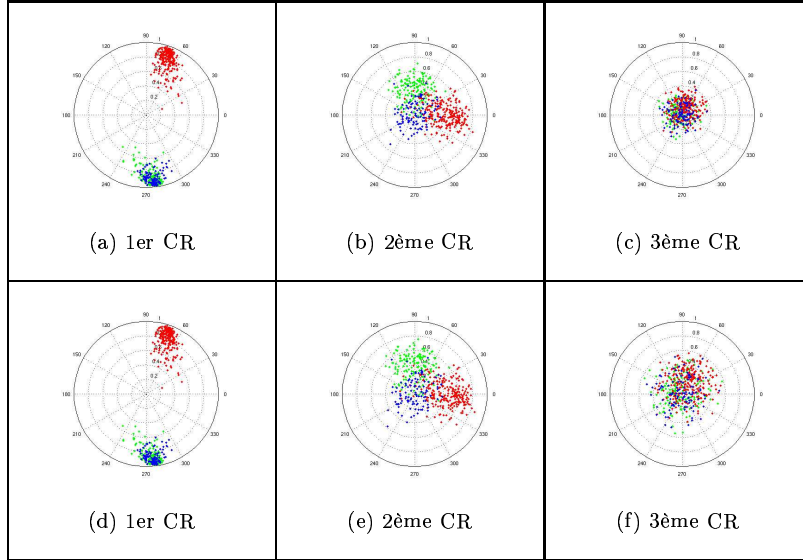


FIG. 6.9 – Exemples de CR calculés (1 à 3) avec 16 récurrences et colorés selon les classes (avec lissage - ligne du haut, et sans lissage - ligne du bas).

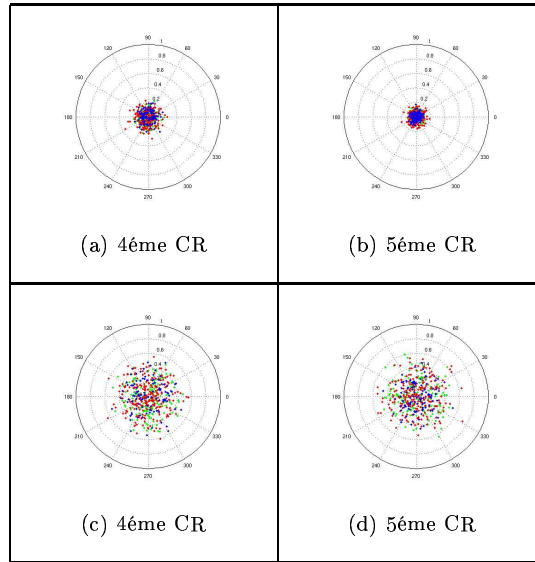


FIG. 6.10 – Exemples de CR calculés (4 et 5) avec 16 récurrences et colorés selon les classes (avec lissage - ligne du haut, et sans lissage - ligne du bas).

La segmentation que nous avons utilisé à partir des données $(\|\mu\|, \bar{\mu})$ repose sur un modèle CMCa, pour lequel nous estimons aussi une dépendance conditionnelle (par la copule gaussienne) pour les normes. Les coefficients estimés pour la copule gaussienne dans les cas lissés et non-lissés sont identiques et sont nuls (en moyenne) pour toutes les classes : nous sommes capables

de détecter l'absence de dépendance conditionnelle. De plus les coefficients de concentration κ des lois de Von Mises-Fisher sont en moyennes égaux à 30 pour toutes les classes, indiquant des classes très homogènes avec des observations proches d'un même profil moyen.

6.2.4.2 Données réelles

Les données que nous utilisons dans cette section sont issues du DLR (*Deutschen Zentrum für Luft - Raumschaft*⁴), et ont été obtenues avec le radar météorologique POLDIRAD [142], qui est un radar en bande C (fréquence à 5,504 GHz, soit une longueur d'onde égale à 5,45 cm). Les radars météorologiques utilisent l'effet Doppler pour estimer les propriétés cinétiques de l'atmosphère et un des objectifs des radars météorologiques est de reconstruire le champ de vent en faisant de la vélocimétrie Doppler. A partir de la densité spectrale et de ses moments, nous pouvons déterminer la réflectivité (moment d'ordre 0), la vitesse radiale moyenne du vent dans la direction de pointage (moment d'ordre 1) et avoir une mesure des turbulences au sein de l'élément d'atmosphère étudié (avec le moment d'ordre 2, parfois appelé largeur spectrale). La vitesse moyenne du vent est estimée avec une précision dépendant de la largeur spectrale, qui dépend elle même de la forme de l'onde émise par le radar et des propriétés de l'atmosphère.

Les mesures ont été faites dans le cadre du projet ATC-Wake (projet européen pour le contrôle du trafic aérien), dont l'objectif est l'amélioration des techniques de vélocimétrie pour la détection et le suivie des turbulences générées dans le sillage des avions. Dans le cas que nous traitons, nous n'avons pas de vortex de ce type, et il s'agit uniquement d'un champ de vent "naturel". L'ensemble de la campagne de mesures a été réalisé à plusieurs élévations et plusieurs azimuts, de telle sorte que nous ayons une couverture volumétrique de l'atmosphère. Une vélocimétrie de l'élément de volume considéré permet de mettre en évidence un cisaillement de vent. Ce dernier est un changement brusque du champ de vent entre différentes couches atmosphériques, illustré par la figure (6.11). Ce graphique montre selon l'altitude des changements dans la vitesse radiale du vent qui indique des modifications brusques de l'intensité et/ou de la direction du vent.

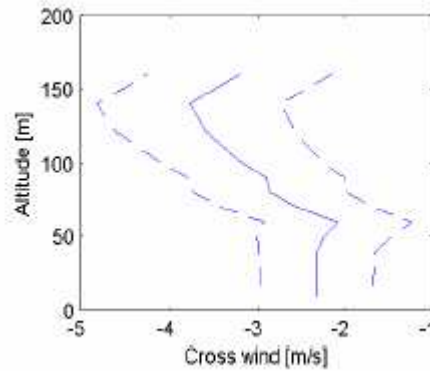


FIG. 6.11 – Estimation de la vitesse du vent en fonction de l'altitude pour un azimut donné.
— : Estimation de la vitesse ; - - : Intervalle de précision de l'estimation

Nous utilisons la segmentation Doppler pour dresser une carte des turbulences à partir des

⁴Centre Allemand d'Aérospatiale, <http://www.dlr.de>

données acquises à une élévation donnée et pour différents azimuts. Nous traitons une image constituée de 32 azimuts, chaque rafale étant constitué de 128 cases distances (résolution de 75 m), soit un ensemble de 4096 observations. Nous avons à notre disposition 64 récurrences ce qui nous permet pas de ne pas régulariser l'algorithme de Burg pour l'analyse haute résolution (pour la segmentation). Comme dans l'exemple précédent sur données simulées, nous utilisons 5 coefficients de réflexion comme a priori pour décrire la complexité de la dynamique des turbulences. L'examen des cartes distance-Doppler permet de mettre en évidence une fluctuation des vitesses à l'élévation considérée (angle de pointage égale à $3,5^\circ$) pour différents azimuts, voir figure (6.12).

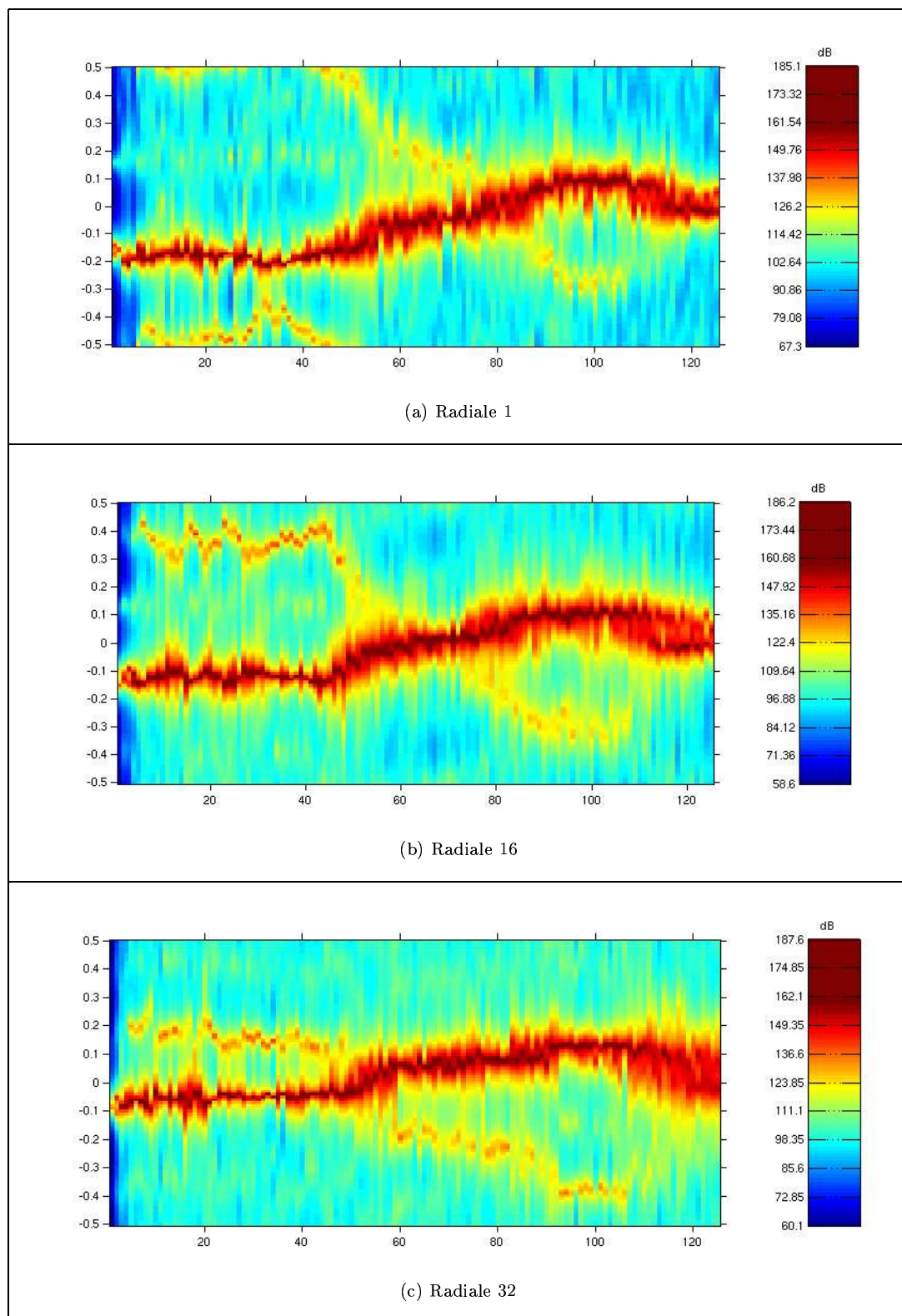


FIG. 6.12 – Cartes distance-Doppler ($20 \times \log_{10}(S(f))$) pour 3 radiales pour une élévation de $3,5^\circ$ (radiales 1, 16 et 32), pour les cases distances entre 10 et 128.

Pour la segmentation Doppler, nous transformons l'image Doppler à cette élévation en une chaîne par un parcours de Peano, que nous pouvons traiter par les algorithmes par chaîne Markov cachée, en utilisant uniquement l'information contenue dans les coefficients de réflexion. L'information de réflectivité portée par la puissance du bruit σ_b^2 n'est pas très discriminante dans ce contexte, parce que les densités spectrales ont sensiblement les mêmes maxima et amplitudes en distance et en azimut, comme le montre les spectres de la figure (6.12)). Malgré l'utilisation de deux modèles différents et deux types d'information distinctes à savoir $\bar{\mu}$ et $(\|\mu\|, \bar{\mu})$, nous obtenons les mêmes segmentations (à quelques pixels près) pour 6 classes voir figure (6.13).

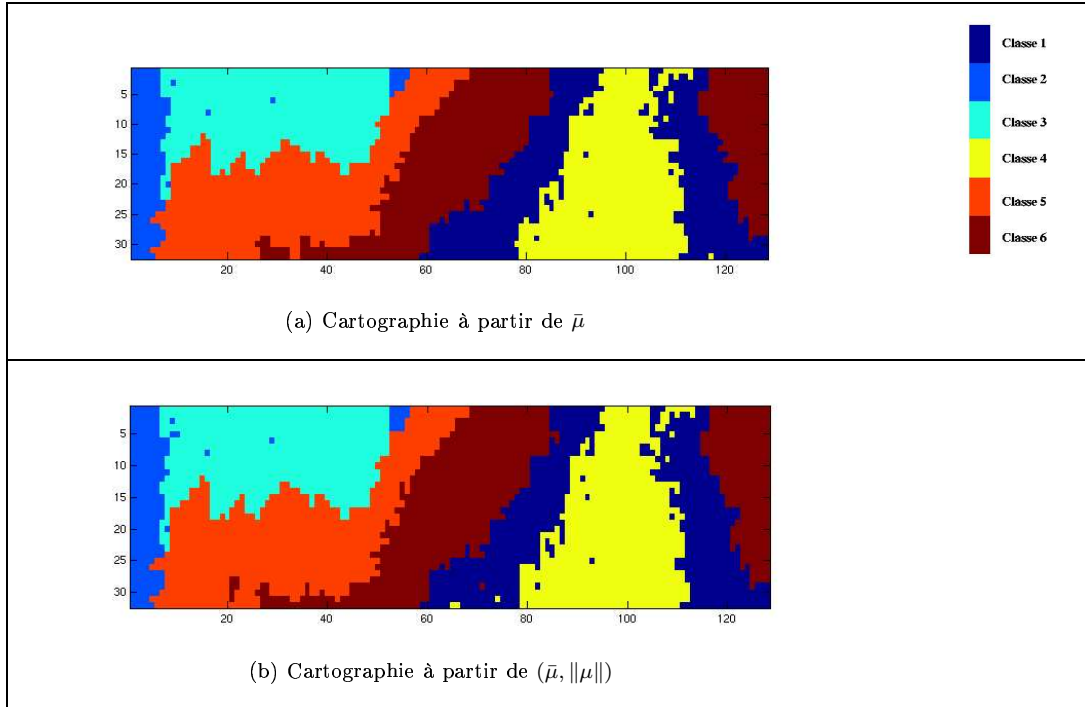


FIG. 6.13 – Cartographies Doppler des turbulences

Comme c'était déjà le cas pour les données simulées, l'information directionnelle est déjà très discriminante pour les profils Doppler, et l'information apportée par la norme $\|\mu\|$ ainsi que la dépendance spatiale estimée par le modèle CMCa (les 6 classes ont une dépendance conditionnelle significative, puisque nous avons des coefficients ρ_{kk} pour la copule gaussienne qui sont tous entre 0,5 et 0,7).

