Analyse et modélisation de données probabilistes par décomposition de mélange de copules et application à une base de données climatologiques

Mathieu Vrac

Université Paris IX - Dauphine
UFR Mathématiques de la Décision

THÈSE

Pour obtenir le grade de

DOCTEUR EN SCIENCES

Spécialité
Mathématiques Appliquées

présentée et soutenue par

Mathieu VRAC

le 6 décembre 2002

Titre

ANALYSE ET MODÉLISATION DE DONNÉES PROBABILISTES PAR DÉCOMPOSITION DE MÉLANGE DE COPULES ET APPLICATION À UNE BASE DE DONNÉES CLIMATOLOGIQUES

Directeurs de thèse : Edwin Diday et Alain Chédin

JURY

M Gérard Mégie	Président
Mme Lynne Billard	Rapporteur
M Gérard Govaert	Rapporteur
M Berthold Schweizer	Rapporteur
M Alain Chédin	Co-directeur
M Edwin Diday	Co-directeur
M Fabrice Rossi	Invité
M Gilbert Saporta	Invité

Thèse préparée au sein du Centre de Recherche de Mathématiques de la Décision
(Université Paris IX Dauphine) et du Laboratoire de Météorologie Dynamique
(CNRS, Ecole Polytechnique, Palaiseau)

A mon père,
Georges, Jean-Paul Vrac

Remerciements

"Ce sont toutes ces petites complicités, ces attentions discrètes, ces centaines de sourires qui ont façonné ma vie. Merci!"
Emilie Delarue

Je tiens tout d'abord à remercier le Professeur Edwin Diday. Il a su sagement orienter mon travail durant ces années de thèse. Ses compétences scientifiques m'ont permis de découvrir le domaine de la recherche sur des bases solides. Nos échanges (scientifiquement parfois vigoureux) ont toujours débouché sur des perspectives et des études passionnantes.

Je tiens également à remercier sincèrement Alain Chédin pour l'intérêt qu'il porte à mon travail et aux statistiques en général. Sa pédagogie et son enthousiasme nous ont permis de "dépasser le mur" séparant la physique atmosphérique et les statistiques.

Je remercie évidemment Gérard Mégie d'avoir bien voulu participer à ce jury et d'en avoir été le président.

Je souhaite remercier vivement Lynne Billard, Gérard Govaert et Berthold Schweizer d'avoir accepté d'être rapporteurs, malgré leurs nombreuses charges respectives.

Merci à Fabrice Rossi et à Gilbert Saporta d'avoir accepté mon invitation à ce jury et d'avoir porté autant d'attention à mes recherches.

Mes remerciements vont évidemment vers Noelle Scott qui m'a accueillit dans le groupe ARA du LMD comme membre à part entière de cette équipe.

Merci à Claude Basdevant, directeur du LMD à mon début de thèse, et à Hervé Le Treut, directeur du LMD à ma fin de thèse, de m'avoir reçu amicalement.

J'adresse bien sur une pensée amicale à toute l'équipe du groupe ARA, et plus particulièrement à Raymond Armante pour son humour et ses compétences informatiques qui m'ont permis de venir à bout de nombre de difficultés ; merci à Gilles Bannerot (qui mériterait presque de faire partie de l'équipe ARA) pour ses envolées partisannes ; merci à Bernard Bonnet de sa bonne humeur ; merci à Chantal Claud de représenter l'Auvergne ; merci à Laurent Crépeau de ses relectures et de son calme légendaire ; merci à Cyril Crevoisier de rire à mes blagues débiles (il ne se force pas), nos échanges théatraux m'ont souvent paru telle une lumière dans le quotidien mathématique (j'en fais des tonnes?) ; merci à Marc Dexet pour son entrain (parfois fatiquant... Je plaisante, t'énerve pas...) à parler de "php", de motos et du Vietnam ; merci à Fadoua Eddounia pour ses gâteaux Marocains (j'aime, tu peux en refaire...) et pour sa gentillesse ; merci à Sarah Franquet d'avoir partagé un bureau de labeur, son écoute et sa douceur m'ont permis de travailler dans une ambiance posée ; merci à Philippe Naveau de son amitié, ses conseils scientifiques ont souvent éclaircit mes perspectives ; merci à Soumia Serrar de sa chaleur, de son aide pour préparer le café et de parler avec tant

d'émotions du Maroc ; merci à Sylvain Heillette, Gaby Radel et Claudia Stubenrauch pour leur sympathie.

Je remercie également (docteur) Karim Ramage et (docteur) Benjamin Sultan pour leur amitié et leur soutien, Philippe Drobinsky pour ses conseils et son attention, l'ancienne équipe LIDAR (Arnaud, Vincent...) ainsi que le secrétariat du LMD X (Eliane, Martine et Stéphane) pour leur aide.

Merci aux gens du Nord (qui ont dans les yeux...) de s'être déplacés et (parfois) de m'accueillir (j'aime la carbonnade flamande...).

(Ça en fait des mercis...)

Je tiens à remercier "Les Bramentombe" (Mathieu Bourgasser, Gérard Bourmeau, Laurent Laravine, Romuald N'Tounta, Jérome Peignot, Karim Smaili). Nos nombreuses aventures sur les planches et en-dehors m'ennivrent toujours autant...

Merci à Annie Vrac (ma maman à moi) et Alexandre Vrac (mon "grand" petit frère) d'être là chaque fois que j'ai besoin d'eux... Merci à Audrey Vrac (ma soeur) de partager des souvenirs...

Je voudrais remercier Hélène Brogniez qui a su mener une part essentielle à l'aboutissement de ce travail : me remonter le moral et me supporter... Bon, c'est un peu solennel, je l'accorde... Disons simplement que je voudrais remercier Hélène Brogniez pour... pour... pour rien. Juste parce que c'est elle...

Je tiens à remercier mon public sans lequel je ne serais pas là...

Table des matières

Introduction **13**

1 Analyse de données symboliques **19**
1.1 Introduction 19
1.2 Analyse de données symboliques 20
1.3 Présentation formelle de l'analyse de données symboliques 21
1.3.1 Objet symbolique booléen 22
1.3.2 Objet symbolique modal 23
1.4 Calcul de l'extension d'un objet symbolique 23
1.5 Généralisation et spécialisation d'une classe d'individus 25
1.5.1 Généralisation 25
1.5.2 Spécialisation 26
1.6 Données "fonctions de répartition" 28
1.7 Histogrammes à modalités nominales 29
1.8 Conclusion 29

2 Mélange de densités **31**
2.1 Introduction 31
2.2 Approche estimation 34
2.2.1 L'algorithme EM 34
2.3 Les variantes de EM 39
2.3.1 SEM 39
2.3.2 SAEM 41
2.3.3 MCEM 41
2.4 Approche classification 42
2.4.1 Décomposition par "nuées dynamiques" 42
2.4.2 CEM 44
2.5 Conclusion 45

3 Distribution de distributions **47**
3.1 Définition des fonctions de distribution de distributions 47
3.2 Estimation des fonctions de distributions de distributions 51
3.2.1 Approche paramétrique 52
3.2.1.1 Lois sur $[0,1]$ 53
3.2.1.2 Lois tronquées 53

3.2.2 Approche non-paramétrique 54
3.2.2.1 Histogramme 54
3.2.2.2 Estimateur naif 54
3.2.2.3 Estimateur par noyaux 56
3.2.2.4 Méthode du plus proche voisin 58
3.2.2.5 Méthode des noyaux variables 60
3.2.2.6 Estimateurs du maximum de vraisemblance pénalisée 60
3.2.2.7 Estimateurs par fonctions de poids générales 62
3.2.2.8 Estimateurs tronqués 62
3.3 Surface de distributions de distributions 63
3.3.1 Surface de densités de distributions 64
3.4 Conclusion 64

4 Copules **67**
4.1 Introduction 67
4.2 Définition et propriétés des copules bivariées 68
4.2.1 Bornes bivariées de Fréchet-Hoeffding 71
4.2.2 t-normes et copules 71
4.3 Définition et propriétés des copules multivariées 73
4.3.1 Bornes multivariées de Fréchet-Hoeffding 77
4.4 Copules Gaussiennes 77
4.5 Copules empiriques 78
4.6 Copules Archimédiennes 79
4.6.1 Copules Archimédiennes 2-dimensions 79
4.6.2 Exemples de copules Archimédiennes 82
4.6.2.1 Famille de Ali-Mikhail-Haq 82
4.6.2.2 Famille de Genest-Ghoudi 82
4.6.2.3 Famille de Frank 82
4.6.3 Copules Archimédiennes multivariées 83
4.7 Estimation de copules Archimédiennes bivariées 84
4.7.1 Estimation non-paramétrique d'une copule Archimédienne 84
4.7.2 Estimation d'une copule Archimédienne par maximum de vraisemblance 86
4.7.3 Estimation d'une copule Archimédienne par "τ" de Kendall ou "ρ" de Spearman 87
4.7.3.1 Coefficient de corrélation du τ de Kendall 88
4.7.3.2 Coefficient de corrélation de Spearman 89
4.8 Estimation de copules Archimédiennes multivariées 91
4.8.1 Estimation d'une copule Archimédienne multivariée par maximum de vraisemblance 91
4.8.1.1 Première méthode par maximum de vraisemblance 91
4.8.1.2 Deuxième méthode par maximum de vraisemblance 92
4.8.1.3 Troisième méthode par maximum de vraisemblance 93
4.8.2 Estimation d'une copule Archimédienne multivariée par τ de Kendall et ρ de Spearman 93
4.8.2.1 Première méthode par τ de Kendall 93

4.8.2.2 Deuxième méthode par τ de Kendall 94
4.8.2.3 Première méthode par ρ de Spearman 94
4.8.2.4 Deuxième méthode par ρ de Spearman 95
4.9 Conclusion 95

5 Décomposition de mélange de copules 97
5.1 Introduction 97
5.2 Objets symboliques associés à un ensemble de fonctions de répartition 99
5.2.1 Objet symbolique par volume d'hyper-cube 100
5.2.2 Objet symbolique par dérivée de copule 101
5.2.2.1 Par densité normée 101
5.2.2.2 Par intégration de densité 102
5.2.3 Objet symbolique par distance entre fonctions de répartition jointes . 102
5.3 Le choix des T_i 103
5.3.1 Méthode des triangles 103
5.3.2 Par estimation de densité et surface de densités 105
5.4 Décomposition de mélange de copules (2-D) 106
5.4.1 Décomposition de mélange de copules Archimédiennes par approche classification 107
5.4.2 Décomposition de mélange de copules Archimédiennes par EM 109
5.4.3 Mélange de copules par SEM 110
5.4.4 Mélange de copules par SAEM 111
5.4.5 Mélange de copules par MCEM 111
5.4.6 Mélange de copules par CEM 112
5.4.7 Décomposition de mélange de copules empiriques 112
5.5 Décompositions en dimension n 113
5.5.1 Copules multidimensionnelles 113
5.5.2 La méthode par couplage 114
5.5.3 Méthode par arbre binaire 115
5.6 Inférence 116
5.7 Convergence de l'algorithme par nuées dynamiques 116
5.8 Comportement asymptotique 118
5.9 Décomposition de mélange de FDD 120
5.10 Généralisation de la décomposition de mélange de densités 122
5.10.1 Propriétés d'une base de distributions de masse unitaire 122
5.10.2 Le mélange de densités comme cas particulier du mélange de copules . 124
5.11 Conclusions 125

6 Application climatique 127
6.1 Introduction 127
6.1.1 La réanalyse de données satellitaires NOAA/TOVS de 1979 à nos jours à des fins d'étude du climat 128
6.1.2 Les données climatiques 131
6.2 Classification par décomposition de mélange de copules - Exemple de 7 classes 131
6.2.1 Classification en température 132

6.2.2 Classification en humidité 136
6.2.3 Résultat de DMC par couplage 140
6.3 Inférence de DMC 143
6.4 Résultats complémentaires par DMC - Exemple de 18 classes 145
6.5 Classification mixte 153
6.5.1 La méthode des "Nuées Dynamiques" (ND) 155
6.5.2 Algorithme des voisins réciproques 156
6.5.3 Résultats 157
6.5.4 Variables discriminantes 159
6.5.5 Classification mixte sur valeurs de fonctions de répartition 161
6.6 Classification par EM 163
6.6.1 Classification sur les données numériques brutes 164
6.6.2 Classification par EM sur les valeurs de fonctions de répartition 165
6.7 Conclusion 167

7 Perspectives **169**

Conclusion **173**

Bibliographie **177**

A Article : "Symbolic Class Descriptions" **185**
A.1 Introduction 187
A.2 The Method 188
A.3 Examples 192
A.4 Conclusion 193

B Article : "Mixture decomposition of copulas and application to clymatology" **195**

C Cartes et graphiques **205**

Table des figures **225**

Liste des tableaux **227**

Introduction

" Comme il y a une infinité de choses sages qui sont menées d'une manière très folle, il y a aussi des folies qui sont conduites d'une manière très sage."
(Montesquieu)

Les tableaux de données de très grandes dimensions sont de plus en plus courants et leurs analyses deviennent extrêmement complexes du fait de la diversité et de la quantité croissante des informations à traiter. La réduction des données en un plus petit nombre de descriptions (associées aux "concepts" de la base, i.e. à des classes cohérentes d'individus issues d'un produit cartésien de variables qualitatives ou d'une classification automatique) est donc une étape devenue indispensable. Ce condensé de l'information est tout d'abord un outil descriptif pour les utilisateurs de grandes bases de données. Il peut également être vu comme une étape intermédiaire vers des analyses qui ne s'appliquent plus aux données initiales mais aux descriptions réduites.

Supposons que nous disposons d'une population de points dans l'atmosphère (chaque point est situé en une longitude, une latitude, une altitude), décrits par deux variables thermodynamiques (température, humidité). Si nous supposons que nous avons les valeurs des variables pour chacun des points 4 fois par jours pendant un an, chaque point est décrit par 2920 données ($2\ (variables) \times 365\ (jours) \times 4\ (fois\ par\ jour)$). Si nous avons 50 niveaux d'altitudes, 360 longitudes, 180 latitudes, une représentation tabulaire utilisée classiquement en analyse de données, contiendrait 9 460 800 000 données ($2920 \times 50\ (niveaux) \times 360\ (longitudes) \times 180\ (latitudes)$). Pour chaque point et pour chaque variable, une réduction des descriptions pourrait être une valeur unique telle qu'une moyenne, une médiane, un écart-type ou une statistique quelconque. Au cours de cette opération, une partie de l'information est perdue. Nous pouvons alors nous interroger sur la manière de garder d'avantage d'informations sur les points. Pour tenter de répondre à cette question, nous utilisons des descriptions d'objets plus complexes avec des indicateurs permettant de reconnaître et de décrire les individus à analyser. Chaque case du tableau de données réduites peut contenir des valeurs multiples, ou un intervalle, ou un histogramme (numérique ou nominal), ou une loi de probabilité (on dit alors qu'on des données probabilistes),... Le tableau de données est alors appelé *tableau de données symboliques*. Il en résulte l'idée d'étendre les méthodes d'analyse de données classiques à ces données plus riches (appelées *objets symboliques*) car elles contiennent comme cas particulier les données classiques (Bock, Diday, 2000, [8]). Ces objets permettent en outre d'ajouter des connaissances supplépentaires telles que des taxonomies, des variables hiérarchiques ou des règles (voir Bock, Diday, 2000, [8]).

Cette thèse a pour objectif principal d'analyser et de modéliser des données probabilistes, en étendant la problématique de la décomposition de mélange fini de densités à des individus décrits par des données probabilistes de type fonctions de répartition (tout au long de cette thèse, nous parlons parfois de "fonctions de distribution" à la place de fonctions de répartition pour éviter les répétitions).

L'intérêt des variables aléatoires à valeurs variables aléatoires date de la fin des années 60 et dès 1975, J. F. C. Kingman (1975, [69]) expliquait :"*In recent years, there are been a good deal of interest in the problem of describing the distributions of random elements which are themselves probability distributions on finite or infinite sets. Much of this has been motivated by problems of Bayesian inference and decision theory, in which the unknow state of nature takes the form of a probability distribution.*" Dans notre travail, nous ne sommes pas motivés par l'approche Bayesienne mais les réalisations des variables aléatoires que nous regardons sont bien des éléments représentant des lois de probabilité. Plus précisément, les individus sont caractérisés par un tableau où chaque colonne (associée à une variable descriptive) contient des réalisations d'une variable aléatoire à valeurs fonctions de répartition et chaque case du tableau contient donc une variable aléatoire (représentée par sa fonction de répartition). Les réalisations sont alors définies sur des espaces mesurés pouvant varier d'une colonne à l'autre. Pour deux individus, les variables aléatoires descriptives peuvent être dépendantes et de lois différentes. Pour caractériser la loi de probabilité de ces éléments aléatoires fonctions de répartition, nous développons la notion de "Fonction de Distribution de Distributions" (FDD) (introduite de façon empirique par Diday dans [33]), correspondant à une fonction de répartition de fonctions de répartition.

Le but de ce travail est également de donner un modèle probabiliste pour représenter et réduire les descriptions aléatoires d'une classe de tels individus par rapport à toutes les variables descriptives en gardant l'aspect probabiliste de la description de chaque individu. Cette réduction (avec un minimum de perte d'information) peut être modélisée par le formalisme des *objets symboliques modaux*. Elle passe par la définition d'une mesure de généralisation opérant sur l'espace des fonctions de répartition et est définie à partir des FDD et de la notion de *fonctions copules*.

Cette notion de copules est au coeur de ce travail en caractérisant la dépendance entre les fonctions de distribution de distributions multidimensionnelles et leurs marginales unidimensionnelles. Bien que de nombreux auteurs les aient découvertes sous différentes formes, les copules ont été formalisées par Abe Sklar en 1959 ([99]). Nous voyons tout au long de ce travail comment le théorème qui porte son nom s'insère dans une extension des méthodes de décomposition de mélange de lois pour données fonctions de répartition. Le principe de base de cette extension est basé sur la notion de fonctions de distribution de distributions et celle de copules. Il est donné par E. Diday dans [33], dans le cas particulier de FDD empiriques, d'un modèle binaire de copule définie par un seuil ϵ et d'un algorithme par "approche classification" en 2 dimensions. Ce principe est le suivant :

- calcul des fonctions de distribution de distributions,
- calcul de leur dépendance à l'aide de copules,
- association d'une copule à chaque classe, puis d'une classe à chaque copule par un

procédé de type nuées dynamiques.

La démonstration du fait que cette méthode de décomposition de mélange de densités pour des données probabilistes, généralise le mélange classique de densités est redonnée en chapitre 5.

De manière non reliée aux objectifs précédents, une autre catégorie de données probabilistes a été également abordée durant cette thèse: les *données probabilistes nominales*. Ces données peuvent être résumées par un tableau, avec chaque case contenant plusieurs réalisations pondérés d'une variable nominale. Le problème est la description d'une classe de tels individus en tenant compte à la fois de l'homogénéité de ses classes internes et de sa "discrimination" par rapport aux classes d'une partition a priori (i.e. par rapport à une variable nominales). Par exemple, si nous disposons de la variable "couleur" pour décrire un ensemble d'individus (tels que des maisons ou des voitures), leur description par rapport à cette variable peut être: rouge (20%), bleue (40%), vert (10%), jaune (30%). Chaque description est donc considérée comme un histogramme dont les modalités sont nominales. Nous abordons le développement d'une méthode permettant de décrire un ensemble d'individus caractérisés par de telles données en couplant un critère d'inertie avec un critère d'impureté, et nous travaillons alors sur une méthode à la fois de classification (algorithme non supervisé) et de discrimination (algorithme supervisé). Cependant, ce type de données ne constitue pas l'essentiel de ce travail et nous développons d'avantage la partie sur la décomposition de mélange de lois pour des données fonctions de répartition.

Le premier chapitre de cette thèse motive l'utilisation des données symboliques en analyse de données. Nous présentons le formalisme utilisé en analyse de données symboliques (ADS) ainsi que les différences entre l'ADS et l'analyse de données classiques. Nous exprimons également deux manières de condenser l'information issue d'un tableau de données symboliques déterministes (par généralisation et spécialisation), avec chaque case du tableau contenant un ensemble de valeurs ou un intervalle de valeurs. Nous présentons ensuite les données traitées dans ce travail, à savoir les données fonctions de distribution, ainsi qu'un type de données probabilistes différents: les histogrammes à modalités nominales.

Dans le deuxième chapitre, nous rappelons les différentes approches (classification et estimation) de la problématique de décomposition de mélange de densités de probabilité, et détaillons les méthodes "classiques" de résolution de cette problématique. Les méthodes données dans ce chapitre servent de point de départ aux extensions que nous réalisons par la suite pour les données probabilistes.

Le troisième chapitre constitue le premier apport original de ce travail en s'intéressant à l'extension du modèle classique de fonction de répartition pour des données probabilistes, à savoir, la modélisation d'une loi de probabilité caractérisant des données fonctions de répartition. Dans son article [33], E. Diday propose la notion de "Fonctions de Distribution de Distributions" (FDD) empiriques, basée sur un ensemble de fonctions de répartition. Dans ce chapitre, nous définissons le cadre mathématique formelle pour travailler sur des fonctions de distribution de distributions continues. Nous prouvons de plus que ces fonctions sont elles-

mêmes des fonctions de répartition, et nous rappelons alors différentes méthodes pour les estimer.

Dans le quatrième chapitre, nous détaillons une théorie insuffisamment utilisée, développée par Sklar (1959, [99]): la théorie des copules. Les copules sont des fonctions permettant de lier la fonction de répartition multidimensionnelle d'un n-uplet de variables aléatoires $(X_1, ..., X_n)$, avec ses marginales unidimensionnelles. Pour plus de clarté, nous présentons tout d'abord les copules bidimensionnelles avant les copules multidimensionnelles ainsi que leurs propriétés respectives. Une méthode de construction de copules est donnée afin d'obtenir une copule associée à la loi Gaussienne multidimensionnelle. Nous rappelons les propriétés d'une classe de copules paramétriques nommée classe des copules Archimédiennes (bivariées et multivariées). Nous présentons ensuite différentes méthodes d'estimation des paramètres de copules Archimédiennes, basées sur la vraisemblance, le coefficient de corrélation de Kendall ou celui de Spearman. En se basant sur une généralisation des copules Archimédiennes, nous proposons trois nouvelles méthodes originales d'estimation basées sur ces mêmes critères.

Le cinquième chapitre s'appuie sur les résultats originaux des chapitres 3 et 4, pour donner un nouvel apport original à ce travail en étendant les méthodes de décomposition de mélange de densités au cas des données de type fonctions de répartition. Ces extensions sont basées sur la notion de fonctions de distribution de distributions et sur la théorie des copules. Pour développer une approche plus générale que celle de [33], nous regardons tout d'abord différentes mesures de généralisation d'une classe d'individus décrits par leurs fonctions de répartition respectives, en définissant des objets symboliques modaux associés à une telle classe. Dès lors, en considérant des FDD continues et dérivables, ainsi que des copules paramétriques Archimédiennes, nous étendons la plupart des méthodes de décomposition de mélange de densités ("approche classification", EM, SEM, etc) au cas des données de type fonctions de répartition. Les extensions sont tout d'abord présentées en 2 dimensions, puis en dimension n avec une approche par copules empiriques, une approche par copules multivariées et deux approches par combinaisons de copules. D'autres aspects théoriques originaux sont ensuite développés, tels que le choix des valeurs T pour définir les FDD, la convergence, le comportement asymptotique, et le mélange de FDD.

Dans le chapitre six, nous mettons en oeuvre l'une des méthodes proposées au chapitre 5 s'appuyant sur notre apport, sur une base (colossale!) de données climatologiques très complexes et dépendantes: des valeurs de profils atmosphériques pour des variables thermodynamiques. Le but est la classification de ces profils en type de masse d'air, en vue d'enrichir une méthode d'inversion de l'équation de transfert radiatif, par des informations probabilistes a priori sur les variables physiques. Les données climatiques sont tout d'abord détaillées, ainsi que le but de l'application. Les résultats climatiques obtenus sont ensuite discutés et nous voyons que la modélisation de ces données de manière probabiliste est particulièrement adaptée, en présentant des comparaisons avec d'autres méthodes plus classiques de décomposition de mélange de densités ou de classification.

Le septième chapitre tente enfin de présenter différentes suites que l'on peut donner à

l'ensemble du travail sur le mélange de lois pour des données probabilistes. Nous donnons quelques éléments qui devraient permettre d'étendre le travail réalisé à des données encore plus générales (i.e. les données fonctionnelles), et de spécialiser les méthodes proposées aux données numériques classiques. Différents points de la méthode sont enfin discutés afin de préciser plusieurs pistes à étudier pour l'amélioration de la méthode et de ses applications.

Une conclusion nous permet ensuite de récapituler les apports de la méthode proposée dans ce travail.

Les annexes sont décomposées en trois parties:
- La première partie aborde les données de type histogrammes à modalités nominales. Le but est de décrire une classe d'individus caractérisés par de telles données, en tenant compte de l'homogénéité des classes le structurant et de sa description par rapport à une partition a priori. Nous présentons une méthode à la fois supervisée et non-supervisée permettant de réaliser ceci par arbre binaire. Cette annexe est un article (Vrac, Diday, Winsberg, Limam, 2002, [112]), publié dans le proceeding de la conférence IFCS 2002.
- La deuxième partie se compose d'un article (Vrac, Diday, Chédin, 2001, [109]) concernant le mélange de copules (ou mélange de distributions de distributions), publié dans le proceeding de la conférence "IPMU 2002".
- La troisième partie est constituée de différentes cartes et différents graphes permettant de détailler les résultats du chapitre 6. Ces tracés sont placés dans cette annexe pour ne pas surcharger ce chapitre.

Nous présentons donc tout d'abord l'analyse de données symboliques, son formalisme et le type de données utilisées.

Chapitre 1

Analyse de données symboliques

"The fact that all Mathematics is Symbolic Logic is one of the greatest discoveries of our age; and when this fact has been established, the remainder of the principles of mathematics consists in the analysis of Symbolic Logic itself."
Russell, Bertrand (1872-1970), Principles of Mathematics. 1903.

1.1 Introduction

Le but de l'analyse de données est d'extraire des connaissances d'un ensemble de données multidimensionnelles. Cet ensemble regroupe un grand nombre d'individus caractérisés par un nombre fini de variables qualitatives (nominales, ordinales) ou quantitatives (discrètes, continues). L'analyse de données utilise des méthodes et des algorithmes de différents types:

- les méthodes de classification qui tentent de former des classes homogènes de l'ensemble des individus;
- les méthodes factorielles qui visent à diminuer le nombre de variables en les résumant par un petit nombre de composantes synthétiques. Dans le cas d'un tableau de données numériques ou qualitatives, on utilisera par exemple l'analyse en composantes principales ou l'analyse des correspondances;
- les méthodes de discrimination qui cherchent à caractériser les classes d'une partition donnée;
- les méthodes de segmentation par arbres, etc.

La plupart des méthodes d'analyse de données structurent les données par une application d'un ensemble d'individus $E = \{e_1, ..., e_N\}$ dans l'ensemble d'observations $O = \{O_1, ..., O_p\}$ de l'ensemble des variables $X = \{X_1, ..., X_p\}$. L'application associe à chaque individu $e_i \in E$ sa description $X_j(e_i)$ à valeurs dans O_j. L'application est généralement représentée par un tableau individus $\times$ variables avec N lignes correspondant aux N individus et p colonnes représentant les p variables (Tableau 1.1). En analyse de données classiques, une case du tableau de données contient une seule valeur: la description d'un individu pour une variable.

	X_1	...	X_j	...	X_p
w_1					
$\vdots$			$\vdots$		
w_i	...	...	$X_j(w_i)$		
$\vdots$					
w_N					

TAB. 1.1: *Tableau de données classiques*

Dans la pratique, les données peuvent être plus complexes et chaque case du tableau peut contenir des valeurs multiples, parfois pondérées, qui peuvent être liées entre elles par des règles ou des taxonomies.

Différentes méthodes de codage ont été créées pour ce genre de données. Ces recodages peuvent entrainer des pertes d'information. La structuration des données complexes implique alors une modélisation loin de la réalité. Le traitement de ces données dans leur format d'origine peut donc s'avérer essentiel.

Des données de cette forme sont dites "symboliques" car elles ne sont pas purement numériques (voir Bock, Diday, 2000, [8]). Concrètement, une case d'un tableau de données symboliques peut contenir différents types de données, par exemple:

- une valeur quantitative ou qualitative,
- des valeurs quantitatives ou qualitatives,
- un intervalle,
- un histogramme ordinal ou nominal,
- une loi de probabilité, etc.

Des connaissances supplémentaires peuvent être fournies pour tenir compte de variables taxonomiques ou de dépendances hiérarchiques.

L'analyse de données classiques est un cas particulier de l'Analyse de Données Symboliques (ADS). L'ADS donne une importance prépondérante aux individus en donnant un cadre dans lequel ils peuvent être représentés et analysés en tenant compte de leur complexité et de leur variation interne de manière proche de la réalité.

1.2 Analyse de données symboliques

Pour décrire la naissance de l'ADS, E. Diday explique que l'idée principale de l'analyse de données symboliques est donnée par Aristote, 4 siècles avant JC. L' "Aristotle Organon" distingue les "individus du premier ordre" (un cheval, un homme) considérés comme des unités associées à des individus, des "individus du second ordre" ou "concepts" (le cheval, l'homme) pris comme unités associées à une classe d'individus. Le but de l'analyse de données symboliques est d'étendre l'analyse de données classiques aux individus du second ordre.

Par exemple, dans une enquête statistique sur un pays, chaque individu de chaque région est décrit par un ensemble de variables numériques ou catégoriques. De tels individus sont nommés "individus du premier ordre". Pour étudier les régions considérées comme "individus du second ordre", nous pouvons décrire chacune en résumant les valeurs prises par ses habitants par des intervalles inter-quartiles, des ensembles de valeurs catégoriques, des histogrammes ou des fonctions de répartition... dépendant de la variable concernée. Nous obtenons une base de données symboliques. Chaque ligne est la description d'une région et chaque colonne est associée à une variable symbolique.

L'ADS extrait des résultats explicatifs (i.e. de la connaissance) par extraction d'"objets symboliques" modélisant un "concept". Un concept est défini par un ensemble de propriétés appelées 'intention" ou "compréhension" et par un ensemble d'individus, appelé "extension", qui satisfont ces propriétés.

Pour décrire la place de l'analyse de données symboliques E. Diday dit ([32]): "L'analyse de données symbolique se situe à un carrefour entre différentes disciplines dont elle s'inspire mais dont elle diffère, que ce soit par les données qu'elle traite, par les méthodes qu'elle utilise ou par les objectifs qu'elle poursuit. Elle étend la problématique de l'analyse de données classiques à des objets plus complexes, en s'intéressant plus aux aspects exploratoires (histogrammes, analyse canonique d'objets symboliques, pyramides,...)."

L'ADS s'appuie donc sur deux notions définies sur le même ensemble de description.

La première est définie par un ensemble d'entités non décomposables du monde réel; ces entités sont nommées *individus* et sont modélisées par des descriptions de propriétés qu'elles satisfont, sous forme de vecteurs.

La seconde est définie par des concepts représentés à l'aide de classes d'individus. Les classes sont décrites sous forme de vecteurs de même dimension que ceux des individus, en prenant en compte leur variation. Les concepts sont modélisés par des objets symboliques, i.e. par des descriptions exprimant des propriétés satisfaites par les individus de la classe qu'il représente et d'un mode de calcul de son extension (voir section 1.4) dans l'ensemble des individus.

1.3 Présentation formelle de l'analyse de données symboliques

L'analyse de données symboliques est définie sur deux espaces: l'espace des individus et l'espace des descriptions.
Nous posons les notations suivantes

- Ω l'ensemble des individus,
- D l'ensemble des descriptions,
- Y une application de Ω dans D, associant à chaque individu de Ω ou à chaque classe d'individus de Ω sa description.

Ainsi D permet de décrire une classe d'individus ou un seul individu considéré comme une classe réduite à un seul élément. Avec ces notations, nous définissons ce qu'est un objet

symbolique.

Définition 1 *Un objet symbolique est un triplet (a, R, d) avec*

- *R un opérateur de comparaison entre 2 descriptions (le résultat peut être à valeurs dans $\{0,1\}$ ou dans $[0,1]$),*
- *d est une description,*
- *a une application de Ω à valeurs dans $L = \{0,1\}$ ou $L = [0,1]$ qui associe à un individu w de Ω un degré d'adéquation entre sa description et la description d. La fonction a est nommée "fonction de reconnaissance".*

Nous notons

$$a(w) = [Y(w)\ R\ d] \tag{1.1}$$

Remarque: Présenté ainsi, nous pouvons nous demander à quoi sert a puisque R et d suffisent à définir un objet symbolique par la relation (1.1). En réalité, cette définition est la plus simple à comprendre lors d'une première approche des objets symboliques. De manière plus détaillée, la fonction de reconnaissance a inclue des filtres h_e sur les individus, h_v sur les variables et h_d sur les descriptions. La relation (1.1) sécrit en réalité

$$a(w) = [h_e(Y(w))\ h_v(R)\ h_d(d)]$$

et $(a, R, d) = (h_e, h_v, h_d, R, d)$. Pour davantage de détails sur la définition formelle des objets symboliques, voir le livre de Bock, Diday (2000, [8]).

Selon que $L = \{0,1\}$ ou $[0,1]$, nous distinguons deux types d'objets symboliques:
Les objets symboliques booléens: lorsque la fonction de reconnaissance a est à valeurs dans $\{0,1\}$. Un individu w vérifie ou ne vérifie pas les propriété de l'objet symbolique.
Les objets symboliques modaux: lorsque la fonction de reconnaissance est à valeurs dans $[0,1]$. Un individu w a une probabilité de vérifier les propriétés de l'objet symbolique.

Nous présentons maintenant ces deux types d'objets symboliques.

1.3.1 Objet symbolique booléen

Si la base de données symboliques contient p variables, nous notons
- $Y(w) = (Y_1(w), ..., Y_p(w))$,
- $D = D_1 \times ... \times D_p$,
- $d = (d_1, ..., d_p) \in D$,
- $R = (R_1, ..., R_p)$ avec R_i la relation sur D_i.

Nous nommons "assertion" une conjonction d'objets symboliques. L'assertion est un cas particulier d'un objet symbolique défini par $s = (a, R, d)$ avec

$$a(w) = \wedge_{i=1}^{p} [Y_i(w) R_i d_i].$$

Pays	âge	taille	continent
w_1	37	1.75	Afrique
w_2	45	1.82	Afrique
w_3	42	1.80	Europe

TAB. 1.2: *Exemple de données pour objet symbolique booléen*

Exemple 1 (Objet symbolique booléen) *Soit une classe $C = (w_1, w_2, w_3)$ de pays décrits par trois variables aléatoires X_1, X_2, X_3 représentant l'âge moyen de la population, la taille moyenne de la population et son continent (Tableau 1.2).*

L'assertion booléenne associée à C peut être:

$$a(w) = [ge(w) \subseteq [37, 45]] \wedge [taille(w) \subseteq [1.75, 1.82]] \wedge [continent \subseteq \{Afrique, Europe\}]$$

Nous posons $d_1 = [1.75, 1.82]$, $d_2 = [37, 45]$, $R_1 = R_2 = R_3 = \subseteq$ et $d_3 = \{Afrique, Europe\}$ et l'objet symbolique associé à C est

$$a(w) = \begin{cases} 1\ (= vrai) & si\ pour\ tout\ i,\ X_i(w) \in d_i \\ 0\ (= faux) & sinon \end{cases}$$

1.3.2 Objet symbolique modal

Les objets symboliques modaux sont définis avec des fonctions d'agrégation et de comparaison entre descriptions modales des individus de Ω et celles de l'espace des descriptions D.