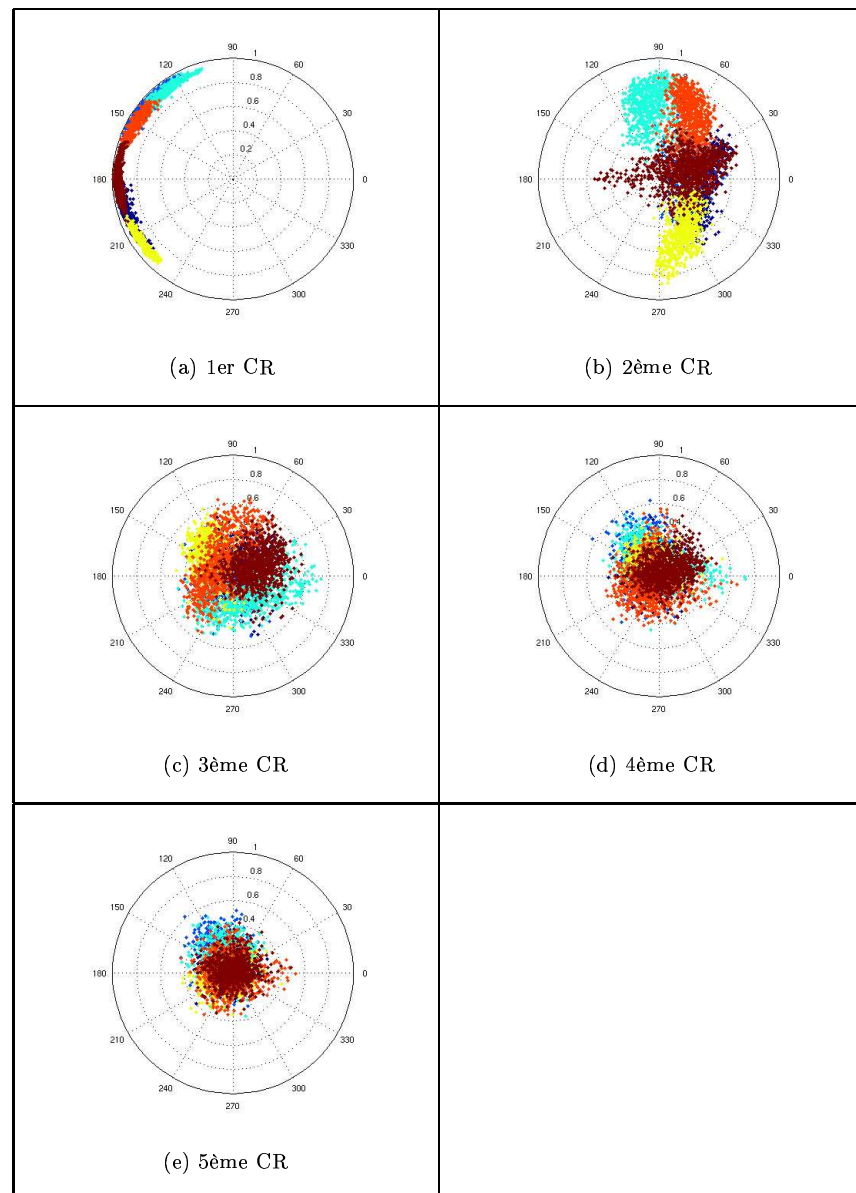


FIG. 6.14 – Les coefficients de réflexion classées par CMCA et MPM

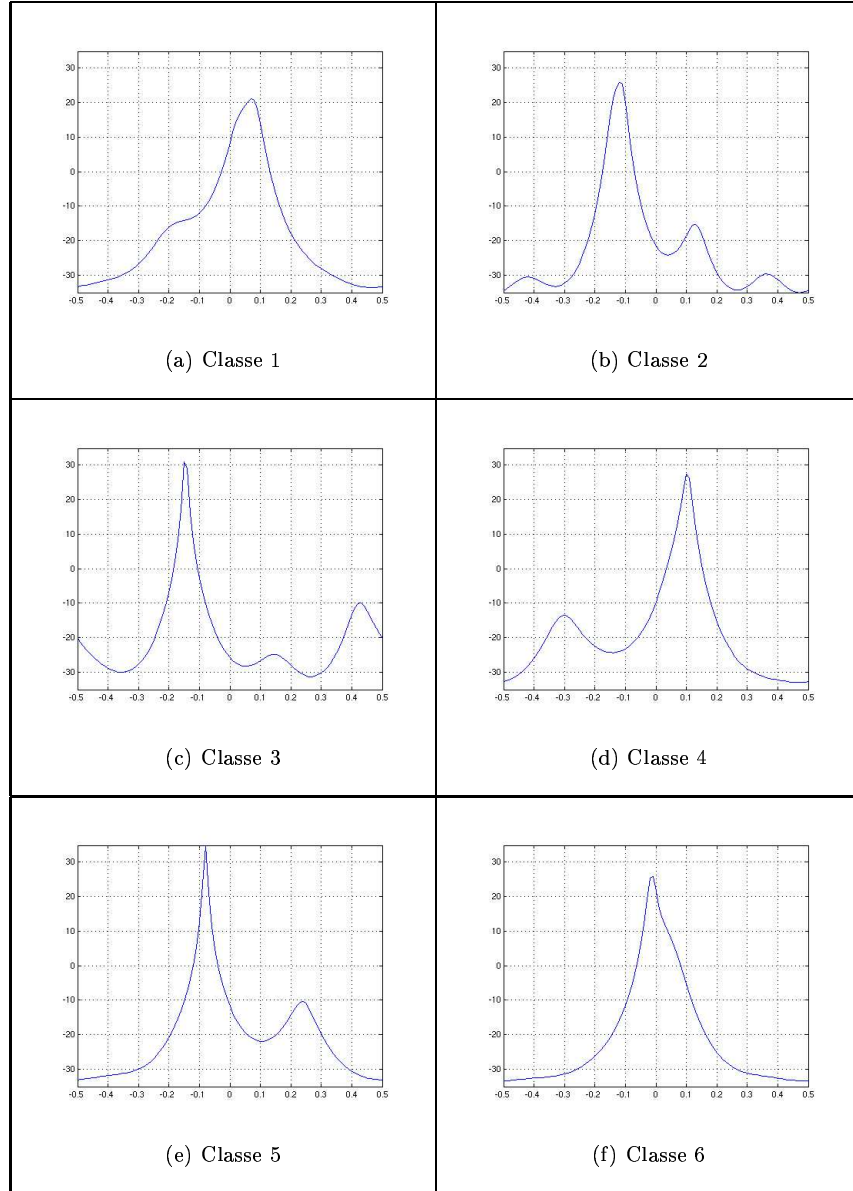


FIG. 6.15 – Graphes des spectres moyens ($20 \times \log_{10}(S(f))$) sur l'intervalle de fréquences $[-\frac{1}{2}, \frac{1}{2}]$.

Les nuages des observations classées en utilisant les informations $(\|\mu\|, \bar{\mu})$ sont rassemblées dans la figure (6.14). Nous pouvons caractériser les classes obtenues en calculant le vecteur μ moyen des observations classées selon l'estimateur MPM, et en représentant le spectre associé (correspondant au modèle autorégressif d'ordre 5), pour lequel nous avons imposé $\sigma_b^2 = 1$. Les 6 “densités spectrales moyennes” ainsi obtenues sont représentées en figure (6.15). Comme nous pouvons le voir dans la figure (6.14), ce sont essentiellement les deux premiers CR qui discriminent les classes, ce qui est bien représentée dans les spectres moyens, qui se caractérisent par la présence et la position de deux pics. Nous avons essentiellement 3 profils Doppler distincts à cette élévation :

- Classe 6 : un seul pic situé au doppler 0,
- Classe 1 et 4 : pic principal en fréquence positive et pic secondaire en fréquence négative,

– Classe 2, 3 et 5 : pic principal en fréquence négative et pic secondaire en fréquence positive. Les classes 1 et 5 apparaissent comme des classes de transition entre la classe 6 (au Doppler nul) et les classes 4 et 3 respectivement. Cette remarque est bien en accord avec la variation continue du spectre et la prise en compte d’une dépendance spatiale autre que celles des états cachées. Cependant notre manière de l’intégrer au modèle de segmentation, n’est pas adéquate pour en tenir compte. Elle ne permet pas entre autres de faire un “lissage des classes”, c’est-à-dire de diminuer le nombre de transitions entre classes dans la carte segmentée par MPM, comme dans le chapitre 4 dans le cas des intensités. La présence de ces comportements différents provient d’un cisaillement du vent et du fait que le lobe d’énergie émis par le radar s’élargit avec l’élévation ce qui fait que nous intégrons des signaux venant de plusieurs altitudes, et que nous mesurons partiellement le champ des vitesses d’altitude supérieur, qui correspond ici au pic secondaire.

En conclusion, nous pouvons constater qu’à partir d’estimation relativement bruitées des CR, nous pouvons séparer effectivement 3 profils distincts de fouillis Doppler. Sur cette exemple, nous voyons que les densités spectrales varient continuellement en distance, ce qui implique une dépendance spatiale entre les formes de celles-ci et contredit donc l’hypothèse d’indépendance conditionnelle des CMCa-BI. Cependant la complexité des observations que nous manipulons ne permet pas de bien décrire une dépendance conditionnelle et d’utiliser des modèles CMCa pertinents. Une extension de ce modèle consisterait donc en l’utilisation de lois de Von Mises bivariées (pour lesquelles il n’existe pas d’expression connue [107]) permettant de modéliser cette dépendance. Ce problème de la modification continue du spectre rend plus difficile la sélection du nombre de classes : le critère BIC n’a pas permis de sélectionner automatiquement le nombre de classes, de sorte que nous avons dû déterminer nous même le nombre de classes. Cependant, il est possible de sélectionner le nombre de classes par BIC de manière correcte lorsque nous n’avons pas une telle déformation continue du spectre sur la distance (voir [25] pour la segmentation Doppler de fouillis de pluie).

6.3 Segmentation polarimétrique

Avant d’introduire la procédure de segmentation polarimétrique, nous rappelons les propriétés polarimétriques de l’onde électromagnétique, et nous donnons différentes représentations de cette information.

6.3.1 Principe physique

6.3.1.1 Structure vectorielle

Une onde plane électromagnétique qui se propage dans l’espace est formée de 2 champs vectoriels (électrique et magnétique) couplés, mutuellement orthogonaux et variant dans le temps (nous pouvons déduire le champ magnétique du champ électrique). Nous désignons par $\vec{r}$ un point de $\mathbb{R}^3$ et par $E(\vec{r})$ le champ vectoriel électrique au point $\vec{r}$. Si $\vec{k}$ est la direction de propagation de l’onde, l’enveloppe complexe du champ électrique s’écrit alors $E(\vec{r}) = E_0 e^{i\vec{r} \cdot \vec{k}}$ (le champ électrique est perpendiculaire à la direction de propagation). A une position $\vec{r}$ et un instant t donnés, la direction et l’amplitude du champ électrique sont données par la partie réelle de $E_0 e^{i(\omega t + \vec{r} \cdot \vec{k})}$.

Nous nous intéressons à l'expression du champ électrique dans un repère orthonormal (e_h, e_v, e_z) , dans lequel $e_z = \vec{k}$. Nous appelons alors polarisation la nature vectorielle de l'onde plane dans le plan (e_h, e_v) (e_h est l'axe horizontal et e_v est l'axe vertical), perpendiculaire à la direction de propagation $\vec{k}$. L'onde s'écrit en fonction de l'ordonnée z , voir figure (6.16) :

$$E(z) = E_h(z)e_h + E_v(z)e_v \quad (6.14)$$

où $E_h(z)$ et $E_v(z)$ sont les coordonnées complexes de l'onde. Elles ont pour expression

$$\begin{cases} E_h(z) = E_{h0}e^{ikz}e^{i\delta_h} \\ E_v(z) = E_{v0}e^{ikz}e^{i\delta_v} \end{cases}$$

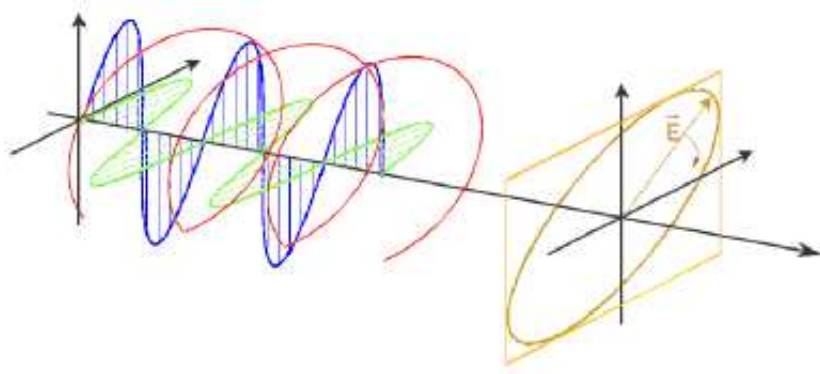


FIG. 6.16 – Propagation du champ électrique dans la direction $\vec{k}$ et ellipse de polarisation dans le plan perpendiculaire à $\vec{k}$

Nous pouvons voir alors par cette écriture que le lieu décrit par l'extrémité du vecteur champ électrique est une ellipse dans le plan (e_h, e_v) . En polarimétrie, cette ellipse est paramétrée par les angles d'inclinaison ψ et d'ellipticité χ . Ces angles sont définis par les relations suivantes

$$\begin{aligned} \tan(2\psi) &= \frac{2E_{h0}E_{v0}}{E_{h0}^2 - E_{v0}^2} \cos(\delta_0) \\ \tan(\chi) &= \pm \frac{a_\xi}{a_\eta} \end{aligned}$$

où $\delta_0 = \delta_v - \delta_h$ est la différence de phase entre les 2 composantes, et $2a_\xi, 2a_\eta$ sont les longueurs (resp.) des grand et petit axes de l'ellipse. Leur interprétation géométrique est donnée par la figure (6.17). L'angle χ caractérise la forme de l'ellipse, ainsi que son sens de parcours suivant son signe (polarisation gauche si $\chi < 0$, polarisation droite sinon). Ces angles s'expriment en fonction des caractéristiques de l'onde :

$$\begin{aligned} \sin(2\chi) &= \sin(2\alpha) \sin(\delta_0) \\ \tan(2\psi) &= \tan(2\alpha) \cos(\delta_0) \end{aligned}$$

L'angle α est défini par $\tan \alpha = \frac{E_{v0}}{E_{h0}}$. En général, nous disons que l'onde est polarisée elliptiquement. Si la différence de phase δ_0 est égale à $\pm \frac{\pi}{2}$ modulo 2π , la polarisation est circulaire, et si δ_0 est nulle, la polarisation est rectiligne.

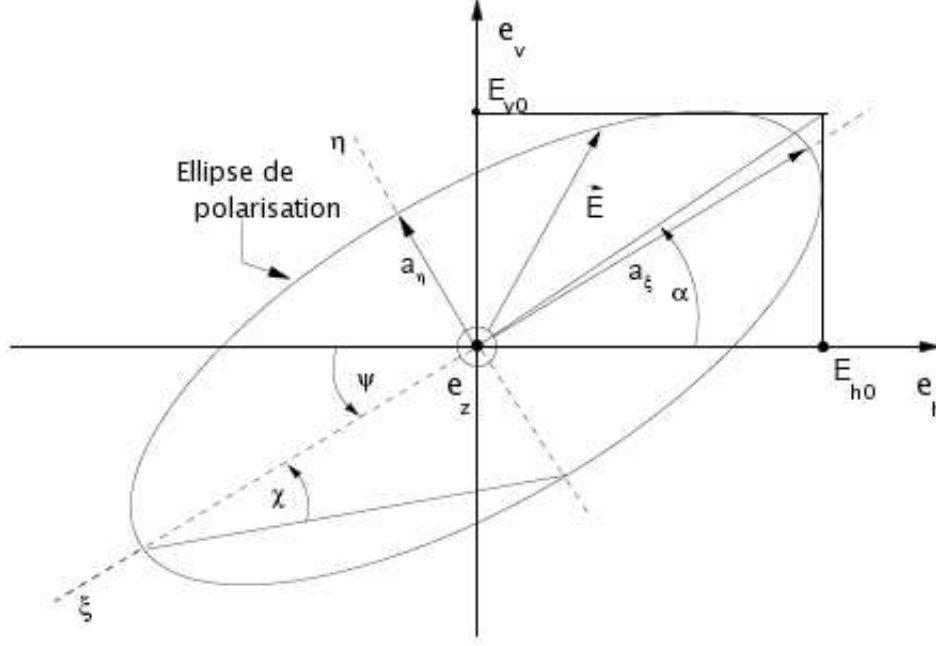


FIG. 6.17 – Paramétrisation de l'ellipse de polarisation par angle d'inclinaison ψ et d'ellipticité χ

6.3.1.2 Vecteur de Stokes

L'onde plane de l'éq. (6.14) peut aussi être représentée par un vecteur $\mathbf{g}$ de $\mathbb{R}^4$ appelé vecteur de Stokes

$$\mathbf{g} = \begin{bmatrix} g_0 \\ g_1 \\ g_2 \\ g_3 \end{bmatrix} = \begin{bmatrix} |E_v|^2 + |E_h|^2 \\ |E_v|^2 - |E_h|^2 \\ 2\Re(E_v E_h^*) \\ 2\Im(E_v E_h^*) \end{bmatrix} = \begin{bmatrix} |E_{v0}|^2 + |E_{h0}|^2 \\ |E_{v0}|^2 - |E_{h0}|^2 \\ 2E_{v0}E_{h0}\cos(\delta) \\ 2E_{v0}E_{h0}\sin(\delta) \end{bmatrix} \quad (6.15)$$

g_0 est l'énergie totale de l'onde, g_1 est la différence entre les énergies polarisées horizontalement et verticalement, et les entrées g_2, g_3 contiennent l'information de différences de phase entre les composantes h et v . Dans le cas d'une onde complètement polarisée (ce qui est presque le cas de l'onde émise par le radar, la dépolarisation étant due aux défauts de l'antenne), nous avons $g_0^2 = g_1^2 + g_2^2 + g_3^2$. Le vecteur de Stokes est relié au paramètre de l'ellipse de polarisation :

$$\mathbf{g} = g_0 \begin{bmatrix} 1 \\ \cos(2\psi) \cos(2\chi) \\ \sin(2\psi) \cos(2\chi) \\ \sin(2\chi) \end{bmatrix} \quad (6.16)$$

où ψ et χ sont les angles d'inclinaison et d'ellipticité. Nous pouvons donc associer à tout vecteur de Stokes un point de la sphère de rayon g_0 (appelée la sphère de Poincaré) voir figure (6.18). Les points qui se situent sur l'équateur correspondent à une polarisation rectiligne (car $\chi = 0$), quant à ceux qui sont aux pôles, ils correspondent à une polarisation circulaire ($\chi = \frac{\pi}{4}$). Enfin, le point $(\chi, \psi) = (0, 0)$ correspond à la polarisation rectiligne horizontale, et le point $(\chi, \psi) = (0, \frac{\pi}{2})$ à la polarisation vorticale.

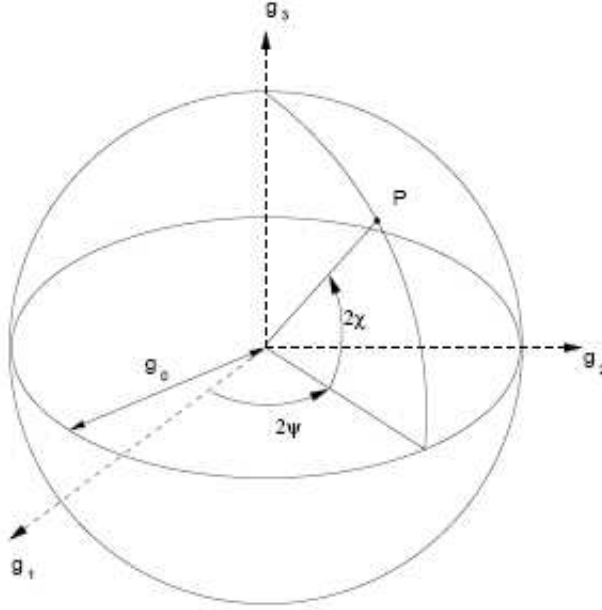


FIG. 6.18 – Représentation d'un vecteur de Stokes sur la sphère de Poincaré

En revanche, l'onde rétrodiffusée est la somme cohérente des ondes rétrodiffusées par les cibles élémentaires dans une même case-distance, qui est donc une variable aléatoire variant au cours du temps et vérifiant l'inégalité

$$g_0^2 \geq g_1^2 + g_2^2 + g_3^2 \quad (6.17)$$

Malgré le caractère aléatoire de l'onde rétrodiffusée, la représentation (6.16) de son état de polarisation sur la sphère de Poincaré est toujours valide, car toute onde plane est la somme d'une onde complètement polarisée et d'une onde dépolarisée (on dit que l'onde est partiellement polarisée). En introduisant le degré de polarisation $d = \frac{\sqrt{g_1^2 + g_2^2 + g_3^2}}{g_0}$, nous avons :

$$\mathbf{g} = g_0 \begin{bmatrix} 1 - d \\ 0 \\ 0 \\ 0 \end{bmatrix} + g_0 \begin{bmatrix} d \\ \cos(2\psi) \cos(2\chi) \\ \sin(2\psi) \cos(2\chi) \\ \sin(2\chi) \end{bmatrix}$$

L'ellipse de la partie polarisée est appelée état de polarisation de l'onde.

Le vecteur de Stokes moyen peut être associé à un point de la sphère de Poincaré, correspondant alors à une ellipse de polarisation moyenne. Les angles d'inclinaison et d'ellipticité sont égaux à :

$$\begin{aligned}\chi &= \frac{1}{2} \arcsin\left(\frac{g_3}{g_0}\right) \\ \psi &= \frac{1}{2} \arctan\left(\frac{g_3}{g_2}\right)\end{aligned}$$

6.3.2 La segmentation polarimétrique

La polarisation est une caractéristique de l'onde électromagnétique qui peut être modifiée par le milieu dans lequel l'onde se propage et sur lequel elle se réfléchit. Ces modifications permettent d'acquérir des informations supplémentaires sur l'environnement radar, et particulièrement sur les éléments fixes qui sont difficiles à caractériser par effet Doppler. Des exemples d'application de l'imagerie polarimétrique SAR pour caractériser finement l'environnement (notamment la végétation, mais aussi la géométrie des réflecteurs) sont donnés dans [106]. Nous proposons dans cette section deux procédures qui utilisent les données obtenues pour une seule voie d'émission, et nous notons comme précédemment les réceptions E_h et E_v . L'objectif est de réaliser une classification des fouillis polarimétriques uniquement selon leur ellipse de polarisation (indépendamment de la réflectivité de la cible) à l'aide de méthode de segmentation bayésienne permettant d'intégrer l'information spatiale.

Remarque 6.3.1. *D'autres méthodes de segmentation utilisent la matrice de covariance entre les réceptions E_h et E_v , voir par exemple [101, 43], mais les méthodes proposées n'utilisent pas l'information spatiale par un modèle CMCA. Ces méthodes exploitent conjointement 3 types d'information : l'état de polarisation, le degré de polarisation et enfin la puissance rétrodiffusée.*

Ainsi en chaque case distance n , nous calculons à partir des n_i récurrences le vecteur de Stokes $\mathbf{g}_n$ sur la rafale, et nous en déduisons l'état de polarisation (χ_n, ψ_n) . A partir de la représentation des états de polarisation sur la sphère de Poincaré, nous proposons un modèle CMCA-BI donc le noyau de transition est

$$p(z_{n+1} | z_n) = p(x_{n+1} | x_n) p((\chi_n, \psi_n) | x_{n+1}) \quad (6.18)$$

avec $p((\chi_n, \psi_n) | x_{n+1})$ densité d'une loi de Von Mises-Fisher sur S^2 , de paramètre $(\kappa, \xi = (\chi, \psi))$. Chaque classe est ainsi caractérisée par un état de polarisation moyen et un paramètre de concentration.

6.3.3 Exemple

Nous montrons sur données réelles une application de la segmentation basée sur l'état de polarisation. Nous reprenons les résultats présentées dans [25], illustrant les techniques de segmentation Doppler et polarimétrique sur les données déjà traitée dans la section 5.3.3.3.

Nous avons à notre disposition les réceptions sur les voies horizontale et verticale des signaux radar (après émission horizontale), grâce auxquelles nous pouvons calculer vecteur de Stokes et état de polarisation. La rafale considérée comporte 256 cases-distances pour lesquelles nous avons 16 récurrences. Par utilisation du critère BIC, nous sélectionnons une CMCA-BI à 3 classes pour réaliser la segmentation, dont nous donnons le résultat dans la figure (6.19).

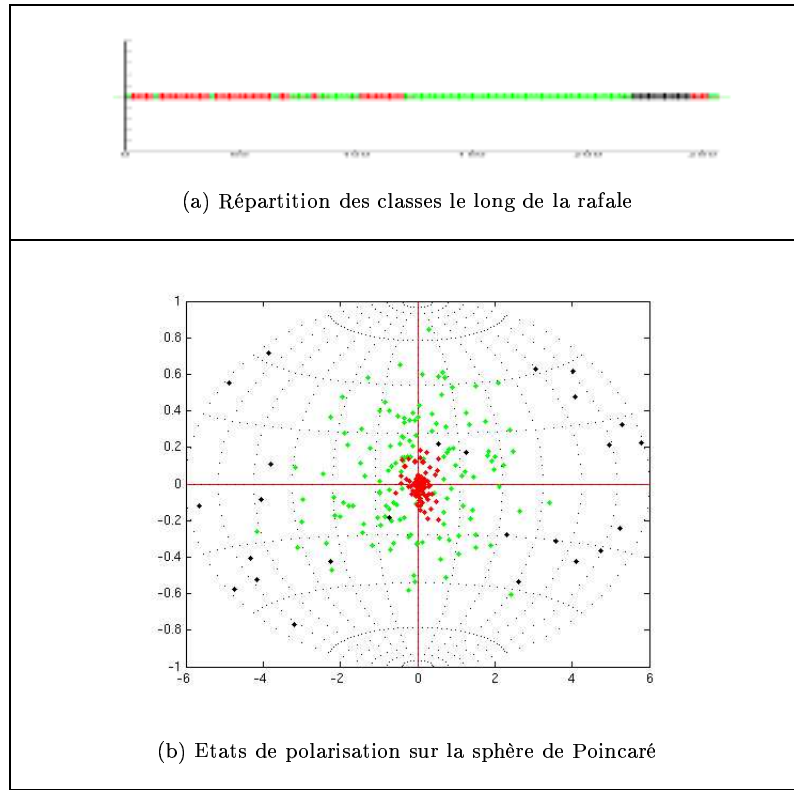


FIG. 6.19 – Segmentation des états de polarisation. La sphère de Poincaré est projetée sur le plan selon la projection de Aitoff-Hammer.

Classe	Etats de Polarisation (χ, ψ) en radians	Concentration	Localisation
1	(0 0)	80,1	3-12, 15-36, 39-63 67-71, 81-83, 102-120
2	(0,03 -0,02)	3,3	1-2, 13-14, 37-38 64-66, 72-80, 84-101 121-219, 253-256
3	(-0,05 1,43)	1,7	220-244

TAB. 6.2 – Paramètres estimées.