Si un individu w de Ω a des descriptions probabilistes $F_1^w, ..., F_p^w$ par rapport à p variables et qu'on veut les comparer aux descriptions probabilistes $G_1, ..., G_p$, l'objet symbolique modal associé est

$$a(w) = f(g(F_1^w, G_1), ..., g(F_p^w, G_p))$$

avec g une fonction de comparaison entre les deux descriptions probabilistes F_i^w et G_i, f est une fonction d'agrégation sur toutes les variables descriptives. La fonction f induit un degré d'adéquation sur toutes les variables. Par exemple, f peut être un produit et g une distance entre fonctions de répartition.

Les objets symboliques permettent donc de comparer des individus à des descriptions. Nous pouvons alors définir la classe des individus vérifiant les descriptions d'objets symboliques: nous parlons d'extension.

1.4 Calcul de l'extension d'un objet symbolique

L'étude des relations entre les ensembles Ω et D passe par la comparaison entre les descriptions des individus de Ω et celles de D.

Dans le cas booléen, l'extension d'un objet symbolique est l'ensemble des individus w tels que $a(w) = vrai$

$$Extent(a) = \{w \in \Omega / a(w) = vrai\}.$$

Dans le cas modal, pour un seuil α donné, l'extension est l'ensemble des individus tels que $a(w) \geq \alpha$

$$Extent(a) = \{w \in \Omega / a(w) \geq \alpha\}.$$

Supposons que l'individu $w \in \Omega$ ait des descriptions probabilistes $(F_1^w, ..., F_p^w) \in D_1 \times ... \times D_p$ par rapport à $X_1, ..., X_p$. Nous voulons comparer $(F_1^w, ..., F_p^w)$ à des descriptions probabilistes $(G_1, ..., G_p)$. Les applications g et f sont définies par

$$g : D_i \times D_i \longrightarrow [0,1]$$

$$g : (F_i^w, G_i) \longmapsto g(F_i^w, G_i)$$

et

$$f : [0,1]^p \longrightarrow [0,1]$$

$$f : ((F_1^w, G_1), ..., (F_p^w, G_p)) \longmapsto f((F_1^w, G_1), ..., (F_p^w, G_p))$$

La valeur $g(F_i^w, G_i)$ est le degré de ressemblance entre la description marginale F_i^w de w et G_i et f est une fonction d'agrégation des différents degrés de ressemblance. Lorsque la fonction g est symétrique, nous pouvons avoir :

- g est une similarité : g est positive, symétrique et

$$g(x,x) = g(y,y) \geq g(x,y) \ \forall x \neq y$$

- g est une dissimilarité : g est symétrique et $g(x,x) = 0 \ \forall x$
- g est une distance : g est une dissimilarité avec inégalité triangulaire et

$$g(x,y) = 0 \Longrightarrow x = y$$

Les deux exemples suivants sont des distances

$$g(F_i^w, G_i) = sup_x |F_i^w(x) - G_i(x)| \text{ dans le cas discret,}$$

$$g(F_i^w, G_i) = \int_{-\infty}^{+\infty} (\sqrt{F_i^w} - \sqrt{G_i})^2 dx \text{ dans le cas continu.}$$

Cependant, g peut être non-symétrique et mesurer l'adéquation ou l'appariement d'un individu avec la description d'une classe.

Nous voyons qu'un objet symbolique décrit de manière exhaustive son extension. Ceci nous amène à la propriété de complétude d'un objet symbolique, en lien étroit avec la théorie des treillis de Galois (Diday, Emilion, 1995, [36], 1997, [37]).

Soient $P(\Omega)$ l'ensemble des parties de Ω et S l'ensemble des objets symboliques. Soit F une application de $P(\Omega)$ dans S qui à une classe C d'individus de Ω associe un objet symbolique $F(C) = s$. Soit G une application de S dans $P(\Omega)$ qui à un objet symbolique associe son extension $G(s) = Extent(s)$. Un objet symbolique complet s est tel que $F(G(s)) = s$.

Pour résumer, toutes les informations sur les individus appartenant à l'extension de s sont dans l'objet symbolique s.

Des notions ont été introduites dans le cadre du traitement des objets symboliques, la notion d'extension, la notion d'intention d'un objet symbolique qui est sa description sous jacente, l'union ou l'intersection d'objets symboliques, etc. Ces notions entrent dans la généralisation ou la spécialisation d'une classe d'individus.

1.5 Généralisation et spécialisation d'une classe d'individus

1.5.1 Généralisation

Soit Ω un ensemble d'individus w_i décrits par p variables $X_1, ..., X_p$. Une description D_j^i est associé à l'individu w_i pour chaque variable X_j. La description D_j^i peut être une valeur unique (dans le cas de données classiques), un ensemble ou un intervalle de valeurs (pour l'analyse de données symboliques). Dans le formalisme de l'ADS, l'objet symbolique associé à w_i est

$$a(w_i) = \wedge_{j=1}^{p}[X_j(w_i)\ R\ D_j^i]$$

avec R un opérateur d'égalité, d'appartenance, d'inclusion.

Soit $C = \{w_1, ..., w_N\}$ une classe d'individus de Ω. La description généralisante de C pour le vecteur $X_j = (X_j^1, ..., X_j^N)$ est

$$D_j = \bigcup_{i=1}^{N} D_j^i.$$

La description de C par rapport au vecteur X_j est

$$a(w) = [X_j(w)\ R\ D_j].$$

Si le processus de généralisation de C se poursuit par rapport à tous les vecteurs $X_1, ..., X_p$, la description généralisante de C est l'objet symbolique intentionnel

$$a(w) = \wedge_{j=1}^{p}[X_j(w)\ R\ D_j].$$

La généralisation résume les lignes d'un tableau de données symboliques autorisant plusieurs valeurs par case, en une seule ligne caractérisée par l'objet symbolique intentionnel donnant une description généralisante de C.

	X_1	...	X_j	...	X_p
w_1					
$\vdots$			$\vdots$		
w_i	...	...	$X_j(w_i) = [a_j^i, b_j^i]$		
$\vdots$					
w_N					

TAB. 1.3: *Tableau de données symboliques*

Soit la classe C des individus décrits dans le tableau 1.3 de données symboliques. La description généralisante de C est donnée par le tableau 1.4 avec $a_j = min_{1 \leq i \leq N} a_j^i$ et $b_j = max_{1 \leq i \leq N} b_j^i$.

Si $R = \subseteq$, l'extension de l'objet symbolique intentionnel généralisant obtenu par union est l'ensemble des individus dont les descriptions expriment des propriétés satisfaites par l'objet symbolique de C.

	X_1	...	X_j	...	X_p
C	$[a_1, b_1]$	...	$[a_j, b_j]$	...	$[a_p, b_p]$

TAB. 1.4: *Généralisation de C*

1.5.2 Spécialisation

Soit $C = \{w_1, ..., w_N\}$ un ensemble d'individus décrits de la même manière que dans la section 1.5.1 par le tableau 1.3. Partant de la classe d'objets symboliques individuels associés aux individus de C, la spécialisation est l'objet symbolique intentionnel associé à la description commune de tous les individus de C. Si D_j^i est le domaine de descriptions de w_i par rapport à X_j, la description spécialisant la classe C par rapport au vecteur $X_j = (X_j^1, ..., X_j^N)$ est

$$D_j = \bigcap_{i=1}^{N} D_j^i.$$

La spécialisation induit l'objet symbolique intentionnel associé à X_j

$$a(w) = [X_j(w)\ R\ D_j]$$

avec R l'opérateur de comparaison ou d'appariement. Le processus de spécialisation des descriptions de C se poursuit par rapport à toutes les variables $X_1, ..., X_p$.

Nous pouvons résumer le cadre mathématique de l'analyse de données symboliques par la figure 1.1

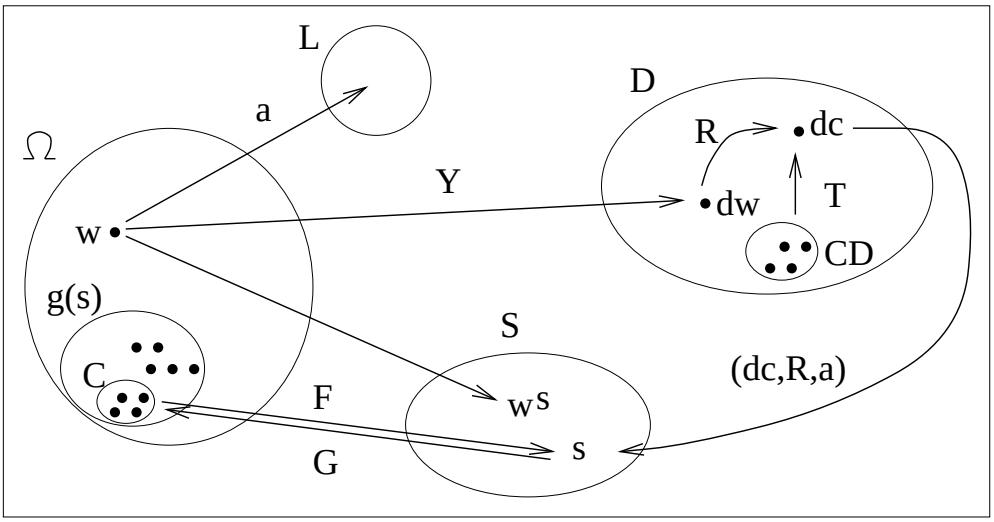


FIG. 1.1: *Schéma du cadre de l'analyse symbolique*

Pour cette figure,

– Ω est l'ensemble des individus,

- D est l'ensemble des descriptions,
- S est l'ensemble des objets symboliques,
- Y est une fonction de description,
- a est une fonction de reconnaissance de Ω dans L,
- $L = \{vrai, faux\}$ ou $L = [0, 1]$ est l'espace d'arrivée de la fonction de reconnaissance,
- R est une relation de comparaison,
- T est une application de généralisation,
- F est une application d'intention,
- G est une application d'extension,
- $d_w = Y(w)$ est une description de w,
- $w^s = F(w) = (a, R, Y(w))$ est un objet symbolique individuel,
- d_C est une description de la classe C,
- s est l'objet symbolique intentionnel donné par $F(C) = (a, R, d_C)$ avec $a = [Y(w)\ R\ d_C]$,
- $G(s)$ est l'extension de s.

Nous observons au moins six avantages à utiliser des objets symboliques:

Tout d'abord, ils donnent un résumé de la base de données symboliques initiale d'une manière explicative (proche du langage de l'utilisateur), en exprimant des descriptions basées sur des propriétés concernant les variables initiales ou les variables significatives.

Deuxièmement, ils peuvent être transformés en terme de requêtes de bases de données et propager des concepts entre bases de données.

Troisièmement, en étant indépendant de la base de données, ils peuvent identifier n'importe quel individu vérifiant leurs descriptions dans n'importe quelle base de données.

Quatrièmement, ils peuvent donner une nouvelle base de données symboliques de niveau supérieur sur laquelle une analyse de données symboliques de second niveau peut être appliquée.

Cinquièmement, pour caractériser un concept, ils peuvent joindre différentes propriétés basées sur différentes variables de différentes relations dans une base de données pour différents échantillons d'une population.

Sixièmement, au lieu d'analyser une gigantesque base de données constituée de plusieurs bases, nous pouvons résumer chaque base de données par des objets symboliques et appliquer l'analyse de données symboliques à l'ensemble des objets symboliques.

Nous avons vu que le formalisme de l'ADS permet de traiter de nombreuses variables symboliques. Dans ce travail, nous nous sommes intéressés aux données probabilistes du type "fonctions de répartition" (nous rappelons que ces données peuvent être appelées "fonctions de distribution" tout au long de cette thèse).

1.6 Données "fonctions de répartition"

Soit $W = (w_1, ..., w_N)$ un ensemble d'individus décrits par p variables à valeurs fonctionnelles et plus précisément à valeurs fonctions de répartition. L'ensemble W est un échantillon de N individus d'une population totale Ω. Pour mémoire, la fonction de répartition F d'une variable aléatoire X est définie par

$$F(x) = \mathbb{P}(X \leq x). \tag{1.2}$$

Nous disposons d'un tableau de données de N lignes et p colonne, avec à l'intersection de la ligne i, colonne j (individu i, variable j) une fonction de répartition (voir Figure 1.2).

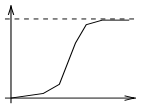
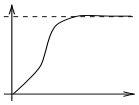
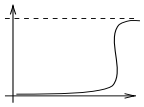
	Variable 1	.....	Variable p
Ind 1		.....	
.....	.....	.....	.....
Ind N		.....	

FIG. 1.2: *Tableau de données fonctions de répartition*

Nous résumons les données par un ensemble $\mathfrak{F} = (F_1, ..., F_N)$ décrivant W, avec $F_i = (F_i^1, ..., F_i^p)$. Chaque F_i^j correspond à la fonction de répartition de l'individu i pour la variable j. L'ensemble $\mathfrak{F}$ (appelé "base de distributions") est un échantillon de N p-uplets de fonctions de distribution provenant de $\Omega_F = \Omega_F^1 \times ... \times \Omega_F^p$, avec Ω_F^j l'ensemble des fonctions de répartition possibles décrivant les individus de Ω pour la variable j. Par ailleurs, nous supposons que F_i^j est une estimation de la fonction de répartition réelle de la variable j, obtenue par un échantillon de valeurs numériques de la variable j pour l'individu i.

Les données dont nous disposons sont donc des données stochastiques: les individus sont décrits par des variables aléatoires caractérisées par leurs lois de probabilité. Dans ce cas, la généralisation et la spécialisation opèrent sur l'espace des fonctions de répartition associé aux variables aléatoires descriptives; on parle de généralisation et spécialisation stochastique. Des opérateurs d'union et d'intersection entre fonctions de distribution ont été définis, ils sont basés sur des fonctions d'ensembles et appartiennent à la classe des capacités de Choquet (Choquet, 1953, [20], Diday et Emilion, 1997, [37], Hillali, 1998, [59]). Les normes et conormes triangulaires développées par Schweizer et Sklar ([95]) dans leur étude des espaces métriques probabilistes, sont des cas particuliers de ces mesures.

Dans notre étude, nous avons regardé quelles méthodes pouvaient réaliser la classification de tels individus en tenant compte de l'information stochastique, et comment caractériser les lois de probabilité de données fonctions de distribution.

Nous introduisons pour cela les notions de "distributions de distribution" ([33], [109]) associées à celle de "copules" (Sklar, [99], Schweizer et Sklar, [95]), pour étendre les méthodes de décomposition de mélange de densités dans le cas de données fonctions de répartition.

De manière plus générale, les méthodes que nous proposons dans cette thèse s'appliquent aux données fonctionnelles dont les fonctions de répartition sont un cas particulier (voir Ramsey et Silverman, 1997, [83] pour plus de détails sur les données fonctionnelles). L'intérêt d'une modélisation fonctionnelle est immédiat quand on dispose par exemple de données non comparables point à point.

1.7 Histogrammes à modalités nominales

Les données stochastiques nominales sont également présentes en ADS, nous les nommons histogrammes à modalités nominales. Elles se constituent de la manière suivante:

supposons que chaque individu soit décrit pour une variable par différentes caractéristiques non opposées. Par exemple une voiture peut être caractérisée par le fait qu'elle soit bleu, rouge et jaune. Nous pouvons préciser dans quelles proportions ces couleurs sont sur la voiture: bleu (25%), rouge (50%) et jaune (25%). Nous avons un histogramme à modalités nominales pour décrire l'individu pour la variable couleur.

Nous donnons dans l'annexe A une méthode permettant de décrire un ensemble de tels individus en tenant compte à la fois de l'homogénéité des sous-classes de l'ensemble et de la discrimination de ces sous-classes par rapport à une partition a priori. Cette approche est donc originale par le fait qu'elle combine un critère d'inertie (homogénéité) avec un critère d'impureté (discrimination), et par les données qu'elle traite (histogrammes à modalités nominales).

1.8 Conclusion

Nous avons vu que le formalisme des objets symboliques est proche du langage utilisé par l'utilisateur. L'ADS diffère donc de l'analyse de données classiques par les descriptions utilisées sur les individus: ces données peuvent être des intervalles, des histogrammes, des fonctions de répartition, des taxonomiques, avec des variables hiérarchiques, des reglès...

De plus, un concept représenté par une classe d'individus (classiques ou symboliques) induit un objet symbolique permettant de le caractériser (par généralisation ou spécialisation) et inversement, un objet symbolique s induit la classe des individus qui vérifient les descriptions de s: l'extension de s.

Par ailleurs, l'ADS étend les méthodes classiques d'analyse aux données symboliques. Nous proposons d'étendre, dans ce travail, les méthodes de décomposition de mélange de densités. Nous présentons donc dans le chapitre suivant la problématique des mélanges et les approches pour effectuer la décomposition.

Chapitre 2

Mélange de densités

"Il y a quatre types idéals: le crétin, l'imbécile, le stupide et le fou. Le normal, c'est le mélange équilibré des quatre."
Umberto Eco, Le pendule de Foucault.

2.1 Introduction

Un des plus vieux problèmes de la statistique inférentielle consiste à estimer une loi de probabilité à partir d'un échantillon observé. Dans beaucoup de cas, pour traiter ce problème, on est amené à considérer que la distribution à estimer est un mélange fini de distributions de formes analytiques simples. C'est le cas lorsque plusieurs sources aléatoires ont été décelées ou soupçonnées, ou lorsque l'assimilation de la distribution à une famille paramétrique de lois de probabilité ne donne pas de résultats satisfaisants.

Une des premières analyses majeures impliquant l'utilisation des modèles de mélanges est due au biométricien Karl Pearson il y a une centaine d'années. Dans son article de 1894 [81], maintenant classique, Pearson crée un mélange de deux densités de probabilité gaussiennes de moyennes μ_1 et μ_2 et de variances σ_1^2 et σ_2^2 dans les proportions π_1 et π_2 pour le jeu de données fournit par Weldon (1892, 1893). La base était composée de mesures de proportions de la taille du crâne par rapport a la longueur du corps de 1000 crabes échantillonnés dans la baie de Naples et enregistrée sous la forme de 29 intervalles. Weldon supposait que l'assymétrie dans l'histogramme des données était un signal signifiant que les données avaient deux sous-espèces. Weldon se tourna vers Karl Pearson pour de l'aide. L'approche de Pearson fût le modèle de mélange. Celle-ci suggérait (comme le faisait Weldon) que deux sous-espèces étaient présentes. Pearson utilisa la méthode des moments pour estimer les cinq paramètres de son mélange de gaussiennes hétéroscedastiques (variances inégales) en résolvant un polynôme de degré neuf (nonic). Les années suivantes, de nombreuses tentatives fûrent faites pour simplifier l'approche de Pearson basée sur les moments.

La formalisation d'un mélange de densités est la suivante. Notons $X_1, ..., X_N$ un échantillon aléatoire de taille N, où X_j est un vecteur aléatoire p-dimensionnel de densité de

probabilité $f(x_j)$ sur $\mathbb{R}^p$. En pratique , X_j contient les variables aléatoires correspondant à p mesures faites pour le j^{eme} individu. Soit $X = (X_1^T, ..., X_N^T)^T$ avec X_j^T, le vecteur transposé de X_j. X représente l'échantillon complet, X est un N-uplet de points dans $\mathbb{R}^p$. Les réalisations d'un vecteur aléatoire sont notées par la lettre minuscule correspondante. Par exemple, $x = (x_1^T, ..., x_N^T)$ est un échantillon aléatoire observé où x_j est la valeur observée du vecteur aléatoire X_j. Le vecteur variable X_j est considéré être un vecteur aléatoire continu. Dans le cas où X_j est discret nous pouvons voir $f(x_j)$ comme une densité par l'adoption d'une mesure de comptage. Le modèle de mélange suppose que la densité $f(x_j)$ de X_j peut être écrite sous la forme

$$f(x_j) = \sum_{i=1}^{k} \pi_i f_i(x_j, a_i) \tag{2.1}$$

où les $f_i(x_j, a_i)$ sont des densités de paramètre a_i de $\mathbb{R}^s$ (s étant le nombre de coordonnées du paramètre) et les π_i sont positifs et de somme 1, π_i étant la probabilité qu'un point de l'échantillon suive la loi de densité $f(., a_i)$,

$$0 \leq \pi_i \leq 1 (i = 1, ..., k) \tag{2.2}$$

et

$$\sum_{i=1}^{k} \pi_i = 1. \tag{2.3}$$

Les quantités $\pi_1, ..., \pi_k$ sont nommées les proportions du mélange. Les fonctions $f_1(x_j), ..., f_k(x_j)$ étant des densités, (2.1) est une densité. Dans la formule (2.1), le nombre de composantes k est fixé. Quelques auteurs ont tenté de définir des critères pour obtenir le nombre optimal de composantes. Nous pouvons citer pour exemple, Celeux et Soromenho (1996, [16]) qui proposent un critère d'entropie appelé NEC (normalized entropy criterion) et Biernackie, Celeux et Govaert (1999, [6]) qui tentent d'améliorer le NEC. Nous renvoyons le lecteur intéressé par le sujet aux différentes approches proposées dans [58], [25], [16], [6], [71], [2], [63], et dans la suite, nous considérons que le nombre de composantes est donné.

Une manière de générer un vecteur aléatoire X_j avec le mélange (2.1) est :
Soit Z_j une variable aléatoire discrète prenant des valeurs dans $\{1, ..., k\}$ avec les probabilités $\pi_1, ..., \pi_k$, respectivement, et supposons que que la densité conditionnelle de X_j, $Z_j = i$ donné, est $f_i(x_j)(i = 1, ..., k)$. La densité de X_j est donc $f(x_j)$. Dans ce contexte, la variable aléatoire Z_j peut être vue comme l'étiquette de la composante (component label) du vecteur X_j car elle définit la composante à laquelle appartient le vecteur X_j. Ultérieurement, nous travaillerons avec un vecteur d'étiquettes de composantes k-dimensionnelle $\boldsymbol{Z}_j$ (au lieu d'une variable discrète Z_j unique), avec le i^{eme} élément de $\boldsymbol{Z}_j$, $Z_{ij} = (\boldsymbol{Z}_j)_i$ égale à 1 ou 0 selon le fait que la composante d'origine de X_j dans le mélange est i ou non $(i = 1, ..., k)$. Nous avons $\boldsymbol{Z}_j$ distribué selon une loi multinomiale consistant en un tirage aléatoire sur k catégories avec probabilités $\pi_1, ..., \pi_k$,

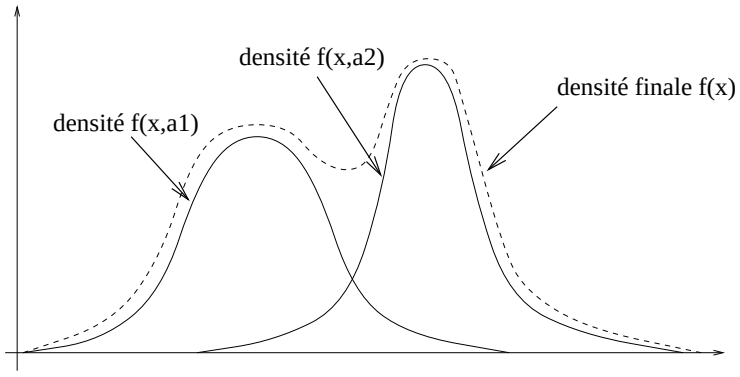


FIG. 2.1: *Mélange de deux densités gaussiennes*

$$\mathbb{P}(\boldsymbol{Z}_j = \boldsymbol{z}_j) = \pi_1^{z_{1j}} \pi_2^{z_{2j}} ... \pi_k^{z_{kj}}. \tag{2.4}$$

Nous écrivons

$$\boldsymbol{Z}_j \sim Mult_k(1, \pi), \tag{2.5}$$

avec $\pi = (\pi_1, ..., \pi_k)^T$.
Dans cette interprétation de modèle de mélange, une situation où le mélange à k composantes (2.1) est applicable est quand X_j est tiré à partir d'une population G constituée de k groupes $G_1, ..., G_k$, en proportions $\pi_1, ..., \pi_k$. Si la densité de X_j dans le groupe G_i est donnée par $f_i(x_j)$ pour $i = 1, ..., k$, la densité de X_j a la forme du mélange à k composante (2.1). Dans cette situation, les k composantes du mélange peuvent être physiquement identifiées avec les k groupes $G_1, ..., G_k$.

Etant donné que n'importe quelle densité continue peut être approchée avec une précision arbitraire par un mélange fini de densités gaussiennes, le modèle de mélange fournit un cadre de travail convenable pour l'estimation de densités. Par exemple, Priebe (1994, [82]) prouve qu'avec $n = 10.000$ observations, une densité log normale peut être approchée par un mélange de 30 normales. Un estimateur par noyaux utilise un mélange de 10.000 normales.

Le mélange de densités a fourni une approche mathématique à la modélisation statistique d'une large variété de phénomènes aléatoires. Grâce à leur flexibilité, les modèles de mélange ont continué à recevoir une attention croissante au fil des ans, aussi bien d'un point de vue pratique que théorique. Dans les dix années passées, les applications des modèles de mélange se sont considérablement développées dans des domaines tels que l'astronomie, la biologie, la génétique, la médecine, l'ingénierie, la psychiatrie, l'économie et le marketing, etc. Dans ces applications, les modèles de mélange servent d'outils à des méthodes incluant la classification et l'analyse de classes latentes, l'analyse discriminante, l'analyse d'images et l'analyse de survie en plus de fournir des modèles descriptifs de distributions et de leur rôle en analyse de données.

Ce n'est qu'au cours des 20 dernières années que des avancées considérables ont été faites dans l'ajustement des modèles de mélanges en particulier par la méthode du maximum de vraisemblance. Même avec l'avènement des gros calculateurs, de nombreuses recherches tentent d'étudier l'ajustement de modèles de mélanges pour des données de plus d'une dimension. Ces efforts sont dus au manque de compréhension des résultats. En effet, les sorties incluent la présence de multiples maxima dans la fonction de vraisemblance et cette fonction n'est pas bornée dans le cas par exemple d'un mélange gaussien à matrices de variance-covariance inégales.

Deux approches foncièrement différentes du problème d'ajustement de modèle de mélange ont fait l'objet d'études. La plus ancienne et la plus répandue (celle de Pearson, 1894, [81]), consiste à y voir un problème d'estimation de paramètres. Nous la désignerons l'"approche estimation". L'autre approche considère qu'il s'agit d'un problème de classification.

2.2 Approche estimation

Il s'agit de l'approche directe qui vise à estimer les paramètres $(a_i, p_i)_{i=1,...,k}$. Elle a donné lieu à beaucoup de publications depuis celle de Pearson (1894, [81]). A tous les points de vue (simplicité, fiabilité, précision, etc.), les méthodes les plus probantes relèvent des techniques d'estimation par le maximum de vraisemblance. Ces méthodes consistent à résoudre itérativement les équations de vraisemblance pour un échantillon $(x_1, ..., x_N)$ donné, le logarithme de la vraisemblance étant:

$$L(x_1, ..., x_N, a_1, ..., a_k, p_1, ..., p_k) = \sum_{j=1}^{N} log \left(\sum_{i=1}^{k} p_i f_i(x_j, a_i) \right). \tag{2.6}$$

Pour d'avantage de clarté, nous noterons $L(\phi)$ le terme de gauche de la formule (2.6), avec $\phi = (a_1, ..., a_k, p_1, ..., p_k)$. Lorsque nous n'avons pas les dérivées de la log-vraisemblance par rapport aux paramètres, les algorithmes les plus efficaces pour résoudre les équations de vraisemblance sont, à des variantes près, les algorithmes de type EM (Estimation, Maximisation) (Dempster, Laird, Rubin, 1977, [30]), (Redner, Walker, 1984, [86]), (Shlezinger, 1968, [97]).

Nous présentons tout d'abord la méthode EM, puis ses variantes.

2.2.1 L'algorithme EM

L'algorithme EM est un algorithme qui estime les paramètres du mélange en calculant le maximum de vraisemblance de données incomplètes. Dans le cadre du modèle de mélange par EM, les données sont appelées incomplètes car nous ne connaissons pas les composantes auxquelles appartiennent les observations. Le terme de "données incomplètes" implique dans sa forme générale l'existence de deux espaces d'échantillonage $\mathcal{X}$ et $\mathcal{Y}$ et d'une application de $\mathcal{Y}$ dans $\mathcal{X}$. Les données observées $\boldsymbol{x} = (x_1, ..., x_N)$ sont des réalisations de $\mathcal{X}$. Les données $\boldsymbol{y} = (y_1, ..., y_N)$ associés de $\mathcal{Y}$ ne sont pas observées directement, mais indirectement à travers $\boldsymbol{x}$. Plus spécifiquement nous supposons qu'il y a une application $\boldsymbol{y} \rightarrow \boldsymbol{x}(\boldsymbol{y})$ de $\mathcal{Y}$ dans $\mathcal{X}$ et que $\boldsymbol{y}$ est dans $\mathcal{Y}(\boldsymbol{x})$, le sous-ensemble de $\mathcal{Y}$ déterminé par l'équation $\boldsymbol{y} = \boldsymbol{x}(\boldsymbol{y})$ pour $\boldsymbol{x}$ les

données observées. Notons $\boldsymbol{y}$ les données complètes.

L'approche traditionnelle pour déterminer l'estimateur du maximum de vraisemblance est d'arriver à un système d'équations de vraisemblance satisfaites par l'estimateur du maximum de vraisemblance. Les équations sont déterminées en considérant les dérivées partielles de la fonction de log-vraisemblance par rapport aux composantes de ϕ. Soit $\phi : (\phi_1, ..., \phi_k, \pi_1, ..., \pi_k)$ un estimateur du maximum de vraisemblance, les équations sont données par

$$\nabla_{\phi_i} L(\phi) = 0, \quad i = 1, ..., k, \tag{2.7}$$

déterminées par les paramètres non contraints $\phi_i,\ i = 1, ..., k$. Le gradient des dérivées partielles premières par rapport aux composantes d'une variable est noté ∇.

Pour obtenir les équations de vraisemblance pour les proportions contraintes positives et de somme 1 nous imitons Redner et Walker (1984, [86]). Soit $\pi = (\pi_1, ..., \pi_k)^T$, on voit

$$0 \geq \nabla_\pi L(\phi)^T (\pi' - \pi) \tag{2.8}$$

pour tout $\pi' = (\pi'_1, ..., \pi'_k)^T$ tel que $\sum_{i=1}^k \pi'_i = 1$ et $\pi'_i \geq 0,\ i = 1, ..., k$. La formule (2.8) est vraie pour tout π' satisfaisant les contraintes si et seulement si

$$0 \geq \nabla_\pi L(\phi)^T (e_i - \pi_i), \quad i = 1, ..., k,$$

avec égalité pour les valeurs de i pour lesquelles $\pi_i > 0$. (e_i est le vecteur $(e_i^1, ..., e_i^k)$ avec $e_i^j = 1$ si $i = j$ et 0 sinon) Donc (2.8) est équivalent à

$$\begin{aligned}
0 &\geq \sum_{m=1}^{k} \frac{\partial}{\partial \pi_m} L(\phi)(e_i^m - \pi_m) \\
&\geq \frac{\partial}{\partial \pi_i} L(\phi) - \pi_i \frac{\partial}{\partial \pi_i} L(\phi) + \sum_{m=1}^{k} -\pi_m \frac{\partial}{\partial \pi_m} L(\phi) - (-\pi_i \frac{\partial}{\partial \pi_i} L(\phi)) \\
&\geq \frac{\partial}{\partial \pi_i} L(\phi) + \sum_{m=1}^{k} \sum_{j=1}^{N} \frac{-\pi_m f_m(x_j|\phi_m)}{f(x_j|\phi)} \\
&\geq \frac{\partial}{\partial \pi_i} L(\phi) - \sum_{j=1}^{N} \frac{f(x_j|\phi)}{f(x_j|\phi)} \\
&\geq \frac{\partial}{\partial \pi_i} L(\phi) - N
\end{aligned}$$

$$\iff \quad 1 \geq \frac{1}{N} \sum_{j=1}^{N} \frac{f_i(x_j|\phi_i)}{f(x_j|\phi)} \tag{2.9}$$

avec égalité pour les valeurs de i telles que $\pi_i > 0$. Finalement en multipliant de chaque côté de (2.9) par π_i pour $i = 1, ..., k$

$$\pi_i = \frac{1}{N} \sum_{j=1}^{N} \frac{\pi_i f_i(x_j|\phi_i)}{f(x_j|\phi)}, \quad i = 1, ..., k. \tag{2.10}$$

La i^{eme} proportion π_i du mélange peut être vue comme la probabilité a priori qu'un individu appartienne à la i^{eme} composante du mélange $(i = 1, ..., k)$.

Pour faciliter l'écriture du résultat général qui suit et qui n'est pas restreint au problème du mélange de densités, nous pouvons écrire les équations de vraisemblance (2.7) et (2.10) sous la forme générale sans contraintes:

$$\nabla_\phi L(\phi) = 0, \tag{2.11}$$

Le théorème suivant établit que, sous des conditions raisonnables, il existe une unique solution *fortement* consistante des équations de vraisemblance (2.11), que cette solution maximise, au moins localement, la fonction de log-vraisemblance et est asymptotiquement normalement distribuée. *Consistant* au sens usuel signifie convergeant avec une probabilité approchant 1 vers le vrai paramètre quand la taille de l'échantillon tend vers l'infini ; *fortement consistant* signifie ayant la même limite avec la probabilité 1. Ce théorème (détaillé par Redner et Walker dans [86]) est un condensé de résultats généralisant les travaux de Cramér [22] concernant un estimateur scalaire du maximum de vraisemblance. Les conditions suivantes, sur lesquelles le théorème est basé, fûrent données par Chanda dans [17].

Pour d'avantage de pratique, nous notons provisoirement $\phi = (\xi_1, ..., \xi_v)$, avec $v = k - 1 + \sum_{i=1}^{k} n_i$ et n_i la dimension du paramètre ϕ_i. Nous posons Θ, l'ensemble des paramètres ϕ et ϕ^* le vrai paramètre. Puisque ces conditions sont locales par nature, nous pouvons restreindre Θ à un voisinage de ϕ^*

Condition 1 (Chanda, [17]) *:*
Pour tout $\phi \in \Theta$, pour tout $x \in \mathbb{R}^n$ et pour $i, j, k = 1, ..., v$, les dérivées partielles $\partial f/\partial\xi_i$, $\partial^2 f/\partial\xi_i\partial\xi_j$ et $\partial^3 f/\partial\xi_i\partial\xi_j\partial\xi_k$ existent et satisfont

$$|\frac{\partial f(x|\phi)}{\partial\xi_i}| \leq r_i(x),\ |\frac{\partial^2 f(x|\phi)}{\partial\xi_i\partial\xi_j}| \leq r_{ij}(x),\ |\frac{\partial^3 log f(x|\phi)}{\partial\xi_i\partial\xi_j\partial\xi_k}| \leq r_{ijk}(x),$$

avec r_i et r_{ij} intégrables et r_{ijk} satisfaisant

$$\int_{\mathbb{R}^n} r_{ijk}(x) f(x|\phi^*) d\mu < \infty.$$

Condition 2 (Chanda, [17]) *:*
La matrice de Fisher $I(\phi)$ donnée par

$$I(\phi) = \int_{\mathbb{R}^n} [\nabla_\phi log f(x|\phi)][\nabla_\phi log f(x|\phi)]^T f(x|\phi) d\mu,$$

est définie positive en ϕ^*.

Théorème 1 ([86]) *:*
Si les conditions 1 et 2 sont satisfaites et qu'un voisinage suffisamment petit de ϕ^* est donné alors avec probabilité 1, avec un N suffisamment grand, il existe une unique solution ϕ^N des équations (2.11) dans ce voisinage et cette solution maximise localement la fonction de log-vraisemblance. De plus, $\sqrt{N}(\phi^N - \phi^*)$ est asymptotiquement distribué avec une moyenne nulle et une matrice de covariance $I(\phi^*)^{-1}$.

Après ce résultat général sur les estimateur par maximum de vraisemblance, revenons à EM. Désignons par $f(\boldsymbol{y}|\phi) = \prod_{j=1}^{N} f(x_j|\phi)$, la vraisemblance des données $\boldsymbol{y}$, dépendant de ϕ et par $g(\boldsymbol{x}|\phi)$, la vraisemblance associée à $\boldsymbol{x}$. La spécification des données complètes $f(.|.)$ est liée à la spécification des données incomplètes $g(.|.)$ par

$$g(\boldsymbol{x}|\phi) = \int_{\mathcal{Y}(x)} f(\boldsymbol{y}|\phi) d\boldsymbol{y}. \tag{2.12}$$

L'algorithme EM donne la valeur de ϕ qui maximise $g(\boldsymbol{x}|\phi)$ pour $\boldsymbol{x}$ donné. Les itérations de EM impliquent deux étapes appelées étape d'estimation (E) et étape de maximisation (M).

Nous introduisons la densité conditionnelle de l'espérance de la vraisemblance de $\boldsymbol{y}$ sachant $\boldsymbol{x}$ et ϕ

$$k(\boldsymbol{y}|\boldsymbol{x}, \phi) = \frac{f(\boldsymbol{y}|\phi)}{g(\boldsymbol{x}|\phi)} \tag{2.13}$$

et la fonction

$$Q(\phi|\phi') = E(log f(\boldsymbol{y}|\phi)|\boldsymbol{x}, \phi') \tag{2.14}$$

qui est supposée exister pour tout (ϕ, ϕ') et nous supposons que $f(\boldsymbol{y}|\phi) > 0$ presque partout dans $\mathcal{Y}$. Supposons que $\phi^{(p)}$ est la valeur courante de ϕ après p itérations de l'algorithme. L'itération de EM qui permet de passer de $\phi^{(p)}$ à $\phi^{(p+1)}$ est

- Etape E: Calculer $Q(\phi|\phi^{(p)})$
- Etape M: Choisir $\phi^{(p+1)}$ une valeur de ϕ qui maximise $Q(\phi|\phi^{(p)})$.

L'idée heuristique est que nous voulons choisir ϕ^* qui maximise $log f(\boldsymbol{y}|\phi)$. Puisque nous ne connaissons pas $log f(\boldsymbol{y}|\phi)$ nous maximisons à la place son espérance courante pour $\boldsymbol{x}$ donné et le $\phi^{(p)}$ courant.
Nous notons

$$H(\phi|\phi') = E(log k(\boldsymbol{y}|\boldsymbol{x}, \phi)|\boldsymbol{x}, \phi'). \tag{2.15}$$

et

$$L(\phi) = log g(\boldsymbol{x}|\phi) \tag{2.16}$$

Nous pouvons dire à partir de (2.16), (2.13) et (2.14) que

$$Q(\phi|\phi') = L(\phi) + H(\phi, \phi'). \tag{2.17}$$

L'algorithme EM utilise une structure de données incomplètes et introduit pour cela le vecteur d'étiquettes de composantes $\boldsymbol{Z}_j$ de variables indicatrices 0-1 pour définir la composante dans le mélange (2.1) à partir de laquelle le vecteur aléatoire $\boldsymbol{X}_j$ est sensé venir. Nous disposons de données $\boldsymbol{x}_1, ..., \boldsymbol{x}_N$, N réalisations de N vecteurs aléatoires indépendants et identiquement distribués (iid) $\boldsymbol{X}_1, ..., \boldsymbol{X}_N$ de densité $f(x_j|\phi)$. Nous écrivons

$$X_1, ..., X_N \sim^{iid} F, \tag{2.18}$$

avec $F(x_j)$ la fonction de répartition associée au mélange de densités $f(x_j)$.
Dans le cadre de EM, les données $\boldsymbol{x}_1, ..., \boldsymbol{x}_N$ sont vues comme des données incomplètes puisque leurs vecteurs d'étiquettes de composantes $\boldsymbol{z}_1, ..., \boldsymbol{z}_N$ ne sont pas connus. Le vecteur données complètes est donc déclaré comme

$$\boldsymbol{y}_c = (\boldsymbol{x}^T, \boldsymbol{z}^T)^T, \tag{2.19}$$

avec

$$\boldsymbol{x} = (\boldsymbol{x}_1^T, ..., \boldsymbol{x}_N^T)^T$$

le vecteur des données observées et

$$\boldsymbol{z} = (\boldsymbol{z}_1^T, ..., \boldsymbol{z}_N^T)^T$$

le vecteur d'étiquettes de composantes. Les vecteurs $\boldsymbol{z}_1, ..., \boldsymbol{z}_N$ sont supposés être des réalisations des vecteurs aléatoires $\boldsymbol{Z}_1, ..., \boldsymbol{Z}_N$ où pour des données indépendantes il est approprié de supposer qu'elles sont identiquement distribuées

$$\boldsymbol{Z}_1, ..., \boldsymbol{Z}_N \sim^{i.i.d} Mult_k(1, \pi).$$

La probabilité a posteriori qu'un individu (avec $\boldsymbol{x}_j$ pour valeur) appartienne à la i^{eme} composante est

$$\begin{aligned} t_i(x_j) &= \mathbb{P}\{individu \ \in i^{eme} \ composante | x_j\} \\ &= \mathbb{P}\{Z_{ij} = 1 \mid \boldsymbol{x}_j\} \\ &= \pi_i f_i(x_j|\phi_i)/f(x_j|\phi) \ \ (i = 1, ..., k) \ et \ (j = 1, ..., N). \end{aligned}$$

Pour simplifier ces notations très générales, nous supposons que nous disposons d'un échantillon de données incomplètes en considérant chaque x_j comme la partie connue d'une observation $y_j = (x_j, i_j)$, avec i_j un entier entre 1 et k indiquant la composante d'origine de la population. Nous avons pu déterminé par la formule (2.10) les proportions optimales du mélange au sens du maximum de vraisemblance. Regardons la forme des équations à résoudre pour déterminer les paramètres $(\phi_i)_{i=1,...,k}$ des densités du mélange.

Pour $\phi = (\pi_1, ..., \pi_k, \phi_1, ..., \phi_k)$, les échantillons $\boldsymbol{x} = (x_1, ..., x_N)$ et $\boldsymbol{y} = (y_1, ..., y_N)$ ont des densités associées $g(\boldsymbol{x}|\phi) = \prod_{j=1}^N f(x_j|\phi)$ et $f(\boldsymbol{y}|\phi) = \prod_{j=1}^N \pi_{i_j} f_{i_j}(x_j|\phi_{i_j})$. Pour $\phi' = (\pi'_1, ..., \pi'_k, \phi'_1, ..., \phi'_k)$ la densité conditionnelle $k(\boldsymbol{y}|\boldsymbol{x}, \phi')$ est donnée par

$$k(\boldsymbol{y}|\boldsymbol{x}, \phi') = \prod_{j=1}^N \frac{\pi'_{i_j} f_{i_j}(x_j|\phi'_{i_j})}{f(x_j|\phi')},$$

et la fonction $Q(\phi|\phi')$ est déterminée par

$$\begin{aligned} Q(\phi|\phi') &= \sum_{i_1=1}^k ... \sum_{i_N=1}^k \sum_{j=1}^N log(\pi_{i_j} f_{i_j}(x_j|\phi_{i_j})) \prod_{j=1}^N \frac{\pi'_{i_j} f_{i_j}(x_j|\phi'_{i_j})}{f(x_j|\phi')} \\ &= \sum_{i=1}^k \sum_{j=1}^N log(\pi_i f_i(x_j|\phi_i)) \frac{\pi'_i f_i(x_j|\phi'_i)}{f(x_j|\phi')} \\ &= \sum_{i=1}^k \left[\sum_{j=1}^N \frac{\pi'_i f_i(x_j|\phi'_i)}{f(x_j|\phi')} \right] log\pi_i \ + \ \sum_{i=1}^k \sum_{j=1}^N log f_i(x_j|\phi_i) \frac{\pi'_i f_i(x_j|\phi'_i)}{f(x_j|\phi')}. \end{aligned} \tag{2.20}$$

La maximisation de $Q(\phi|\phi')$ sur $(\phi_i)_{i=1,...,k}$ est donc la maximisation du deuxième terme de (2.20).

Pour les mélanges de densités, à partir d'une solution initiale $((\pi_i^0, \phi_i^0)_{i=1,...,k})$, l'algorithme EM est le suivant:
Itération n: $(n \geq 1)$

Etape E (estimation)

Pour $i = 1, ..., k$; $j = 1, ..., N$, calcul des

$$t_i^n(x_j) = p_i^n f_i(x_j, \phi_i^n) / \sum_{i=1}^{k} p_i^n f_i(x_j, \phi_i^n). \tag{2.21}$$

Les t_i^n sont les probabilités a posteriori d'appartenance de x_j à la composante i à la n^{eme} itération.

Etape M (maximisation)

Pour $i = 1, ..., k$, calcul de

$$p_i^{n+1} = \frac{1}{N} \sum_{j=1}^{N} t_i^n(x_j)$$

(estimateur du maximum de vraisemblance des proportions du mélange) et résolution des équations de vraisemblance pour les paramètres $(\phi_i = \phi_{mi}, m = 1, ..., s) \in \mathbb{R}^s$:

$$\forall i = 1, ..., k,\ m = 1, ..., s: \quad \sum_{j=1}^{N} t_i^n(x_j) \frac{\partial Log(f_i(x_j, \phi_i^{n+1})}{\partial \phi_{mi}} = 0.$$

L'algorithme EM fonctionne pour un grand nombre de composantes et dans le cas multidimensionnel; EM fournit en général de bons résultats si le nombre de composantes est connu. De plus, sous des conditions assez larges, les estimations des paramètres convergent presque sûrement vers les vraies valeurs lorsque N tend vers l'infini (Redner, Walker, 1984, [86]). Malheureusement, EM peut converger lentement et cette lenteur peut être rédhibitoire dans le cas où la solution initiale est éloignée de la solution limite accessible.

2.3 Les variantes de EM

Les algorithmes de reconnaissance de mélanges évoqués présentent les limitations suivantes:

- le nombre k de composantes est supposé connu,
- la solution obtenue dépend de la position initiale de l'algorithme.

2.3.1 SEM

L'extension de EM que nous présentons ici répond en grande partie à ces limitations. Il s'agit d'un algorithme EM auquel est ajouté une étape d'apprentissage probabiliste, d'où son nom, l'algorithme SEM: Stochastique,Estimation, Maximisation (Celeux, Diebolt, 1986, [12]).

Initialisation:

On fixe le paramètre k majorant supposé du nombre de composantes et un seuil $c(N,p)$ compris entre 0 et 1.

En chaque point $(x_j, j = 1, ..., N)$ on choisit (en général au hasard) les probabilités initiales d'appartenance à l'une des composantes:

$$t_i^0(x_j),\ i = 1, ..., k \text{ avec } 0 < t_i^0(x_j) < 1 \text{ et } \sum_{i=1}^{k} t_i^0(x_j) = 1.$$

Itération $n(n \geq 1)$**:**

Etape S (stochastique)

On tire en chaque point x_j la variable aléatoire multinomiale $e^n(x_j) = (e_i^n(x_j), i = 1, ..., k)$ d'ordre un et de paramètres $(t_i^n(x_j), i = 1, ..., k)$. Les réalisations $e^n(x_j)$ définissent une partition $P^n = (P_1^n, ..., P_k^n)$ de l'échantillon avec $P_i^n = \{x_j,\ tels\ que\ e_i^n(x_j) = 1\}$. Si pour un i, $card(P_i^n)$ est plus petit que $Nc(N,p)$ l'algorithme est ré-initialisé.

Etape M (maximisation)

Calcul des estimations du maximum de vraisemblance des paramètres

$$(q_i^{n+1} = (p_i^{n+1}, a_i^{n+1}),\ i = 1, ..., k)$$

du mélange sur la base des sous-échantillon $(P_i^n,\ i = 1, ..., k)$.

$$\text{On a: } p_i^{n+1} = \frac{1}{N} \sum_{j=1}^{N} e_i^n(x_j).$$

L'estimation des a_i^{n+1} dépend de la famille paramétrique, posée a priori, des composantes du mélange.

Etape E (estimation)

A partir des $(q_i^{n+1} = (p_i^{n+1}, a_i^{n+1}),\ i = 1, ..., k)$ on calcule pour $i = 1, ..., k$ et $j = 1, ..., N$

$$t_i^{n+1}(x_j) = p_i^{n+1} f_i(x_j, a_i^{n+1}) / \sum_{i=1}^{k} p_i^{n+1} f_i(x_j, a_i^{n+1}).$$

A chaque itération, l'algorithme construit une partition en classes dont les enveloppes convexes peuvent se couper. A la stabilité de l'algorithme, on obtient une famille de partitions statistiquement admissibles pour les estimations des paramètres du mélange. Ces estimations sont précises et asymptotiquement sans biais (Celeux, Diebolt, 1986, [12]). Le type de convergence

obtenue est une convergence en loi correspondant à la stationnarité de la suite des estimés $q^n = (p^n, a^n)$. Les perturbations introduites à chaque itération par les tirages aléatoires empêchent la convergence vers un maximum local instable de la vraisemblance comme cela peut être le cas pour EM.

L'algorithme SEM permet de reconnaître les composantes d'un mélange même si elles sont imbriquées et fournit le nombre exact de composantes pourvu qu'il soit initialisé avec un majorant de ce nombre. L'algorithme SEM converge plus rapidement que EM quelle que soit la configuration initiale, les tirages aléatoires l'empêchent de stationner loin de la solution limite. Pour de petits échantillons, il risque de sous-estimer le nombre de composantes, les aléas introduits prenant trop d'importance.

2.3.2 SAEM

L'algorithme SAEM (Simulated Annealing EM ou Stochastic Approximation EM) (Celeux et Diebolt, 1992, [13]) est une modification de l'algorithme SEM telle que la convergence en loi peut être remplacée par la convergence presque sûre. Le comportement erratique possible de SEM pour de petits ensembles de données peut être atténué sans sacrifier la nature stochastique de l'algorithme. Ceci est réalisé en utilisant une suite de réels positifs (γ^n) décroissante vers 0 (avec $\gamma^0 = 1$). Plus précisément, si ϕ^n est le paramètre courant estimé par SAEM, l'approximation ϕ^{n+1} de ϕ est

$$\phi^{n+1} = (1 - \gamma^{n+1})\phi_{EM}^{n+1} + \gamma^{n+1}\phi_{SEM}^{n+1}, \tag{2.22}$$

avec ϕ_{EM}^{n+1} l'approximation de ϕ par EM et ϕ_{SEM}^{n+1} l'approximation de ϕ par SEM.

SAEM progresse d'un pur SEM au début vers un pur EM à la fin. Le choix du taux de convergence vers 0 de γ^n est important. Un faible taux de convergence est nécessaire pour de bons résultats. Il est important que γ^n reste près de $\gamma^0 = 1$ durant les premières itérations pour éviter les valeurs stationnaires de $L(\phi)$.

2.3.3 MCEM

Les implémentations Bayésiennes de la décomposition de mélange de lois sont nombreuses et reposent toutes sur le calcul de distributions à posteriori par augmentation de données (Data augmentation) proposé dans [106] par M.A. Tanner et W. Hung Wong en 1987. La méthode de référence est l'algorithme MCEM (Wei et Tanner, 1990, [114]). Wei et Tanner proposent une méthode de Monte-Carlo pour l'étape E. En remplaçant le calcul de $Q(\phi|\phi^n)$ par celui d'une version empirique $Q_{n+1}(\phi|\phi^n)$ basée sur m (m$\gg$1) réalisations de $\boldsymbol{y}$ à partir de $k(\boldsymbol{y}|\boldsymbol{x}, \phi^n)$.

Plus précisément, la n^{eme} étape est:
(a) Générer un échantillon indépendant et identiquement distribué $(\boldsymbol{z}_{(1)}^n, ..., \boldsymbol{z}_{(m)}^n)$ à partir de $k(\boldsymbol{y}|\boldsymbol{x}, \phi^n)$ et
(b) calculer l'approximation courante de $Q(\phi|\phi^n)$ par

$$Q_{n+1}(\phi|\phi^n) = \frac{1}{m}\sum_{j=1}^{m} log(f(\boldsymbol{x}, \boldsymbol{z}_{(j)}^n|\phi). \tag{2.23}$$

(c) L'étape M est donnée par $\phi^{n+1} = argmax_{\phi} Q_{n+1}(\phi|\phi^n)$.

Si $m = 1$, MCEM est réduit à SEM. Si m est grand, MCEM marche approximativement comme EM et il a les mêmes résultats que EM.

Wei et Tanner ont motivé l'introduction de l'algorithme MCEM comme une alternative remplaçant le calcul analytique de l'intégrale dans (2.12) par un calcul numérique d'une approximation de Monte-Carlo. Wei et Tanner recommande de démarrer avec une petite valeur de m et de l'augmenter quand ϕ^n est près du du maximiseur de $L(\phi)$. Plus précisément, si nous sélectionnons une suite (m^n) d'entiers tels que $m^0 = 1$ et m^n croit vers l'infini quand n tend vers l'infini, nous allons d'un pur SEM ($m^0 = 1$) vers un pur EM ($m = \infty$) quand n tend vers l'infini.

Les variantes de EM sont donc nombreuses et sont encore à l'étude. En effet, la plupart des résultats théoriques obtenus (e.g. de convergence) concernent la famille de lois exponentielles (lois gaussiennes, lois exponentielles).

2.4 Approche classification

Dans cette approche, les vecteurs d'étiquettes, $\boldsymbol{z}_i = (z_{i,k}, k = 1, ..., K)$ avec $z_{ik} = 1$ ou 0 selon que x_i viennent de la composante k ou d'une autre, sont considérés comme étant des paramètres inconnus.

L'approche classification (Diday, Ok, Schroeder, 1974, [38], Scott et Symons, 1971, [96], Symons, 1981, [105]) consiste à rechercher une partition $P = (P_1, ..., P_K)$ telle que chaque classe P_l soit assimilable à un sous-échantillon suivant la loi de densité $f(., a_i)$, le nombre K de composantes du mélange étant supposé connu. Dans ce cadre, la plupart des algorithmes utilisés sont de type "nuées dynamiques". Ainsi nous présentons tout d'abord un algorithme détaillé dans [38] et [92], valide pour toute famille paramétrique de densités dont les paramètres admettent des estimateurs du type maximum de vraisemblance.

2.4.1 Décomposition par "nuées dynamiques"

L'espace de représentation $\boldsymbol{L}$ d'une classe étant l'espace de définition des paramètres a dont dépendent les densités $f(., a)$, la méthode vise à maximiser le critère de *vraisemblance classifiante* ou de *log-vraisemblance classifiante*, qui mesure l'adéquation d'une partition et de sa représentation. Notons $A = (a_1, ..., a_K)$ un élément de $\boldsymbol{L}_K = \boldsymbol{L} \times ... \times \boldsymbol{L}$ (K fois). Dans notre cas, a_i est le paramètre d'une densité $f_i(., a_i)$ associée à la classe P_i. La log-vraisemblance classifiante est:

$$lvc_r(A, P) = \sum_{i=1}^{K} Log(v(P_i, a_i)) \quad A \in \boldsymbol{L}_K, P \in \boldsymbol{P}_K \tag{2.24}$$

où $\boldsymbol{P}_K$ est l'espace des partitions possibles en K classes de l'échantillon et

$$v(P_i, a_i) = \prod_{x_j \in P_i} f_i(x_j, a_i) \tag{2.25}$$

est la vraisemblance du sous-échantillon P_i pour la loi de densité $f_i(x, a_i)$. Nous avons donc

$$lvc_r(A, P) = \sum_{i=1}^{K} \sum_{x_j \in P_i} Log(f_i(x_j, a_i)).$$

La log-vraisemblance classifiante (2.24) est la version restrictive de ce critère. Cette version s'applique quand l'échantillon $\{x_1, ..., x_N\}$ des données observées est supposé formé en prenant n_i observations de la i^{eme} composante, avec n_i fixé avant échantillonage. Les proportions p_i sont implicitement supposées égales. Nous parlons de schéma d'échantillonage séparé.

La log-vraisemblance non-restrictive est

$$lvc(A, P) = \sum_{i=1}^{K} \sum_{x_j \in P_i} Log(p_i f_i(x_j, a_i)), \tag{2.26}$$

elle peut s'écrire

$$lvc = lvc_r + \sum_{i=1}^{K} card(P_i) Log(p_i). \tag{2.27}$$

Cette version s'applique quand $\{x_1, ..., x_N\}$ vient du mélange (2.1). Nous parlons de schéma d'échantillonage par mélange.

En ayant donné le critère, la méthode de décomposition de mélange par nuées dynamiques se déroule de la manière suivante:

A partir d'une partition en K classes de l'échantillon, on estime les paramètres associés à chaque classe par maximisation de la vraisemblance. On obtient ainsi une représentation pour chaque classe: c'est la fonction g. A l'aide des fonctions de densité induites par cette représentation, on affecte les individus à la classe de plus grande probabilité: c'est la fonction h. On recommence le procédé jusqu'à convergence. Regardons plus précisément ce que sont les fonctions g et h.

$g : \boldsymbol{P}_K \longrightarrow \boldsymbol{L}_K$, fonction de représentation est définie par:

$$g(P) = g(P_1, ..., P_K) = (a_1, ..., a_K)$$

où a_i est l'estimation du maximum de vraisemblance du paramètre de la densité associée au sous-échantillon P_i.

Par exemple, dans le cas d'un mélange de lois de densités gaussiennes, le paramètre de chaque composante est constitué de sa moyenne et de sa matrice de covariance. L'estimation du maximum de vraisemblance des paramètres s'écrit:

$$\forall i = 1, ..., K \quad a_i = (m_i, \Gamma_i) \text{ avec } m_i = \frac{1}{card P_i} \sum_{x_j \in P_i} x_j$$

et

$$\Gamma_i = \frac{1}{card P_i} \sum_{x_j \in P_i} (x_j - m_i)^T (x_j - m_i).$$

$h : \boldsymbol{L}_K \longrightarrow \boldsymbol{P}_K$, la fonction d'affectation est définie par:

$$h(A) = h(a_1, ...a_K) = (P_1, ..., P_K)$$

où

$$\forall i = 1, ..., K \; P_i = \{x \in \Omega / f_i(x, a_i) \geq f_m(x, a_m), \; \forall m \neq i, \quad \text{avec } i < m \text{ en cas d'égalité}\}.$$

On a le résultat suivant (Diday, Schroeder, 1976, [39]):

Théorème 2 (Diday, Schroeder, [39]) *:*
Sous l'hypothèse que la famille de densités $f(x, a)$ soit bornée supérieurement pour tout $x \in \mathbb{R}^p$ et pour tout $a \in \boldsymbol{L}$, la suite $v_n = (A^n, P^n)$ converge dans $\boldsymbol{L}_K \times \boldsymbol{P}_K$ en un nombre fini d'itérations et atteint sa limite; la suite réelle $u_n = lvc_r(v_n)$ converge en croissant vers un maximum local.

A la convergence, on obtient une partition $P = (P_1, ..., P_K)$ et une estimation des paramètres $(a_1, ..., a_K)$. Les paramètres $(p_1, ..., p_K)$ sont estimés par les quantités $(card P_1/N, ..., card P_K/N)$.

Une autre approche par classification existe: la méthode "CEM" (Classification EM).

2.4.2 CEM

La méthode CEM de Celeux et Govaert ([14]) est une approche intéressante, rajoutant à EM une étape de classification. A partir d'une solution initiale $(p_1^0, ..., p_K^0, a_1^0, ..., a_k^0)$, l'algorithme est le suivant:
Itération $n \geq 1$

- Etape E (Estimation). Calcul des probabilités conditionnelles courantes $t_i^n(x_j)$ $(1 \leq j \leq N, \, 1 \leq i \leq K)$ selon (2.21),
- Etape C (Classification). Affectation de chaque observation x_j à la classe P_i^n qui donne la probabilité conditionnelle courante $t_i^n(x_j)$ maximale,
- Etape M (Maximisation). Calcul des estimateurs (p_i^n, a_i^n) des paramètres par maximum de vraisemblance sur la classe P_i^n comme sous-ensemble.

La méthode des nuées dynamiques comme la méthode CEM sont plus rapides qu'EM. Cependant, elles produisent des estimations biaisées des paramètres si les classes ne sont pas bien séparées. Pour d'avantage de détails sur la méthode CEM, voir [14] et [15].

2.5 Conclusion

Nous avons détaillé différentes méthodes permettant de décomposer une densité comme une somme de densités paramétriques pondérées. Nous pouvons remarquer que les deux approches existantes (estimation et classification) n'ont pas nécessairement les mêmes applications du fait de leurs formulations respectives. L'approche "estimation" permet (comme son nom l'indique) d'obtenir une estimation de la densité, suivant le modèle de mélange (dont on peut se servir, après convergence, pour faire une classification), alors que l'approche "classification" se sert de cette estimation , étape par étape, pour trouver une partition des observations. Quoi qu'il en soit, les deux approches reposent sur la notion de densité des observations. Dans ce travail, nous disposons de données fonctions de répartition et la question qu'on peut se poser est: Quelle densité modéliser pour des observations de ce type, afin détendre les méthodes de décomposition de mélange de densités?

Dans le chapitre suivant, nous proposons une modélisation de lois de probabilité pour ce type de données probabilistes. Les extensions des méthodes de décomposition vues dans ce chapitre, sont basées sur cette modélisation.

Chapitre 3

Distribution de distributions

"Distributions are the number of the future."
B. Schweizer (1984)

Depuis la fin des années 60 s'est développé un intérêt pour la description de lois de probabilité d'éléments aléatoires qui sont eux-mêmes des lois de probabilité sur des ensembles finis ou infinis. La plupart de ces lois ont été motivées par l'inférence Bayésienne et la théorie de la décision, dans lesquelles "l'état inconnu" prend la forme de lois de probabilité (voir Good, 1967, [57] et Ericson, 1969, [42]). Un bon exemple de recherches dans ce domaine est l'article de Ferguson (1974, [44]), traitant de lois sur des ensembles assez généraux, ou l'article de Kingman (1975, [69]) sur les lois discrètes aléatoires.

Dans cette thèse, nous tentons d'étendre les méthodes de décomposition de mélange de densités au cas des données probabilistes de type fonctions de distribution (nous appelons parfois "fonctions de distribution" les fonctions de répartition pour eviter les répétitions). Pour cela, nous tentons donc également de définir une notion de lois de lois de probabilité et nous développons la notion de "Fonction de Distribution de Distributions" (FDD) initialement introduite de manière empirique par E. Diday (2001, [33]), sur un ensemble de fonctions de répartition. Dans cet article, E. Diday utilise un ensemble de fonctions de répartition $\mathfrak{F} = \{F_1, ..., F_N\}$ appelé "base de distributions" (avec $F_i(x) = Pr(X_i \leq x)$), pour définir un "point de distribution de distributions" associé à une valeur T par :
$G_T(x) = card(\{F_i \in \mathfrak{F} \ / \ F_i(T) \leq x\})/card(\mathfrak{F})$.
Nous développons cette définition dans un cadre probabiliste général (pouvant inclure la modélisation empirique de [33]).

3.1 Définition des fonctions de distribution de distributions

Soit Ω une population d'individus w décrits par p variables continues dont le domaine est inclu dans $\mathbb{R}$. Soient V_j l'ensemble des valeurs possibles pour la variable j et $V = V_1 \times ... \times V_p$. L'ensemble V_j étant un sous-ensemble de $\mathbb{R}$, nous notons ν_j la σ-algèbre des boréliens sur V_j et $\nu = \nu_1 \times ... \times \nu_p$ la σ-algèbre produit sur V. Soit l'ensemble $\Omega_F = \Omega_F^1 \times ... \times \Omega_F^p$ avec

$$\Omega_F^j = \{F : \ F \text{ est une fonction de répartition unidimensionnelle sur } (V_j, \nu_j)\}.$$

Nous définissons la σ-algèbre $\mathcal{A}^j$ sur Ω_F^j $(j = 1, ..., p)$ par la σ-algèbre engendrée par les ensembles de la forme $A_T^x = \{F \in \Omega_F^j / \ F(T) \leq x\}$ pour tout $x \in [0,1]$ et $T \in V_j$. Nous notons,

$$\mathcal{A}^j = \sigma^j(A_T^x).$$

Nous définissons $\mathcal{A}$ la σ-algèbre de sous-ensembles de Ω_F par la σ-algèbre produit :

$$\mathcal{A} = \mathcal{A}^1 \times ... \times \mathcal{A}^p.$$

Nous disposons alors de l'espace mesurable $(\Omega_F, \mathcal{A})$. Dans notre étude, nous disposons d'une variable aléatoire X, qui à tout w de Ω associe le p-uplet de fonctions de répartition $X(w) = F_w = (F_w^1, ..., F_w^p) \ \in \Omega_F$:

$$\begin{array}{rccc} X: & (\Omega, \mathcal{M}, \mathbb{P}) & \longrightarrow & (\Omega_F, \mathcal{A}) \\ & w & \mapsto & F_w \in \Omega_F, \end{array}$$

avec $\mathcal{M}$ une σ-algèbre de Ω et $\mathbb{P}$ une mesure de probabilité sur $(\Omega, \mathcal{M})$.

Remarque : Notons $T = (T_1, ..., T_p) \ \in \ V$, $x = (x_1, ..., x_p) \in [0,1]^p$ et définissons $F(T) \leq x$, pour $F \in \Omega_F$, par $F_w^j(T_j) \leq x_j$, $\forall j = 1, ..., p$. L'ensemble $\{w \in \Omega / F_w(T) \leq x\}$ est mesurable par la mesure $\mathbb{P}$ car image réciproque de $\{F \in \Omega_F / F(T) \leq x\}$ par la fonction mesurable X,

$$X^{-1}(\{F \in \Omega_F / F(T) \leq x\}) = \{w \in \Omega / F_w(T) \leq x\} \in \mathcal{M}.$$

Soit $\mathfrak{F} = (F_1, ..., F_N)$ un échantillon de N réalisations indépendantes et identiquement distribuées de la variable aléatoire à valeurs dans Ω_F. L'ensemble $\mathfrak{F}$ décrit donc $W = (w_1, ..., w_N)$, N individus de Ω avec $F_i = (F_i^1, ..., F_i^p)$. Chaque F_i^j correspond à la fonction de répartition de l'individu i pour la variable j. L'ensemble $\mathfrak{F}$ (appelé "base de distributions") est un échantillon de N p-uplets de fonctions de répartition provenant de Ω_F.

Pour plus de clarté, nous supposons que nous ne disposons que d'une variable. Dans ce cas, $F_w = F_w^1$ est la fonction de distribution de l'individu w pour cette variable, $V = V_1$ et $\Omega_F = \Omega_F^1$.

Dans toute la suite nous utilisons la définition d'une fonction de répartition donnée ainsi :

Définition 2 *:*
Une application F de $\mathbb{R}$ dans $[0,1]$ est une fonction de répartition ssi

1. *F est croissante,*
2. *$lim_{x \to -\infty} F(x) = 0$,*
3. *$lim_{x \to +\infty} F(x) = 1$,*
4. *F est continue à droite.*

Remarque: Cette définition diffère de celle donnée par exemple dans le livre de Nelsen (1998, [80]) par le point 4. Dans l'ouvrage de Nelsen, la continuité à droite n'est pas mentionnée.

Cependant, nous utilisons par la suite la définition 2, plus conforme avec l'usage habituel des fonctions de répartition.

Définition 3 (Fonction de distribution de distributions) *:*
Une "fonction de distribution de distributions" (FDD) au point T est la fonction définie par:

$$\begin{array}{rccc} G_T: & \overline{\mathbb{R}} & \longrightarrow & [0,1] \\ & x & \mapsto & G_T(x) \end{array}$$

avec

$$G_T(x) = \mathbb{P}(\{w \in \Omega \ / F_w(T) \leq x\}).$$

Si cette fonction est modélisée de manière empirique à partir de $\mathfrak{F}$, la FDD est:

$$G_T^{emp}(x) \quad = \quad \frac{card(\{F_i \in \mathfrak{F} \ / F_i(T) \leq x\})}{card(\mathfrak{F})}. \tag{3.1}$$

Par exemple, dans la Figure 3.1, si x=0.4, $G_{T_1}(x)$ est le pourcentage d'individus dont la fonction de répartition prend une valeur inférieure à 0.4 au point T_1, soit $G_{T_1}(0.4) = 3/5$ (3 individus sur 5).

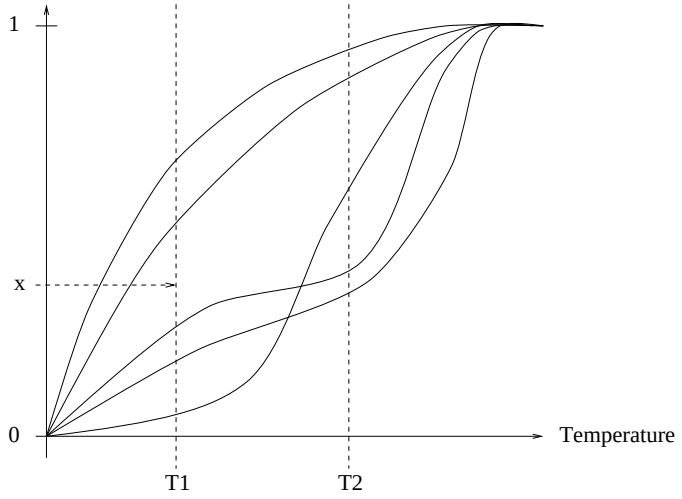


FIG. 3.1: *Fonction de distribution de distributions pour 2 valeurs de T*

En faisant pivoter la figure 3.1 de 90 degrés dans le sens anti-trigonométrique, nous pouvons voir que la fonction de distribution de distribution G_{T_1} est la fonction de répartition des valeurs des fonctions de répartition des individus en T_1 (voir figure 3.2). Nous pouvons définir la notion similaire en dimension n.