Les paramètres du modèle permettent d'avoir une caractérisation des classes : la première classe (rouge) est une classe dont la réponse polarimétrique est très concentrée, et dont l'état de polarisation est rectiligne horizontale (i.e. similaire à l'émission). Elle est constituée essentiellement de la zone de pluie située dans la première partie de la radiale (la pluie s'étend des cases 1 à 135, et 241 à 256)⁵. Nous retrouvons ainsi que la pluie ré-émet de manière très polarisée, et c'est sa réponse que nous mesurons essentiellement dans les zones où elle se situe. La seconde classe (verte) a aussi un état de polarisation moyen rectiligne uniforme, mais a une émission beaucoup plus hétérogène : le paramètre de concentration est beaucoup plus faible. Elle comprend la grande zone de verdure et d'habitat dispersé (sans pluie) à la réponse moins homogène, ainsi que quelques

⁵La segmentation Doppler utilisant l'information $\bar{\mu}$ sépare la rafale en trois profils différents : une classe pluie et deux classes correspondant à des fouillis de sol [25].

constructions ou reliefs dans la première partie de la radiale qui ont des réponses polarimétriques assez fortes pour “dépasser” la réponse de la pluie. Enfin la dernière classe (noire), qui contient peu d’observations, est très dépolarisée et l’état de polarisation moyen est difficilement interprétable. Cette classe correspond à une zone de hauteur et de relief qui n’a pas de propriété polarimétrique unique. Les degrés de polarisation moyen des 3 classes sont $d_1 = 0,97$, $d_2 = 0,87$ et $d_3 = 0,85$, et viennent confirmer les conclusions précédentes (ils sont une mesure de l’homogénéité temporelle de la polarisation) de faible homogénéité polarimétrique des classes 2 et 3, et de la très forte polarisation de la classe 1.

La segmentation par CMCa-BI et loi de Von Mises-Fisher apparaît comme un modèle particulièrement simple qui permet néanmoins de retrouver des comportements très différents (ainsi que leur nombre par utilisation du critère BIC), et de les caractériser par une état moyen de polarisation et une mesure d’homogénéité, tout en étant insensible aux fluctuations d’intensité (puisque celle-ci n’est pas prise en compte). Comme pour la segmentation Doppler, l’utilisation de lois bivariées sur la sphère possédant des marges de type Von Mises - Fisher permettrait d’utiliser des CMCa au lieu de CMCa-BI, et donc de détecter et d’exploiter une dépendance spatiale supplémentaire entre les observations.

Chapitre 7

CONCLUSION

Bilan

Le travail que nous avons développé s'inscrit dans le cadre de la segmentation bayésienne non-supervisée des signaux et des images. Nous avons apporté un certain nombre de contributions à cette approche par l'utilisation d'un modèle original, ce dernier faisant partie de la famille générale des chaînes de Markov couples. Ce modèle permet une meilleure modélisation de la loi du processus des observations, en prenant en compte la dépendance conditionnelle à l'aide des copules, ce qui fournit entre autres une meilleure description de la fonction d'autocovariance du processus observé. Malgré la complexité accrue du modèle proposé, les procédures de segmentation non-supervisée restent rapides. En effet, d'une part nous restons dans le cadre des modèles couples et d'autre part nous utilisons des méthodes originales d'estimation des paramètres, conçues pour simplifier les calculs, et basées sur la projection de fonctions estimantes.

Nous avons aussi généralisé l'usage des copules aux chaînes de Markov cachées multivariées afin de procéder à une meilleure modélisation des lois des bruits qui peuvent être utiles en traitement d'image ou du signal. Nous proposons un algorithme simple et rapide d'estimation, fondé encore sur la projection de fonctions estimantes. Nous avons aussi introduit l'utilisation des lois T et K en segmentation d'image et du signal, et nous donnons les formules analytiques de l'algorithme EM qui permettent d'en faire l'estimation par maximum de vraisemblance.

Par ailleurs, nous avons utilisé des modes de représentation des informations Doppler et polarimétrique qui permettent de cartographier l'environnement radar de telle sorte que ces informations soient utilisées indépendamment de la réflectivité pour la segmentation. Les méthodes proposées ont été testées sur des données simulées, ainsi que sur des données réelles. Concernant la segmentation utilisant l'information de réflectivité, nous montrons que la prise en compte de la dépendance conditionnelle, rendue possible par les nouveaux modèles proposés, peut constituer une réelle amélioration. Enfin, nous apportons quelques précisions sur le travail informatique ayant permis les simulations et expérimentations réalisées dans cette thèse. Les programmes ont été développés en langage MATLAB et C, et implémentent les méthodes d'estimation que nous avons présentées dans ce mémoire (EM, SEM, IFM-ECI version déterministe et stochastique), les différentes lois d'émission (SIRV, familles exponentielles, copules) et les modèles CMCa-BI, CMCa et CMCo. Ces programmes seront intégrés à la plate-forme de simulation en cours de développement au sein du

groupe THALES Air Defence.

Perspectives

Nous identifions essentiellement 3 axes de prolongements possibles qui concernent chacun un des domaines abordés dans la thèse : la statistique, la modélisation en traitement d'image et les applications spécifiques au radar. Le premier axe est théorique et concerne l'étude des hypothèses sous lesquelles nous pouvons assurer la consistance des estimateurs obtenus par projection des fonctions estimantes. En suivant le raisonnement heuristique présenté dans la section 2.5.1, la nécessité du bon comportement de l'estimateur sur données complètes apparaît clairement. L'objectif est alors de déterminer des hypothèses en termes de propriétés du processus complet sous lesquelles l'existence et la consistance d'une racine de la fonction peut être conservée, pour pouvoir appliquer ce résultat aux chaînes et aux champs de Markov cachés, ou plus généralement aux processus markoviens partiellement observés. Pour résoudre les fonctions estimantes, nous avons aussi proposé un algorithme intuitif qui se comporte "bien" dans les applications, mais pour lequel il manque encore des hypothèses vérifiables pour garantir son existence et sa convergence dans des situations pratiques. Un prolongement de notre travail est donc le développement de critères plus faciles à vérifier pour s'assurer de la convergence de la méthode proposée (ECI). Un autre prolongement des aspects relatifs à l'estimation des modèles à données manquantes consiste en le développement de variantes de ECI similaires à celles de l'algorithme EM, telles que l'algorithme *Expectation Conditional Maximization* de Meng et Rubin [110] ou l'algorithme *Stochastic Approximation* EM de Delyon, Lavielle et Moulines [51], afin d'accélérer la procédure d'estimation ou de limiter l'influence de la valeur initiale.

Nous avons développé l'utilisation des copules dans le cadre des chaînes de Markov pour modéliser la dépendance conditionnelle, afin de pouvoir utiliser n'importe quelle loi paramétrique pour les marginales. Nous avons vu dans le chapitre 1 que la modélisation couple pouvait aussi être faite dans les champs de Markov : il est donc possible d'étendre la méthode de classification des données dépendantes développées pour les chaînes de Markov cachées (dans le chapitre 4) aux champs de Markov cachés. La possibilité de prendre en compte d'une part des marginales particulières et d'autre part de choisir des structures de dépendance de champ de Markov spécifiées par copules doit permettre d'améliorer significativement les procédures de segmentation non-supervisée (comme le montrent déjà les travaux de D. Benboudjema dans lesquels la loi des observations conditionnellement aux observations est un champ gaussien markovien [15]). Il est vraisemblable que l'estimation par projection de fonctions estimantes et l'algorithme ECI constitue un moyen pratique de construire (et d'étudier) un estimateur simple de ces modèles.

Dans le cadre de la modélisation du signal radar, nous avons proposé dans la section 1.3 un modèle permettant de tenir compte de la corrélation de la texture, tout en garantissant que la loi marginale des observations appartiennent aux modèles SIRV. Ce modèle peut être étendu pour la modélisation d'un environnement hétérogène, en introduisant une chaîne de Markov représentant les classes, de telle sorte que l'on puisse utiliser un modèle validé sur les données pour représenter le signal radar dans les procédures de segmentation. Nous avons proposé dans [27] une méthode d'estimation des processus cachés (classe et texture) utilisant un algorithme de filtrage particulière. Un prolongement consisterait alors en le développement de l'estimation des paramètres par

ECI ou EM, reposant sur l'utilisation d'un filtrage particulière pour calculer les probabilités a posteriori nécessaires. Ceci permettrait, premièrement, le développement de procédures de segmentation non-supervisée rapides, que l'on pourrait comparer avec la méthode présentée dans [29] et qui néglige la corrélation de la texture. Deuxièmement, cela donnerait une caractérisation simple de l'environnement physique sous forme d'une décomposition texture et speckle. Un autre prolongement concerne l'application de la cartographie à l'amélioration de la détection. Pour cela, nous avons donné dans la section 6.1.2 différentes exploitations de la carte qu'il serait intéressant de tester et de comparer sur des données réelles.

Enfin, bien que les travaux de cette thèse aient été développés dans le contexte du traitement d'image et du signal, les modèles et méthodes d'estimation que nous proposons, ainsi que les perspectives de recherche que nous envisageons sont susceptibles d'être appliquées dans les domaines ayant déjà adopté la modélisation de modèles de Markov cachés et leurs extensions, comme l'économétrie (avec les modèles à volatilité stochastique [143]) ou la génomique (avec les modèles markoviens et semi-markoviens [64]).

Annexe A

Vecteurs Gaussiens Complexes

Les signaux radar sont des variables aléatoires dans $\mathbb{C}$ ou $\mathbb{C}^d$. Nous rappelons les définitions de vecteur aléatoire complexe, de circularité¹ et de vecteur gaussien complexe, utilisées communément en traitement du signal (voir le chapitre 15 de [98], notamment sur l'adaptation des procédures statistiques classiques des données réelles aux données complexes). Nous donnons aussi la généralisation complexe des vecteurs sphériquement invariants.

A.1 Généralités

Définition A.1.1. Vecteur aléatoire Complexe

Soient X et Y deux vecteurs aléatoires dans $\mathbb{R}^d$, la variable aléatoire $Z = X + iY \in \mathbb{C}^d$ est appelée vecteur aléatoire complexe.

- Si les espérances $m_X = E[X]$ et $m_Y = E[Y]$ existent, l'espérance de Z est définie par $E[X] + iE[Y]$, et est notée $E[Z]$ ou m_Z .
- Si les espérances $E[XX']$, $E[YY']$ et $E[XY']$ existent, nous définissons les matrices C_Z et S_Z par

$$\begin{cases} S_Z &= E[(Z - E[Z])(Z - E[Z])'] \\ C_Z &= E[(Z - E[Z])(Z - E[Z])^*] \end{cases} \quad (\text{A.1})$$

* désigne l'opérateur de transposition-conjugaison (transposition hermitienne), et ' l'opérateur de transposition.

La matrice C_Z est toujours semi-définie positive (contrairement à S_Z) et est appelée covariance de Z . Les matrices C_Z et S_Z peuvent être exprimées en fonction des matrices de covariance C_X et C_Y des vecteurs aléatoires réels X et Y , ainsi que de leur covariance croisée $C_{XY} = E[(X - E[X])(Y - E[Y])']$. Nous avons

¹voir [121] pour une discussion de la notion de circularité pour les signaux complexes.

$$\begin{cases} C_Z &= C_X + C_Y + i (C'_{XY} - C_{XY}) \\ S_Z &= C_X - C_Y + i (C'_{XY} + C_{XY}) \end{cases} \quad (\text{A.2})$$

Nous pouvons aussi considérer que le vecteur complexe $Z = X + iY$ est un vecteur aléatoire réel $\tilde{Z} = \begin{pmatrix} X' & Y' \end{pmatrix}'$ à valeurs dans $\mathbb{R}^{2d}$, obtenu en concaténant parties réelle et imaginaire. La moyenne de $\tilde{Z}$ est donc $m_{\tilde{Z}} = \begin{pmatrix} E[X] \\ E[Y] \end{pmatrix}$, et sa matrice de covariance se décompose en blocs

$$C_{\tilde{Z}} = \begin{pmatrix} C_X & C_{XY} \\ C_{YX} & C_Y \end{pmatrix}$$

Ainsi la connaissance des matrices C_Z et S_Z est équivalente à la connaissance de la matrice $C_{\tilde{Z}}$ (qui contient plus d'informations que la seule matrice C_Z en général).

Définition A.1.2. Circularité

Un vecteur aléatoire complexe Z est (fortement) circulaire si Z et $e^{i\theta}Z$ ont même loi, pour tout $\theta \in \mathbb{R}$.

Un vecteur aléatoire complexe Z est circulaire d'ordre deux si nous avons $S_Z = 0$.

Ainsi un vecteur aléatoire qui est circulaire d'ordre 2 est telle que

$$\begin{cases} C_{XY} = -C_{YX} \\ C_X = C_Y \end{cases} \quad (\text{A.3})$$

soit

$$C_Z = 2(C_X + iC_{YX}) \quad (\text{A.4})$$

Nous avons donc équivalence de la connaissance de C_Z et de la connaissance de $C_{\tilde{Z}}$. La circularité forte implique la circularité d'ordre 2.

A.2 Vecteurs gaussiens et sphériquement invariants

Définition A.2.1. Vecteur gaussien complexe

Soit $Z = X + iY$ un vecteur aléatoire complexe dans $\mathbb{C}^d$. Il est dit normal ou gaussien si $\tilde{Z} = \begin{pmatrix} X' & Y' \end{pmatrix}'$ est un vecteur gaussien de loi $N(m_{\tilde{Z}}, C_{\tilde{Z}})$. Nous dirons que Z suit une loi $CN(m_Z, C_Z)$.