Définition 4 (Fonction de distribution jointe de n distributions) *:*
Une "fonction de distribution jointe de n distributions" (FDJD) au point $T = (T_1, ..., T_n)$ est la fonction définie par:

$$\begin{array}{rccc} H_T: & \overline{\mathbb{R}}^n & \longrightarrow & [0,1] \\ & x = (x_1, ..., x_n) & \mapsto & H_T(x) \end{array}$$

avec

$$H_T(x_1, ..., x_n) = \mathbb{P}(\{w \in \Omega \ / \ F_w(T_1) \leq x_1; ...; F_w(T_n) \leq x_n\}).$$

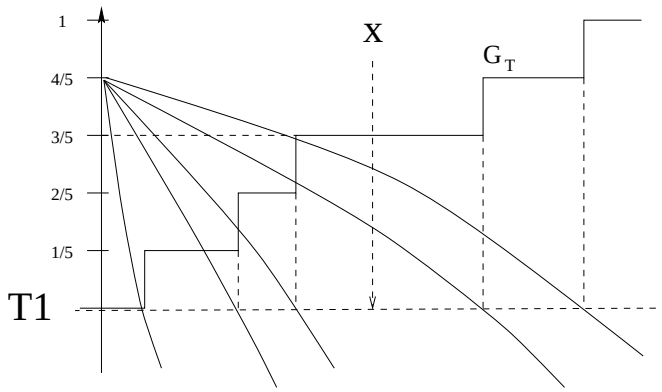


FIG. 3.2: *Fonction empirique de distribution de distributions*

Les notions de FDD et de FDJD sont les notions essentielles dans la méthode de décomposition de mélange que nous proposons au chapitre 5 pour les données fonctions de répartition. Leur définition nous amène à la proposition suivante:

Proposition 1 *:*

- G_{T_i}, $i = 1, ..., n$ *est une fonction de répartition.*
- $H_{T_1,...,T_n}$ *est une fonction de répartition jointe n-dimensionnelle de marginales* $G_{T_1}, ..., G_{T_n}$.

Démonstration :
La preuve résulte des définitions de fonctions de répartition unidimensionnelles et n-dimensionnelles. Une fonction de répartition G est définie par quatre propriétés:

- G est croissante,
- $\lim_{x\to-\infty} G(x) = 0$,
- $\lim_{x\to+\infty} G(x) = 1$,
- G est continue à droite.

Nous voyons clairement que G_{T_i} est croissante, $G_{T_i}(-\infty) = 0$ et $G_{T_i}(+\infty) = 1$. De plus $H_{T_1,...,T_n}$ est croissante, $H_{T_1,...,T_n}(x_1, ..., x_n) = 0$ pour tout $x = (x_1, ..., x_n) \in \mathbb{R}^n$ tel que $x_i = 0$ pour un $i \in \{1, ..., n\}$ (les T_i sont croissants et T_1 tel que $min_{i=1,...,N}\{F_i(T_1)\} > 0$) et $H_{T_1,...,T_n}(+\infty, ..., +\infty) = H_{T_1,...,T_n}(1, ..., 1) = 1$. Avec la définition de fonction de répartition donnée par Nelsen (1998, [80]), la démonstration est faite.
Avec la définition 2 que nous utilisons, nous devons regarder la continuité à droite de G_T. Soit $(x_n)_{n\in\mathbb{N}}$ une suite de réels tendant supérieurement vers x : $\lim_{n\to+\infty} x_n = x$ et $\forall n\ x_n \geq x$. Soit la suite $(y_p)_{p\in\mathbb{N}}$ avec $y_p = \sup_{n\geq p} x_n$. Nous avons donc $lim_{p\to+\infty} y_p = x$, $x \leq y_{p+1} \leq y_p$ et $x_p \leq y_p$. Soient les trois ensembles A_p, B_p et A_∞ de Ω, définis par :

- $A_p = \{w \in \Omega \ / \ F_w(T) \leq y_p\}$,
- $B_p = \{w \in \Omega \ / \ F_w(T) \leq x_p\}$,

– $A_\infty = \{w \in \Omega \ / \ F_w(T) \leq x\}$.

Or, $x \leq x_p \leq y_p$, donc $A_\infty \subseteq B_p \subseteq A_p$ et nous avons

$$\mathbb{P}(A_p) \geq \mathbb{P}(B_p) \geq \mathbb{P}(A_\infty). \tag{3.2}$$

La suite $(A_p)_p$ est suite décroissante d'ensembles $(A_{p+1} \subseteq A_p)$, donc d'après la théorie des probabilités,

$$\lim_{p \to +\infty} \mathbb{P}(A_p) = \mathbb{P}(\bigcap_{p \in \mathbb{N}} A_p).$$

Or, $\bigcap_p A_p = A_\infty$, donc $\lim_{p\to+\infty} \mathbb{P}(A_p) = \mathbb{P}(A_\infty)$ et avec l'encadrement (3.2),

$$\lim_{p \to +\infty} \mathbb{P}(B_p) = \mathbb{P}(A_\infty).$$

Ceci est vérifié pour toute suite $(x_n)_n$ tendant supérieurement vers x, donc G_T est continue à droite, et $\forall T$, G_T est une fonction de répartition. La démonstration est similaire pour $H_{T_1,...,T_n}$.

Nous voyons que la FDD G_T est la fonction de répartition des valeurs de fonctions de répartition en T appartenant à Ω_F, c'est-a-dire que pour un individu décrit par la fonction de distribution F, la variable aléatoire $F(T)$ est de loi G_T,

$$\{F_1(T), ..., F_N(T)\} \sim G_T$$

et $(F(T_1), ..., F(T_n))$ est de loi $H_{T_1,...,T_n}$,

$$\{(F_1(T_1), ..., F_1(T_n)), ..., (F_N(T_1), ..., F_N(T_n))\} \sim H_{T_1,...,T_n}.$$

Les définitions 3 et 4 sont posées pour une variable. Celles-ci peuvent être généralisées pour p variables par la définition suivante:

Définition 5 (Fonction de Distribution Jointe de Distributions pour p variables) *Soit $T = ((T_1^1, ..., T_1^{n_1}), (T_2^1, ..., T_2^{n_2}), ..., (T_p^1, ..., T_p^{n_p}))$ et $s = \sum_{j=1}^{p} n_j$. Une "fonction de distributions jointes de distributions" (FDJD) pour p variable au point T est la fonction définie par:*

$$\begin{array}{rccc} H_T : & \overline{\mathbb{R}}^s & \longrightarrow & [0,1] \\ & x = (x_1^1, ..., x_1^{n_1}, ..., x_p^1, ..., x_p^{n_p}) & \mapsto & H_T(x) \end{array}$$

avec

$$\begin{aligned} H_T(x) \quad = \quad & \mathbb{P}(\{w \in \Omega \ / \ F_w^1(T_1^1) \leq x_1^1; ...; F_w^1(T_1^{n_1} \leq x_1^{n_1}; ... \\ & ...; F_w^p(T_p^1) \leq x_p^1; ...; F_w^p(T_p^{n_p}) \leq x_p^{n_p}\}). \end{aligned}$$

3.2 Estimation des fonctions de distributions de distributions

D'après la proposition 1, l'estimation des FDD se résume à l'estimation d'une fonction de répartition. Nous pouvons donc appliquer les méthodes d'estimation de densité, puis intégrer

ces densités pour avoir les FDD. Les techniques de détermination sont très nombreuses. Précisons les méthodes d'estimation de densités.

La fonction de densité de probabilité est un concept fondamental en statistiques. Considérons une variable aléatoire X ayant une densité f. La fonction f donne une description naturelle de la loi de X et permet d'associer des probabilités à X avec la relation:

$$\mathbb{P}(a < X < b) = \int_a^b f(x)dx \quad \forall a < b. \tag{3.3}$$

Supposons que nous ayons un échantillon de données observées suivant une fonction de densité inconnue. L'estimation de densité est la construction d'une densité à partir des données observées. La première approche pour cette estimation est paramétrique. Nous supposons que les données sont tirées suivant une densité appartenant à une famille paramétrique de densités, par exemple la densité normale de moyenne μ et de variance σ^2. La densité f des données peut être estimée en trouvant les paramètres μ et σ à partir des données et en remplaçant ces estimations dans la formule de la densité normale. Nous regardons cette approche dans une première partie. La seconde approche est non-paramétrique. Les méthodes liées à cette approche font que les données parlent plus d'elles mêmes que dans le cas d'une famille paramétrique de densité. Nous présentons certaines de ces méthodes dans une seconde partie.

La littérature sur les densités estimées est vaste et de nombreux ouvrages présentent des résultats que nous n'évoquont pas dans cette thèse. Prakasa Rao (1983, [84]) donne un aspect théorique du sujet. Silverman (1986, [98]) donne une liste de méthodes et d'applications de l'estimation de densité.

Introduisons les notations que nous utilisons dans la suite. Nous supposons que nous avons un échantillon de N réalisations $(x_1, ..., x_N)$ de densité f à estimer par $\tilde{f}$.

Nous utilisons un jeu de données climatiques pour illustrer les méthodes. Il s'agit de températures de l'atmosphère terrestre, au niveau du sol, issues du modèle numérique de l'ECMWF (European Center for Medium range Weather Forcasting), pour la journée du 15 décembre 1998 à 0H. Les températures sont fournies à une résolution spatiale de 1 degré de longitude par 1 degré de latitude. Nous n'utiliserons qu'un point sur deux en longitude et un sur deux en latitude soient: $360/2 \times 180/2 = 16200$ valeurs.

3.2.1 Approche paramétrique

Cette approche est la plus ancienne. La loi la plus étudiée est la loi normale. Son domaine de définition est $\mathbb{R}$, elle n'est donc pas applicable directement sur nos données dans $[0, 1]$. Les FDD étant des fonctions de répartition sur des valeurs de fonctions de répartition, les lois les modélisant doivent être sur $[0, 1]$. Nous donnons deux méthodes permettant d'avoir des lois sur $[0, 1]$.

3.2.1.1 Lois sur $[0,1]$

Une des lois les plus connues travaillant sur $[0,1]$ est la loi béta. Celle-ci correspond à la loi de Dirichlet en dimension 1. La fonction de densité est:

$$f_{a,b}(x) = \frac{x^{a-1}(1-x)^{b-1}}{\int_0^1 y^{a-1}(1-y)^{b-1}dy},$$

avec $a > 0$ et $b > 0$ les deux paramètres de la loi béta. Sa particularité essentielle pour notre étude est d'aller de [0,1] dans [0,1]. La modélisation des FDD suivant cette loi est donc l'estimation des paramètres $\alpha = (a,b)$ par la méthode du maximum de vraisemblance par exemple. Cette méthode est la maximisation de $L(\alpha)$ la vraisemblance donnée par:

$$L(\alpha) = \prod_{i=1}^{N} f(x_i, \alpha),$$

avec $(x_1, ..., x_N)$ un échantillon d'une variable aléatoire X de densité supposée f. Dans le cas de la loi béta, maximiser $L(\alpha)$ est trouver a et b tels que:

$$\frac{\partial L(\alpha)}{\partial a} = 0$$

et

$$\frac{\partial L(\alpha)}{\partial b} = 0$$

ou de manière équivalente,

$$\frac{\partial log(L(\alpha))}{\partial a} = 0$$

et

$$\frac{\partial log(L(\alpha))}{\partial b} = 0.$$

Suivant la loi, les solutions peuvent être analytiques ou pas.

3.2.1.2 Lois tronquées

Pour travailler sur l'intervalle $[0,1]$, nous pouvons également utiliser des lois classiques travaillant sur n'importe quel intervalle (telle que la loi normale), puis normaliser ces densités pour réduire le domaine de définition. Considérons une densité f dont le domaine de définition est $[a,b]$ ($\mathbb{R}$ pour une densité gaussienne). La densité f_{01} correspondant sur $[0,1]$ est définie par:

$$f_{01}(x) = \begin{cases} f(x)/\int_0^1 f(y)dy & \text{si } x \in [0,1]. \\ 0 & \text{sinon.} \end{cases} \tag{3.4}$$

La fonction de répartition F_{01} associée à f_{01} est:

$$F_{01}(x) = \begin{cases} 0 & \text{si } x \leq 0, \\ \int_0^x f(y)dy/\int_0^1 f(y)dy & \text{si } x \in [0,1]. \\ 1 & \text{si } x \geq 1. \end{cases} \tag{3.5}$$

La méthode d'estimation par loi tronquée permet d'avoir de nombreuses densités sur $[0,1]$. Cette méthode s'applique sur toute loi dont le domaine de définition contient $[0,1]$.

3.2.2 Approche non-paramétrique

Les estimateurs non-paramétriques ont été proposés en premier par Fix et Hodges (1951, [48]) comme une manière de se libérer des hypothèses rigides de densités en analyse discriminante et ont été utilisés dans des contextes variés. Une utilisation naturelle de ces densités estimées est l'étude probabiliste d'un ensemble de données. Les densités non-paramétriques peuvent donner des indications sur la multimodalité des données.

3.2.2.1 Histogramme

La plus ancienne et la plus répandue des méthodes d'estimation d'une densité est l'histogramme. Etant donnée une origine x_0 et une largeur d'intervalle h, les intervalles des histogrammes sont $[x_0 + mh, x_0 + (m+1)h[$ pour un entier m. L'histogramme est défini par:

$$\tilde{f}(x) = \frac{1}{Nh} card\{x_i \text{ dans le même intervalle que } x\}. \tag{3.6}$$

Pour construire l'histogramme nous devons choisir une origine et une largeur d'intervalle. La largeur contrôle principalement la qualité de lissage inhérente à la procédure. L'histogramme peut être généralisé en autorisant la largeur d'intervalle à varier. Supposons que la droite réelle est coupée en intervalles, la densité estimée est:

$$\tilde{f}(x) = \frac{1}{N} \frac{card\{x_i \text{ dans le même intervalle que } x\}}{\text{Largeur de l'intervalle contenant } x}. \tag{3.7}$$

La fonction de répartition associée est

$$\tilde{F}(x) : \sum_{\{i/x-ih \geq min_j x_j\}} f(x - ih).$$

Un exemple se trouve Figure 3.3. Le découpage peut être a priori ou dépendre des observations. Une question revient régulièrement: pourquoi utiliser des méthodes plus élaborées que le simple histogramme? Le choix des méthodes dépend du contexte. En terme de description mathématique des fréquences, l'histogramme peut être amélioré. La forme mathématique de l'histogramme traduit l'utilisation inefficace des données si l'histogramme est utilisé comme densité estimée dans des procédures de classification ou d'analyse discriminante non paramétrique. La discontinuité de l'histogramme crée des difficultés si la dérivée de l'estimation est nécessaire. Quand les estimations sont des composantes intermédiaires pour d'autres méthodes, l'utilisation d'alternatives aux histogrammes est importante.

3.2.2.2 Estimateur naif

A partir de la définition d'une densité (basée sur la dérivée de la fonction de répartition), si la variable aléatoire X a une densité continue f, alors:

$$f(x) = \lim_{h \longrightarrow 0} \frac{1}{2h} \mathbb{P}(x - h < X < x + h).$$

Pour tout h, nous pouvons estimer $\mathbb{P}(x - h < X < x + h)$ par la proportion de l'échantillon dans l'intervalle $(x - h, x + h)$. Un estimateur naturel $\tilde{f}$ de la densité est donc donné en

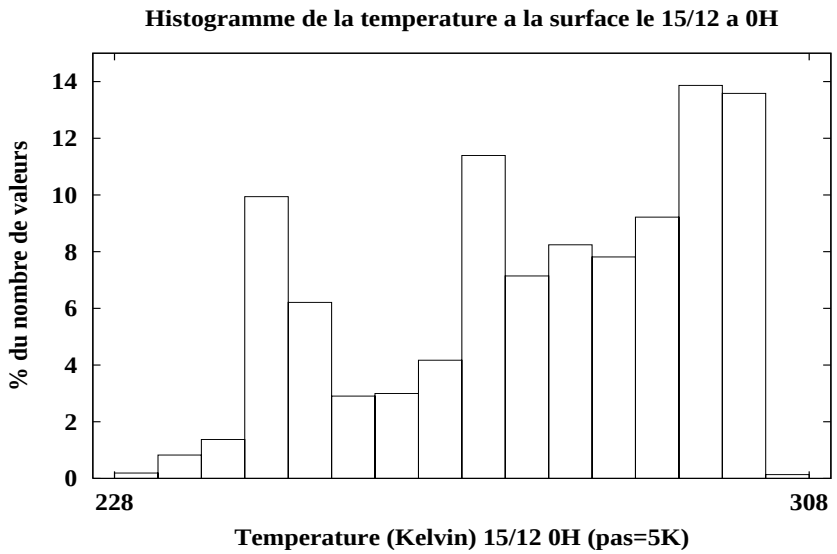


FIG. 3.3: *Exemple d'histogramme (h=5 K)*

choisissant un petit h par

$$\tilde{f}(x) = \frac{1}{2hN} card\{x_i \ dans \ (x-h, x+h)\},$$

$\tilde{f}$ est appelé l'estimateur naif.

Pour exprimer l'estimateur de manière plus transparente, définissons la fonction poids w par:

$$w(x) = \begin{cases} \frac{1}{2} & si|x| < 1, \\ 0 & sinon. \end{cases} \tag{3.8}$$

L'estimateur naif peut être écrit:

$$\tilde{f}(x) = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{h} w\left(\frac{x - x_j}{h}\right). \tag{3.9}$$

Il découle de (3.9) que l'estimation est construite en plaçant une "boite" de largeur $2h$ et de hauteur $(2Nh)^{-1}$ sur chaque observation et en sommant. Un exemple d'estimateur naif se trouve figure 3.4.

L'estimateur naif n'est pas totalement satisfaisant. La fonction $\tilde{f}$ n'est pas continue mais présente des sauts aux points $x_j \pm h$ et a des dérivées nulles partout ailleurs. En partie pour passer outre cette difficulté et en partie pour d'autres raisons techniques, nous considérons maintenant la généralisation de l'estimateur naif.

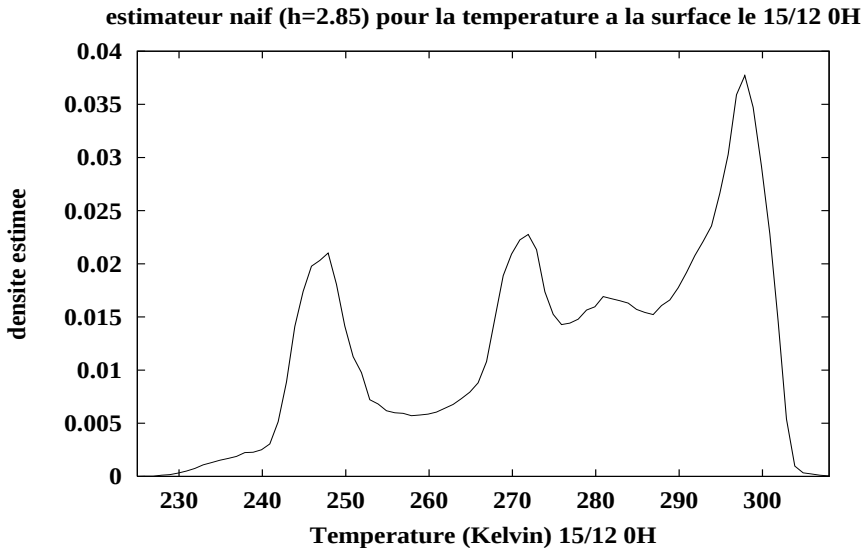


FIG. 3.4: *Exemple d'un estimateur naif*

3.2.2.3 Estimateur par noyaux

Il est facile de généraliser l'estimateur naif. Remplaçons la fonction poids w par une fonction noyau K (kernel) satisfaisant la condition

$$\int_{-\infty}^{+\infty} K(x)dx = 1. \tag{3.10}$$

Généralement (non systématiquement), K est une densité de probabilité, par exemple la densité normale. Par analogie avec la définition de l'estimateur naif, l'estimateur par noyaux, de noyau K est:

$$\tilde{f}(x) = \frac{1}{Nh} \sum_{j=1}^{N} K\left(\frac{x - x_j}{h}\right) \tag{3.11}$$

où h est la largeur de fenêtre ou paramètre de lissage.

L'estimateur naif peut être considéré comme une somme de "boites" centrées sur les observations, l'estimateur par noyaux est une somme de "bosses" placées sur les observations. La fonction noyau K détermine la forme des bosses et la largeur de fenêtre h détermine leurs largeurs.

Les effets de changements de largeur de fenêtre sont illustrés figures 3.5 et 3.6 pour $h = 2.85$, $h = 0.25$ et des noyaux gaussiens.

La limite de $\tilde{f}$ quand h tend vers 0 est une somme de delta de Dirac placés aux observations. Quand h est trop petit, les structures trop fines deviennent visibles. Quand h devient grand, tous les détails s'effacent.

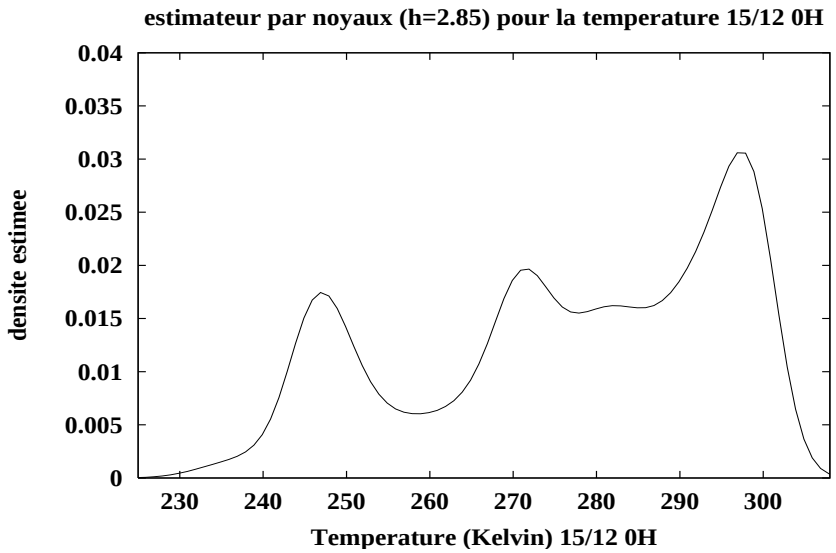


FIG. 3.5: *Exemple d'un estimateur par noyaux (h=2.85)*

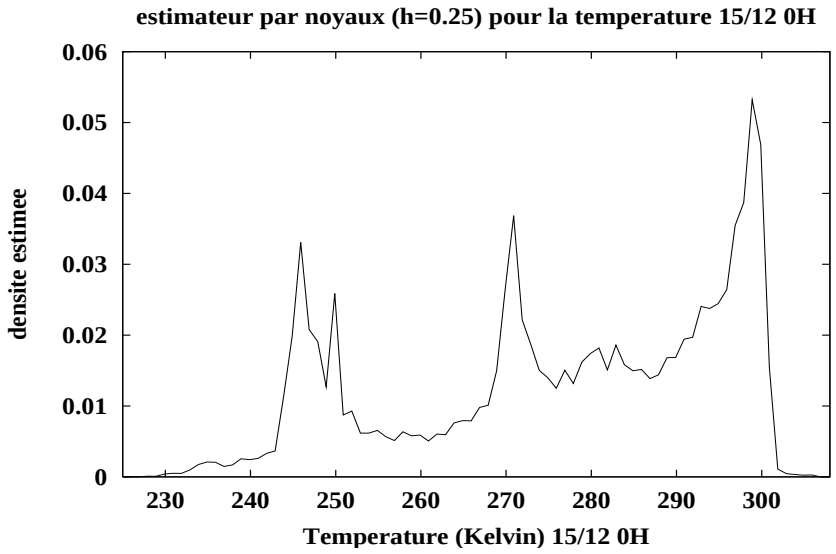


FIG. 3.6: *Exemple d'un estimateur par noyaux (h=0.25)*

Des propriétés élémentaires des estimateurs par noyaux viennent des formules (3.10) et (3.11). Si le noyau K est positif et satisfait (3.10) (i.e. est une densité), $\tilde{f}$ sera une densité et $\tilde{f}$ héritera des propriétés de continuité et de différentiabilité du noyau K. Si le noyau K est fonction de densité normale, $\tilde{f}$ sera une fonction lissée ayant des dérivées de tous ordres. A part les histogrammes, l'estimateur par noyaux est l'estimateur le plus répandu et le plus étudié mathématiquement.

3.2.2.4 Méthode du plus proche voisin

La classe d'estimateurs du plus proche voisin répond au souci d'adapter le degré de lissage à la densité locale. Le degré de lissage est contrôlé par un entier k choisi plus petit que la taille de l'échantillon ($k \approx N^{1/2}$). Nous définissons la distance de deux réalisations x et y $d(x,y)$, par $|x-y|$ et pour chaque valeur de température t ($t \in [0,\infty)$) nous définissons:

$$d_1(t) \leq d_2(t) \leq ... \leq d_N(t)$$

les distances dans l'ordre croissant de t aux points de l'échantillon. L'estimateur du k^{ieme} plus proche voisin est:

$$\tilde{f}(t) = \frac{k}{2Nd_k(t)}. \tag{3.12}$$

Supposons que la densité soit f. Pour un N-échantillon, nous attendons environ $2rNf(t)$ observations dans $[t-r, t+r]$, $\forall r > 0$. Nous avons exactement k observations dans l'intervalle $[t-d_k(t), t+d_k(t)]$ (par définition), une estimation de la densité en t peut être obtenue en posant:

$$k = 2d_k(t)N\tilde{f}(t)$$

qui nous donne(3.12). Nous donnons deux exemples en Figures 3.7 et 3.8 pour $k = 100$ et $k = 900$.

L'estimateur naif est basé sur le nombre d'observations tombant dans une boite de taille donnée, centrée au point x considéré; alors que l'estimateur du plus proche voisin est inversement proportionnel à la taille de la boite nécessaire pour contenir un nombre donné d'observations. Dans les queues de distribution, la distance $d_k(t)$ sera plus grande que dans la partie principale de la densité. La difficulté de sous-lissage est donc réduite dans les queues.

L'estimateur du plus proche voisin (3.12) n'est pas régulier. La fonction $d_k(t)$ peut être continue mais sa dérivée a une discontinuité en chaque point de la forme $\frac{1}{2}(X_{(j)} + X_{(j+k)})$, où $X_{(j)}$ sont les statistiques d'ordre de l'échantillon. La fonction $\tilde{f}$ est donc positive et continue partout mais a des discontinuités aux mêmes points que d_k. Elle n'est pas une densité car son intégrale n'est pas 1. Pour t inférieur au plus petit des points, $d_k(t) = X_{(k)} - t$ et pour $t > X_{(N)}$, $d_k(t) = t - X_{(n-k+1)}$. En remplaçant $d_k(t)$ dans (3.12) on a $\int_{-\infty}^{+\infty} \tilde{f}(t)dt$ infinie et les queues meurent à un taux t^{-1} (lentement). L'estimateur du plus proche voisin n'est pas approprié si une estimation de la densité entière est nécessaire.

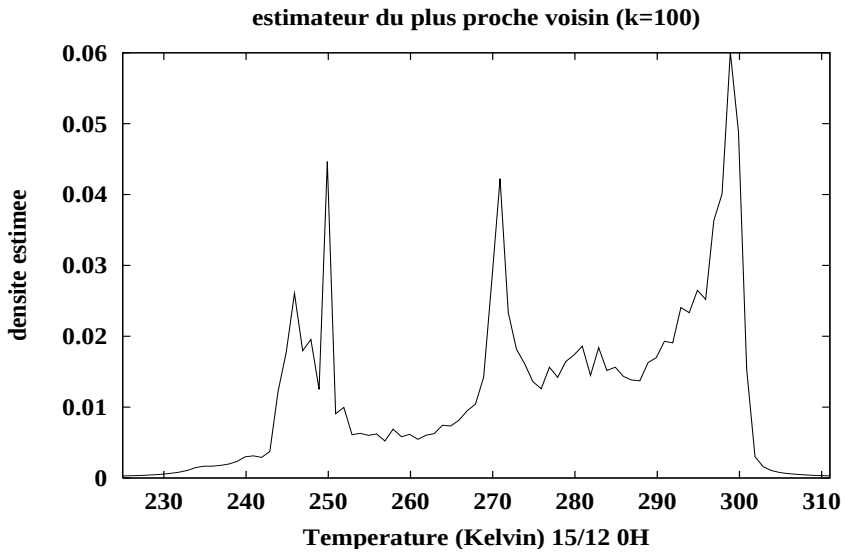


FIG. 3.7: *Exemple d'un estimateur du plus proche voisin (k=100)*

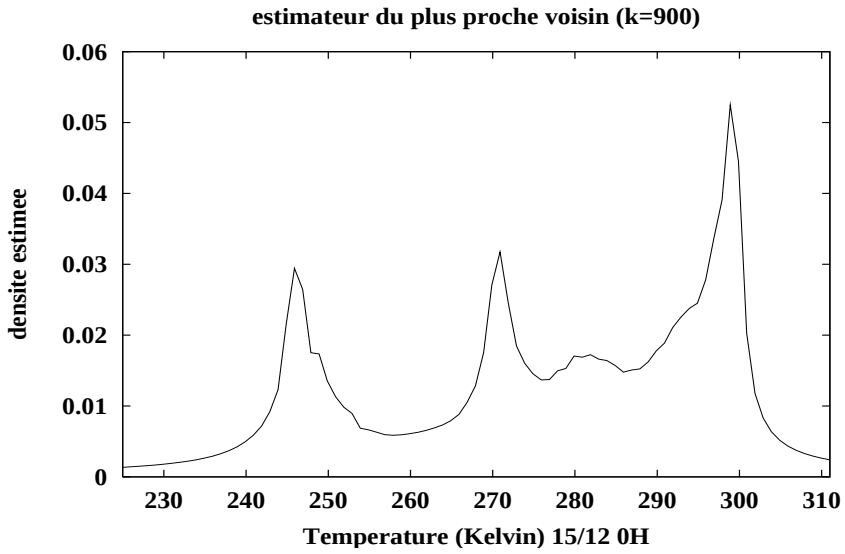


FIG. 3.8: *Exemple d'un estimateur du plus proche voisin (k=900)*

On peut généraliser l'estimateur du plus proche voisin pour obtenir un estimateur proche de l'estimateur par noyaux. Soit $K(x)$ une fonction noyau dont l'intégrale est 1. L'estimateur généralisé du k^{ieme} plus proche voisin est:

$$\tilde{f}(t) = \frac{1}{N d_k(t)} \sum_{j=1}^{N} K\left(\frac{t - x_j}{d_k(t)}\right). \tag{3.13}$$

On peut voir que $\tilde{f}(t)$ est l'estimateur par noyaux évalué en t avec la largeur de fenêtre $d_k(t)$. Le degré de lissage est gouverné par le choix de l'entier k et la largeur de fenêtre utilisée en un point particulier dépend de la densité des observations proches du point. Les dérivées de l'estimateur généralisé du plus proche voisin sont discontinues aux points où la fonction $d_k(t)$ a sa dérivée discontinue.

3.2.2.5 Méthode des noyaux variables

Comme la méthode du plus proche voisin, la méthode des noyaux variables adapte le degré de lissage à la densité locale des données. L'estimateur est construit de manière similaire à l'estimateur classique par noyaux mais le paramètre de lissage h est autorisé à varier d'une observation à une autre.

Soit K une fonction noyau et k un entier positif. Définissons $d_{j,k}$, la distance de x_j au k^{ieme} plus proche point de l'ensemble comprenant les $N-1$ autres points. L'estimateur par noyaux variables de paramètre de lissage h est défini par :

$$\tilde{f}(t) = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{h d_{j,k}} K\left(\frac{t - x_j}{h d_{j,k}}\right). \tag{3.14}$$

La largeur de fenêtre du noyau placé sur le point x_j est proportionnelle à $d_{j,k}$ (i.e. les données dans les régions de densité faible auront des noyaux plus plats). Pour k donné, le degré total de lissage dépend du paramètre h. Le choix de k détermine l'importance que la largeur de fenêtre a pour les détails locaux. Nous donnons en Figure 3.9 un exemple avec $k = 100$ et $h = 2.85$.

Dans (3.13) la largeur de fenêtre utilisée dépend de la distance de t aux observations dans l'échantillon; dans (3.14) les largeurs de fenêtres sont indépendantes du point t auquel la densité est estimée et dépend uniquement des distances entre les données.

Par contraste avec l'estimateur généralisé du plus proche voisin, l'estimateur par noyaux variables est une densité si le noyau K l'est. De plus, comme pour l'estimateur classique par noyaux, toutes les propriétés de régularités locales du noyau seront héritées par l'estimateur à noyaux variables.

3.2.2.6 Estimateurs du maximum de vraisemblance pénalisée

Les méthodes discutées précédemment proviennent toutes de la définition d'une densité (basée sur celle de la dérivée). Est-il possible d'appliquer les techniques statistiques classiques

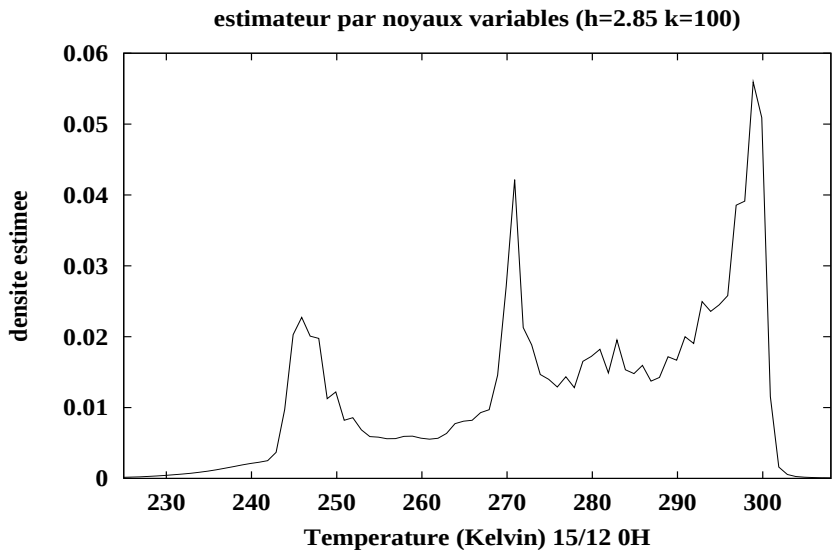


FIG. 3.9: *Exemple d'un estimateur par noyaux variables*

comme le maximum de vraisemblance? Nous voyons dans cette partie que c'est le cas si nous posons des contraintes. La vraisemblance d'une densité g d'une variable aléatoire ayant N réalisations indépendantes et identiquement distribuées est:

$$L(g|x_1, ..., x_N) = \prod_{j=1}^{N} g(x_j). \tag{3.15}$$

La vraisemblance n'a pas de maximum fini sur la classe de toutes les densités. En effet, soit $\tilde{f}_h$ l'estimateur naif de largeur de fenêtre $\frac{1}{h}$, pour tout j,

$$\tilde{f}_h(x_j) \geq \frac{1}{Nh}$$

et donc

$$\prod_{j=1}^{N} \tilde{f}_h(x_j) \geq N^{-N} h^{-N} \longrightarrow \infty \quad quand\ h \longrightarrow 0.$$

Il n'est donc pas possible de choisir le maximum de vraisemblance pour estimer la densité sans placer de restrictions sur la classe des densités sur laquelle la vraisemblance est maximisée.

Il y a néanmoins des approches possibles liées au maximum de vraisemblance. Une méthode est d'incorporer dans la vraisemblance un terme qui décrit la rugosité de la densité. Supposons que $R(g)$ soit une fonctionnelle mesurant la rugosité de g. Un choix possible de R est:

$$R(g) = \int_{-\infty}^{\infty} (g'')^2.$$

Nous définissons la log-vraisemblance pénalisée avec α le paramètre positif de lissage, par:

$$l_\alpha(g) = \sum_{j=1}^{N} \log(g(x_j)) - \alpha R(g). \tag{3.16}$$

La log-vraisemblance pénalisée peut être vue comme une manière de mesurer le conflit entre la régularité et l'adéquation aux données, le terme $\sum logg(x_j)$ mesurant l'adéquation de g aux données. La densité $\tilde{f}$ est un estimateur par maximum de vraisemblance pénalisée si elle maximise $l_\alpha(g)$ sur la classe de toutes les fonctions g qui satisfont $\int_{-\infty}^{\infty} g = 1$, $g(x) \geq 0$ pour tout x, $R(g) < \infty$. Le paramètre α contrôle la quantité de régularité car il détermine le "taux d'échange" entre la régularité et l'adéquation. Plus α est petit, plus la rugosité (pour $R(\tilde{f})$) est grande. Les estimations obtenues par la méthode du maximum de vraisemblance pénalisée sont donc des densités de probabilité.

3.2.2.7 Estimateurs par fonctions de poids générales

Il est possible de définir une classe générale d'estimateurs de densités qui inclue plusieurs des estimateurs définis plus haut. Supposons que $w(x, y)$ soit une fonction à deux variables qui (dans la plupart des cas) satisfait les conditions

$$\int_{-\infty}^{\infty} w(x, y)dy = 1 \quad \forall x \tag{3.17}$$

et

$$w(x, y) \geq 0 \quad pour\ tout\ x\ et\ y. \tag{3.18}$$

Un estimateur de densité peut être:

$$\tilde{f}(t) = \frac{1}{N} \sum_{j=1}^{N} w(x_j, t). \tag{3.19}$$

Les estimateurs de la forme (3.19) sont des estimateurs par fonction de poids générale. A partir de (3.19), les conditions (3.17) et (3.18) sont suffisantes pour que $\tilde{f}$ soit une densité de probabilité et hérite des propriétés de régularité de $w(x, .)$. Cette classe d'estimateurs peut être vue de deux manières:

- C'est un concept unifiant permettant d'obtenir des résultats théoriques applicables à une variété d'estimateurs apparemment distincts.
- Il est possible de définir des estimateurs utiles qui ne sont pas dans les classes présentées plus haut et ayant quand même la forme (3.19).

3.2.2.8 Estimateurs tronqués

Les méthodes de la forme (3.19) sont applicables à des données sur $[0, 1]$ avec une modification. A partir d'un échantillon $(X_1, ..., X_n)$ de taille n, une estimation de la densité peut être:

$$\hat{f}(x) = \frac{1}{c_n} \frac{1}{n} \sum_{i=1}^{n} w(x_j, t),$$

avec c_n tel que $\int \hat{f} = 1$. Cette approche permet de travailler sur $[0,1]$ avec une simple normalisation.

3.3 Surface de distributions de distributions

Nous pouvons donc définir une fonction de distribution de distributions en chaque T. En évaluant ces fonctions régulièrement sur le domaine de définition de la variable étudiée, nous pouvons tracer une "surface de distributions de distributions". En utilisant les notations définies au début de ce chapitre, une surface de distributions de distributions est définie formellement par :

Définition 6 *:*
La surface S de distributions de distributions, associée à la population Ω et à une variable donnée de domaine V est

$$S = \{(T, x, z) \ / \ T \in V;\ x \in [0,1];\ z = G_T(x)\}.$$

Exemple 2 *Nous disposons de données de températures atmosphériques du globe pour la journée du 15 décembre 1998 (voir chapitre 6). Ces données sont réparties tridimensionnellement, pour chaque degré de longitude et de latitude et sur 50 niveaux d'altitude. En calculant la fonction de répartition associée aux données d'une longitude sur deux et d'une latitude sur deux pour 26 niveaux d'altitude, nous avons une base de 16200 distributions.*

Avec une modélisation empirique des FDD, nous pouvons tracer la surface de distributions (Figure 3.10).

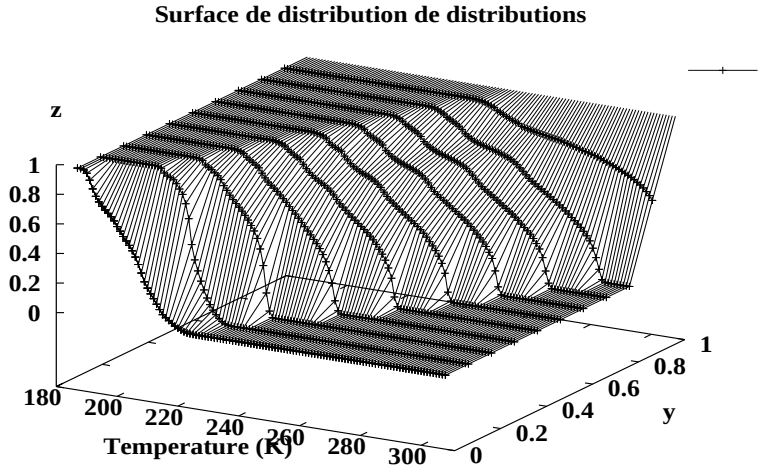


FIG. 3.10: *Exemple de surface de distributions de distributions*

3.3.1 Surface de densités de distributions

De la même manière que pour la surface de distributions de distributions, nous pouvons tracer la surface de densités de distributions en évaluant les densités associées aux FDD G_T pour des T réguliers. Si les FDD $G_T(x)$ admettent des densités $g_T(x)$, une surface de densités de distributions est définie par:

Définition 7 *:*
La surface S de densités de distributions, associée à la population Ω et à une variable donnée de domaine V est

$$S = \{(T, x, z) \ / \ T \in V; \ x \in [0,1]; \ z = g_T(x)\}.$$

Exemple 3 *En reprenant les données de l'exemple ci -dessus et en modélisant les densités par la méthode "tronquée" par noyaux, nous avons la surface de densités de distributions (Figure 3.11).*

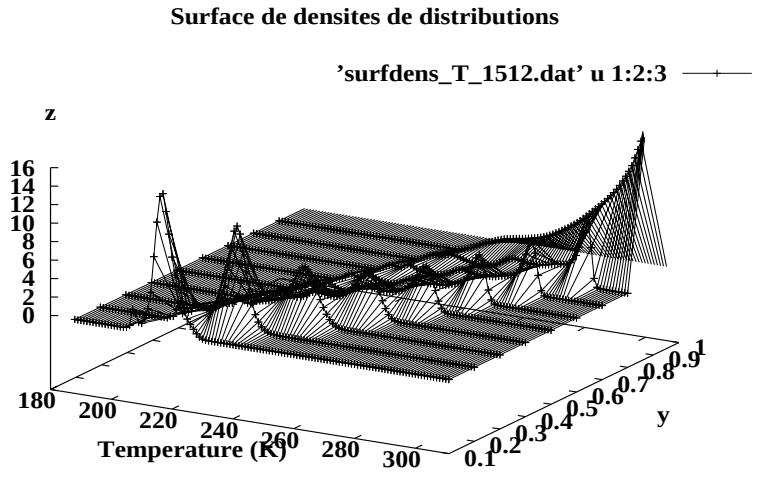


FIG. 3.11: *Exemple de surface de densités de distributions, associée à la surface de distribution de distributions*

3.4 Conclusion

Nous avons proposé une modélisation de lois de probabilité pour des variables probabilistes. La notion de fonctions de distribution de distributions nécessite de donner des valeurs T pour lesquels cette fonction est définie. Nous avons également vu que ces fonctions sont elles-mêmes des fonctions de répartition (sur $[0,1]$) et nous avons alors donné quelques méthodes pour les estimer. De plus, en estimant ces FDD unidimensionnelles pour différents T régulièrement, nous pouvons tracer une surface de distributions de distributions ou une surface de densités de distributions. Nous voyons chapitre 5 que ces surfaces peuvent nous

permettre de déterminer des T qui ont un sens pour la décomposition de mélange de lois pour données probabilistes.

Supposons que nous disposons d'un ensemble de fonctions de répartition $\mathfrak{F} = \{F_1, ..., F_N\}$ et que nous connaissons la FDD G_{T_1} à la valeur T_1 et la FDD G_{T_2} à la valeur T_2. Nous avons alors une connaissance de la répartition probabiliste des F_i $(1 \leq i \leq N)$ selon les axes T_1 et T_2 (voir Figure 3.1, page 49). Comment lier ces informations à la fonction de distribution jointe de distributions aux valeurs T_1 et T_2? Autrement dit, à partir de G_{T_1}, G_{T_2}, et $\mathfrak{F}$, peut-on caractériser la FDJD en T_1 et T_2 et avoir ainsi une connaissance de la répartition statistique des fonctions de répartition de $\mathfrak{F}$ en différents T? De manière plus générale, quels sont les liens entre une loi multidimensionnelle de probabilité et ses marginales unidimensionnelles? Dans le chapitre suivant, nous regardons une approche permettant de répondre à ces questions: les fonctions copules.

Chapitre 4

Copules

"Il est faux que l'égalité soit une loi de la nature. La nature n'a rien fait d'égal. Sa loi souveraine est la subordination et la dépendance."
Vauvenargues, Réflexions.

4.1 Introduction

L'étude des copules ("copulas" en anglais) et de leurs applications en statistiques est un phénomène récent qui date de la fin des années 50. Il y a encore 20 ans, il était malgré tout difficile de trouver le mot "copule" dans la littérature statistique. Que sont les copules? Les copules sont des fonctions qui couplent les fonctions de répartition multivariées avec les fonctions de répartition à une dimension (marginales). Alternativement, les copules sont des fonctions de distribution multivariées dont les marginales unidimensionnelles sont uniformes sur l'intervalle $[0,1]$. Comme le dit Fisher (1997, [46]), les copules intéressent les statisticiens pour deux raisons. Premièrement comme manière de mesurer la dépendance et deuxièmement comme point de départ pour construire des familles de lois bivariées. Le mot "copula" est latin et signifie "un lien". En grammaire et en logique, "copula" est utilisé pour décrire la partie de la proposition qui connecte le sujet au prédicat. Le mot "copule" a été utilisé dans un sens mathématique en premier par Abe Sklar (1959, [99]) dans le théorème qui porte son nom. Le lecteur intéressé par quelques uns de ceux qui ont participé à l'évolution du sujet peuvent consulter les articles de Dall'Aglio (1991, [27]), Schweizer (1991, [94]) ou Sklar (1996, [100]).

Comme le dit Sklar, les copules existaient avant leur nom, apparaissant dans le travail de Fréchet [50], Féron [45], Dall'Aglio [26], dans l'étude des lois multivariées de marginales univariées fixées. De nombreux résultats basiques sur les copules sont liés au travail de Hoeffding. Hoeffding (1940 [61], 1941 [62]) trouve des "distributions standardisées" ("standardized distributions") bivariées dont le support est contenu dans $[-1/2, 1/2]$ et dont les marginales sont uniformes sur $[-1/2, 1/2]$. Schweizer (1991, [94]) écrit: "had Hoeffding chosen the unit square $[0,1]^2$ instead of $[-1/2, 1/2]^2$ for his normalization, he would have discovered copulas". Hoeffding a obtenu les meilleures bornes possibles pour ces fonctions et étudia les mesures

de dépendance qui sont "scale-invariant", i.e. invariantes sous transformations strictement croissantes. Jusqu'à récemment ce travail n'avait pas reçu l'attention qu'il mérite, principalement car les articles étaient publiés dans un journal allemand relativement obscur au sortir de la deuxième guerre mondiale. Ils ont récemment été traduits en anglais et ont été publiés par Fisher et Sen (1994, [47]). Ignorant le travail de Hoeffding, Fréchet (1951, [50]) obtient plusieurs des mêmes résultats qui ont menés aux termes tels que "bornes de Fréchet" ("Fréchet bounds") et "classes de Fréchet" ("Fréchet classes"). En reconnaissant la responsabilité partagée de ces notions, Nelsen (1998, [80]) les nomme "bornes de Fréchet-Hoeffding" et "classes de Fréchet-Hoeffding". De nombreux auteurs redécouvrirent les copules. Kimeldorf et Sampson (1975, [68]) les nomment "uniform representations" et Galambos (1978, [51]) et Deheuvels (1978, [29]), "dependence functions".

Quand Sklar écrit son article de 1959 ([99]) avec le terme "copula" il collaborait avec Berthold Schweizer au développement de la théorie des espaces métriques probabilistes (probabilistic metric spaces) ou espace PM ([95]). De 1958 à 1976, la plupart des résultats sur les copules ont été obtenus dans le cadre de l'étude des espaces PM. De manière non formelle, un espace métrique consiste en un ensemble S et une distance d qui mesure les distances entre deux point p et q de S. Dans un espace métrique probabiliste, la distance d est remplacée par une fonction de répartition F_{pq} dont la valeur $F_{pq}(x)$ pour tout x, est la probabilité que la distance entre p et q soit inférieure à x. La première difficulté dans la construction d'un espace PM est d'essayer de trouver une analogie probabiliste à l'inégalité $d(p,r) \leq d(p,q)+d(q,r)$. Quelle est la relation correspondante parmi les fonctions de répartition F_{pr}, F_{pq} *et* F_{qr} $\forall$ p, q *et* r de S? Menger (1942, [76]) propose $F_{pr}(x+y) \geq T(F_{pq}(x), F_{qr}(y))$, avec T une "norme triangle" (t-norme). Comme une copule, une t-norme est une application de $[0,1]^2$ dans $[0,1]$. Quelques t-normes sont des copules et inversement quelques copules sont des t-normes. Dans un sens il était donc inévitable que les copules arrivent dans l'étude des espaces PM. Pour des détails sur les espaces PM, consulter Schweizer et Sklar (1983, [95]) et Schweizer (1991, [94]).

Un des plus importants résultats sur les espaces PM est la classe des t-normes Archimédiennes. Ces t-normes T satisfont $T(u,u) < u$ $\forall u \in [0,1]$. Les t-normes Archimédiennes sont des copules nommées copules Archimédiennes. Grâce à leur forme simple, leur facilité de construction et leurs propriétés, les copules Archimédiennes apparaissent fréquemment dans les discussions sur les lois multivariées (Genest et MacKay [55], Marshall et Olkin [74]), Joe [66]).

Pour amener progressivement le lecteur vers les copules multivariées, ainsi que vers les copules Archimédiennes, nous présentons en premier lieu les copules bivariées et leurs propriétés élémentaires.

4.2 Définition et propriétés des copules bivariées

Considérons un couple de variables aléatoires X et Y de fonction de répartition

$$F(x) = \mathbb{P}(X \leq x) \; et \; G(y) = \mathbb{P}(Y \leq y)$$

et de fonction de répartition jointe

$$H(x,y) = \mathbb{P}(X \leq x,\ Y \leq y).$$

Pour chaque couple (x,y), nous pouvons associer 3 nombres: $F(x)$, $G(y)$ *et* $H(x,y)$. Notons que chacun des nombres appartient à $[0,1]$. Chaque paire (x,y) mène donc à un point $(F(x),G(y))$ dans le carré unité $[0,1]\times[0,1]$, et à cette paire correspond à une valeur $H(x,y)$ de $[0,1]$. La correspondance associant la valeur de la fonction de distribution jointe à chaque fonction de distribution unidimensionnelle est une fonction copule.

Nous notons $\mathbb{R}$ l'ensemble des réels $(-\infty,+\infty)$, $\overline{\mathbb{R}}$ l'ensemble des réels étendus $[-\infty,+\infty]$. Un rectangle B de $\overline{\mathbb{R}}^2$ est le produit cartésien de deux intervalles fermés, $B = [x_1,x_2]\times[y_1,y_2]$. Les sommets de B sont les points $(x_1,y_1),(x_1,y_2),(x_2,y_1),(x_2,y_2)$. Le carré unité est le produit $I\times I$ avec $I=[0,1]$. Une fonction bidimensionnelle H réelle est une fonction dont le domaine, DomH, est un sous ensemble de $\overline{\mathbb{R}}^2$ et dont l'image, RanH, est un sous ensemble de $\mathbb{R}$.

Définition 8 *:*
Soit S_1 et S_2 deux sous-ensemble non vides de $\overline{\mathbb{R}}$ et H une fonction réelle bivariée telle que DomH= $S_1\times S_2$. Soit $B=[x_1,x_2]\times[y_1,y_2]$ un rectangle dont les sommets sont dans DomH. Le H-volume de B est :

$$V_H(B) = H(x_2,y_2) - H(x_2,y_1) - H(x_1,y_2) + H(x_1,y_1). \tag{4.1}$$

Si nous définissons les différences d'ordre un de H,

$$\Delta_{x_1}^{x_2}H(x,y) = H(x_2,y) - H(x_1,y) \text{ et } \Delta_{y_1}^{y_2}H(x,y) = H(x,y_2) - H(x,y_1),$$

$V_H(B)$ est la différence du deuxième ordre de H sur B :

$$V_H(B) = \Delta_{y_1}^{y_2}\Delta_{x_1}^{x_2}H(x,y).$$

Définition 9 *:*
Une fonction réelle bidimensionnelle H est 2-croissante ("2 increasing") si $V_H(B)\geq 0$ pour tout rectangle B dont les sommets sont dans DomH.

Soit S_1 *et* S_2 des sous ensembles non vides de $\overline{\mathbb{R}}$. Supposons que S_1 a un plus petit élément a_1 et S_2 un plus petit élément a_2. Nous disons que la fonction H est "grounded" si $H(x,a_2) = 0 = H(a_1,y)$ pour tout (x,y) dans $S_1\times S_2$. Soient b_1 et b_2 les plus grand éléments de S_1 et S_2. Nous disons que la fonction de $S_1\times S_2$ dans $\mathbb{R}$ a des marginales F et G définies par

$$DomF = S_1 \text{ et } F(x) = H(x,b_2)\ \forall x\in S_1,$$
$$DomG = S_2 \text{ et } G(y) = H(b_1,y)\ \forall y\in S_2.$$

Avant de définir précisément les sous copules et copules bivariées, nous donnons 2 lemmes dont les preuves peuvent être trouvées dans Schweizer et Sklar (1983).

Lemme 1 *:*
Soient S_1 et S_2 des sous-ensembles non vides de $\overline{\mathbb{R}}$ et H une fonction "grounded", 2-croissante et dont le domaine est $S_1\times S_2$. Alors H est croissante en chaque argument.

Lemme 2 (Condition de Lipschitz) :
Soient S_1 et S_2 des sous-ensembles non vides de $\overline{\mathbb{R}}$ et H une fonction "grounded", 2-croissante, de marginales F et G et dont le domaine est $S_1 \times S_2$. Soient (x_1, y_1) et (x_2, y_2) des points de $S_1 \times S_2$. Alors :

$$|H(x_2, y_2) - H(x_1, y_1)| \leq |F(x_2) - F(x_1)| + |G(y_2) - G(y_1)|.$$

Nous définissons en premier les sous-copules comme une classe de fonctions 2-croissantes et "grounded" puis les copules comme des sous copules de domaine I^2.

Définition 10 :
Une sous-copule bidimensionnelle est une fonction C' avec les propriétés:

1. *$DomC' = S_1 \times S_2$ avec S_1 et S_2 des sous-ensembles de I contenant 0 et 1;*
2. *C' est "grounded" et 2-croissante;*
3. *Pour tout u de S_1 et tout v de S_2,*

$$C'(u, 1) = u \text{ et } C'(1, v) = v. \tag{4.2}$$

Pour tout (u, v) dans DomC', $0 \leq C'(u, v) \leq 1$. RanC' est donc un sous-ensemble de I.

Définition 11 :
Une copule bidimensionnelle est une sous-copule de domaine I^2
$\Longleftrightarrow$ une copule est un fonction de I^2 dans I avec les propriétés:

1. *Pour tout u, v de I:*

$$C(u, 0) = 0 = C(0, v) \text{ ("grounded")} \tag{4.3}$$

et

$$C(u, 1) = u \text{ et } C(1, v) = v. \tag{4.4}$$

2. *Pour tout u_1, u_2, v_1, v_2 de I tels que $u_1 \leq u_2$ et $v_1 \leq v_2$:*

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

.

La différence entre copules et sous copules parait mineure mais est importante dans le théorème de Sklar.

Théorème 3 (Sklar, 1959, [99]) :
Soit H une fonction de répartition bidimensionnelle de marginales F et G. Alors il existe une copule C telle que pour tout x, y de $\overline{\mathbb{R}}$:

$$H(x, y) = C(F(x), G(y)). \tag{4.5}$$

Si les fonctions F et G sont continues, C est unique; sinon C est uniquement déterminé sur $RanF \times RanG$. Inversement si C est une copule et F et G des fonctions de répartition, la fonction H définie par (4.5) est une fonction de répartition jointe de marginales F et G.

Ce théorème est apparu en premier dans l'article de Sklar (1959, [99]). Le nom "copule" a été choisit pour exprimer la manière avec laquelle une copule couple une jointe et les marginales univariées. Ce théorème est le plus important résultat concernant les copules et est utilisé dans toutes les applications des copules. La démonstration peut être trouvée dans l'article de Sklar de 1959, [99], ou de 1996 [100].

4.2.1 Bornes bivariées de Fréchet-Hoeffding

Il est intéressant de considérer le théorème suivant :

Théorème 4 *:*
Soit C' une sous-copule. Pour tout (u, v) dans le $DomC'$

$$max(u + v - 1, 0) \leq C'(u, v) \leq min(u, v).$$

Démonstration (voir Nelsen, 1998, [80]):

Soit (u, v) un point de $DomC'$. Le fait que $C'(u, v) \leq C'(u, 1) = u$ et que $C'(u, v) \leq C'(1, v) = v$ implique $C'u, v) \leq min(u, v)$. De plus, $V_{C'}([u, 1] \times [v, 1]) \geq 0$ implique $C'(u, v) \geq u + v - 1$, qui combiné avec $C'(u, v) \geq 0$ implique $C'(u, v) \geq max(u + v - 1, 0)$.

Les copules étant des sous-copules, l'inégalité est vérifiée pour les copules. Les bornes du théorème 4 sont elles mêmes des copules notées $M(u, v) = min(u, v)$ et $W(u, v) = max(u + v - 1, 0)$. On a donc pour toute copule C et tout u, v de I,

$$W(u, v) \leq C(u, v) \leq M(u, v). \tag{4.6}$$

La copule W est la borne inférieure de Fréchet-Hoeffding et M est la borne supérieure de Fréchet-Hoeffding. Une troisième importante copule rencontrée est la copule "produit" $\Pi(u, v) = uv$. Celle-ci traduit l'indépendance des marginales. Les graphes des bornes de Fréchet-Hoeffding et de la copule Π se trouvent Figure 4.1.

4.2.2 t-normes et copules

En introduction de ce chapitre, nous avons précisé que quelques t-normes sont des copules et inversement quelques copules sont des t-normes. Rappelons la définition d'une t-norme.

Définition 12 (Schweizer et Sklar, [95]) *:*
Une opération binaire $T : [0, 1] \longrightarrow [0, 1]$ est une t-norme si elle vérifie les conditions suivantes: pour tout x, y, z, x', $y' \in [0, 1]$:

- $T(1, x) = x$,
- $T(x, y) = T(y, x)$,
- $x \leq y$, $x' \leq y' \Longrightarrow T(x, x') \leq T(y, y')$,
- $T(x, T(y, z)) = T(T(x, y), z)$.

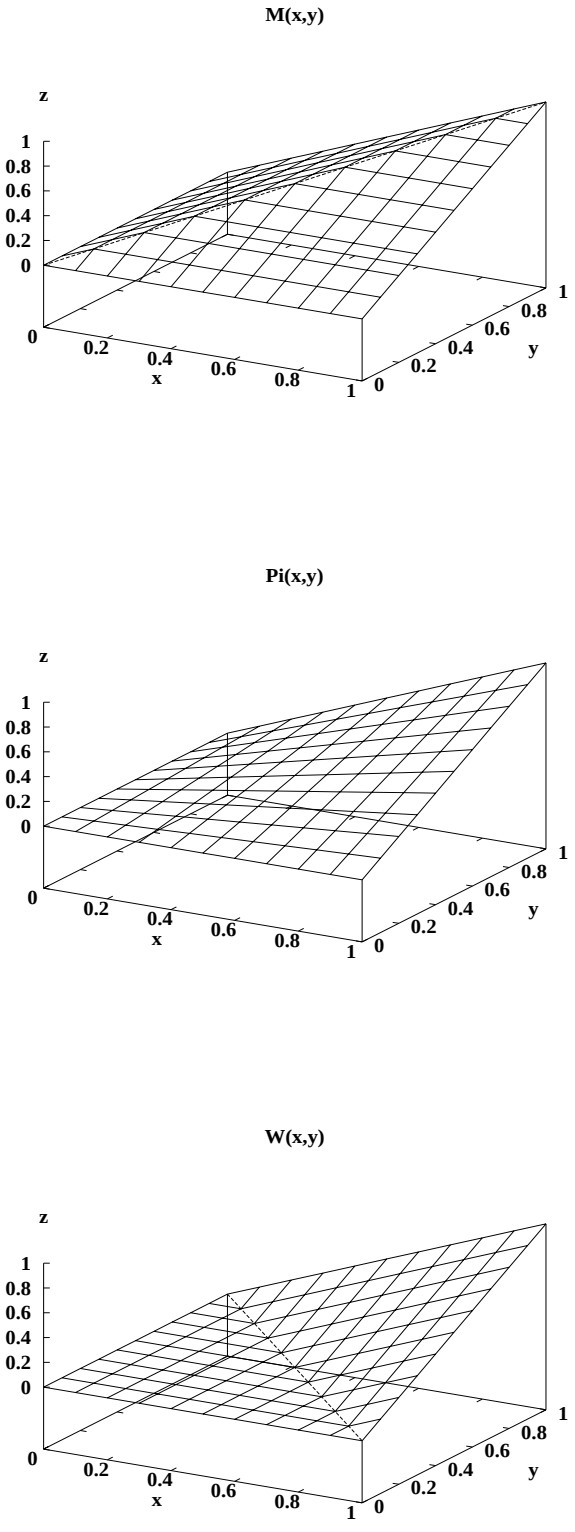


FIG. 4.1: *Graphes de copules*

En comparant les propriétés des normes triangulaires et des copules, nous voyons qu'elles se différencient par deux caractéristiques:

1. La condition d'associativité des t-normes (définition 12) n'est pas nécessairement vérifiée pour les copules,
2. La condition de Lipschitz (lemme 2) est vérifiée par les copules mais pas nécessairement vérifiée pour les t-normes.

Le lien entre t-normes et copules a été étudié par Schweizer et Sklar (1983, [95]). Ils nous donnent la proposition suivante:

Proposition 2 (Schweizer et Sklar [95]) *:*
Une copule à 2 dimensions est une norme triangulaire si et seulement si elle est associative.

Le lecteur intéressé par les correspondances entre copules et t-normes peut consulter les travaux Schweizer, 1983, [95] et Hillali, 1998, [59].

4.3 Définition et propriétés des copules multivariées

Nous étendons les résultats de la section 4.2 au cas multivarié. Alors que de nombreuses définitions et de nombreux théorèmes ont des versions analogues multivariées, toutes n'en ont pas et nous devons procéder avec attention. Pour plus de clarté, nous posons la plupart des définitions et théorèmes dans leur version multivariée.

Nous notons $\overline{\mathbb{R}}^n$ l'espace n-dimensionnel $\overline{\mathbb{R}} \times ... \times \overline{\mathbb{R}}$, nous utilisons la notation vectorielle pour les points de $\overline{\mathbb{R}}^n$, $x = (x_1, ..., x_n)$. Nous écrivons $a \leq b$ (resp. $a < b$) si $a_k \leq b_k$ (resp. $a_k < b_k$) pour tout k. Pour tout $a \leq b$ nous notons $[a, b]$ la boite de dimension n $B = [a_1, b_1] \times ... \times [a_n, b_n]$.

Définition 13 *:*
Soit S_1, ...,S_n des sous-ensemble non vides de $\overline{\mathbb{R}}$ et H une fonction réelle n-dimensionnelle telle que DomH= $S_1 \times ... \times S_n$. Soit $B = [a, b]$ un rectangle dont les sommets sont dans DomH. Le H-volume de B est :

$$V_H(B) = \sum sgn(c) H(c), \tag{4.7}$$

avec la somme sur tous les sommets c de B ($c = (c_1, ..., c_n)$ et $c_k = a_k ou b_k$) et $sgn(c)$ donné par :

$$sgn(c) = \begin{cases} 1 & si\ c_k = a_k\ pour\ un\ nombre\ pair\ de\ k, \\ -1 & si\ c_k = a_k\ pour\ un\ nombre\ impair\ de\ k. \end{cases}$$

Si nous définissons les n différences d'ordre un de H par :

$$\Delta_{a_k}^{b_k} H(t) = H(t_1, ..., t_{k-1}, b_k, t_{k+1}, ..., t_n) - H(t_1, ..., t_{k-1}, a_k, t_{k+1}, ..., t_n),$$

$V_H(B)$ est la différence d'ordre n de H sur B :

$$V_H(B) = \Delta_a^b H(t) = \Delta_{a_n}^{b_n} \Delta_{a_{n-1}}^{b_{n-1}} ... \Delta_{a_1}^{b_1} H(t).$$

Définition 14 :
Une fonction réelle n-dimensionnelle H est n-croissante ("n increasing") si $V_H(B) \geq 0$ *pour toute n-boite B dont les sommets sont dans DomH.*

Par ailleurs, supposons que le domaine d'une fonction réelle n-dimensionnelle soit $DomH = S_1 \times ... \times S_n$ avec chaque S_k ayant un plus petit élément a_k. La fonction H est dite "grounded" si $H(t) = 0$ pour tout t dans $DomH$ tel que $t_k = a_k$ pour au moins un k.

De plus, si chaque S_k est non vide et a un plus grand élément b_k, alors nous disons que H a des marginales. Les marginales unidimensionnelles de H sont les fonctions H_k données par $DomH_k = S_k$ et

$$H_k(x) = H(b_1, ..., b_{k-1}, x, b_{k+1}, ..., b_n) \text{ pour tout } x \in S_k. \tag{4.8}$$

Les marginales de dimensions supérieures sont définies en fixant moins de variables dans H. Dans la suite, les marginales unidimensionnelles sont appelées les marginales.

Avant de définir précisément les sous copules et copules multivariées, nous donnons les versions mutlidimensionnelles des deux lemmes de la section 4.2 dont les preuves peuvent être trouvées dans Schweizer et Sklar, 1983, [95].

Lemme 3 :
Soient $S_1, ..., S_n$ *des sous-ensembles non vides de* $\overline{\mathbb{R}}$ *et H une fonction "grounded", n croissante et dont le domaine est* $S_1 \times ... \times S_n$. *Alors H est croissante en chaque argument: si* $(t_1, ..., t_{k-1}, x, t_{k+1}, ..., t_n)$ *et* $(t_1, ..., t_{k-1}, y, t_{k+1}, ..., t_n)$ *sont dans DomH et* $x < y$ *alors*

$$H(t_1, ..., t_{k-1}, x, t_{k+1}, ..., t_n) \leq H(t_1, ..., t_{k-1}, y, t_{k+1}, ..., t_n).$$

Cette version n-dimensionnelle du lemme 1 est nécessaire à la démonstration de la continuité uniforme des n-copules et du théorème de Sklar n-dimensionnel (voir Schweizer, 1983, [95]).

Lemme 4 :
Soient $S_1, ..., S_n$ *des sous-ensembles non vides de* $\overline{\mathbb{R}}$ *et H une fonction "grounded", n croissante, avec des marginales et dont le domaine est* $S_1 \times ... \times S_n$. *Soient* $x = (x_1, ..., x_n)$ *et* $y = (y_1, ..., y_n)$ *des points de* $S_1 \times ... \times S_n$. *Alors :*

$$|H(x) - H(y)| \leq \sum_{k=1}^{n} |H_k(x_k) - H_k(y_k)|.$$

Nous définissons les sous-copules de dimension n comme une classe de fonctions n croissantes et grounded puis nous définissons les copules comme des sous-copules de domaine I^n.

Définition 15 :
Une sous-copule n-dimensionnelle (ou n sous copule) est une fonction C' avec les propriétés:

1. $DomC' = S_1 \times ... \times S_n$ *avec* S_k *des sous ensembles de I contenant 0 et 1;*

2. *C' est "grounded" et n croissante;*

3. *C' a des marginales (unidimensionnelles), C'_k, $k = 1, ..., n$ telles que :*

$$C'_k(u) = u \text{ pour tout } u \in S_k. \tag{4.9}$$

Pour tout u dans $DomC'$, $0 \leq C'(u) \leq 1$. $RanC'$ est donc un sous-ensemble de I.

Définition 16 *:*
Une copule n-dimensionnelle (ou n-copule) est une n-sous-copule de domaine I^n $\Longleftrightarrow$ une copule est une fonction de I^n dans I avec les propriétés:

1. *Pour tout u de I^n,*

$$C(u) = 0 \text{ si au moins une coordonnée de } u \text{ est } 0, \tag{4.10}$$

$$\text{et si toutes les coordonnées de } u \text{ sont } 1 \text{ sauf } u_k \text{ alors } C(u) = u_k; \tag{4.11}$$

2. *Pour tout a et b de I^n tels que $a \leq b$, $V_C([a, b]) \geq 0$.*

De cette définition, découle le fait que chaque marginale k-dimensionnelle d'une copule n-dimensionnelle est une k-copule. Le lemme 4 nous amène à la continuité uniforme des n-sous-copules (et donc des n-copules):

Théorème 5 *:*
Soit C' une n-sous-copule. Alors pour tout u et v dans $DomC'$:

$$|C'(v) - C'(u)| \leq \sum_{k=1}^{n} |v_k - u_k|. \tag{4.12}$$

La n-sous-copule C' est donc uniformément continue sur son domaine.

Avant de donner le théorème de Sklar dans sa version n-dimensionnel, rappelons ce qu'est une fonction de répartition en dimension n.

Définition 17 *:*
Une fonction de répartition n-dimensionnelle est une fonction H de domaine $\overline{\mathbb{R}}^n$ telle que :

1. *H est n croissante,*

2. *H est continue à droite en toutes ses composantes,*

3. *$H(t) = 0$ pour tout t dans $\overline{\mathbb{R}}^n$ tel que $t_k = -\infty$ pour au moins un k et $H(\infty, ..., \infty) = 1$.*

La fonction H est donc "grounded" et puisque $DomH = \overline{\mathbb{R}}^n$, nous déduisons du lemme 3 que les marginales unidimensionnelles données par (4.8), d'une fonction de répartition de dimension n, sont des fonctions de répartition, que nous appelons pour $n \geq 3$, $F_1, F_2, ..., F_n$.

Théorème 6 (Sklar, 1959, [99]) *:*
Soit H une fonction de répartition n-dimensionnelle de marginales $F_1, F_2, ..., F_n$. Alors il existe une n-copule C telle que pour tout x, y de $\overline{\mathbb{R}}^n$:

$$H(x_1, ..., x_n) = C(F_1(x_1), ..., F_n(x_n)). \tag{4.13}$$

Si $F_1, ..., F_n$ sont continues, C est unique; sinon C est uniquement déterminé sur $RanF_1 \times ... \times RanF_n$. Inversement si C est une n-copule et $F_1, ..., F_n$ des fonctions de répartition, alors la fonction H définie par (4.13) est une fonction de répartition n-dimensionnelle de marginales $F_1, ..., F_n$.