Pour un vecteur gaussien, les propriétés de circularité forte et circularité à l'ordre 2 sont équivalentes. Dans ce cas, sa densité relativement à la mesure de Lebesgue sur $\mathbb{R}^d \times \mathbb{R}^d$ s'écrit alors :

$$f_Z(z_1, \dots, z_d) = \frac{1}{\pi^d |C_Z|} \exp \left(-(\mathbf{z} - m_Z)^* C_Z^{-1} (\mathbf{z} - m_Z) \right)$$

Nous pouvons aussi écrire sa densité comme un vecteur de $\mathbb{R}^{2d}$, en notant $\forall k, z_k = x_k + iy_k$:

$$f_Z(z_1, \dots, z_d) = \frac{1}{(2\pi)^d |C_{\tilde{Z}}|^{1/2}} \exp \left(-\frac{1}{2} \left((x - m_X)' (y - m_Y)' \right) C_{\tilde{Z}}^{-1} \begin{pmatrix} x - m_X \\ y - m_Y \end{pmatrix} \right)$$

Définition A.2.2. Vecteur complexe sphériquement invariant

Soit $Z = X + iY$ un vecteur aléatoire complexe dans $\mathbb{C}^d$. Il est dit sphériquement invariant si nous pouvons l'écrire

$$Z = m_Z + U\epsilon$$

avec $\epsilon \sim CN(0, \Sigma)$, U variable aléatoire réelle telle que $U > 0$, $m_Z \in \mathbb{C}^d$.

Les SIRV complexes sont donc une généralisation des SIRV réels (voir section 4.2). Si nous adoptons l'écriture $Z = m_Z + U^{-1/2}\epsilon$, alors la densité de Z est :

$$f_Z(\mathbf{z}) = \frac{|\Sigma|^{-1}}{\pi^d} \int_0^\infty u^d e^{-uq(\mathbf{z})} g(u) du \quad (\text{A.5})$$

qui est la version complexe de l'équation (4.12). Nous avons noté $q(\mathbf{z}) = (\mathbf{z} - m_Z)^* \Sigma^{-1} (\mathbf{z} - m_Z)$ et g pour la densité de la texture U . Nous avons $C_Z = E[(Z - E[Z])(Z - E[Z])^*] = E[U^{-1/2}] \Sigma$.

La loi K complexe a pour densité

$$f_K(\mathbf{z}, (m_Z, \Sigma, a)) = \frac{2a^a |\Sigma|^{-1}}{\pi^d \Gamma(a)} \left(\frac{q(\mathbf{z})}{a} \right)^{\frac{a-d}{2}} K_{a-d} \left(2\sqrt{aq(\mathbf{z})} \right) \quad (\text{A.6})$$

La loi T complexe a pour densité

$$f_T(\mathbf{z}(m_Z, \Sigma, \nu)) = \frac{2^d \Gamma(\frac{\nu}{2} + d) |\Sigma|^{-1}}{(\pi\nu)^d \Gamma(\frac{\nu}{2})} \left(1 + \frac{2q(\mathbf{z})}{\nu} \right)^{-(\frac{\nu}{2} + d)} \quad (\text{A.7})$$

Annexe B

Processus stationnaires et coefficients de réflexion

Les propriétés au second ordre d'un processus stationnaire complexe centré $\mathbf{X} = (X_n)_{n \in \mathbb{Z}}$ sont décrites par la fonction d'autocovariance définie pour tout $n \in \mathbb{Z}$ par $R(n) = E[X_0 X_n^*]$. Il est équivalent de connaître sa densité spectrale $S(f) = \sum_{n \in \mathbb{Z}} R(n) e^{2i\pi n f}$ (définie sur le segment $]-\frac{1}{2}, \frac{1}{2}[$). Nous rappelons ici la définition des coefficients de réflexion μ_n que nous utilisons pour la segmentation Doppler (section 6.2), ainsi que leur lien avec le problème de prédiction linéaire. Nous donnons les relations de passage entre ces paramétrages en termes de décomposition de la matrice de covariance du processus. Finalement, nous rappelons l'algorithme de Burg et sa version régularisée (développée par Barbaresco [10]).

B.1 Prédiction linéaire

Le problème de prédiction linéaire du processus $\mathbf{X}$ à l'ordre p est la recherche de la meilleure approximation de X_n (au sens de L^2) par $X_{n-1}, \dots, X_{n-p}$, ce qui revient à chercher le vecteur de $\mathbb{C}^p$ atteignant

$$\min_{(\alpha_1, \dots, \alpha_p) \in \mathbb{C}^p} \left\| X_n - \sum_{k=1}^p \alpha_k X_{n-k} \right\| \quad (\text{B.1})$$

La solution (unique) à (B.1) est notée $A_p = \left(a_1^{(p)} \dots a_p^{(p)} \right)' \in \mathbb{C}^p$ et $\sigma_p^2 = \left\| X_n - \sum_{k=1}^p a_k^{(p)} X_{n-k} \right\|^2$ est la variance de l'erreur de prédiction. Nous notons par extension $\sigma_0^2 = R(0)$, et les coefficients $a_k^{(p)}$ sont appelés coefficients autorégressifs (AR). La prédiction linéaire est équivalente à la recherche de la projection orthogonale de X_n sur l'espace vectoriel engendré par $X_{n-1}, \dots, X_{n-p}$.

Soient $n, m \in \mathbb{Z}$ ($n \leq m$) et $P_{n,m}$ la projection orthogonale sur l'espace vectoriel engendré par $\{X_n, \dots, X_m\}$. Nous avons donc $P_{n-p, n-1}(X_n) = \sum_{k=1}^p a_k^{(p)} X_{n-k}$. Par définition, le coefficient $a_p^{(p)}$ est le p -ième coefficient de réflexion, noté μ_p . Il est aussi appelé coefficient de corrélation partielle parce qu'il est égal au coefficient de corrélation entre les variables $e_p^r = X_1 - P_{2,p}(X_1)$ et $e_p^d = X_{p+1} - P_{2,p}(X_{p+1})$ (correspondant aux erreurs de prédiction rétrograde et directe).

Nous donnons les liens entre les premiers termes de la fonction d'autocovariance $(R(n))_{n \geq 0}$, les vecteurs $(A_p)_{p \geq 1}$ et les coefficients de réflexions $(\mu_n)_{n \geq 1}$ sous forme matriciel. Ceux-ci mettent en évidence les propriétés algébriques des matrices Toeplitz et de la décomposition de Choleski.

B.2 Relations de Passage

Nous renvoyons à [88, 86] pour une présentation détaillée et des applications des algorithmes permettant de faire le lien entre la matrice de covariance R_p (de taille $p + 1$), les paramètres autorégressifs A_p et les coefficients de réflexion $\mu = (\mu_1, \dots, \mu_p)$.

B.2.1 Prédiction linéaire et autocovariance : $R_p \leftrightarrow A_p$

Nous pouvons exprimer A_p à partir de la fonction d'autocovariance. Pour cela, nous notons R_p la matrice d'autocovariance ($\in \mathcal{SDP}(p + 1)$) du vecteur $(X_0, \dots, X_p)$, et C_p pour le vecteur $(R(1) \dots R(p))' \in \mathbb{C}^p$, pour tout $p \geq 1$. Nous supposons que pour tout p , la matrice R_p est inversible. Le vecteur A_p est déterminé alors par les équations de **Yule-Walker** :

$$\begin{cases} R_{p-1} A_p &= C_p \\ R(0) - C_p' A_p &= \sigma_p^2 \end{cases} \quad (\text{B.2})$$

La relation inverse de (B.2), i.e. le passage des coefficients AR à la matrice R_p est donnée par la formule de **Gohberg-Semencul** :

$$R_p^{-1} = \frac{1}{\sigma_p^2} (K_1 K_1^* - K_2 K_2^*) \quad (\text{B.3})$$

où nous avons :

$$K_1 = \begin{bmatrix} 1 & & & & \\ -a_1^{(p)} & 1 & & & (0) \\ \vdots & \ddots & \ddots & & \\ \vdots & & \ddots & 1 & \\ -a_p^{(p)} & \dots & \dots & -a_1^{(p)} & 1 \end{bmatrix} \quad \text{et} \quad K_2 = \begin{bmatrix} 0 & & & & \\ -a_p^{(p)} & 0 & & & (0) \\ \vdots & \ddots & \ddots & & \\ \vdots & & \ddots & 0 & \\ -a_1^{(p)} & -a_2^{(p)} & \dots & -a_p^{(p)} & 0 \end{bmatrix}$$

B.2.2 Prédiction linéaire et coefficients de réflexion : $A_p \leftrightarrow \mu$

Les coefficients AR jouent un rôle de pivot entre fonction d'autocovariance et coefficients de réflexion. Les algorithmes de Levinson direct et inverse sont des algorithmes récursifs permettant de lier prédiction linéaire avec coefficients de réflexion.

B.2.2.1 Algorithme de Levinson

L'algorithme de Levinson permet de résoudre itérativement les équations de Yule-Walker (B.2) à partir des CR $(\mu_1, \dots, \mu_p)$. Pour tout vecteur $x = (x_1 \dots x_p)' \in \mathbb{C}^p$, nous désignons par $x^{(-)} =$

$(x_p \dots x_1)^*$, le vecteur “renversé” et transconjugué, et A_0 est le vecteur vide. Nous avons alors

$$\forall 0 \leq k \leq p-1, A_{k+1} = \begin{bmatrix} A_k \\ 0 \end{bmatrix} + \mu_{k+1} \begin{bmatrix} A_k^{(-)*} \\ 1 \end{bmatrix} \quad (\text{B.4})$$

Les puissances d’erreur sont obtenues elles aussi récursivement :

$$\forall 0 \leq k \leq p-1, \sigma_{k+1}^2 = \sigma_k^2(1 - |\mu_{k+1}|^2) \quad (\text{B.5})$$

B.2.2.2 Algorithme de Levinson inverse

A partir de la seule connaissance du vecteur A_p et de σ_p^2 , nous pouvons retrouver tous les vecteurs A_k pour $k \leq p-1$, et par conséquent tous les coefficients de réflexion $(\mu_k)_{1 \leq k \leq p-1}$ correspondants. Cette procédure récursive est possible si tous les coefficients de réflexion ont un module inférieur à 1. Dans ce cas-là, le passage de A_p à A_{p-1} se fait de la manière suivante :

$$\forall 1 \leq k \leq p-1, a_k^{(p-1)} = \frac{1}{1 - |\mu_p|^2} (a_k^{(p)} - \mu_p a_{p-k}^{(p)}) \quad (\text{B.6})$$

En particulier, le coefficient de réflexion d’ordre inférieur s’écrit :

$$\mu_{p-1} = \frac{1}{1 - |\mu_p|^2} (a_{p-1}^{(p)} - \mu_p a_1^{(p)})$$

Remarque B.2.1. Si les CR $(\mu_k)_{1 \leq k \leq p}$ sont tous de module strictement inférieur à 1, alors le polynôme autorégressif $1 - \sum_{k=1}^p a_k^{(p)} z^{-k}$ est à minimum de phase, i.e. toutes ses racines sont dans le disque unité.

B.2.3 Factorisation de Choleski de $R_p^{-1} : \mu \leftrightarrow R_p$

Les coefficients de réflexion sont liés à la matrice de covariance, par la décomposition de Choleski de l’inverse de R_p . Nous avons

$$R_p^{-1} = \mathbf{A}_p \Sigma_p^{-1} \mathbf{A}_p^*$$

avec

$$\mathbf{A}_p = \begin{bmatrix} 1 & & & & & \\ a_1^{(p)} & 1 & & & & \\ a_2^{(p)} & a_1^{(p-1)} & \ddots & & & \\ \vdots & & \ddots & \ddots & & \\ a_{p-1}^{(p)} & & & & 1 & \\ a_p^{(p)} & a_{p-1}^{(p-1)} & \dots & \dots & a_1^{(1)} & 1 \end{bmatrix} \quad \text{et} \quad \Sigma_p = \begin{bmatrix} \sigma_p^2 & & & & & \\ & \sigma_{p-1}^2 & & & & \\ & & \sigma_{p-2}^2 & & & \\ & & & \ddots & & \\ & & & & \sigma_1^2 & \\ & & & & & \sigma_0^2 \end{bmatrix}$$

La décomposition de Choleski de R_p^{-1} est alors $R_p^{-1} = TT^*$, avec $T = \mathbf{A}_p \Sigma_p^{-1/2}$. Nous remarquons que la dernière ligne de $\mathbf{A}_p$ est constituée de la suite des coefficients de réflexion.

Remarque B.2.2. *Processus autorégressifs*

Lorsque nous supposons que $\mathbf{X}$ est un processus autorégressif d'ordre p , le vecteur A_p est alors le paramètre du modèle autorégressif proprement dit. Nous avons alors l'expression de l'inverse de la matrice de covariance pour tout n en fonction de ces paramètres, en exploitant le fait que $\forall k \geq p+1, \mu_k = 0$. Nous pouvons écrire la décomposition de Choleski de R_n^{-1} pour tout $n \geq p$:

$$\forall n \geq p, R_n^{-1} = \mathbf{A}_n \Sigma_n^{-1} \mathbf{A}_n^* \quad (\text{B.7})$$

avec $\mathbf{A}_n = \begin{bmatrix} 1 & & & & \\ A_p & 1 & & & \\ 0 & \ddots & 1 & & \\ \vdots & 0 & A_p & \ddots & \\ & \vdots & & \ddots & 1 \\ 0 & 0 & & A_1 & 1 \end{bmatrix} \quad (0)$ et $\Sigma_n = \begin{bmatrix} \sigma_p^2 & & & & \\ & \ddots & & & \\ & & \sigma_p^2 & & \\ & & & \ddots & \\ (0) & & & & \sigma_0^2 \end{bmatrix}.$

B.3 Estimation des Coefficients de Réflexion

Burg a proposé un algorithme direct d'estimation des coefficients de réflexion qui permet d'avoir des CR de module inférieur à 1. Cet algorithme a de plus l'avantage d'être récursif, et de pouvoir être modifié pour intégrer une contrainte de douceur spectrale (et diminuer ainsi le nombre de pics d'interférence dans le spectre).