Le théorème de Sklar est central dans la théorie des copules et est le fondement de la plupart des applications de cette théorie. Il élucide le rôle que jouent les copules dans la relation entre les fonctions de distribution multivariées et leurs marginales univariées.

De plus, le théorème de Sklar donne une méthode de construction des copules. Pour cela, nous devons définir le quasi-inverse d'une fonction de répartition.

Définition 18 *:*
Soit F une fonction de répartition. Le quasi-inverse de F est une fonction $F^{(-1)}$ de domaine I telle que :

1. *Si t est dans $RanF$, alors $F^{(-1)}(t)$ est un nombre x tel que $F(x) = t$, c'est-à-dire*

$$\text{pour tout } t \in RanF, \quad F(F^{(-1)}(t)) = t;$$

2. *si t n'est pas dans $RanF$, alors:*

$$F^{(-1)}(t) = inf\{x|F(x) \geq t\} = sup\{x|F(x) \leq t\}.$$

Si F est strictement croissante, alors son quasi-inverse est unique et est l'inverse classique noté F^{-1}.

Corollaire 1 *:*
Soient H, C, F_1, ..., F_n comme dans le théorème 6, et $F_1^{(1)}$, ..., $F_n^{(1)}$ les quasi-inverses de F_1, ..., F_n. Alors pour tout u de I^n :

$$C(u_1, ..., u_n) = H(F_1^{(1)}(u_1), ..., F_n^{(-1)}(u_n)). \tag{4.14}$$

4.3.1 Bornes multivariées de Fréchet-Hoeffding

Les extensions des copules bivariées M, Π et W en dimension n sont notées M^n, Π^n et W^n et sont données par :

$$\begin{aligned} M^n(u) &= min(u_1, ..., u_n); \\ \Pi^n(u) &= u_1 u_2 ... u_n; \\ W^n(u) &= max(u_1 + ... + u_n - n + 1, 0). \end{aligned} \tag{4.15}$$

Les fonctions M^n et Π^n sont des n-copules pour tout $n \geq 2$ alors que W^n n'est pas une n-copule pour $n > 2$. Malgré cela, nous avons la version n-dimensionnelle des bornes de Fréchet-Hoeffding vues en (4.6).

Théorème 7 *:*
Si C' est une n-sous-copule, alors pour tout u dans $DomC'$:

$$W^n(u) \leq C'(u) \leq M^n(u). \tag{4.16}$$

La démonstration découle directement des lemmes 3 et 4.
Bien que la borne inférieure de Fréchet-HGoeffding ne soit pas une copule pour $n > 2$, le membre gauche de (4.16) est le "meilleur possible" dans le sens que pour tout $n \geq 3$ et tout u de I^n, il existe une copule C telle que $C(u) = W^n(u)$.

Théorème 8 *:*
Pour tout $n \geq 3$ et tout u de I^n, il existe une n-copule C (qui dépend de u) telle que

$$C(u) = W^n(u).$$

La preuve peut être trouvée dans [80].

4.4 Copules Gaussiennes

A partir de la formule (4.14) permettant la construction d'une copule, nous définissons ce qu'est la copule associée à une loi Gaussienne multivariée.

Soit $(X_1, ..., X_n)$ un vecteur de variables aléatoires continues de loi jointe H, Gaussienne de dimension n de moyenne μ, de matrice de covariance Σ et de marginales Gaussiennes $F_1, ..., F_n$. Nous supposons que H est standardisée (moyennes nulles et les variances égales à 1) et nous considérons donc R comme étant la matrice des coefficients de corrélation linéaires usuels,

$$R = \begin{pmatrix} \rho_{11} & \cdots & \rho_{1n} \\ \rho_{21} & \cdots & \rho_{2n} \\ \vdots & & \vdots \\ \rho_{n1} & \cdots & \rho_{nn} \end{pmatrix}$$

avec

$$\rho_{ij} = \frac{Cov(X_i, X_j)}{\sqrt{Var(X_i)}\sqrt{Var(X_j)}}.$$

Proposition 3 :
Dans le cas bivarié, la copule Gaussienne C_G associée à la loi jointe H Gaussienne, de marginales F_1, F_2 Gaussiennes s'écrit :

$$C_G(u,v) = \int_{-\infty}^{F_1^{-1}(u)} \int_{-\infty}^{F_2^{-1}(v)} \frac{1}{2\pi(1-\rho_{12}^2)^{1/2}} \exp\left(-\frac{s^2+2\rho_{12}st+t^2}{2(1-\rho_{12}^2)}\right) ds\ dt.$$

Démonstration :
La copule Gaussienne C_G est la copule vérifiant :

$$H(x_1,...,x_n) = C_G(F_1(x_1),...,F_n(x_n)).$$

En posant $u_i = F_i(x_i)$, $i = 1,...,n$, nous avons la copule Gaussienne définie par

$$C_G(u_1,...,u_n) = H(F_1^{-1}(u_1),...,F_n^{-1}(u_n)).$$

Dans le cas bivarié, C_G s'écrit donc telle que dans la proposition 3.

4.5 Copules empiriques

Une copule peut également s'exprimer de manière empirique. En effet, la fonction de répartition F d'une variable aléatoire X peut s'écrire de manière empirique à partir d'un échantillon $(x_1,...,x_N)$ de N réalisations de X par la fonction

$$F_e = \frac{nb\ de\ x_i\ tels\ que\ x_i \leq x}{N}.$$

De la même manière, la répartition empirique bivariée H_e d'un couple de variables aléatoires (X,Y) s'écrit

$$H_e(x,y) = \frac{nb\ de\ (x_i,y_i)\ tels\ que\ x_i \leq x\ et\ y_i \leq y}{N}.$$

Supposons que la v.a. X ait une fonction de distribution F et Y une fonction de distribution G. La copule C bidimensionnelle de (X,Y) étant la fonction de répartition des marginales F et G, nous pouvons dire qu'à partir de l'échantillon $(x_i,y_i)_{i=1,...,N}$, la copule empirique $C_e$de (X,Y) est

$$C_e(u,v) = \frac{nb\ de\ (x_i,y_i)\ tels\ que\ F(x_i) \leq u\ et\ G(y_i) \leq v}{N}.$$

De manière plus formelle, nous donnons la définition d'une copule empirique (Nelsen, 1998, [80]).

Définition 19 :
Soit $(x_k,y_k)_{k=1,...,N}$ un échantillon d'une variable aléatoire de loi bivariée. La copule empirique est la fonction C_e donnée par

$$C_e(\frac{i}{N},\frac{j}{N}) = \frac{nb\ de\ (x,y)\ de\ l'échantillon\ tels\ que\ x \leq x_{(i)}\ et\ y \leq y_{(j)}}{N}$$

avec $x_{(i)}$ et $y_{(j)}$ $1 \leq i, j \leq N$, les statistiques d'ordre de l'échantillon.

La fréquence de la copule empirique est donnée par

$$c_e(\frac{i}{N}, \frac{j}{N}) = \begin{cases} 1/N & si\ (x_{(i)}, y_{(j)})\ est\ un\ element\ de\ l'echantillon \\ 0 & sinon \end{cases}$$

La relation entre les fonctions C_e et c_e peut être trouvée dans le livre de Nelsen (1998, [80]) et est donnée par:

$$C_e(\frac{i}{N}, \frac{j}{N}) = \sum_{p=1}^{i} \sum_{q=1}^{j} c_e(\frac{p}{N}, \frac{q}{N})$$

et

$$c_e(\frac{i}{N}, \frac{j}{N}) = C_e(\frac{i}{N}, \frac{j}{N}) - C_e(\frac{i-1}{N}, \frac{j}{N}) - C_e(\frac{i}{N}, \frac{j-1}{N}) + C_e(\frac{i-1}{N}, \frac{j-1}{N}).$$

4.6 Copules Archimédiennes

Comme nous le voyons dans le chapitre 5, les extensions des méthodes de décomposition de mélange de densités que nous proposons, reposent, entre autres, sur la notion de copules. Nous nous intéressons particulièrement à une classe importante de copules paramétriques connues sous le nom de copules Archimédiennes. Ces copules sont utilisées dans nos extensions (et dans de nombreuses applications) pour plusieurs raisons:

- La facilité avec laquelle elles peuvent être construites;
- La grande variété de familles qui appartiennent à cette classe;
- Les propriétés des membres de cette classe.

Comme mentionné dans l'introduction de ce chapitre, les copules Archimédiennes sont apparues dans l'étude des espaces métriques probabilistes en étant étudiées dans le développement de la version probabiliste de l'inégalité triangulaire.

Pour faciliter l'approche des copules Archimédiennes multidimensionnelles, nous présentons tout d'abord les copules Archimédiennes bivariées.

4.6.1 Copules Archimédiennes 2-dimensions

Définition 20 *Soit φ une fonction strictement décroissante et continue de I dans $[0, \infty]$ telle que $\varphi(1) = 0$. Le pseudo-inverse de φ est la fonction $\varphi^{(-1)}$ avec $Dom\varphi^{(-1)} = [0, \infty]$ et $Ran\varphi^{(-1)} = I$ donnée par:*

$$\varphi^{(-1)}(t) = \begin{cases} \varphi^{-1}(t), & 0 \leq t \leq \varphi(0), \\ 0, & \varphi(0) \leq t \leq \infty. \end{cases} \tag{4.17}$$

Remarquons que $\varphi^{(-1)}$ est continue et non décroissante sur $[0, \infty]$, strictement croissante sur $[0, \varphi(0)]$, $\varphi^{(-1)}(\varphi(u)) = u$ sur I et que

$$\begin{aligned} \varphi(\varphi^{(-1)}(t) &= \begin{cases} t, & 0 \leq t \leq \varphi(0) \\ \varphi(0), & \varphi(0) \leq t \leq \infty \end{cases} \\ &= min(t, \varphi(0)). \end{aligned}$$

Lemme 5 *:*
Soit φ une fonction strictement décroissante et continue de I dans $[0,\infty]$ telle que $\varphi(1)=0$ et $\varphi^{(-1)}$ le pseudo-inverse de φ. Soit C la fonction de I^2 dans I définie par:

$$C(u,v)=\varphi^{(-1)}(\varphi(u)+\varphi(v)). \tag{4.18}$$

La fonction C satisfait les conditions (4.3) et (4.4).

Observons que ces conditions - nécessaires et suffisantes pour que $C_\varphi(x,y)$ soit une fonction de répartition (Schweizer et Sklar, 1983, [95], thm 5.4.8) - sont équivalentes à la condition suivante: la fonction $v\mapsto 1-\varphi^{-1}(v)$ est une fonction de répartition unimodale sur $[0,\infty)$ de mode 0.

Lemme 6 *:*
Soit φ, $\varphi^{(-1)}$ et C qui satisfont les hypothèses du lemme 5, C est "2-croissante" si et seulement si $\forall\ u_1\leq u_2$,

$$C(u_2,v)-C(u_1,v)\leq u_2-u_1. \tag{4.19}$$

La preuve peut être trouvée dans [80].

A partir de telles fonctions φ nous pouvons définir des copules.

Théorème 9 *:*
Soit φ une fonction strictement décroissante et continue de I dans $[0,\infty]$ telle que $\varphi(1)=0$ et $\varphi^{(-1)}$ le pseudo-inverse de φ. La fonction C de (4.18) est une copule, si et seulement si φ est convexe.

Les copules de la forme (4.18) sont nommées copules Archimédiennes. La fonction φ est nommée générateur de la copule. Si $\varphi(0)=\infty$, φ est dite générateur stricte. Dans le cas $\varphi^{(-1)}=\varphi^{-1}$ et $C(u,v)=\varphi^{-1}(\varphi(u)+\varphi(v))$, C est dite copule Archimédienne stricte. Donnons un théorème concernant les propriétés algébriques des copules Archimédiennes.

Théorème 10 *:*
Soit C une copule Archimédienne de générateur φ.

- *C est symétrique, $C(u,v)=Cv,u)\ \forall\ u,v\ \in\ I$,*
- *C est associative, $C(C(u,v),w)=C(u,C(v,w))\ \forall\ u,v,w\ \in\ I$,*
- *si $c>0$ est une constante, $c\varphi$ est aussi générateur.*

Remarque: Une copule Archimédienne est associative, nous pouvons donc déduire de la proposition 2 qu'une telle copule est également une t-norme.

Donnons le sens du terme "Archimédienne" (introduit par Ling en 1965 dans [73]). Pour cela rappelons l'axiome Archimédien pour les réels positifs. Si a, b sont des réels positifs il existe un entier n tel que $na>b$. Une copule Archimédienne se comporte comme une

opération binaire sur l'intervalle $[0,1]$. Une copule C donne pour chaque paire u, v de I une valeur $C(u,v)$ dans I. Quelque soit u de I nous pouvons définir les C-puissances u_C^n de u récursivement, $u_C^1 = u$ et $u_C^{n+1} = C(u, u_C^n)$. La version de l'axiome Archimédien pour le semi-groupe Abélien ordonné (I, C) est: pour tout u, v de I il existe un entier positive n tel que $u_C^n < v$. Le théorème qui suit nous dit que les copules Archimédiennes satisfont cette version de l'axiome Archimédien et méritent leur nom.

Théorème 11 *:*
Soit C une copule Archimédienne générée par φ. Pour tout u, v de I il existe un entier positif n tel que $u_C^n < v$.

La preuve peut être trouvée dans [80]. De plus, le théorème qui suit nous donne une condition suffisante pour avoir une copule Archimédienne.

Théorème 12 (Critère d'Abel) *:*
Une copule C est Archimédienne si elle possède deux dérivées partielles et s'il existe une fonction intégrable f de $[0,1]$ dans $[0,\infty[$ telle que :

$$f(u)\frac{\partial}{\partial v}C(u,v) = f(v)\frac{\partial}{\partial u}C(u,v)$$

pour tout $0 \leq u, v \leq 1$.

En utilisant la définition d'une copule bidimensionnelle, nous pouvons donner une méthode de simulation de réalisations d'un vecteur aléatoire (X, Y) de copule C engendrée par une application φ. Cette méthode est proposée par Nelsen dans [79]:

- Etape 1. Générer deux variables aléatoires indépendantes S et U de lois uniformes sur $[0,1]$,
- Etape 2. Calcul de $Z = \varphi'^{-1}(\varphi'(S)/U)$,
- Etape 3. $X = S$ et $Y = \varphi^{-1}(\varphi(Z) - \varphi(X))$.

Démonstration:

Soient U et S deux variables aléatoires indépendantes de lois uniformes sur $[0,1]$ et $X = S$. Soit $Z = C(X,Y)$. La fonction de répartition de Z sachant X est:

$$\begin{aligned}
\mathbb{P}(Z \leq z | X = x) &= \mathbb{P}(C(X,Y) \leq z | X = x) \\
&= \mathbb{P}(\varphi^{-1}(\varphi(X) + \varphi(Y)) \leq z | X = x) \\
&= \mathbb{P}(Y \leq \varphi^{-1}(\varphi(z) - \varphi(X)) | X = x) \\
&= \lim_{t \to 0} \mathbb{P}(Y \leq \varphi^{-1}(\varphi(z) - \varphi(X)) | X \in [x-t, x+t])
\end{aligned}$$

Soit y une observation de Y. En posant $y = \varphi^{-1}(\varphi(z) - \varphi(X))$ on a

$$\begin{aligned}
\mathbb{P}(Z \leq z | X = x) &= \lim_{t \to 0} \frac{\mathbb{P}(Y \leq y, X \leq x+t) - \mathbb{P}(Y \leq y, X \leq x-t)}{\mathbb{P}(X \in [x-t, x+t])} \\
&= \frac{\partial H(x,y)}{\partial x} \\
&= \frac{\varphi'(x)}{\varphi'(z)}
\end{aligned}$$

En posant $U(x) = \mathbb{P}(Z \leq z | X = x)$ et $x \in [0,1]$, U est une variable aléatoire de loi uniforme sur $[0,1]$. On a donc $\frac{\varphi'(x)}{\varphi'(z)} = U$ qui implique $Z = \varphi\prime^{-1}(\frac{\varphi'(X)}{U})$ et donc $Y = \varphi^{-1}(\varphi(Z) - \varphi(X))$ telle que $\mathbb{P}(X \leq x, Y \leq y) = C(x,y)$.

4.6.2 Exemples de copules Archimédiennes

Nous présentons ici trois familles de copules Archimédiennes qui ont fait le sujet d'études dans la littérature scientifique.

4.6.2.1 Famille de Ali-Mikhail-Haq

Pour tout u, $v \in [0,1]$ et $\beta \in [-1,1]$, la famille de copules de Ali-Mikhail-Haq (1978, [3]) est définie par :

$$C_\beta(u,v) = \frac{uv}{1-\beta(1-u)(1-v)}.$$

Elle a la caractéristique de ne jamais atteindre les bornes de Fréchet et est monotone par rapport au paramètre β. La fonction génératrice φ_β est :

$$\varphi_\beta(t) = (1-\beta)^{-1} log(\frac{1+\beta(t-1)}{t}).$$

4.6.2.2 Famille de Genest-Ghoudi

Elle permet de modéliser tous les degrés de dépendance possibles entre deux variables aléatoires réelles, absolument continues par rapport à la mesure de Lebesgue. La famille de Genest-Ghoudi (1994, [54]) est définie pour u, $v \in [0,1]$ et $\beta \in [0,1]$:

$$C_\beta(u,v) = \varphi_\beta^{-1}(min(1, \varphi_\beta(u) + \varphi_\beta(v)))$$

avec

$\varphi_\beta(t) = \varphi_\beta^{-1}(t) = (1-t^\beta)^{1/\beta}$, $t \in [0,1]$,
$C_0(u,v) = min(u,v)$,
$C_1(u,v) = max(0, u+v-1)$.

La fonction C_β n'atteint pas la loi d'indépendance (la copule "produit") et est une fonction croissante de β. La densité correspondant à C_β s'annule sur

$$E_\beta = \{(x,y) \in [0,1]^2 \ : \ \varphi_\beta(x) + \varphi_\beta(y) > 1\}.$$

4.6.2.3 Famille de Frank

La famille Frank (1979, [49], [52]) est définie pour $\beta \neq 1$ et strictement positif par :

$$C_\beta(u,v) = \frac{log(1 + \frac{(\beta^u - 1)(\beta^v - 1)}{(\beta - 1)})}{log(\beta)}.$$

L'application φ_β génératrice est :

$$\varphi_\beta(t) = -log(\frac{1-\beta^t}{1-\beta}).$$

La famille C_β atteint les bornes de Fréchet et la loi d'indépendance asymptotiquement :
$lim_{\beta\to 0}C_\beta(u,v) = min(u,v)$,
$lim_{\beta\to 1}C_\beta(u,v) = uv$,
$lim_{\beta\to \infty}C_\beta(u,v) = max(0,u+v-1)$.

4.6.3 Copules Archimédiennes multivariées

La copule produit s'écrit $\Pi(u,v) = uv = exp\,(-((-lnu)+(-lnv)))$. De la même manière la copule produit en dimensions n s'écrit pour $u = (u_1,...,u_n)$:

$$\Pi^n(u) = u_1...u_n = exp(-((-lnu_1)+...+(-lnu_n))).$$

Cette écriture nous amène à la généralisation de (4.18):

$$C^n(u) = \varphi^{(-1)}(\varphi(u_1)+...+\varphi(u_n)), \tag{4.20}$$

nommée "itéré en série" (serial iterates) (Schweizer, 1983, [95]) de la copule bidimensionnelle Archimédienne générée par φ. En posant $C^2(u_1,u_2) = C(u_1,u_2) = \varphi^{(-1)}(\varphi(u_1)+\varphi(u_2))$ nous avons pour tout $n \geq 3$:

$$C^n(u_1,...,u_n) = C(C^{n-1}(u_1,...,u_{n-1}),u_n).$$

Cette méthode ne donne pas des copules n-dimensionnelles pour tous les φ continues, strictement décroissantes et convexes. En utilisant $\varphi(t) = 1-t$ dans (4.20), nous avons W^n et W^n n'est pas une copule si $n > 2$. Quelles propriétés supplémentaires de φ (et de $\varphi^{(-1)}$) nous disent que C^n de (4.20) est une copule? Pour répondre à cette question, nous utilisons la définition d'une fonction complètement monotone donnée par Widder en 1941 dans [115].

Définition 21 (Widder (1941), [115]) :
Une fonction $\beta(t)$ est complètement monotone sur un intervalle J si elle est continue dessus et a ses dérivées de tous ordres qui alternent en signe

$$(-1)^k\frac{d^k}{dt^k}\beta(t) \geq 0 \tag{4.21}$$

pour tout t dans l'intérieur de J et $k = 0,1,2,...$

Le théorème 13 nous donne des conditions nécessaires et suffisantes pour un générateur stricte φ pour générer des copules Archimédiennes de dimension n.

Théorème 13 :
Soit φ une fonction continue strictement décroissante de I dans $[0,\infty)$ telle que $\varphi(0) = \infty$ et φ^{-1} l'inverse de φ. Si C^n est la fonction de I^n dans I de (4.20) C^n est une copule pour $n \geq 2$ si et seulement si φ^{-1} est complètement monotone sur $[0,\infty)$.

Les lois multidimensionnelles (4.20) ont un intérêt limité du fait que chaque élément de la famille Archimédienne est défini par la même application φ. Un haut degré de symétrie est donc présent. Hillali ([59]) propose une autre généralisation de familles Archimédiennes pour

ne pas avoir la symétrie.

Soit C_2 la copule bivariée, la généralisation à l'ordre n est :

$$C_n(u_1, ..., u_n) = \varphi_n^{-1}(\varphi_n(C_{n-1}(u_1, ..., u_{n-1})) + \varphi(u_n)) \tag{4.22}$$

avec

$$C_{n-1}(u_1, ..., u_{n-1}) = \varphi_{n-1}^{-1}(\varphi_{n-1}(C_{n-2}(u_1, ..., u_{n-2})) + \varphi_{n-1}(u_{n-1}))$$

avec $0 \leq u_1, ..., u_n \leq 1$ et les applications φ_i strictement décroissantes, continues et convexes, $1 \leq i \leq n$.

Dans le cas d'une famille de copules à 3 dimensions, nous avons donc 2 fonctions φ_{β_1} et φ_{β_2} dépendant de 2 paramètres β_1 et β_2, définissant la copule $C_{\beta_1\ \beta_2}$.

Pour illustrer les copules (4.22), nous donnons une méthode permettant de générer une réalisation du vecteur aléatoire (X, Y, Z) de marginales uniformes sur $[0, 1]$, de copule C de type (4.22), engendrée par φ_1 et φ_2:

- Etape 1. On génère trois variables aléatoires indépendantes X, U et T de lois uniformes sur $[0, 1]$,
- Etape 2. Calcul de $W_1 = (\varphi_1^{-1})'(\varphi_1(X)/U)$,
- Etape 3. ON pose $Y = \varphi_1^{-1}(\varphi_1(W_1) - \varphi_1(X))$,
- Etape4. Calcul de $W_2 = F^{-1}(T)$, avec F la fonction de répartition conditionnelle de $W_2 = C(X, Y, Z)$ sachant X et Y,
- Etape 5. On pose $Z = \varphi_2^{-1}(\varphi_2(W_2) - \varphi_2(\varphi_1^{-1}(\varphi_1(X) + \varphi_1(Y))))$.

La preuve de cette méthode de simulation se trouve dans [59].

Les copules Archim'ediennes sont donc définies à partir de fonctions génératrices φ dépendant d'un ou de plusieurs paramètres β. Comment estimer ces paramètres?

4.7 Estimation de copules Archimédiennes bivariées

Dans le cas d'une copule Archimédienne avec deux variables, la connaissance de la copule équivaut donc à la connaissance de son paramètre (pour une famille de copules fixée). Nous présentons différentes méthodes pour estimer le paramètre d'une copule ou même la copule sans le paramètre.

4.7.1 Estimation non-paramétrique d'une copule Archimédienne

Pour estimer une copule Archimédienne, Genest et Rivest [56] tirent partie du fait que $C_\varphi(x, y) = \varphi^{-1}(\varphi(x) + \varphi(y))$ est uniquement déterminé par la fonction $K(v) = v - \varphi(v)/\varphi'(v)$ définie sur l'intervalle unité. Ceci vient de la proposition suivante.

Proposition 4 (Genest, Rivest, [56]) *:*
Soient X et Y des variables aléatoires uniformes dont la copule $C(x,y)$ est de la forme $\varphi^{-1}(\varphi(x)+\varphi(y))$ pour φ convexe, décroissante et définie sur $(0,1]$ avec $\varphi(1)=0$.

$$\textit{Soient } U=\frac{\varphi(X)}{\varphi(X)+\varphi(Y)},\ V=C(X,Y) \textit{ et } \lambda(v)=\frac{\varphi(v)}{\varphi'(v)} \textit{ pour } 0<v\leq 1.$$

Alors
(a) U est uniformément distribuée sur $[0,1]$,
(b) V est distribuée selon la loi $K(v)=v-\lambda(v)$ sur $(0,1)$ et
(c) U et V sont des variables aléatoires indépendantes.

Connaissant K nous pouvons estimer φ en résolvant l'équation différentielle

$$\varphi(v)/\varphi'(v)=v-K(v).$$

Ceci nous donne

$$\varphi(v)=\exp\left\{\int_{v_0}^{v}\frac{1}{\lambda(t)}dt\right\}, \tag{4.23}$$

avec $0<v_o<1$ une constante choisie arbitrairement.

La proposition suivante définit la "projection" de (presque) toutes les copules C dans la classe des copules Archimédiennes.

Proposition 5 (Genest, Rivest, [56]) *:*
Soient X et Y des variables aléatoires uniformes de copule $C(x,y)$. Pour $0\leq v\leq 1$, soit $K(v)=\mathbb{P}(C(X,Y)\leq v)$ et définissons $K(v^-)=\lim_{t\nearrow v}K(t)$. La fonction $\varphi(v)$ définie par (4.23) est convexe et décroissante et satisfait $\varphi(1)=0$ si et seulement si $K(v^-)>v$ pour tout $0<v<1$.

Il est clair que parmi les copules bivariées $C(x,y)$ pour lesquelles la distribution de $V=C(X,Y)$ satisfait $K(v^-)>v$ sur son domaine, les copules Archimédiennes $C_\varphi(x,y)=\varphi^{-1}(\varphi(x)+\varphi(y))$ sont les seules pour lesquelles (a) et (b) de la Proposition 4 sont vraies.

Regardons la procédure d'estimation non-paramétrique de C. Soit $\{(X_1,Y_1),...,(X_N,Y_N)\}$ un échantillon d'un couple de variables aléatoires de loi bivariée $H(x,y)$ avec marginales continues $F(x)$ et $G(y)$ et de copule $C(x,y)$. Supposons que nous voulons estimer C sous l'hypothèse que C est Archimédienne. La procédure présentée ici étant indépendante des marginales, nous pouvons supposer sans perte de généralités que celles-ci sont uniformes sur l'intervalle unité. Ainsi H et C seront confondues puisque des confusions ne peuvent arriver.

D'après la proposition 4, les copules Archimédiennes sont caractérisées par le comportement stochastique de la variable aléatoire $V=H(X,Y)$ et donc une manière de procéder est d'estimer la fonction de répartition univariée $K(v)=\mathbb{P}(H(X,Y)\leq v)=\mathbb{P}(C(F(X),G(Y))\leq v)$ sur l'intervalle $(0,1)$. Ceci peut être effectué en deux étapes, que la représentation uniforme de H soit Archimédienne ou pas:

1. Construire la fonction de répartition bivariée empirique $H_N((x,y)$ associée à H.

2. Calculer $H_N(X_i, Y_i)$ pour $i = 1, ..., N$ et utiliser ces pseudo-observations pour construire une fonction de répartition uni-dimensionnelle empirique pour K.

En pratique il n'est pas nécessaire de construire H_N pour avoir un estimateur de K. A la place on peut utiliser que $H_N(X_i, Y_i)$ représente la proportion d'observations de l'échantillon inférieures ou égales à (X_i, Y_i). Puisque $H_N(X_i, Y_i)$ est plus grand que $1/N$ et qu'il est souhaitable d'avoir un estimateur prenant des valeurs sur $(0, 1)$, il est plus pratique par la suite d'utiliser les variables :

$$V_i = \sharp\{(X_j, Y_j);\ X_j < X_i,\ Y_j < Y_i\}/(N-1),\ 1 \leq i \leq N, \tag{4.24}$$

avec $\sharp$ le cardinal d'un ensemble. Soit $\delta(t)$ la distribution de masse à l'origine,

$$\delta(t) = \begin{cases} 1 & si\ t = 0, \\ 0 & sinon. \end{cases}$$

Un estimateur non-paramétrique de K est donné par :

$$K_N(v) = \sum_{i=1}^{N} \delta(v - V_i)/N. \tag{4.25}$$

Sous l'hypothèse que la copule associée à H est Archimédienne, un estimateur de la fonction $\lambda(v) = \varphi(v)/\varphi'(v)$ peut être $\lambda_N(v) = v - K_N(v)$, $0 < v < 1$. En assurant $K(v^-) > v$ sur son domaine, la formule (4.23) fournit un estimateur de C dans la classe des copules Archimédiennes, que la copule associée à la fonction de répartition bivariée $H(x, y)$ soit Archimédienne ou pas.

Nous pouvons également fixer une famille de copules Archimédiennes et chercher à estimer son paramètre β.

4.7.2 Estimation d'une copule Archimédienne par maximum de vraisemblance

Dans le cas où les les lois marginales sont connues, la méthode du maximum de vraisemblance est la plus utilisée. Elle est facile de mise en oeuvre et a des propriétés d'optimalité asymptotique.

Soient X et Y deux variables aléatoires de copule C_β, de fonction de répartition bivariée $H(x, y)$ et de marginales $F(x)$ et $G(y)$. Soient $x = (x_1, ..., x_N)$ et $y = (y_1, ..., y_N)$ deux réalisations de taille N de X et de Y. L'estimateur $\hat{\beta}$ de β, est obtenu par maximisation de la vraisemblance ou de la log-vraisemblance. Soient $L_{x,y}(\beta)$ la vraisemblance des observations, $l_{x,y}(\beta)$ son logarithme et h_β la densité correspondant à C_β,

$$L_{x,y}(\beta) = \prod_{i=1}^{N} h_\beta(x_i, y_i), \tag{4.26}$$

$$l_{x,y}(\beta) = \sum_{i=1}^{N} log(h_\beta(x_i, y_i)). \tag{4.27}$$

D'après le principe de translation de Nataf (1962, [78]), dans les équations (4.26) et (4.27), nous pouvons remplacer les observations x_i et y_i par les valeurs de leur fonction de répartition $F(x_i)$ et $G(y_i)$ pour l'estimation de $\hat{\beta}$.

Par ailleurs, dans le cas de la copule de Frank (voir section 4.6.2.3), pour des marginales F et G uniformes, nous pouvons démontrer par un calcul rigoureux et fastidieux que la densité $h_\beta(x,y)$ associée à C_β et correspondant à

$$h_\beta(x,y) = \frac{\partial^2 C}{\partial x \partial}(x,y),$$

est donnée par

$$h_\beta(x,y) = \frac{(\beta-1)log(\beta)\beta^{x+y}}{[(\beta-1)+(\beta^x-1)(\beta^y-1)]^2} \quad (0<x,\ y<1), \tag{4.28}$$

et $h_\beta(x,y)$ tend vers 1 quand β tend vers 1.

La méthode du maximum de vraisemblance consiste à estimer la valeur du paramètre β par

$$\hat{\beta} = argmax_\beta\ L_\beta(x,y)$$

ou

$$\hat{\beta} = argmax_\beta\ l_\beta(x,y)$$

L'estimateur $\hat{\beta}$ a - sous des conditions de régularité - certaines propriétés d'optimalité lorsque le nombre d'observations est grand. Le calcul de $\hat{\beta}$ est la résolution de l'équation

$$\frac{\partial l_\beta(x,y)}{\partial \beta} = 0,$$

qui ne possède pas de solution explicite. L'équation se résout donc de manière numérique.

4.7.3 Estimation d'une copule Archimédienne par "τ" de Kendall ou "ρ" de Spearman

Lorsque les marginales sont inconnues, des méthodes d'estimation non paramétriques doivent être utilisées. Deux méthodes liées, au coefficient de corrélation de rangs de Kendall et à celui de Spearman, existent. Ces deux coefficients sont basés sur la notion de concordance des réalisations de variables aléatoires. Revenons sur cette notion par la définition que l'on peut trouver dans Nelsen (1998, [80]).

Définition 22 *:*
Soient (x,y) et $(\tilde{x},\tilde{y})$, deux observations d'un vecteur (X,Y) de variables aléatoires continues. Alors (x,y) et $(\tilde{x},\tilde{y})$ sont dits concordants si $(x-\tilde{x})(y-\tilde{y}) > 0$ et discordants si $(x-\tilde{x})(y-\tilde{y}) < 0$.

Le théorème suivant peut également être trouvé dans Nelsen (1998, [80] p. 127).

Théorème 14 *:*
Soient (X,Y) et $(\tilde{X},\tilde{Y})$ des vecteurs indépendants de variables aléatoires continues avec fonctions de distribution jointes respectives H et $\tilde{H}$, et de marginales communes F (pour X et $\tilde{X}$) et G (pour Y et $\tilde{Y}$). Soient C et $\tilde{C}$ les copules de (X,Y) et $(\tilde{X},\tilde{Y})$ respectivement telles que $H(x,y) = C(F(x),G(y))$ et $\tilde{H}(x,y) = \tilde{C}(F(x),G(y))$. Notons Q la différence entre la probabilité de concordance et de discordance de (X,Y) et $(\tilde{X},\tilde{Y})$

$$Q = \mathbb{P}\{(X-\tilde{X})(Y-\tilde{Y}) > 0\} - \mathbb{P}\{(X-\tilde{X})(Y-\tilde{Y}) < 0\}.$$

Alors,

$$\begin{aligned} Q = Q(C,\tilde{C}) &= 4\iint_{Dom(X)\times Dom(Y)} \tilde{C}(F(x),G(y))dC(F(x),G(y)) \; - 1 \\ &= 4\iint_{[0,1]^2} \tilde{C}(u,v)dC(u,v) \; - 1 \end{aligned}$$

Nous regardons tout d'abord le coefficient τ de Kendall et sa relation avec Q, puis nous faisons de même pour le ρ de Spearman.

4.7.3.1 Coefficient de corrélation du τ de Kendall

Les deux définitions et le théorème suivants sont empruntés au livre de Nelsen (1998, [80]).

Définition 23 (τ de Kendall) *:*
Soit (X,Y) un couple de variables aléatoires continues. Le τ de Kendall est défini par :

$$\tau(X,Y) = \mathbb{P}\{(X-\tilde{X})(Y-\tilde{Y}) > 0\} - \mathbb{P}\{(X-\tilde{X})(Y-\tilde{Y}) < 0\},$$

avec $(\tilde{X},\tilde{Y})$ un couple de variables aléatoires continues de même loi que (X,Y) et indépendant de (X,Y).

Le τ de Kendall est donc la probabilité de concordance moins la probabilité de discordance de vecteurs aléatoires continues de même loi et indépendant entre eux.

Théorème 15 *:*
Soit (X,Y) un couple de variables aléatoires continues de copule C. Le τ de Kendall de (X,Y) est donné par :

$$\tau(X,Y) = Q(C,C) = 4\iint_{[0,1]^2} C(u,v)dC(u,v) \; - 1.$$

Nous pouvons voir que l'intégrale ci-dessus est l'espérance de la variable aléatoire $C(U,V)$ avec U et V de loi uniforme sur $[0,1]$ et de fonction de répartition jointe C. Nous avons donc $\tau(X,Y) = 4\mathbb{E}(C(U,V)) - 1$.

De plus, si la copule C est une copule Archimédienne C_β de paramètre β, le τ de Kendall $\tau(X,Y)$ peut être noté $\tau(C_\beta)$ ou $\tau(\beta)$ car la fonction $\tau(\beta)$ est monotone en β. De plus, $\tau(\beta)$ peut prendre n'importe quelle valeur dans l'intervalle $[-1,1]$.

Nous pouvons définir le τ de Kendall empirique associé à un échantillon de réalisations du couple de variables aléatoires (X,Y).

Définition 24 *:*
Soient $(x_1,...,x_N)$ un échantillon de N réalisations de X et $(y_1,...,y_N)$ un échantillon de N réalisations de Y. L'estimateur empirique τ_{emp} du τ de Kendall est

$$\tau_{emp} = \frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} X_{ij}Y_{ij}}{\begin{pmatrix} N \\ 2 \end{pmatrix}}$$

avec

$$X_{ij} = \begin{cases} 1 & six_i \leq x_j \\ -1 & six_i < x_j \end{cases}$$

et Y_{ij} défini de la même manière.

A partir de la définition 24 et du théorème 15, nous pouvons donner un estimateur $\hat{\beta}$ du paramètre β de la copule C_β:

$$\hat{\beta} = \tau^{-1}(\tau_{emp}).$$

D'après un résultat de Genest et Mackay ([55]), le τ de Kendall s'écrit:

$$\tau(C_\beta) = 4\int_0^1 \frac{\varphi_\beta(t)}{\varphi'_\beta(t)}dt + 1.$$

En se basant sur ce résultat, et en considérant la copule de Genest-Ghoudi (voir section 4.6.2) de fonction génératrice $\varphi_\beta(t) = \varphi_\beta^{-1}(t) = (1-t^\beta)^{1/\beta},\ t \in [0,1]$, nous avons :

$$\tau(C_\beta) = \frac{3\beta - 2}{\beta - 2}.$$

L'estimateur $\hat{\beta}$ de β est donc dans ce cas :

$$\hat{\beta} = \tau^{-1}(\tau_{emp}) = \frac{2 - 2\tau_{emp}}{3 - \tau_{emp}}.$$

Une estimation analogue basée sur le ρ de Spearman peut être définie.

4.7.3.2 Coefficient de corrélation de Spearman

Soient X et Y deux variables aléatoires de copule C_β et de fonctions de répartition $F(x)$ et $G(y)$. La mesure ρ de Spearman est le coefficient de corrélation usuel des variables aléatoires $U = F(X)$ et $V = G(Y)$ (U et V sont deux variables aléatoire uniformes sur $[0,1]$ et donc $E(U) = E(V) = \frac{1}{2}$ et $Var(U) = Var(V) = \frac{1}{12}$). Nous avons donc :

$$\rho(X,Y) = \frac{Cov(U,v)}{\sqrt{Var(U)}\sqrt{Var(V)}}.$$

Plus précisément, le ρ de Spearman se définit de la manière suivante:

Définition 25 (ρ de Spearman) :
Soit (X,Y) un vecteur de variables aléatoires continues. Le ρ de Spearman est défini par:

$$\rho(X,Y) = 3(\mathbb{P}\{(X-\tilde{X})(Y-Y') > 0\} - \mathbb{P}\{(X-\tilde{X})(Y-Y') < 0\}),$$

avec (X,Y), $(\tilde{X},\tilde{Y})$ et (X',Y') 3 copies indépendantes.

En utilisant le théorème 14, en notant que $\tilde{X}$ et Y' sont indépendants et que Q est symétrique ($Q(C,C') = Q(C',C)$), nous avons le résultat suivant emprunté au livre de Nelsen (1998, [80]:

Théorème 16 :
Soit (X,Y) un vecteur de variables aléatoires continues de copule C. Le ρ de Spearman est donné par:

$$\rho(X,Y) = 3Q(C,\Pi) = 12 \iint_{[0,1]^2} uvdC(u,v) - 3 = 12 \iint_{[0,1]^2} C(u,v)dudv - 3.$$

Si X est de loi F et Y de loi G, avec $U = F(X)$ et $V = G(Y)$

$$\begin{aligned}
\rho(X,Y) &= 12 \iint_{[0,1]^2} uvdC(u,v) - 3 \\
&= 12\mathbb{E}(UV) - 3 \\
&= \frac{\mathbb{E}(UV) - 1/4}{1/12} \\
&= \frac{Cov(U,V)}{\sqrt{Var(U)}\sqrt{Var(V)}}
\end{aligned}$$

De même que pour le coefficient de Kendall, si (X,Y) est de copule Archimédienne C_β de paramètre β, le ρ de Spearman $\rho(X,Y)$ peut être noté $\rho(C_\beta)$ ou $\rho(\beta)$. Le calcul de $\rho(\beta)$ est compliqué et le recours aux méthodes d'approximations numériques d'intégrales est nécessaire.

Nous pouvons également définir le ρ de Spearman empirique associé à un échantillon, donné dans le livre de Stuart et al. (1999, [104]).

Définition 26 :
Soient $(x_1,...,x_N)$ un échantillon de N réalisations de X et $(y_1,...,y_N)$ un échantillon de N réalisations de Y. L'estimateur empirique ρ_{emp} de Kendall est

$$\rho_{emp} = 1 - 6\sum_{i=1}^{N} \frac{D_i^2}{N(N^2-1)}$$

où $D_i$est la différence entre le rang de x_i et le rang de y_i.

A partir de la définition 26 et du théorème 16, nous pouvons donner un estimateur $\hat{\beta}$ du paramètre β de la copule C_β:

$$\hat{\beta} = \rho^{-1}(\rho_{emp}).$$

4.8 Estimation de copules Archimédiennes multivariées

Comme pour les copules Archimédiennes bivariées, les méthodes d'estimation des paramètres d'une n-copule Archimédienne (4.22) sont paramétriques ou non paramétriques. Dans les deux cas, les calculs sont complexes vu la forme des copules multivariées. Pour plus de clarté et sans perte de généralité, nous travaillons avec des copules tridimensionnelles (4.22):

$$\begin{aligned} C_{\beta_1,\beta_2}(u,v,w) &= \varphi_{\beta_2}^{-1}(\varphi_{\beta_2}(C_{\beta_1}(u,v)) + \varphi_{\beta_2}(w)) \\ &= \varphi_{\beta_2}^{-1}(\varphi_{\beta_2}(\varphi_{\beta_1}^{-1}(\varphi_{\beta_1}(u) + \varphi_{\beta_1}(v))) + \varphi_{\beta_2}(w)). \end{aligned}$$

Nous regardons tout d'abord des méthode d'estimation paramétrique basées sur la vraisemblance, puis nous présentons des méthodes non-paramétriques s'appuyant sur les coefficient de corrélation de Kendall et de Spearman.

4.8.1 Estimation d'une copule Archimédienne multivariée par maximum de vraisemblance

Dans le cas où les marginales unidimensionnelles sont connues, l'estimation est basée sur le maximum de vraisemblance. Elle est plus compliquée qu'en dimension 2. Les deux sections suivantes reprennent des résultats indiqués dans Hillali (1998, [59]), puis nous donnons dans la troisième section une méthode originale d'estimation.

4.8.1.1 Première méthode par maximum de vraisemblance

La première étape est l'initialisation de l'un des paramètres, β_1 ou β_2 par maximisation de la vraisemblance des lois marginales bidimensionnelles C_{β_1} ou C_{β_2}. L'initialisation de β_1 est :

Soit C'_{β_1} la densité de C_{β_1} absolument continue par rapport à la mesure de Lebesgue et concentrée sur le carré unité $[0,1]^2$. Soient $(u_1,...,u_N)$ et $(v_1,...v_N)$ deux réalisations de taille N des variables aléatoires U et V de copule C_{β_1} et de marginales uniformes. L' initialisation de β_1 se fait par maximisation de la vraisemblance ou de la log-vraisemblance comme dans le paragraphe 4.7.2. L'estimateur du maximum de vraisemblance noté β_1^0 est obtenu numériquement.

La deuxième étape consiste à estimer le paramètre β_2 en maximisant la vraisemblance des réalisations $(u_1,...,u_N)$, $(v_1,...v_N)$ et $(w_1,...,w_N)$ de loi $C_{\beta_1\ \beta_2}$. La densité de $C_{\beta_1\ \beta_2}$ est notée $C''_{\beta_1\ \beta_2}$ et le logarithme de la densité $C''_{\beta_1\ \beta_2}$ est noté $l(\beta_1,\beta_2)$. L'estimateur du maximum de vraisemblance de β_2 noté β_2^1 est solution de l'équation

$$\frac{\partial l(\beta_1^0,\beta_2)}{\partial \beta_2} = 0.$$

Lorsque β_2^1 est obtenu, le calcul de β_1^1 est réalisé par résolution de l'équation

$$\frac{\partial l(\beta_1,\beta_2^1)}{\partial \beta_1} = 0.$$

La procédure s'arrête quand on obtient les estimateurs du maximum de vraisemblance $\hat{\beta}_1$ et $\hat{\beta}_2$ solutions de de l'équation

$$\frac{\partial^2 l(\beta_1, \beta_2)}{\partial \beta_1 \partial \beta_2} = 0.$$

La méthode est coûteuse du fait du nombre d'itérations nécessaires à la convergence et des formules complexes.

Une autre méthode existe (donné dans Hillali, 1998, [59]), basée sur une procédure d'estimation itérative du maximum de vraisemblance. Nous rappelons cette méthode.

4.8.1.2 Deuxième méthode par maximum de vraisemblance

Les estimateurs du maximum de vraisemblance sont obtenus par les équations

$$\frac{\partial l(\beta_1, \beta_2)}{\partial \beta_1} = 0, \text{ et } \frac{\partial l(\beta_1, \beta_2)}{\partial \beta_2} = 0.$$

Soit $\hat{\theta}$ l'estimateur du maximum de vraisemblance de $\theta = (\beta_1, \beta_2)$ et $\theta^j = (\beta_1^j, \beta_2^j)$ la j^{ieme} solution des équations ci-dessus. La formule itérative s'écrit

$$\theta^{j+1} = \theta^j + I(\hat{\theta})^{-1} gradl(\theta^j)$$

avec la matrice d'information

$$I(\hat{\theta}) = \begin{pmatrix} -\frac{\partial^2 l(\beta_1,\beta_2)}{\partial^2 \beta_1} & -\frac{\partial^2 l(\beta_1,\beta_2)}{\partial \beta_1 \partial \beta_2} \\ -\frac{\partial^2 l(\beta_2,\beta_1)}{\partial \beta_2 \partial \beta_1} & -\frac{\partial^2 l(\beta_1,\beta_2)}{\partial^2 \beta_2} \end{pmatrix}$$

et $gradl(\theta) = (\frac{\partial l(\beta_1,\beta_2)}{\partial \beta_1}, \frac{\partial l(\beta_1,\beta_2)}{\partial \beta_2})$. La matrice d'information $I(\hat{\theta})$ est estimée par $I(\theta^j)$.

L'algorithme est le suivant:

1. Initialisation de β_1 et β_2 par β_1^0 et β_2^0.
   $\theta^0 = (\beta_1^0, \beta_2^0)$ et $j = 0$
2. Calculer $gradl(\theta^j)$
3. Calculer $I(\theta^j)$ (estimation de $I(\hat{\theta})$)
4. Calculer $\theta^{j+1} = \theta^j + I(\theta^j)^{-1} gradl(\theta^j)$
5. $j \longrightarrow j+1$ et retour à étape 2.
6. En cas de non convergence ou de convergence vers un maximum local, retour à l'étape 1.

Cette méthode est complexe à mettre en pratique. Nous proposons donc dans la section suivante, une méthode originale d'estimation des paramètres d'une n-copule. Cette méthode est utiliser en section 5.5 afin de considérer le cas multidimensionnel des mélanges de copules, par copules Archimédiennes multivariées.

4.8.1.3 Troisième méthode par maximum de vraisemblance

Soient U, V et W trois variables aléatoires de copule de la forme (4.22). A partir de la forme (4.22) (page 84) des n-copules, nous pouvons considérer la copule C_{β_1} de paramètre β_1 comme liant les deux variables aléatoires U et V. Nous avons donc β_1 assimilé à un coefficient d'association des variables aléatoires U et V. Nous pouvons donc estimer β_1, à partir des réalisations $(u_1, ..., u_N)$ et $(v_1, ..., v_N)$, de la même manière que dans la section 4.7.2 d'estimation de copules Archimédiennes bivariées par maximum de vraisemblance. L'estimation de la copule C_{β_2} se fait alors en considérant β_2 comme un coefficient d'association des variables aléatoires $C_{\beta_1}(U, V)$ et W. En effet, à la vue des copules (4.22), la copule C_{β_2} couple $C_{\beta_1}(U, V)$ et W. Disposant de la valeur $\hat{\beta}_1$ estimant β_1, nous pouvons calculer l'échantillon $(C_{\hat{\beta}_1}(u_1, v_1), ...C_{\hat{\beta}_1}(u_N, v_N))$ et l'utiliser avec $(w_1, ..., w_N)$ pour estimer β_2 par maximum de vraisemblance.

Regardons les méthodes d'estimation par coefficients de corrélation.

4.8.2 Estimation d'une copule Archimédienne multivariée par τ de Kendall et ρ de Spearman

Les mesures de dépendance de Kendall et de Spearman s'étendent à toutes les familles finies de variables aléatoires $X_1, ..., X_n$. Soit nous utilisons $2^n - 1$ mesures de dépendance pour tenir compte de l'ensemble des couples de variables aléatoires, soit nous utilisons une unique mesure collective. Nous nous limitons à ce dernier cas. Soit C_n la copule de dimension n du vecteur aléatoire $(X_1, ..., X_n)$ de lois marginales unidimensionnelles $F_1, ..., F_n$ et de loi jointe H telle que:

$$H(x_1, ..., x_n) = C_n(F_1(x_1), ..., F_n(x_n)). \tag{4.29}$$

4.8.2.1 Première méthode par τ de Kendall

Le τ de Kendall d'une copule bidimensionnelle se généralise à une copules multidimensionnelle ainsi:

$$\begin{aligned} \tau(C_n) &= \frac{1}{2^{n-1}-1}(2^n \int C_n(F_1(x_1), ..., F_n(x_n)) dC_n(F_1(x_1), ..., F_n(x_n)) \ - 1) \\ &= \frac{1}{2^{n-1}-1}(2^n \int C_n(u_1, ..., u_n) dC_n(u_1, ..., u_n) \ - 1). \end{aligned} \tag{4.30}$$

Pour plus de pratique, nous étudions le cas de l'estimation de paramètres de copules Archimédiennes à 3 dimensions du type (4.22). La procédure suivante est proposée dans Hillali (1998, [59]).

Soit T le coefficient de Kendall généralisé estimé à partir des N réalisations des trois variables aléatoires U, V et W, $(u_1, ..., u_N)$, $(v_1, ..., v_N)$ et $(w_1, ..., w_N)$ de copule C_{β_1,β_2}. Le coefficient T est calculé par le coefficient moyen de (U, V), (U, W) et (V, W):

$$T = \frac{1}{3}(\tau_{Emp}(UV) + \tau_{Emp}(UW) + \tau_{Emp}(VW)). \tag{4.31}$$

Estimer β_1, β_2 est trouver deux paramètres $\hat{\beta}_1$, $\hat{\beta}_2$ tels que

$$\tau(C_{\hat{\beta}_1,\hat{\beta}_2}) = T. \tag{4.32}$$

L'équation (4.32) ne peut être résolue directement car $\hat{\beta}_1$ et $\hat{\beta}_2$ sont inconnus. Cependant, le paramètre $\hat{\beta}_1$ peut être interprété comme le coefficient d'association de U et V ayant pour copule C_{β_1}. En se basant sur le chapitre 4.7.3, un estimateur non paramétrique $\hat{\beta}_1$ de β_1 est :

$$\hat{\beta}_1 = \tau^{-1}(\tau_{Emp}(UV)), \tag{4.33}$$

avec $\tau_{Emp}(UV)$ le coefficient de Kendall entre U et V estimé à partir des réalisations $(u_1, ..., u_N)$ et $(v_1, ..., v_N)$. L'estimateur $\hat{\beta}_2$ de β_2 est donc la solution de l'équation :

$$\tau(C_{\hat{\beta}_1,\beta_2}) = T.$$

Dans cette méthode, le coefficient $\tau(C_{\beta_2})$ ne dépend que de β_2, il est indépendant du choix des lois de (U, W) ou (V, W). On ne peut donc pas estimer β_2 à partir de $\tau(C_{\beta_2})$ et ensuite estimer β_1 par l'équation $\tau(C_{\beta_1,\hat{\beta}_2}) = T$. Le coefficient de Kendall $\tau(C_{\hat{\beta}_1})$, estimé à partir des réalisations dépend des variables aléatoires choisies.

4.8.2.2 Deuxième méthode par τ de Kendall

Nous proposons une méthode originale d'estimation des paramètres β_1 et β_2, basée sur le τ de Kendall. Comme dans la première méthode d'estimation basée sur le τ de Kendall, le paramètre β_1 est interprété comme coefficient d'association des deux variables aléatoires U et V de copule C_{β_1}. Le paramètre β_1 peut donc être estimé par l'équation (4.33). Le paramètre β_2 peut être interprété comme le coefficient d'association des variables aléatoires $Z = H_{\beta_1}(U, V)$ et W. En posant $z_i = H_{\beta_1}(u_i, v_i)$, $i = 1, ..., N$, l'estimation $\hat{\beta}_2$ de β_2 est solution de l'équation :

$$\tau(C_{\beta_2}) = \tau_{emp}(ZW) \tag{4.34}$$

La méthode présentée est déduite de la méthode de généralisation des copules Archimédiennes multidimensionnelles. Elle est donc mieux adaptée à l'estimation des paramètres des copules de la forme (4.22). La méthode se généralise aux vecteurs aléatoires $(U_1, ..., U_n)$ de dimension $n \geq 3$ en interprétant les paramètres β_i comme des coefficients d'association des variables aléatoires $C_{\beta_1,...,\beta_{i-1}}(U_1, ..., U_i)$ et U_{i+1}.

4.8.2.3 Première méthode par ρ de Spearman

Le coefficient de corrélation ρ de Spearman se généralise en dimension n:

$$\begin{aligned} \rho(C_n) &= \frac{1}{(n+1)^{-1} - 2^{-n}} \left(\int F_1(x_1)...F_n(x_n) dC_n(F_1(x_1), ..., F_n(x_n)) \; - 2^{-n} \right) \\ &= \frac{1}{(n+1)^{-1} - 2^{-n}} \left(\int_{[0,1]^n} u_1...u_n dC_n(u_1, ..., u_n) \; - 2^{-n} \right) \end{aligned} \tag{4.35}$$

La méthode par ρ de Spearman ne diffère pas de la première méthode par τ de Kendall.

Soient $(u_1, ..., u_N)$, $(v_1, ..., v_N)$ et $(w_1, ..., w_N)$ des réalisations des variables aléatoires U, V et W de copule C_{β_1,β_2}). Le coefficient de corrélation de Spearman est :

$$\rho_{emp} = \frac{n^{-1} \sum_i R_i^1 R_i^2 R_i^3 - ((n+1)/2)^3}{n^{-1} \sum_i i^3 - ((n+1)/2)^3}, \tag{4.36}$$

avec R_i^1, R_i^2 et R_i^3 les rangs respectifs de u_i, v_i et w_i (voir Joe, 1993, [65]).

Soit $\rho(C_{\beta_1,\beta_2}$ le coefficient de corrélation de Spearman théorique. des variables aléatoires U, V et W. Les estimateurs $\hat{\beta}_1$, $\hat{\beta}_2$ de β_1, β_2 sont solutions de l'équation :

$$\rho(C_{\beta_1,\beta_2}) = \rho_{emp}. \tag{4.37}$$

Il n'est pas possible de résoudre cette équation, les deux paramètres β_1 et β_2 étant inconnus. Les estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$ sont calculés en faisant appel aux lois marginales de C_{β_1,β_2}.

Soit $\rho(C_{\beta_1}$ le coefficient de corrélation de Spearman associé à U et V de copule $C_{\beta_1})$.Un estimateur $\hat{\beta}_1$ de β_1 est solution de l'équation

$$\rho(C_{\beta_1}) = \rho_{emp}(UV), \tag{4.38}$$

avec $\rho_{emp}(UV)$ le coefficient de corrélation de Spearman calculé à partir des réalisations de U et V. Un estimateur de β_2 est donc solution de :

$$\rho(C_{\hat{\beta}_1,\beta_2}) = \rho_{emp}.$$

4.8.2.4 Deuxième méthode par ρ de Spearman

Nous proposons une méthode d'estimation originale des paramètres β_1 et β_2 par le coefficient de corrélation de Spearman.

Le paramètre β_1 est vu comme coefficient d'association de U et V. L'estimateur $\hat{\beta}_1$ est donc solution de (4.38). Le paramètre β_2 peut être vu comme coefficient d'association des variables aléatoires $Z = C_{\beta_2}(U, V)$ et W. L'estimateur $\hat{\beta}_2$ de β_2 est donc solution de :

$$\rho(C_{\beta_2}) = \rho_{emp}(ZW),$$

avec $\rho_{emp}(ZW)$ le coefficient de corrélation de Spearman calculé à partir des $(z_1, ..., z_N)$ et $(w_1, ..., w_N)$ et $z_i = C_{\beta_2}(u_i, v_i)$.

La méthode proposée se généralise à la dimension n sur le vecteur $(U_1, ..., U_n)$ par le calcul des ρ de Spearman des variables aléatoires $C_{\beta_1,...,\beta_{i-1}}(U_1, ..., U_i)$ et U_{i+1}.

4.9 Conclusion

Nous avons vu que les fonctions copules permettent de caractériser le lien entre la fonction de répartition multidimensionnelle et les marginales d'un n-uplet de variables aléatoires. Ce lien est défini dans le théorème de Sklar. Nous avons vu également que les propriétés des copules bivariées ne sont pas tout à fait les mêmes que celles des copules de dimension supérieure à trois. Ces dernières sont d'ailleurs souvent délicates à traiter. La classe des copules Archimédiennes autorise cependant deux généralisations en dimension n. Cette famille est paramétrique et nous avons présenté différentes méthodes pour l'estimation du paramètre d'une copule Archimédienne. Ces méthodes sont basées soit sur la vraisemblance, soit sur le coefficient de corrélation de Kendall ou celui de Spearman. A partir d'un type de généralisation de copules Archimédiennes (la forme 4.22), nous avons proposé trois méthodes d'estimation

des paramètres associés (voir les sections 4.8.1.3, 4.8.2.2 et 4.8.2.4).

Nous pouvons donc définir la relation entre une loi n-dimensionnelle et ses marginales. A partir de cette connaissance (due au théorème de Sklar) et de la notion de distribution de distributions, nous pouvons étendre la décomposition de mélange de lois aux données fonctions de répartition. Le chapitre suivant propose une approche le permettant.

Chapitre 5

Décomposition de mélange de copules

" Plus les mathématiques évoluent plus elles deviennent abstraites - et par conséquent peut-être aussi plus pratiques."
(Eric Temple Bell, 1883-1960)

5.1 Introduction

Les méthodes de décomposition de mélange de lois de probabilité cherchent à estimer la densité d'une variable aléatoire à partir d'un échantillon de N réalisations. La densité est supposée être une somme pondérée de K densités appartenant à une famille paramétrique de densités (voir le chapitre 2) :

$$f(x) = \sum_{k=1}^{K} p_k f_k(x, \alpha_k).$$

L'approche que nous avons ici est l'extension de cette problématique au cas où les variables aléatoires décrivant les individus sont à valeurs dans un espace de fonctions de répartition.

Notre but est donc d'estimer la loi de probabilité d'un ensemble d'individus décrits par des fonctions de distribution et de classer ces individus.

La méthode que nous proposons, travaille donc sur la notion de "fonction de distribution de distributions" développée au chapitre 3. Cette méthode permet de tenir compte des dépendances qui existent entre les variables et des liens qui existent à l'intérieur même d'une variable entre différents points de sa fonction de répartition. Ces dépendances intra et inter variables sont prises en compte à l'aide des fonctions copules (ou fonctions de dépendance, voir le chapitre 4).

La première méthode étendant la décomposition de mélange de densités est proposée par Diday dans son article de 2001, [33]. Cette méthode travaille sur des FDD empiriques avec un modèle binaire de copule C ($C = Min$ ou $C = M$) défini par un seuil ϵ. L'approche de Diday est du type "approche classification" et est développée dans le cas d'une unique

variable aléatoire à valeurs fonctions de répartition, avec 2 valeurs T_1 et T_2 données sur 5 individus. Le principe de cette méthode est le suivant (le nombre K de classes étant donné) :

- calcul des fonctions de distribution de distributions,
- calcul de leur dépendance à l'aide de copules,
- association d'une copule à chaque classe, puis d'une classe à chaque copule par un procédé de type nuées dynamiques.

Une autre méthode travaillant sur le même type de données est développée par Emilion (2002, [41]). Cette approche consiste en une discrétisation des fonctions de répartition (i.e. en histogrammes classiques). Le nombre K de classes est donné et les données discrétisées sont ensuite utilisées dans un algorithme de type "classification" (nuées dynamiques) ou "estimation" (EM et ses variantes) avec une loi de Dirichlet multidimensionnelle. Cette loi, travaillant sur $[0,1]$, s'applique bien aux données histogrammes. En sortie, la méthode donne les paramètres de la loi de Dirichlet ainsi que la classification des individus en K classes.

Dans ce chapitre, nous proposons de généraliser le principe de Diday (2001, [33]) sur des modèles continues de FDD (tel que la loi béta), sur des modèles paramétriques non binaires de copules continues (les copules Archimédiennes), ainsi qu'à la plupart des méthodes de décomposition de mélange de densités (EM, SEM, etc). Par ailleurs, les méthodes sont également développées dans le cas multidimensionnel (pour plus d'une unique variable de type fonctions de répartition" et/ou avec plus de 2 valeurs T_1 et T_2). Pour réaliser cette généralisation, nous proposons une méthode par copule empirique, une méthode par copules Archimédiennes multivariées, et deux méthodes par couplage de copules Archimédiennes bivariées.

Nous disposons d'un échantillon appelé "base de distributions" et noté $\mathfrak{F} = (F_1, ..., F_N)$ avec $F_i = (F_i^1, ..., F_i^p)$ et F_i^j la fonction de répartition décrivant l'individu i pour la variable j. L'échantillon peut être résumé par un tableau de données où chaque case (intersection d'un individu i et d'une variable j) contient une fonction de répartition F_i^j (voir Figure 5.1). Chaque F_i^j correspond à une estimation de la fonction de répartition de l'individu i

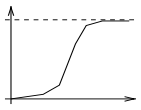
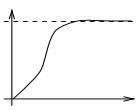
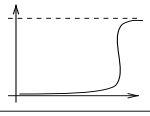
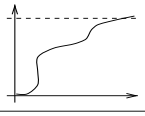
	Variable 1	.....	Variable p
Ind 1		.....	
.....	.....	.....	.....
Ind N		.....	

FIG. 5.1: *Tableau de données fonctions de distribution*

pour la variable j et $F_i = (F_i^1, ..., F_i^p)$ est donc l'ensemble des estimations des fonctions de répartition décrivant l'individu i. La j^{eme} variable peut donc être vue comme une variable aléatoire à valeurs fonctions de répartition, autrement dit comme une variable aléatoire à valeurs variables aléatoires.

Nous voulons décomposer la loi de probabilités des données (et obtenir ainsi une classification en K classes des individus de $\mathfrak{F}$) et modéliser les dépendances qui existent dans une variable et entre les variables. Nous obtiendrons donc:

- une partition $P = (P_1, ..., P_K)$ de l'échantillon,
- la densité estimée des variables aléatoires à valeurs fonctions de répartition dont nous donnons la forme,
- une mesure de dépendance entre les variables utilisées.

Dans une première partie, nous regardons comment associer un objet symbolique à un ensemble de données fonctions de répartition. Nous proposons ensuite des extensions de la plupart des méthodes de décomposition de mélange de densités pour de telles données dans un cas bidimensionnel, puis nous regardons les développements d'approches multidimensionnelles. Différents points théoriques sont enfin discutés, tels que le fait que la méthode proposée de décomposition de mélange est une généralisation des méthodes classiques.

5.2 Objets symboliques associés à un ensemble de fonctions de répartition

Dans cette section, nous supposons que nous disposons d'un ensemble $\Omega = (w_1, ..., w_N)$ d'individus décrit par une base de distributions $F = (F_1, ..., F_N)$. Ayant définit n valeurs T_1,...,T_n, nous pouvons définir n fonctions de distribution de distributions G_{T_1},...,G_{T_n} (voir chapitre 3). Les variables aléatoires X_1,...,X_n, caractérisées par G_{T_1},...,G_{T_n}, sont de loi jointe H.

La proposition suivante est la clé de la relation entre les FDD et les copules.

Proposition 6 *:*
Soit une fonction de distribution jointe de distributions en $T = T_1, ..., T_n$, de FDD marginales $G_{T_1}, ..., G_{T_n}$.
Il existe une copule C telle que pour tout x_1,...,x_n appartenant à $[0, 1]$,

$$H(x_1, ..., x_n) = C(G_{T_1}(x_1), ..., G_{T_n}(x_n)).$$

La démonstration de cette proposition vient directement du théorème de Sklar, en utilisant le fait que les FDD sont des fonctions de répartition.

La fonction de distribution jointe de distributions H (caractérisée par C, et G_{T_1},..., G_{T_n}) est une mesure de généralisation stochastique (c'est-à-dire pour des données probabilistes). En effet, à partir de H (voir la définition 4), plusieurs objets symboliques modaux associés à des données probabilistes peuvent être définis.

Ces objets formalisent la notion de concept que nous rappelons pour mémoire. Un concept est défini par une intention (également appelée sa description, i.e. ses propriétés caractéristiques) et une extension (les individus qui vérifient ces propriétés). Dans notre étude, les individus définissent la base de distributions et sont supposés satisfaire les propriétés d'un concept donné. Par exemple les individus sont des profils atmosphériques décrits par des variables thermodynamiques (e.g. la fonction de répartition de la température, de la pression, etc) ; le concept est le type de masse d'air de ces profils atmosphériques. Plus formellement, si C est un type de masse d'air, $Extent(C)$ est l'ensemble des profils vérifiant les propriétés de cette masse d'air, $Intent(C) = d_C$ est la description des propriétés la masse d'air. Un objet symbolique (voir chapitre 1), est un modèle mathématique pour un concept C, et est défini par un triplet $s = (a, R, d)$ avec:

- d_c est la fonction de distribution jointe de n distributions : $G_{T_1,...,T_n}$,
- R est une relation binaire entre les descriptions telle que la valeur $[dRd'] \in [0,1]$ mesure le degré avec lequel d' est en relation avec d,
- a mesure l'adéquation entre un individu w et le concept C. Il s'agit d'une application de Ω (l'ensemble des individus) dans $[0,1]$ telle que $a(x) = [Y(w)\ R\ d_C]$ avec w un individu et $Y(w)$ la fonction de distribution jointe de n distributions pour un ensemble d'individus réduit à $\{w\}$.

L'adéquation entre w et C peut être mesurée par la "densité des distributions" autour de $Y(w)$ parmi l'ensemble $\{Y(w') \ / \ Y(w') \in Extent(C)\}$. Dans ce cadre, la densité peut se définir de deux manières : en utilisant une approximation de la dérivée d'une copule ou en utilisant la dérivée d'une copule quand la dérivée existe. L'adéquation entre w et C peut aussi être mesurée par comparaison de la fonction de distribution jointe de n distributions associée à l'individu w et celle associée au concept C.

L'approximation de la dérivée d'une copule C passe par le calcul d'un C-volume (défini au chapitre 4) que nous pouvons nommer "volume d'hyper-cube". Pour des questions pratiques nous considérons le cas $n = 2$.

5.2.1 Objet symbolique par volume d'hyper-cube

Cette modélisation d'un objet symbolique nous est donnée dans Diday, 2001, [33].
Nous définissons un objet symbolique $s = (a, R(\eta), d)$ tel que

$$a(w, \eta) = [Y(w)\ R(\eta)\ C(G_{T_1}, G_{T_2})] \in \mathbb{R}^+$$

mesurant l'adéquation entre une fonction de répartition $L = Y(w)$ et la copule $C(G_{T_1}, G_{T_2}) = H_{T_1,T_2}$ associée à base de distribution $\mathfrak{F}$. Avec $x_i = L(T_i)$ (la valeur de la fonction de répartition L en T_i), et $\eta = (\epsilon_1, \epsilon_2)$, la relation $R(\eta)$ est définie par le volume de l'hyper cube $[x_1 - \epsilon_1, x_2 - \epsilon_2] \times [x_1 + \epsilon_1, x_2 + \epsilon_2]$ pour $C(G_{T_1}, G_{T_2})$:

$$\begin{aligned} L\ R(\eta)\ C(G_{T_1}, G_{T_2}) \quad = \quad & C(G_{T_1}(x_1 + \epsilon_1), G_{T_2}(x_2 + \epsilon_2)) \\ & -C(G_{T_1}(x_1 + \epsilon_1), G_{T_2}(x_2 - \epsilon_2)) \\ & -C(G_{T_1}(x_1 - \epsilon_1), G_{T_2}(x_2 + \epsilon_2)) \\ & +C(G_{T_1}(x_1 - \epsilon_1), G_{T_2}(x_2 - \epsilon_2)). \end{aligned} \tag{5.1}$$

Cette définition entraine la proposition suivante.

Proposition 7 (Diday, [33]) *:*
$a(w,\eta) = [Y(w)\ R(\eta)\ C(G_{T_1}, G_{T_2})]\ \in\ [0,1]$.

Démonstration :
Nous avons :

$$\begin{aligned} a(w,\eta) &= H(x_1+\epsilon_1, x_2+\epsilon_2) - H(x_1+\epsilon_1, x_2-\epsilon_2) \\ &\quad -H(x_1-\epsilon_1, x_2+\epsilon_2) - H(x_1-\epsilon_1, x_2-\epsilon_2) \\ &= \mathbb{P}(\{F_i\ \in\ F\ /\ F_i(T_1) \leq x_1+\epsilon_1\}\ \cap\ \{F_i\ \in\ F\ /\ F_i(T_2) \leq x_2+\epsilon_2\}) \\ &\quad -\mathbb{P}(\{F_i\ \in\ F\ /\ F_i(T_1) \leq x_1+\epsilon_1\}\ \cap\ \{F_i\ \in\ F\ /\ F_i(T_2) \leq x_2-\epsilon_2\}) \\ &\quad -\mathbb{P}(\{F_i\ \in\ F\ /\ F_i(T_1) \leq x_1-\epsilon_1\}\ \cap\ \{F_i\ \in\ F\ /\ F_i(T_2) \leq x_2+\epsilon_2\}) \\ &\quad +\mathbb{P}(\{F_i\ \in\ F\ /\ F_i(T_1) \leq x_1-\epsilon_1\}\ \cap\ \{F_i\ \in\ F\ /\ F_i(T_2) \leq x_2-\epsilon_2\}) \\ &= \mathbb{P}(\{F_i\ \in\ F\ /\ x_1-\epsilon_1 \leq F_i(T_1) \leq x_1+\epsilon_1\} \cap \\ &\quad \{F_i\ \in\ F\ /\ x_2-\epsilon_2 \leq F_i(T_i)x_2+\epsilon_2\} \end{aligned}$$

et donc $a(w,\eta)\ \in\ [0,1]$.