B.3.1 Algorithme de Burg

Les CR sont déterminés récursivement par minimisation de la somme des puissances des erreurs de prédiction directes $\sigma_p^{2,d}$ et rétrogrades $\sigma_p^{2,r}$, égales (respectivement) à $\|e_p^r\|^2 = \|X_1 - P_{2,p}(X_1)\|^2$ et $\|e_p^d\|^2 = \|X_{p+1} - P_{2,p}(X_{p+1})\|^2$. Si nous remplaçons les erreurs rétrogrades et directes par leur expression empirique (à partir d'un échantillon $x_1, \dots, x_N$), l'algorithme de Burg est initialisée par

$$e_{0,N}^d = e_{0,N}^r = x_N$$

et les erreurs sont calculées récursivement par

$$\text{pour } p \geq 1, \begin{cases} e_{p,n}^d = e_{p-1,n}^d + \mu_n e_{p-1,n-1}^r \\ e_{p,n}^r = e_{p-1,n}^r + \mu_n^* e_{p-1,n}^d \end{cases}$$

La minimisation de la version empirique des puissances d'erreur $\sigma_p^{2,d}$ et $\sigma_p^{2,r}$ donne l'estimateur suivant de μ_p :

$$\forall p \geq 2, \hat{\mu}_p = \frac{2 \sum_{n=p+1}^N e_{p-1,n}^{d*} e_{p-1,n-1}^r}{\sum_{n=p+1}^N |e_{p-1,n}^d|^2 + |e_{p-1,n-1}^r|^2} \quad (\text{B.8})$$

où $e_{p-1,n}^d$ et $e_{p-1,n-1}^r$ sont respectivement les erreurs de prédiction directe et rétrograde à l'ordre $p-1$ de x_n et x_{n-1} . Cela correspond aussi à une version empirique du coefficient de corrélation partiel entre erreur de prédiction directe et erreur de prédiction rétrograde.

B.3.2 Algorithme de Burg régularisé

Pour chaque entier p , l'algorithme de Burg cherche le coefficient qui minimise l'erreur du "modèle autorégressif" d'ordre p approchant le vrai processus $\mathbf{X}$. Nous savons que lorsqu'un processus est effectivement autorégressif, sa densité spectrale est égal à $\frac{1}{|A_p(f)|^2}$, où $A_p(f) = 1 - \sum_{k=1}^p a_k^{(p)} e^{-2i\pi k f}$. La régularité de la densité spectrale de $\mathbf{X}$ peut être mesurée alors par l'intégrale $J_p^2 = \int_{-\pi/2}^{\pi/2} \left| \frac{\partial A(f)}{\partial f} \right|^2 df = \sum_{k=1}^p (2\pi k)^2 \left| a_k^{(p)} \right|^2$.

Pour déterminer des CR correspondants à un spectre autorégressif qui soit lisse, on remplace la quantité à minimiser dans l'algorithme de Burg

$$\min_{\mu_p} \{ \sigma_p^{2,d} + \sigma_p^{2,r} \} \quad (\text{B.9})$$

par la fonctionnelle régularisée [10]

$$\min_{\mu_p} \{ \sigma_p^{2,d} + \sigma_p^{2,r} + \lambda J_p^2 \} \quad (\text{B.10})$$

où λ est un hyperparamètre. Cette minimisation a une solution analytique permettant de plus une estimation récursive :

$$\forall p \geq 2, \hat{\mu}_p = \frac{2 \sum_{n=p+1}^N e_{p-1,n}^{d*} e_{p-1,n-1}^r + \lambda (2\pi)^2 \sum_{k=1}^{p-1} k^2 a_k^{(p-1)} a_{p-k}^{(p-1)}}{\sum_{n=p+1}^N |e_{p-1,n}^d|^2 + |e_{p-1,n-1}^r|^2 + \lambda (2\pi)^2 \sum_{k=1}^{p-1} k^2 a_k^{(p-1)} a_{p-k}^{(p-1)}} \quad (\text{B.11})$$

Annexe C

Chaînes de Markov

Nous rappelons ici les concepts importants de la théorie des chaînes de Markov, et des résultats connus de convergence. Ces rappels sont basés essentiellement sur les ouvrages de Brémaud [23] et l'article introductif à la théorie des chaînes de Markov de Tierney [105].

C.1 Quelques notations

Soit $\mathbf{X} = (X_n)_{n \geq 0}$ une chaîne de Markov à valeurs dans $(E, \mathcal{E})$ espace mesuré. Elle est homogène si les probabilités de transition $P(X_{n+1} | X_n)$ sont toutes les mêmes pour tout n . Le noyau de transition $P(x, A)$ est la fonction telle que $\forall x \in E, \forall A \in \mathcal{E}, P(X_{n+1} \in A | X_n = x) = P(x, A)$. Trois opérations peuvent être définies à partir du noyau P

- Opération sur les mesures : si ν est une mesure sur $(E, \mathcal{E})$, alors nous notons νP la mesure définie par $\nu P(A) = \int P(x, A) \nu(dx)$;
- Opération sur les fonctions : si h est une fonction $h : E \rightarrow \mathbb{R}$, nous définissons la fonction $Ph(x) = \int P(x, dy) h(y) = E[h(X_1) | X_0 = x]$;
- Opération sur les noyaux de transition : si Q est un autre noyau de transition, nous définissons le noyau de transition $PQ(x, A) = \int P(x, dy) Q(y, A)$. Si nous itérons l'opération, nous obtenons $P(X_n \in A | X_0 = x) = P \dots P(x, A) = P^n(x, A)$.

La distance en variation totale entre 2 mesures ν_1 et ν_2 , utile pour l'étude de la convergence des chaînes de Markov est définie par :

$$\|\nu_1 - \nu_2\|_{TV} = 2 \sup_{A \subset E} |\nu_1(A) - \nu_2(A)|$$

Enfin, si $x \in E$, P_x désigne la probabilité du processus $\mathbf{X}$ lorsque $P_{X_0} = \delta_x$.

C.2 Définitions et propriétés

Nous donnons maintenant les premières définitions importantes pour l'étude de la stabilité des chaînes de Markov stationnaires.

Définition C.2.1. Premier temps de retour

Soit $A \in \mathcal{E}$, le premier temps de retour en A est la variable aléatoire $\tau_A = \inf \{n \geq 1 : X_n \in A\}$. Nous notons $\tau_A = \infty$ lorsque la chaîne ne retourne jamais en A .

Définition C.2.2. Irréductibilité

Soit φ une mesure de probabilité sur $(E, \mathcal{E})$, une chaîne de Markov $\mathbf{X}$ est φ -irréductible (ou irréductible) si

$$\varphi(A) > 0 \implies \forall x \in E, P_x(\tau_A < \infty) > 0$$

La probabilité φ est appelée distribution d'irréductibilité de la chaîne.

Une définition équivalente de la φ -irréductibilité, utilisant les noyaux de transition et ne faisant plus intervenir le temps de retour est la suivante

$$\exists n \geq 1, \forall x \in E, \forall A \text{ tq } \varphi(A) > 0, P^n(x, A) > 0$$

Une condition suffisante d'irréductibilité d'un noyau P est l'existence d'une densité f relativement à φ qui soit strictement positive et telle que pour $n \geq 1$:

$$\exists n \geq 1, \forall x \in E, \forall A \in \mathcal{E}, P^n(x, A) = \int_A f(x, y) \varphi(dy)$$

Il suffit alors de prouver que le noyau P^n admet une densité relativement à φ .

Remarque C.2.1. Irréductibilité des chaînes à espace d'état discret

Pour les chaînes de Markov à état discrets, une chaîne est dite irréductible si nous avons :

$$\forall x, x', \exists n P^n(x, x') > 0$$

Nous avons alors équivalence entre l'existence d'une probabilité stationnaire qui charge tous les points de l'espace et l'irréductibilité de la matrice de transition P . Lorsque n est uniforme en les états, i.e. si $\exists n, \forall x, x', P^n(x, x') > 0$, la matrice de transition est dite primitive. Nous avons alors A primitive si et seulement si la chaîne de Markov est ergodique, avec une unique probabilité stationnaire positive chargeant tous les états.

Parmi toutes les distributions d'irréductibilité que possède une chaîne irréductible, il est possible de montrer qu'il existe une distribution irréductible maximale, par rapport à laquelle toutes les autres distributions d'irréductibilité admettent une densité. Les distributions maximales sont alors toutes équivalentes, i.e. possèdent les mêmes ensembles négligeables.

L'irréductibilité signifie que tout ensemble de φ -mesure non nulle peut être atteint depuis n'importe quel point $x \in E$. Il est intéressant de savoir si ces ensembles peuvent être atteints un nombre fini ou infini de fois. Cette propriété, qui entraîne la définition du concept de récurrence, s'intéresse aux probabilités des ensembles $\{X_n \in A \text{ i.s.}\} = \bigcap_{n \geq 0} \bigcup_{k \geq n} \{X_k \in A\}$, où i.s. signifie "infiniment souvent".

Définition C.2.3. Récurrence

Une chaîne irréductible (avec distribution d'irréductibilité maximale ψ) est récurrente si pour tout ensemble A tel que $\psi(A) > 0$, les deux conditions suivantes sont vérifiées

- (i) $\forall x \in E, P_x(X_n \in A \text{ i.s.}) > 0$.

(ii) $P_x(X_n \in A \text{ i.s.}) = 1, \psi - ps.$

De plus, une chaîne irréductible récurrente est dite positive récurrente si elle possède une probabilité invariante. Dans le cas contraire, elle est dite nulle-récurrente.

Remarque C.2.2. Récurrence des chaînes discrètes

Pour les chaînes discrètes, la récurrence est définie pour les états. La définition de la récurrence pour un espace quelconque est alors une conséquence de l'irréductibilité de la chaîne. Pour une chaîne discrète récurrente, il est équivalent d'être positive récurrente et de posséder une probabilité invariante.

Le résultat suivant permet de faire le lien entre irréductibilité et récurrence :

Proposition C.2.1. Soit $\mathbf{X}$ une chaîne de Markov est irréductible, possédant une distribution stationnaire π . Alors $\mathbf{X}$ est π - irréductible, π est une distribution maximale et est l'unique distribution invariante de la chaîne. De plus, $\mathbf{X}$ est positive récurrente.

La propriété de récurrence positive est suffisante pour avoir l'existence d'une loi forte des grands nombres :

Théorème C.2.1. Soit $\mathbf{X}$ est une chaîne de Markov irréductible de distribution invariante π , et $f : (E, \mathcal{E}) \rightarrow \mathbb{R}$ telle que $E_\pi(|f|) = \int |f(x)| \pi(dx) < \infty$. Alors

$$P_x \left(\left\{ \frac{1}{n} \sum_{k=1}^n f(X_k) \rightarrow E_\pi(f) \right\} \right) = 1 \quad \pi - ps \quad (\text{C.1})$$

Pour avoir des résultats de convergence plus fort, il est nécessaire de contrôler les phénomènes de périodicité. Nous disons qu'une chaîne est cyclique si il existe $(E_1, \dots, E_p)$ partition finie de E , telle que

$$\forall i \leq p-1, \forall x \in E_{i-1}, P(x, E_i) = 1$$

La période est le plus grand cycle que la chaîne peut décrire. Si la période est égale à 1, la chaîne est dite apériodique. En pratique, il est plus facile de montrer la propriété d'apériodicité forte qui implique alors l'apériodicité.

Définition C.2.4. Apériodicité forte

La chaîne irréductible $\mathbf{X}$, de distribution stationnaire π est dite fortement apériodique si

$$\exists \nu \text{ probabilité sur } E, \beta > 0, C \subset E \text{ t.q. } \nu(C) > 0 \text{ et } \forall x \in C, \forall A, P(x, A) > \beta \nu(A)$$

Nous avons alors pour une chaîne de Markov $\mathbf{X}$ irréductible, avec noyau de transition P et une distribution invariante π un théorème ergodique pour la chaîne $\mathbf{X}$.

Théorème C.2.2. Convergence en Variation d'une chaîne de Markov

Si $\mathbf{X}$ est une chaîne de Markov irréductible, apériodique, de noyau de transition P et de distribution stationnaire π , alors

$$\|P^n(x, \cdot) - \pi(\cdot)\|_{TV} \rightarrow 0 \quad (\text{C.2})$$

pour $\pi - ps$ tout x .

Une faiblesse de ce résultat est que la convergence n'est pas assurée sur un ensemble de π -mesure nulle. Ceci est dû au fait que la condition (ii) d'irréductibilité est vraie seulement presque-sûrement. Pour pallier ce défaut, la propriété d'Harris-réurrence a été introduite :

Définition C.2.5. Harris-réurrence

Une chaîne irréductible avec une distribution d'irréductibilité maximale ψ est dite Harris-récurrente si

$$\forall A \subset E, \psi(A) > 0, \forall x \in E, P_x(X_n \in A \text{ i.o.}) = 1$$

En remplaçant la condition de récurrence par celle d'Harris-réurrence, les convergences (C.1) et (C.2) sont vraies pour tout $x \in E$, et plus seulement presque sûrement.

C.3 Propriétés de mélanges des chaînes de Markov

Lorsqu'une chaîne de Markov est irréductible, apériodique et Harris-récurrente, alors nous avons des lois fortes des grands nombres, pour cette raison elles sont dites ergodiques. Cependant, cela ne suffit pas pour avoir un théorème central limite, et il est nécessaire de faire des hypothèses plus fortes, sur la vitesse de convergence de $\|P^n(x, \cdot) - \pi(\cdot)\|_{TV}$. Cette dernière est reliée aux propriétés de mélange des processus stationnaires (section 5.2.2). Une condition de convergence souvent utilisée est la propriété d'ergodicité géométrique :

Définition C.3.1. Ergodicité géométrique

Une chaîne de Markov ergodique, de distribution invariante π est géométriquement ergodique si il existe une fonction à valeurs réelles M , π -intégrable et une constante $0 < r < 1$ tel que

$$\forall x, \forall n, \|P^n(x, \cdot) - \pi(\cdot)\|_{TV} \leq M(x)r^n \quad (\text{C.3})$$

Une chaîne de Markov géométriquement ergodique ayant pour distribution initiale sa distribution stationnaire π est α -mélangeante à une vitesse géométrique :

$$\alpha_n = \sup_{A, B \subset E} |P_\pi(X_0 \in A, X_n \in B) - \pi(A)\pi(B)| = O(r^n), r < 1$$

Le majorant de (C.3) dépend de la valeur initiale, et il est alors souvent intéressant de supposer que cette vitesse de convergence est vérifiée uniformément sur l'espace des états, ce qui aboutit à la notion d'ergodicité géométrique uniforme, appelée directement ergodicité uniforme.

Définition C.3.2. Ergodicité uniforme

Une chaîne ergodique de distribution invariante π est uniformément ergodique si

$$\exists M > 0 \text{ et } r < 1 \text{ tel que } \forall x, \forall n, \|P^n(x, \cdot) - \pi(\cdot)\|_{TV} \leq Mr^n$$

L'uniforme ergodicité est alors équivalente à la condition de Doeblin¹ et à la ϕ -mélangeance exponentielle, i.e.

¹condition de Doeblin globale : $\exists \mu$ mesure de probabilité, $\exists \delta > 0, \exists \epsilon < 1, \exists m, \forall B \subset E, \mu(B) > \epsilon \implies \inf_{x \in X} P(x, B) > \delta$

$$\phi_n = \sup_{A, B \subset E} |P_\pi(X_n \in B | X_0 \in A) - \pi(B)| = O(r^n), r < 1$$

Elle est le type de convergence le plus fort le plus communément utilisé pour les chaînes de Markov. Dans ce cas, la chaîne “mélange suffisamment rapidement” pour garantir un théorème central limite :

Théorème C.3.1. *Théorème Central Limite*

Soit f une fonction à valeurs réelles et soit $\mathbf{X}$ une chaîne ergodique de distribution invariante π vérifiant l’une des deux conditions suivantes :

- (i) *la chaîne est géométriquement ergodique et il existe $\epsilon > 0$ tel que $f^{2+\epsilon}$ soit π -intégrable*
- (ii) *la chaîne est uniformément ergodique et f est carré π -intégrable*

Alors la variance de $(f(X_n))_{n \geq 0}$ est finie et vaut

$$\sigma_f^2 = E_\pi \left[(f(X_0) - \pi f)^2 \right] + 2 \sum_{k=1}^{\infty} E_\pi [(f(X_0) - \pi f) (f(X_k) - \pi f)]$$

De plus, $\sqrt{n}(\frac{1}{n} \sum_{k=0}^n f(X_k) - \pi f)$ converge en loi vers une variable aléatoire normale $\mathcal{N}(0, \sigma_f^2)$.

Bibliographie

- [1] R.J. Adler, R.E. Feldman, and M.S. Taquq, editors. *A practical guide to heavy tails : statistical techniques and applications*. Birkhauser, Boston, 1998.
- [2] P. Ailliot. *Modèles auto-régressifs à changements de régimes markoviens. Applications aux séries temporelles de vent*. PhD thesis, Université Rennes 1, 2004.
- [3] R.J McKay Altman. Assessing the goodness-of-fit of hidden Markov models. *Biometrics*, 60 :444–450, 2004.
- [4] B.D.O. Anderson and T.B. Moore. *Optimal Filtering*. Information and System sciences. Prentice-Hall, 1979.
- [5] Andrews, Askey, and Roy. *Special Functions, Encyclopedia of Mathematics and its applications*. Cambridge University Press, 2000.
- [6] R. Azencott, editor. *Simulated annealing : parallelization techniques*. Wiley, 1992.
- [7] A. Banerjee, I. Dhillon, J. Ghosh, and S. Sra. Expectation maximization for clustering on hyperspheres. Technical Report 03-07, Departement of Computer Science, University of Texas, 2003. [http ://www.cs.utexas.edu/users/suvrit/pubs/](http://www.cs.utexas.edu/users/suvrit/pubs/).
- [8] A. Banerjee, I.S. Dhillon, J. Ghosh, and S. Sra. *Journal of Machine Learning Research*, 6 :1345–1382, september 2005.
- [9] R. Barakat. Direct derivation of intensity and phase statistics of speckle produced by a weak scatterer from the random sinusoid model. *Journal of the Optical Society of America*, 71(1) :86–90, 1981.
- [10] F. Barbaresco. Algorithme de burg régularisé FSDS, comparaison avec l’algorithme de burg MFE. In *Proceedings of XVème colloque GRETSI*, septembre 1995.
- [11] F. Barbaresco. 3D echographic data segmentation and carotid artery turbulence mapping by Doppler velocimetry by a common approach based on calculus of variations. In *IEEE Conference ICIP’01*, Thessalonique, october 2001.
- [12] O.E. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. Wiley, 1978.
- [13] L.E. Baum and T.P. Petrie. Statistical inference for probabilistic functions of finite states Markov chains. *Annals of Mathematical Statistics*, 37 :1554–1563, 1966.
- [14] D. Benboudjema and W. Pieczynski. Parameter estimation in pairwise Markov random field. In *Proceeding of Advanced Concepts for Intelligent Vision Systems (ACIVS 04)*, Brussels, 2004.

- [15] D. Benboudjema and W. Pieczynski. Unsupervised image segmentation using triplet Markov fields. *Computer Vision and Image Understanding*, 99(3) :476–498, 2005.
- [16] D. Benboudjema and W. Pieczynski. Unsupervised image segmentation using triplet Markov fields. *Computer Vision and Image Understanding*, 2005. à paraître.
- [17] A. Bendjebbour, Y. Delignon, L. Fouque, V. Samson, and W. Pieczynski. Multisensor images segmentation using Dempster-Shafer fusion in Markov fields context. *IEEE Transactions on Geoscience and Remote Sensing*, 39(8) :1789–1798, 2001.
- [18] B. Benmiloud and W. Pieczynski. Estimation des paramètres dans les chaînes de Markov cachées et segmentation d’images. *Traitement du Signal*, 12(5) :389–399, 1995.
- [19] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society B*, 36(2) :192–236, 1974.
- [20] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society B*, 48 :259–302, 1986. with discussion.
- [21] P.J. Bickel, Y. Ritov, and T. Ryden. Asymptotic normality of the maximum likelihood estimator for general hidden Markov models. *Annals of statistics*, 26(4) :1614–1635, 1998.
- [22] C. Biernacki. *Choix de Modèles en classification*. PhD thesis, Université de Technologie de Compiègne, 1997.
- [23] P. Brémaud. *Markov chains : Gibbs fields, Monte Carlo simulation, and queues*. Number 31 in Texts in applied mathematics. Springer-Verlag New-York, 1999.
- [24] P.J. Brockwell and R.A. Davis. *Time series : Theory and Methods*. Springer series in Statistics. Springer-Verlag, 2nd edition, 1991.
- [25] N. Brunel and F. Barbaresco. Doppler and polarimetric statistical segmentation of radar clutter environment based on pairwise Markov chains. In *Proceedings of International Conference on Radar Systems (Radar 2004)*, Toulouse, 18-22 Octobre 2004.
- [26] N. Brunel and W. Pieczynski. Copulas in vectorial hidden markov chains for multicomponent image segmentation. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing, (ICASSP’05)*, Philadelphia, Pennsylvania, march 18-23 2005.
- [27] N. Brunel and W. Pieczynski. Modeling temporal dependence of spherically invariant random vectors with triplet Markov chains. In *Proceedings of IEEE Workshop on Statistical Signal Processing, SSP’05*, Bordeaux, 17-20 juillet 2005.
- [28] N. Brunel and W. Pieczynski. Unsupervised signal restoration using hidden Markov chains with copulas. *Signal Processing*, 2005.
- [29] N. Brunel, W. Pieczynski, and F. Barbaresco. Chaînes de Markov multivariées à bruit corrélé non gaussien. In *20ème Colloque Traitement du Signal, GRETSI’05*, Louvain-la-Neuve, 6-9 septembre 2005.
- [30] K.P. Burnham and D.R. Anderson. *Model Selection and Inference*. Springer-Verlag, 1998.
- [31] C. Genest, J.J.Quesada Molina, J. A. Rodriguez Lallena, and C. Sempi. De l’impossibilité de construire des lois à marges multidimensionnelles données à partir de copules. *Comptes rendus de l’Académie des Sciences*, 320 :723–726, 1995.

- [32] O. Cappé, E. Moulines, and T. Ryden. *Inference in Hidden Markov Models*. Springer-Verlag, 2005.
- [33] C. Carincotte, S. Derrode, G. Sicot, and J-M. Boucher. Unsupervised image segmentation based on a new fuzzy HMC model,. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04)*, Philadelphia, Pennsylvania, May 2004.
- [34] G. Celeux, D. Chauveau, and J. Diebolt. Stochastic versions of the EM algorithm : an experimental study. *Journal of Statistical Computation and Simulation*, 55 :287–314, 1996.
- [35] 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.
- [36] G. Celeux, J.S. Marques, and J. Nascimento. Learning switching dynamic models for objects tracking. Technical Report RR-4863, INRIA, Juin 2003.
- [37] B. Chalmond. An iterative Gibbsian technique for reconstruction of m-ary images. *Pattern Recognition*, 22(6) :747–761, 1989.
- [38] B. Chalmond. *Éléments de modélisation pour l'analyse d'images*, volume 33 of *Mathématiques et applications*. Springer-Verlag, 2000.
- [39] F. Chatelain and J-Y Tourneret. Composite likelihood estimation for multivariate poisson distribution. In *Proceedings of Statistical Signal Processing (SSP'05)*, Bordeaux, 17-20 juillet 2005.
- [40] B. Chen, P.K. Varshney, and J. Michels. Adaptive CFAR detection for clutter-edge heterogeneity using Bayesian inference. *IEEE Transactions on Aerospace and Electronic Systems*, 39(4), 2003.
- [41] X. Chen and Y. Fan. Estimation of copula-based semi-parametric time series models. Technical report, Department of Economics, Vanderbilt University, 2004.
- [42] J-P. Cocquerez and S. Phillip. *Analyse d'images : filtrage et segmentation*. Masson, 1995.
- [43] K. Conradsen, A. Nielsen, J. Schou, and H. Skriver. A test statistic in the complex Wishart distribution and its application to change detection in polarimetric SAR data. *IEEE Transactions on Geoscience and Remote Sensing*, 41(1) :632 – 647, 2003.
- [44] E. Conte, A. De Maio, and C. Galdi. Statistical validation of the compound-Gaussian model on clutter data from IPIX radar. In *Proceedings of International Conference on Radar Systems (Radar 2004)*, Toulouse, 18-22 Octobre 2004.
- [45] D. Dacunha-Castelle and M. Duflo. *Probabilités et statistiques, tome 2 - problèmes à temps mobile*. Masson, 1983.
- [46] W.F. Darsow, B. Nguyen, and E.T. Olsen. Copulas and Markov processes. *Illinois Journal of Mathematics*, 36(4) :600–642, 1992.
- [47] P. Deheuvels. La fonction de dépendance empirique et ses propriétés. *Académie Royale de Belgique - Bulletin de la Classe des Sciences*, 65 :274–292, 1979.
- [48] Y. Delignon, A. Marzouki, and W. Pieczynski. Estimation of generalized mixture and its application in image segmentation. *IEEE Transactions on Image Processing*, 6(10) :1364–1475, 1997.

- [49] Y. Delignon and W. Pieczynski. Modeling non-Rayleigh speckle distribution in SAR images. *IEEE Transactions on Geoscience and Remote Sensing*, 40(6) :1430–1435, 2002.
- [50] J-P. Delmas. An equivalence of the EM and ICE algorithm for exponential family. *IEEE transactions on Signal Processing*, 45(10), 1997.
- [51] B. Delyon, M. Lavielle, and E. Moulines. On a stochastic approximation version of the EM algorithm. *Annals of Statistics*, 27(1) :94–128, 1999.
- [52] 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 (B)*, 39 :1–38, 1977.
- [53] S. Derrode and W. Pieczynski. Signal and image segmentation using pairwise markov chains. *IEEE Transactions on Signal Processing*, 52(9) :2477–2489, 2004.
- [54] F. Desbouvries and W. Pieczynski. Particle filtering with pairwise Markov processes. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'03)*. IEEE, 2003.
- [55] F. Destremes and M. Mignotte. A statistical model for contours in images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5) :626–638, 2004.
- [56] L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*. Applications of Mathematics. Springer, Berlin, 1996.
- [57] J. Diebolt and E. Ip. *Markov Chain Monte Carlo in Practice*, chapter Stochastic EM : Method and application. Interdisciplinary Statistics. Chapman and Hall / CRC, 1996.
- [58] R. Douc and C. Matias. Asymptotics of the maximum likelihood estimator for general hidden Markov models. *Bernoulli*, 7(3) :381–420, 2001.
- [59] R. Douc, E. Moulines, and T. Ryden. Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *Annals of Statistics*, 14(4) :1643–1665, 2004.
- [60] A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte-Carlo in Practice*. Statistics for Engineering and Information Science. Springer-Verlag, 2001.
- [61] A. Doucet, A. Logothetis, and V. Krishnamurthy. Stochastic sampling algorithms for state estimation of jump Markov linear systems. *IEEE Transactions on Automatic Control*, 45(2) :188–202, 2000.
- [62] P. Doukhan. *Theory and applications of long-range dependence*, chapter Models, Inequalities, and Limit Theorems for stationary sequences. Birkhauser, 2002.
- [63] J-B. Durand. *Modèles à structure cachée : inférence, sélection de modèle et applications*. PhD thesis, Université Joseph Fourier, 2003.
- [64] R. Durbin, S.R. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis : Probabilistic models of proteins and nucleic acids*. Cambridge Univserity Press, 1998.
- [65] E. Lebarbier and T. Mary-Huard. Le critère bic : fondements théoriques et interprétation. Technical Report RR-5315, INRIA, Septembre 2004.
- [66] 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 Transactions on Medical Imaging*, 24(2) :263–276, 2005.

- [67] H.-B. Fang, K.-T. Fang, and S. Kotz. The Meta-elliptical distributions with given marginals. *Journal of Multivariate Analysis*, 82(1) :1–16, 2002.
- [68] J.-D. Fermanian and O. Scaillet. Some statistical pitfalls in copula modeling for financial applications. Technical Report 108, FAME - International Center for Financial Asset Management and Engineering, University of Geneva, March 2004.
- [69] C. Fraley and A. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458), 2002.
- [70] B. J. Frey and N. Jojic. A comparison of algorithms for inference and learning in probabilistic graphical models. *IEEE Trans. on Pattern and Machine Intelligence*, 27(9), 2005.
- [71] E. Gassiat. Likelihood ratio inequalities with application to various mixtures. *Annales de l'Institut Poincaré*, 38(6, pages = 897-906, OPTmonth = , OPTnote = , OPTannote =), 2002.
- [72] E. Gassiat and C. Keribin. The likelihood ratio test for the number of components in a mixture with Markov regime. *ESAIM : Probability and statistics*, 4, 2000.
- [73] D. Geman and S. Geman. Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6 :721–741, 1984.
- [74] C. Genest and B.J.M. Werker. *Conditions for the asymptotic semi-parametric efficiency of an omnibus estimator of dependence parameters in copula models*, chapter Proceedings of the Conference on distributions with given marginals and statistical modelling. Kluwer Academic Publishers, 2002.
- [75] Z. Ghahramani and M.I. Jordan. Factorial hidden Markov models. *Machine Learning*, 29 :245–273, 1997.
- [76] W.R. Gilks, S. Richardson, and D.J Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice*, chapter General State Space Markov Chains. Interdisciplinary Statistics. Chapman and Hall / CRC, 1996.
- [77] F. Gini and M. Greco. Covariance matrix estimation for CFAR detection in correlated heavy tailed clutter. *Signal Processing*, 82(12) :1847–1859, 2002.
- [78] N. Giordana and W. Pieczynski. Estimation of generalized multisensor HMC and unsupervised image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(5) :465–475, 1997.
- [79] V.P. Godambe. *Estimating Functions*. Oxford University Press, 1991.
- [80] P.J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4) :711–732, 1995.
- [81] U. Grenander and A. Srivastava. Probability models for clutter in natural images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(4), 2001.
- [82] A.K. Gupta and T. Varga. *Elliptically contoured models in statistics*. Mathematics and its applications. Kluwer Academic, 1993.
- [83] X. Guyon. *Champs aléatoires sur un réseau : modélisations, statistique et applications*. Collection Techniques Stochastiques. Masson, 1992.

- [84] Y. Guédon. Estimating hidden semi-Markov chains from discrete sequences. *Journal of Computational and Graphical Statistics*, 12(3) :604–639, 2003.
- [85] J. D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.
- [86] S. Haykin. *Adaptive filter theory*. Prentice Hall International Editions, 3rd edition, 2000.
- [87] S. Haykin, R. Bakker, and B. Currie. Uncovering nonlinear dynamics : the case study of sea clutter. *Proceedings of the IEEE*, 90(5), 2002.
- [88] S. Haykin and A. Reinhardt, editors. *Adaptive Radar Detection and Estimation*. Wiley-Interscience, 1992.
- [89] C.C. Heyde. *Quasi-likelihood and its application : a general approach to optimal parameter estimation*. Springer series in Statistics. Springer-Verlag, New-York, 1997.
- [90] C.C. Heyde and R. Morton. Quasi-likelihood and generalizing the EM algorithm. *Journal of the Royal Statistical Society (B)*, 58(2) :317–327, 1996.
- [91] P.J. Huber. *Robust Statistics*. Wiley, New York, 1981.
- [92] S. Ikeda, T. Tanaka, and S-I. Amari. Information geometry of turbo and low-density parity-check codes. *IEEE Transactions on Information Theory*, 50(6) :1097–1114, 2004.
- [93] E. H. Ip. A stochastic EM estimator in the presence of missing data - theory and applications. Technical Report 304, Department of Statistics, Stanford University, 1994.
- [94] H. Joe. *Multivariate Models and Dependence Concepts*, volume 73 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, 1997.
- [95] H. Joe. Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94(2), 401-419 2004.
- [96] R.E. Kass and A.E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430) :773–795, 1995.
- [97] Z. Kato, J. Zerubia, and M. Berthod. Unsupervised parallel image classification using a hierarchical Markovian model. In *Proceedings of Fifth International Conference on Computer Vision*, pages 169–174, 20-23 Juin 1995.
- [98] S.M. Kay. *Fundamentals of statistical signal processing - Estimation theory*. International editions. Prentice Hall, 1993.
- [99] P. Lanchantin and W. Pieczynski. Unsupervised non-stationary image segmentation using triplet Markov chains. In *Proceedings of Advanced Concepts for Intelligent Vision Systems (ACIVS'04)*, Bruxelles, 31 aout - 3 septembre 2004.
- [100] S.L. Lauritzen. *Graphical Models*. Oxford university press, 1996.
- [101] J-S. Lee, M.R. Grunes, T.L. Ainsworth, L-J Du, and D.L. Schuler. Unsupervised classification using polarimetric decomposition and the complex Wishart classifier. *IEEE Transactions on Geoscience and Remote Sensing*, 37(5) :2249–2258, 1999.
- [102] B. G. Leroux. Maximum-likelihood estimation for hidden Markov models. *Stochastic processes and their applications*, 40 :127–143, 1992.
- [103] G. Lindgren. Markov regime models for mixed distributions and regressions. *Scandinavian Journal of Statistics*, 5 :81–91, 1978.

- [104] P. Lombardo and A. Farina. Coherent radar detection against K-distributed clutter with partially correlated texture. *Signal Processing*, 48(1) :1–15, 1996.
- [105] L.Tierney. *Markov Chain Monte Carlo in Practice*, chapter General State Space Markov Chains. Interdisciplinary Statistics. Chapman and Hall / CRC, 1996.
- [106] H. Maitre, editor. *Traitement des images RSO*. Traitement du signal et de l'image. Hermès, 2001.
- [107] K.V. Mardia and P. Jupp. *Directional Statistics*. Wiley series in Probability and Statistics. Wiley, 1999.
- [108] G. J. McLachlan and T. Khrishnan. *The EM algorithm and extensions*. Wiley series in probability and statistics. Wiley Interscience, 1996.
- [109] G.J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley series in Probability and Statistics. Wiley, 2000.
- [110] X. Meng and D. Rubin. Maximum likelihood via the ECM algorithm : a general framework. *Biometrika*, 80 :267–278, 1993.
- [111] X-L. Meng and D. Van Dyk. The EM algorithm - an old folk song sung to a fast new tune. *Journal of the Royal Statistical Society (B)*, 59 :511–567, 1997.
- [112] L. Mevel. *Statistique Asymptotique pour les modèles de Markov Cachés*. PhD thesis, Université de Rennes 1, 1997.
- [113] M. Mignotte, C. Collet, and P. Perez annd P. Bouthemy. Unsupervised Markovian segmentation of sonar images. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP'97)*.
- [114] R. B. Nelsen. *An introduction to Copulas*. Number 139 in Lecture notes in Statistics. Springer-Verlag, 1998.
- [115] S.F. Nielsen. The stochastic em algorithm : estimation and asymptotic results. *Bernoulli*, 6(3) :457–489, 2000.
- [116] B. North, A. Blake, M. Isard, and J. Rittscher. Learning and classification of complex dynamics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(9), 2000.
- [117] C. Oliver and S. Quegan. *Understanding Synthetic Aperture Radar Images*. Artech House, 1998.
- [118] F. Pascal, J.P. Ovarlez, P. Forster, and P. Larzabal. Constant false alarm rate detection in spherically invariant random processes. In *Proceedings of the European Signal Processing Conference (EUSIPCO 2004)*, Vienne, Autriche, 6-10 Septembre 2004.
- [119] X. Pennec. Probabilities and statistics on Riemanian manifolds : a geometric approach. Technical Report 5093, INRIA, 2004.
- [120] P. Perez and J. Vermaak. Visual tracking and auxiliary discrete processes. In *Proceedings of the International Symposium on Applied Stochastic Models and Data Analysis (ASMDA)*, Brest, 17-20 Mai 2005.
- [121] B. Picinbono. On circularity. *IEEE Transactions on Signal Processing*, 42(12), 1994.
- [122] W. Pieczynski. Champs de Markov cachés et estimation conditionnelle itérative. *Traitement du Signal*, 11(2) :141–153, 1994.

- [123] W. Pieczynski. Triplet Markov chains. *Comptes Rendus de l'Académie des sciences - Mathématique, série I*, 335(3) :275–278, 2002.
- [124] W. Pieczynski. Pairwise Markov chains. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5) :634–639, 2003.
- [125] W. Pieczynski. Triplet Markov chains and theory of evidence. *International Journal of Approximate Reasoning*, 2003. soumis.
- [126] W. Pieczynski. Triplet partially Markov chains and trees. In *Proceedings of the International Symposium on Image/Video Communications (ISIVC'04)*, Brest, 7-9 Juillet 2004.
- [127] W. Pieczynski. Modeling non-stationary hidden semi-Markov chains with triplet Markov chains and theory of evidence. In *Proceedings of Statistical Signal Processing (SSP'05)*, Bordeaux, 17-20 juillet 2005.
- [128] W. Pieczynski and F. Desbouvries. Kalman filtering using pairwise Gaussian models. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'03)*. IEEE, 2003.
- [129] W. Pieczynski and F. Desbouvries. On triplet Markov chains. In *Proceedings of the International Symposium on Applied Stochastic Models and Data Analysis (ASMDA)*, Brest, May 2005.
- [130] W. Pieczynski and P. Lanchantin. Restoring hidden non stationary process using triplet partially Markov chain with long memory noise. In *Proceedings of IEEE Workshop on Statistical Signal Processing (SSP'05)*, Bordeaux, 17-20 Juillet 2005.
- [131] W. Pieczynski and A-N. Tebbache. Pairwise Markov random fields and segmentation of textured images. *Machine Graphics and Vision*, 9(3) :705–718, 2000.
- [132] J-N. Provost, C. Collet, P. Perez, and P. Bouthemy. Hierarchical Markovian segmentation of multispectral images for the reconstruction of water depth maps. *Computer Vision and Image Understanding*, 93(2) :155–174, 2004.
- [133] B. Prum. *Processus sur un réseau et mesures de Gibbs*. Collection Techniques Stochastiques. Masson, 1986.
- [134] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2) :257–286, 1989.
- [135] M. Rangaswamy, D. Weiner, and A. Ozturk. Non-Gaussian random vector identification using spherically invariant random process. *IEEE Transactions on Aerospace and Electronic Systems*, 29(1) :111–124, 1993.
- [136] S. Richardson and P.J. Green. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society (B)*, 59(4) :731–792, 1997. with discussion.
- [137] C. P. Robert. *The Bayesian Choice : a decision theoretic approach*. Springer Texts in Statistics. Springer-Verlag, New-York, 1994.
- [138] C.P. Robert and G. Casella. *Monte-Carlo Statistical Methods*. Springer texts in Statistics. Springer-Verlag, New-York, 1999.

- [139] W.J.J. Roberts and S. Furui. Maximum likelihood estimation of K-distribution parameters via the Expectation-Maximization algorithm. *IEEE transactions on Signal Processing*, 48(12) :3303–3306, 2000.
- [140] D.B. Rubin. Iteratively reweighted least squares. In S. Kotz, N.L. Johnson, and C.B. Read, editors, *Encyclopedia of Statistical Sciences*, volume 4, pages 272–275. Wiley, New York, 1983.
- [141] T. Ryden. Consistent and asymptotically normal parameter estimates for hidden markov models. *Annals of Statistics*, 22(4) :1884–1895, 1994.
- [142] A.C. Schroth, M.S. Chandra, and P.F. Meischner. A C band coherent polarimetric radar for propagation and cloud physics research. *Journal of Atmospheric and Oceanic Technology*, 5(6) :803–822, 1988.
- [143] N. Sheppard, editor. *Stochastic volatility - selected readings*. Advanced texts in econometrics. Oxford University Press, 2005.
- [144] J.H. Shih and T.A. Louis. Inference on the association parameter in copula models for bivariate survival data. *Biometrics*, 51(4) :1384–1399, 1995.
- [145] A. Srivastava, X. Liu, and U. Grenander. Universal analytical forms for modeling image probabilities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9), 2002.
- [146] T. Roncalli. *Gestion des risques multiples*. Groupe de recherche Opérationnelle - Crédit Lyonnais, 2002.
- [147] M. A. Tanner. *Tools for Statistical Inference : Methods for the exploration of posterior distributions and likelihood functions*. Springer series in Statistics. Springer-Verlag, New-York, 3rd edition, 1996.
- [148] H. Teicher. Identifiability of mixtures of product measures. *Annals of Mathematical Statistics*, 38 :1330–1302, 1967.
- [149] V.N. Vapnik. *Statistical Learning Theory*. John Wiley and sons, 1998.
- [150] A. Wald. Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, 20 :595–601, 1949.
- [151] K.D. Ward, C.J. Baker, and S. Watts. Maritime surveillance radar I. radar scattering from the ocean surface. *IEE Proceedings of Radar and Signal Processing*, 137(2) :51–62, 1990.
- [152] S. Watts. Cell-averaging CFAR gain in spatially correlated K-distributed clutter. *IEE Proceedings of Radar, Sonar and Navigation*, 143(5) :321–327, 1996.
- [153] S. Watts. Tutorial on radar clutter and CFAR detection. Conference Radar 2004, 17-22 Octobre 2004.
- [154] G. Wei and M. Tanner. A Monte-Carlo implementation of the EM algorithm and the poor man’s data augmentation algorithms. *Journal of the American Statistical Association*, 95 :699–704, 1990.
- [155] Y. Weiss and T. Freeman. Correctness of belief propagation in Gaussian graphical models of arbitrary topology. In *Proceedings of Neural Information Processing Systems (NIPS)*, 1999.
- [156] G. Winkler. *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods*. Number 27 in Stochastic Modelling and applied probability. Springer, 2nd edition, 2000.

- [157] J.S. Yedidia, W.T. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. Technical report, Mitsubishi electric research laboratories, 2002. <http://www.merl.com/publications/TR2001-022/>.
- [158] L. Younes. Parametric inference for imperfectly observed Gibbsian fields. *Probability Theory and Related Fields*, 82 :625–645, 1989.
- [159] L. Younes. *Proceedings of Stochastic models, "statistical methods and algorithms in image analysis"*, chapter Parameter estimation for imperfectly observed Gibbs fields and some comments on Chalmond's EM Gibbsian algorithm. Lecture Notes in Statistics. Springer, 1992.