Ce formalisme implique cependant de donner un seuil $\eta = (\epsilon_1, \epsilon_2)$.

Dans le cas où la dérivée de la copule C existe, nous proposons une autre manière de calculer l'objet symbolique, qui se passe de seuil.

5.2.2 Objet symbolique par dérivée de copule

L'idée principale est d'utiliser la densité de la copule décrivant l'ensemble d'individus pour mesurer le degré d'adéquation entre l'ensemble et un nouvel individu.

Soit h la densité associée à la copule C:

$$h(u,v) = \frac{\partial^2 C(u,v)}{\partial u \partial v}.$$

En effet, les copules étant des fonctions de répartition, nous pouvons définir les densités leurs correspondant.

La densité h induit deux méthodes pour le calcul d'un objet symbolique.

5.2.2.1 Par densité normée

Nous posons

$$M = max_{x_1,x_2} h(x_1, x_2),$$

et

$$a(w) = \frac{h(F_w(T_1), F_w(T_2))}{M},$$

avec w un nouvel individu. Cette normalisation nous permet d'avoir $a(w) \in [0,1]$ et qui vaut 1 quand les valeurs des fonctions de répartition en (T_1, T_2) de w sont dans le maximum de la densité de l'ensemble.

5.2.2.2 Par intégration de densité

Nous étendons l'idée de la section 5.2.1 de calcul d'objet symbolique par volume d'hypercube. La densité de la copule étant connue, nous pouvons poser un seuil $\eta = (\nu_1, \nu_2)$. Avec un nouvel individu w de valeurs $x_1 = F_w(T_1)$ et $x2 = F_w(T_2)$ nous pouvons calculer la probabilité d'avoir des valeurs dans $[x_1 - \nu_1, x_1 + \nu_1] \times [x_2 - \nu_2, x_2 + \nu_2]$ par :

$$\mathbb{P}(\{X_1 \in [x_1 - \nu_1, x_1 + \nu_1]\} \cap \{X_2 \in [x_2 - \nu_2, x_2 + \nu_2]\}) = \int_{x_2-\nu_2}^{x_2+\nu_2} \int_{x_1-\nu_1}^{x_1+\nu_1} h(x_1, x_2) dx_1 dx_2$$

Pour η donné, nous posons

$$a(w) = \mathbb{P}(\{X_1 \in [x_1 - \nu_1, x_1 + \nu_1]\} \cap \{X_2 \in [x_2 - \nu_2, x_2 + \nu_2]\}).$$

L'objet symbolique peut également être calculé sans la dérivée de la copule, en regardant la distance entre deux fonctions de répartition.

5.2.3 Objet symbolique par distance entre fonctions de répartition jointes

Notons $H_{T_1,...,T_n,w}$ la distribution jointe de n distributions associée à la base $F^* = \{Y(w)\}$ réduite à $Y(w)$) notée F_w. Soit $G_T = (G_{T_1}, ..., G_{T_n})$. Nous pouvons définir l'objet symbolique associé à $\mathfrak{F}$ par $s = (a, R, C(G_T))$ avec $a(w) = [H_{T_1,...,T_k,w} RC(G_T)]$ avec R mesurant le degré avec lequel la distribution jointe de n distributions $H_{T_1,...,T_n,w}$ est en relation avec la distribution jointe de n distributions associée à $\mathfrak{F}$. La mesure R peut être choisi parmi les dissimilarités entre fonctions de répartition étendues aux jointes (voir Kullback-Leibler [71], ou Tassy-Legait (1990), [107]).

Proposition 8 (Diday, [33]) *:*
La n-copule C associée à $H_{T_1,...,T_n,w}$ satisfait les propriétés

- *le domaine de $H_{T_1,...,T_n,w}$ est $\{0, 1\}$,*
- *$C = Min$ ou $C = \Pi$.*

Démonstration :
Par définition d'une "distribution jointe de n distributions" :

$$H_{T_1,...,T_n,w}(x_1, ..., x_n) = \mathbb{P}(\{F_w(T_1) \leq n_1\} \cap ... \cap \{F_w(T_n) \leq x_n\}) \tag{5.2}$$

vaut 1 si pour tout i nous avons $F_w(T_i) \leq x_i$ et 0 s'il existe i tel que $F_w(T_i) > x_i$. Autrement dit,

$$H_{T_1,...,T_n,w}(x_1, ..., x_n) = \begin{cases} 1 & si\ pour\ tout\ i,\ G_{T_i}(x_i) = 1, \\ 0 & si\ \exists\ i\ tel\ que\ G_{T_i}(x_i) = 0. \end{cases}$$

De plus, la copule C est la copule produit ou la copule Min, qui sont équivalentes dans ce cas.

Dans les méthodes proposées, il est nécessaire de donner différents T pour l'estimation des fonctions de distribution de distributions. Le choix des T est complexe car il est évident que les objets symboliques (et les résultats de la méthode détaillée plus tard en chapitre 5.4) ne sont pas le mêmes selon les T. Nous présentons deux approches permettant d'avoir un aperçu de la qualité des T.

5.3 Le choix des T_i

Il est évident qu'un T donné est mauvais si tous les F_j de $\mathfrak{F}$ ont la même valeur en T_i. Le T_i est aussi mauvais si tous les $F_j(T_i)$ sont uniformément distribués sur $[0,1]$. Nous pouvons dire d'un T_i est bon si des classes distinctes de valeurs structurent l'ensemble des valeurs $\{F_j(T_i) \ / \ j = 1, ..., N\}$. Jain et Dubes (1988, [64]) présentent différentes méthodes pour faire apparaître des tendances de "clustering". Dans cette partie, nous proposons deux méthodes pour un ensemble de valeurs appartenant à $[0,1]$.

5.3.1 Méthode des triangles

Dans [33] est donnée une méthode basée sur le nombre de triangles dont les sommets sont des points de $[0,1]$. L'ensemble des triangles dont les deux côtés de longueurs les plus proches sont les plus grands (respectivement plus petit) que le côté restant, est noté A (respectivement B). Nous définissons l'hypothèse H^0 qu'il y a une tendance de clustering par la distribution d'une variable aléatoire X^0 qui associe à N points aléatoirement distribués sur $[0,1]$ la valeur :

$$X^0 = \frac{|A| - |B|}{C_N^3} = \frac{6(|A| - |B|)}{N(N-1)(N-2)}$$

appartenant à $[-1,1]$, le cardinal de A étant $|A|$. Pour l'ensemble $U = \{F_j(T_i) \ / \ j = 1, ..., N\}$ de points, l'article [33] suggère de calculer le nombre de triangles appartenant à $A(U)$ ou $B(U)$, les sommets des triangles étant les points de U. A partir de la loi de X^0, la valeur $(|A(U)| - |B(U)|)/C_N^3$ peut rejeter ou accepter l'hypothèse nulle pour un niveau donné.

Nous pouvons cependant remarquer que les triangles formés par l'ensemble $\{F_j(T_i) \ / \ j = 1, ..., N\}$ sont des triangles plats et donc les trois sommets des triangles sont alignés. A partir de cette remarque nous déduisons :

Proposition 9 *:*
La variable aléatoire $|A|$ est de loi binomiale de paramètres $(C_N^3, \frac{2}{3})$. L'espérance de la variable aléatoire X^0 est $\frac{1}{3}$ et sa variance est $\frac{8}{9C_N^3}$.

Démonstration :
Soient a_1, a_2 et a_3 trois points de $[0,1]$ dont on veut savoir si le triangle formé appartient à A ou B. L'appartenance à A ou B dépend de la position du point intermédiaire par rapport aux points des extrémités (Figure 5.2).

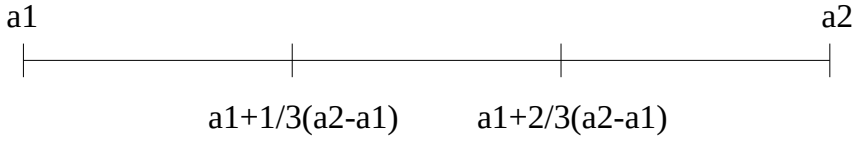


FIG. 5.2: *Position des points*

- Si $a_3 \in [a_1, a_1 + \frac{1}{3}(a_2 - a_1)]$, les deux côtés de longueurs les plus proches sont a_1a_2 et a_2a_3. Ces côtés sont plus grands que le côté a_1a_3 : le triangle est dans A.

- Si $a_3 \in [a_1 + \frac{1}{3}(a_2 - a_1), a_1 + \frac{2}{3}(a_2 - a_1)]$, les deux côt'es de longueurs les plus proches sont a_1a_3 et a_2a_3. Ces côtés sont plus petits que le côté a_1a_2 : le triangle est dans B.
- Si $a_3 \in [a_1 + \frac{2}{3}(a_2 - a_1), a_2]$, les deux côtés de longueurs les plus proches sont a_1a_3 et a_1a_2. Ces côtés sont plus grands que le côté a_2a_3 : le triangle est dans A.

Un triangle a donc $\frac{2}{3}$ de chances d'être dans A et $\frac{1}{3}$ de chances d'être dans B. Nous pouvons donc définir la variable aléatoire T caractérisant l'évènement "tirer un triangle appartenant à A", de loi de Bernoulli de paramètre $p = \frac{2}{3}$ =probabilité d'avoir un triangle dans A :

$$T = \begin{cases} 1 & \textit{si le triangle est dans } A, \\ 0 & \textit{sinon.} \end{cases}$$

Nous avons donc C_N^3 tirages aléatoires de triangles et autant de variables aléatoires $(T_i)_{i=1,\dots C_N^3}$. Le cardinal de A est $|A| = \sum_{i=1}^{C_N^3} T_i$. La variable aléatoire $|A|$ est donc de loi binomiale de paramètres (C_N^3, p). Son espérance est donc $\mathbb{E}(|A|) = C_N^3 p$, sa variance $var(|A|) = C_N^3 p(1-p)$. L'espérance de X^0 est donc:

$$\begin{aligned} \mathbb{E}(X^0) &= \mathbb{E}(\frac{|A| - |B|}{C_N^3}) \\ &= \mathbb{E}(\frac{|A| - (C_N^3 - |A|)}{C_N^3}) \\ &= \mathbb{E}(\frac{2|A|}{C_N^3}) - 1 \\ &= \frac{2}{C_N^3}\mathbb{E}(|A|) - 1 \\ &= 2p - 1 \\ &= \frac{1}{3} \end{aligned} \tag{5.3}$$

et sa variance est :

$$\begin{aligned} var(X^0) &= var(\frac{|A| - (C_N^3 - |A|)}{C_N^3}) \\ &= var(\frac{2|A|}{C_N^3}) \\ &= \frac{4}{(C_N^3)^2} var(|A|) \\ &= \frac{4}{(C_N^3)^2} C_N^3 p(1-p) \\ &= \frac{4}{C_N^3}\frac{2}{3}\frac{1}{3} \\ &= \frac{8}{9C_N^3}. \end{aligned} \tag{5.4}$$

Cette méthode implique une combinatoire importante et n'a malheureusement pas donnée d'excellents résultats pour le moment. Le choix des T utilisés dans l'application climatique du chapitre 6 a été effectué avec plus de succées par la méthode qui suit. Celle-ci est basée sur le calcul des fonctions de distribution de distributions et de leur dérivée.

5.3.2 Par estimation de densité et surface de densités

La répartition des valeurs de fonctions de répartition $\{F_i(T)\}_{i=1,\ldots,N}$ selon différents T peut donner des informations sur les T à prendre. Les méthodes non-paramétriques d'estimation de densité peuvent donc nous fournir une visualisation de la qualité des T. Si la densité des valeurs pour un T donné est bimodale, nous pouvons supposer que les données se scindent en deux classes selon T. La méthode par estimation de densité est une méthode visuelle qui ne donne pas d'indice numérique de la qualité de T.

De manière plus générale, nous pouvons aussi tracer la surface de densités de distributions de l'ensemble de données par estimation non paramétrique. Le tracé de la surface 3-D peut faire émerger des classes pour différents T. De plus, la méthode donne des informations d'ensemble pour tous les T, la surface étant tracée sur tout le domaine de la variable à l'étude. De même que pour la méthode par estimation de densité, le résultat est visuel et ne donne pas d'indice de la qualité de T. Cependant, les "pics" présents dans la surface de densités de distributions nous indiquent des valeurs "clés" pour la variable considérée et donc un choix de T intéressant. Par exemple, regardons la surface de densités de distributions décrivant des valeurs d'humidité pour différents sites géographiques.

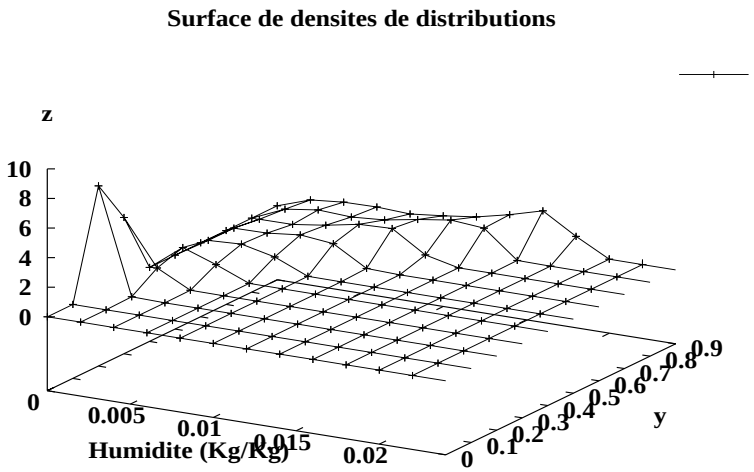


FIG. 5.3: *Exemple de surface de densités de distributions d'humidité*

La Figure 5.3 nous permet de voir que l'humidité globale est plutôt bien répartie. Cependant nous observons un pic pour des valeurs faibles d'humidité. Cette information nous permet par exemple de dire qu'il est nécessaire de choisir au moins l'un des T avec une faible valeur (ce qui est fait avec $T_1 = 0.000003$ Kg/Kg, valeur de l'humidité pour le pic). Cet exemple montre bien l'intérêt de considérer chaque application comme un cas particulier, la seconde valeur T_2 doit ici être déterminée par des connaissances a priori d'experts sur la structure des données (ce qui est fait avec $T_2 = 0.006$ Kg/Kg, voir section 6.2.2 du chapitre 6).

Nous avons définit les notions importantes de fonctions de distribution de distributions et de copules. Ces notions nous permettent d'étendre, pour des données probabilistes, les différentes approches de la décomposition de mélange de lois (présentées au chapitre 2). Ces extensions sont tout d'abord proposées en 2 dimensions pour simplifier les notations puis développées en dimension n. Deux variantes pour traiter le cas multidimensionnel sont également proposées.

5.4 Décomposition de mélange de copules (2-D)

Définissons deux valeurs T_1 et T_2. Pour chaque valeur, une FDD (Fonction de Distribution de Distributions) est définie, G_{T_1} en T_1, G_{T_2} en T_2. Les FDD sont des fonctions de répartition de variables aléatoires sur $[0,1]$ que nous notons X_1 (pour G_{T_1}) et X_2 (pour G_{T_2}). Nous notons H la loi jointe de (X_1, X_2). Nous avons vu au paragraphe 3.2 que l'estimation des FDD se résume à l'estimation d'une fonction de répartition. Pour modéliser la dépendance entre les FDD, nous utilisons les fonctions copules (voir le chapitre 4). Une copule (Schweizer et Sklar ,1983, [95]) est une fonction mettant en relation la fonction de répartition multidimensionnelle avec chacune des fonctions de répartitions marginales d'un n-uplet de variables aléatoires.
La modélisation des dépendances par copules est basée sur la proposition suivante, découlant du théorème de Sklar.

Proposition 10 ([33], [109]) *:*
Soit G_{T_1} la fonction de répartition d'une variable aléatoire X_1, G_{T_2} la fonction de répartition d'une variable aléatoire X_2 et H la loi jointe de (X_1, X_2). Il existe une 2-copule C telle que $\forall\ (x_1, x_2) \in \overline{\mathbb{R}}^2$ ($\overline{\mathbb{R}} = [-\infty, +\infty]$),

$$H(x_1, x_2) = C(G_{T_1}(x_1), G_{T_2}(x_2)).$$

De plus, C est uniquement déterminé sur $Dom(G_{T_1}) \times Dom(G_{T_2})$.

La décomposition de mélange peut s'appliquer du fait que H est une fonction de répartition jointe et nous avons

$$H(x_1, x_2) = \sum_{k=1}^{K} p_k H_k(x_1, x_2, \alpha_k) \tag{5.5}$$

avec $\forall k = 1, ..., K,\ 0 < p_k < 1$ et $\sum_{k=1}^{K} p_k = 1$, où $H_k(..., \alpha_k)$ est une fonction de répartition de paramètre α_k appartenant à $\mathbb{R}^d$ (d est le nombre de coordonnées de α_k) et p_k est la probabilité qu'un point de l'échantillon suive la loi $H(..., \alpha_k)$.
A partir de la proposition précédente, en remarquant que $H_k(x_1, x_2, \alpha_k)$ est la fonction de distribution jointe de paramètre α_k de la classe k, et en choisissant une famille de copules paramétrique C_{β_k}, H_k s'écrit :

$$H_k(x_1, x_2) = C_{\beta_k}(G^k_{T_1}(x_1, b^k_1), G^k_{T_2}(x_2, b^k_2)), \tag{5.6}$$

avec

- $G^k_{T_i}(., b)$ la FDD au point T_i de paramètre b,

– $G^k_{T_i}(.,b)$ est la marginale unidimensionnelle de $H_k(.,.,\alpha_k))$,

– b^k_i est donc le paramètre de la FDD au point T_i pour la classe k.

La décomposition devient :

$$H(x_1,x_2)=\sum_{k=1}^{K} p_k C_{\beta_k}(G^k_{T_1}(x_1,b^k_1),G^k_{T_2}(x_2,b^k_2)), \tag{5.7}$$

avec β_k le paramètre de la copule de la classe k.

Posons h la densité correspondant à H et h_k la densité correspondant à H_k:

$$h(x_1,x_2)=\frac{\partial^2 H}{\partial x_1 \partial x_2}(x_1,x_2)$$

et

$$h_k(x_1,x_2)=\frac{\partial^2 H_k}{\partial x_1 \partial x_2}(x_1,x_2).$$

L'équation (5.6) peut alors s'écrire:

$$h_k(x_1,x_2)=(\prod_{i=1}^{2}\frac{dG^k_{T_i}}{dx}(x_i,b^k_i)) \times \frac{\partial^2 C_{\beta_k}}{\partial u \partial v}(G^k_{T_1}(x_1,b^k_1),G^k_{T_2}(x_2,b^k_2)). \tag{5.8}$$

En travaillant sur les densités et non sur les fonctions de répartition, à partir des équations (2.1) et (5.8), $h(x_1,x_2)$ s'écrit :

$$\begin{aligned} h(x_1,x_2) &= \sum_{k=1}^{K} p_k h_k(x_1,x_2) \\ &= \sum_{k=1}^{K} p_k (\prod_{i=1}^{2}\frac{dG^k_{T_i}}{dx}(x_i,b^k_i)) \times \frac{\partial^2 C_{\beta_k}}{\partial u \partial v}(G^k_{T_1}(x_1,b^k_1),G^k_{T_2}(x_2,b^k_2)). \end{aligned} \tag{5.9}$$

Dans la suite, nous disposons de la base de distributions $\mathfrak{F}=\{F_1,...,F_N\}$ des individus $\{w_1,...,w_N\}$ et nous calculons pour chaque i la valeurs des fonctions de répartition en T_1 et T_2. Nous avons donc l'ensemble $\{(x^1_1,x^1_2),...,(x^N_1,x^N_2)\}$ avec $x^j_i=F_j(T_i)$ et $h_k(x^j_1,x^j_2)$ est notée $h_k(w_j)$.

5.4.1 Décomposition de mélange de copules Archimédiennes par approche classification

Nous proposons une extension aux données fonctions de répartition de la méthode de décomposition de mélange de lois par nuées dynamiques (Diday, Ok, Schroeder, 1974, [38]). A chaque étape, nous déterminons les paramètres des densités h_i associées aux distributions jointes de distributions H_i qui décrivent au mieux les classes de la partition courante, au sens d'un critère de qualité choisi. Pour cela, nous fixons un modèle de copules paramétriques (voir pour exemples la section 4.6.2) ou non-paramétriques. Nous considérons, dans cette partie,

une famille de copules Archimédiennes. Dans la section 5.4.7, nous regardons la version non-paramétrique de la méthode.

Par ailleurs, nous avons besoin d'un critère d'adéquation entre une partition $(P_k)_{k=1,\dots,K}$ et les paramètres des densités $(h_k)_{k=1,\dots,K}$ associées à chacune des classes. Pour l'algorithme des nuées dynamiques, de nombreux critères d'adéquation existent, basés sur la notion de vraisemblance:

- la vraisemblance
$$v(P,\beta) = \prod_{i=1}^{N}(\sum_{k=1}^{K} p_k h_k(w_i))$$
- la vraisemblance classifiante
$$vc(P,\beta) = \prod_{k=1}^{K} \prod_{w_i \in P_k} h_k(w_i)$$
- la log-vraisemblance
$$lv(P,\beta) = \sum_{i=1}^{N} log(\sum_{k=1}^{K} p_k h_k(w_i))$$
- la log-vraisemblance classifiante
$$lvc(P,\beta) = \sum_{k=1}^{K} \sum_{w_i \in P_k} log(h_k(w_i)).$$

La vraisemblance ou log-vraisemblance sont appelées classifiantes quand elles s'intéressent d'avantage à l'estimation des lois internes de chacune des classes de la partition qu'à l'estimation de la loi générale sur l'ensemble des classes, facilitant ainsi la classification des individus. De plus, les versions "classifiantes" ont l'avantage de pouvoir se calculer classe par classe. L'utilisation de ces deux critères est donc privilégiée dans l'approche classification.

La méthode de décomposition de mélange de copules (DMC) par nuées dynamiques se résume ainsi:

- Etape 0. Initialisation d'une partition aléatoire en K classes,
- Etape 1. Estimation des paramètres du mélange (5.7) (équivalent au mélange (5.9)),
- Etape 2. Affectation des individus dans les nouvelles classes $(P_k)_{k=1,\dots,K}$:
$$P_k = \{\text{individus } w \text{ tq } p_k h_k(w) \geq p_m h_m(w)\ \forall\ m\},\ \text{avec } k \leq m \text{ en cas d'égalité}\},$$
- Etape 3. Retour à l'Etape 1 jusqu'à convergence du critère.

L'étape essentielle est évidemment l'étape 1 d'estimation des paramètres. Dans le cas ou le critère d'adéquation entre une partition et sa représentation par densités jointes de distributions est basé sur la vraisemblance, les marginales univariées doivent être connues ([80]) et l'étape 1 d'estimation se résume donc par les trois étapes:

1. Estimation des proportions du mélange $(p_1, ..., p_K)$ par

$$p_k = \frac{card(P_k)}{card(\mathfrak{F})},$$

2. Estimation des paramètres b_j^k des FDD $G_{T_j}^k$ pour chacune des classes $k = 1, ..., K$ (dépend de la modélisation des FDD),

3. Estimation des paramètres $(\beta_1, ..., \beta_K)$ de copules qui maximisent le critère pour les FDD calculées. Cette estimation est effectuée par l'une des méthodes de la section 4.7.2, basées sur le maximum de vraisemblance .

Dans la suite de ce chapitre, nous estimons les paramètres avec une approche similaire. C'est-à-dire que les FDD marginales sont tout d'abord calculées puis utilisées pour l'estimation des paramètres de copules quand le critère est du type vraisemblance.

5.4.2 Décomposition de mélange de copules Archimédiennes par EM

L'extension de la méthode EM à la décomposition de mélange de copules est similaire à l'extension de la méthode par nuées dynamiques. L'idée principale est de travailler sur les densités h_k définies par (5.8), dérivées des fonctions de distribution jointes de distributions H_k. A chaque étape, nous déterminons donc les paramètres de copules qui décrivent au mieux les classes de la partition courante, au sens d'un critère de qualité donné et pour une famille de copules choisie. La méthode consiste donc à résoudre itérativement les équations de vraisemblances, le logarithme de la vraisemblance étant

$$L(a_1, ..., a_K, p_1, ..., p_K) = \sum_{j=1}^{N} log\left(\sum_{k=1}^{K} p_k h_k(x_j, a_k)\right),$$

avec $a_k = (\beta_k, b_{T_1}^k, b_{T_2}^k)$ le paramètre de h_k.

Pour le mélange de copules en K classes par EM, à partir d'une solution initiale $(p_k^0, a_k^0)_{k=1,...,K}$ avec $a_k^0 = (\beta_k^0, b_{T_1}^{k,0}, b_{T_2}^{k,0})$, l'algorithme à l'itération n $(n \geq 1)$ est donc le suivant :

– Etape E (estimation). Pour $k = 1, ..., K$; $j = 1, ..., N$, calcul des

$$t_k^n(w_j) = p_k^n h_k(x_1^j, x_2^j, a_k^n) / \sum_{k=1}^{K} p_k^n h_k(x_1^j, x_2^j, a_k^n). \tag{5.10}$$

Les t_k^n sont les probabilités a posteriori d'appartenance de w_j à la composante k à la n^{eme} itération.

– Etape M (maximisation). Pour $k = 1, ..., K$, calcul de

$$p_k^{n+1} = \frac{1}{N} \sum_{j=1}^{N} t_k^n(w_j)$$

(estimateur du maximum de vraisemblance des proportions du mélange) et résolution des équations de vraisemblance pour les paramètres $a_k = (\beta_k, b_1^k, b_2^k)$ notés $a_k = (a_{k,1}, a_{k,2}, a_{k,3})$:

$$\forall k = 1, ..., K,\ m = 1, 2, 3 : \quad \sum_{j=1}^{N} t_k^n(w_j) \frac{\partial Log(h_k(x_1^j, x_2^j, a_k^{n+1})}{\partial a_{k,m}} = 0.$$

Les deux étapes d'estimation et de maximisation sont itérées jusqu'à convergence du critère de vraisemblance.

5.4.3 Mélange de copules par SEM

Comme pour EM, le mélange de copules peut s'étendre à la méthode SEM en travaillant sur les densités h des distributions jointes de distributions modélisées avec des copules. L'initialisation consiste à fixer le paramètre K, majorant du nombre supposé de composantes, ainsi qu'un seuil $c(N, p) \in [0, 1]$. Pour chaque individu w, les probabilités initiales d'appartenance à l'une des composantes sont choisies telles que:

$$Pour\ k = 1, ..., K,\ 0 < t_k^0(w) < 1 \text{ et } \sum_{k=1}^{K} t_k^0(w) = 1.$$

L'algorithme à l'itération n est donc le suivant:

– Etape S (stochastique). Pour chaque individu w, la variable aléatoire multinomiale $e^n(w) = (e_k^n(w), k = 1, ..., K)$ d'ordre un et de paramètres $(t_k^n(w), k = 1, ..., k)$ est simulée. Une partition $P^n = (P_1^n, ..., P_K^n)$ de l'échantillon est définie par les réalisations $e^n(w)$ avec $P_k^n = \{w,\ tels\ que\ e_k^n(w) = 1\}$. Si pour un k, $card(P_k^n)$ est plus petit que $Nc(N, p)$ l'algorithme est ré-initialisé.

– Etape M (maximisation). Estimation des paramètres du mélange

$$(p_k^{n+1}, a_k^{n+1}),\ k = 1, ..., K$$

par maximum de vraisemblance sur les sous-échantillon $(P_k^n,\ k = 1, ..., K)$ avec

$$p_k^{n+1} = \frac{1}{N} \sum_{j=1}^{N} e_k^n(w).$$

L'estimation des a_k^{n+1} dépend de la famille paramétrique de copules utilisée et de la modélisation des FDD.

– Etape E (estimation). Pour chaque individu w et chaque $k = 1, ..., K$, le calcul des $t_k^{n+1}(w)$ est effectué par

$$t_k^{n+1}(w) = p_k^{n+1} h_k(w, a_k^{n+1}) / \sum_{k=1}^{K} p_k^{n+1} h_k(w, a_k^{n+1}),$$

avec h_k la densité associée à la distribution jointe de distributions H_k de la classe k.

Comme dans l'algorithme SEM classique, une partition recouvrantre est obtenue à chaque itération et à la convergence de l'algorithme, nous avons une famille de partitions statistiquement admissibles. Les perturbations introduites à chaque itération par les tirages aléatoires empêchent la convergence vers un maximum local instable de la vraisemblance comme cela peut être le cas pour EM.

5.4.4 Mélange de copules par SAEM

L'algorithme SAEM étendu au mélange de copules est une modification de l'algorithme SEM par copules. Cette modification répond, comme dans le cas de mélange de densités, aux besoins de contrôler le comportement de SEM pour des ensembles de données relativement petits, sans sacrifier la partie stochastique de la méthode. Une suite de réels positifs (γ^n) décroissants vers 0 (avec $\gamma^0 = 1$) est utilisée. Si α^n est le paramètre estimé par SAEM à l'étape n, le paramètre α^{n+1} est calculé à l'étape $n+1$ par :

$$\alpha^{n+1} = (1 - \gamma^{n+1})\alpha_{EM}^{n+1} + \gamma^{n+1}\alpha_{SEM}^{n+1}, \tag{5.11}$$

avec α_{EM}^{n+1} l'approximation de α par EM (pour le mélange de copules) et α_{SEM}^{n+1} l'approximation de α par SEM (pour le mélange de copules).

Comme SAEM classique, la méthode SAEM étendue part d'un pur SEM au début et tend vers un pur EM. De même que dans la section 2.3.2, la taux de convergence de γ^n vers 0 doit être faible pour obtenir de bons résultats : les premières itérations doivent être effectuées avec des γ^n près de 1.

5.4.5 Mélange de copules par MCEM

L'algorithme MCEM (Wei et Tanner, 1990) de la section 2.3.3, travaillant avec une méthode de Monte-Carlo, s'applique évidemment au traitement des données probabilistes par la méthode des copules. Comme pour les autres extensions, les modifications principales reposent sur le fait qu'on travaille sur h, densités jointes de distributions, modélisées par copules. L'algorithme reste similaire à celui de la section 2.3.3. En gardant les mêmes notations, si ϕ est le paramètre à estimer et ϕ^n son approximation à l'étape n, le calcul de $Q(\phi|\phi^n)$ est remplacé par celui de la version empirique $Q_{n+1}(\phi|\phi^n)$ basée sur m (m»1) réalisations de données complètes y à partir de $\boldsymbol{k}(y|x, \phi^n)$ (La densité conditionnelle est notée en gras pour

ne pas la confondre avec le numéro k de la classe) :

- Etape 1. On génère un échantillon iid $(z^n_{(1)}, ..., z^n_{(m)})$ à partir de $\boldsymbol{k}(y|x, \phi^n)$, avec

$$\boldsymbol{k}(y|x,\phi) = \prod_{j=1}^{N} \frac{\pi'_{k_j} h_{k_j}(x_j|\phi'_{k_j})}{h(x_j|\phi')}.$$

  Dans la définition de $\boldsymbol{k}$, k_j est le numéro de la composante à laquelle appartient l'observation x_j, et h_{k_j} est la densité associée à la FJDD H_{k_j} aux points T_1 et T_2 dans la composante k_j.

- Etape 2. On calcule

$$Q_{n+1}(\phi|\phi^n) = \frac{1}{m} \sum_{j=1}^{m} log(h(x, z^n_{(j)}|\phi). \tag{5.12}$$

- Etape 3. Estimation de ϕ^{n+1} par $\phi^{n+1} = argmax_\phi Q_{n+1}(\phi|\phi^n)$.

Les étapes sont itérées jusqu'à convergence du critère.

5.4.6 Mélange de copules par CEM

Comme dans la méthode CEM "classique" de Celeux et Govaert ([14]), nous pouvons combiner l'approche "estimation" avec l'approche "classification" de DMC pour définir une méthode CEM par copules. A partir d'une solution initiale $(p^0_1, ..., p^0_K, a^0_1, ..., a^0_k)$, l'algorithme à l'itération n est le suivant:

- Etape E (Estimation de EM). Calcul des probabilités conditionnelles courantes $t^n_k(x_j)$ $(1 \leq j \leq N, 1 \leq k \leq K)$ selon (5.10),

- Etape C (Classification). Affectation de chaque observation x_j à la classe P^n_k qui donne la probabilité conditionnelle courante $t^n_k(x_j)$ maximale,

- Etape M (Maximisation). Calcul des estimateurs (p^n_k, a^n_k) des paramètres par maximum de vraisemblance sur la classe P^n_k comme sous-ensemble.

Cette extension de CEM présente les avantages et les défauts de CEM "classique" : la méthode CEM est plus rapide qu'EM ; cependant, CEM produit des estimations biaisées des paramètres si les classes ne sont pas bien séparées.

5.4.7 Décomposition de mélange de copules empiriques

Dans le cas de figure où nous nous intéressons plus à la classification finale qu'aux paramètres la décrivant, la modélisation de copules par des méthodes non-paramétriques peut être appropriée.

L'estimation des FDD G^k_T de la classe k peut être effectuée par une modélisation continue (telle que la loi béta) ou de manière empirique avec un calcul de leur dérivée par méthode des noyaux de Parzen.

Indépendemment de l'estimation des FDD, l'estimation de la dérivé seconde de C_k (dans la classe k) peut se faire en remarquant que C_k n'est autre que la fonction de répartition des valeurs de distributions de distributions $(G^k_{T_1}(x^1_w), G^k_{T_2}(x^2_w))$ avec $(x^1_w, x^2_w) = (F_w(T_1), F_w(T_2))$, les valeurs pour l'individu w en T_1 et T_2 (dans le cas 2-D). La dérivé seconde de C_k par rapport à u et v est donc la densité bi-dimensionnelle des valeurs de distribution de distributions dans la classe k. Les méthodes "classiques" d'estimation non-paramétriques de densité sont donc applicables en considérant les couples de valeurs de distribution de distributions comme données de base. Cette méthode par modélisation non-paramétrique a l'avantage de pouvoir travailler sur n valeurs de T au lieu des 2 T_1 et T_2 des méthodes ci-dessus. Cependant, la convergence de cette méthode est rédhibitoirement lente.

5.5 Décompositions en dimension n

Mis à part la méthode non paramétrique, les méthodes de décomposition de mélange de copules de la section 5.4 ne travaillent que sur deux valeurs T_1 et T_2. La question qu'on peut se poser est de savoir comment travailler sur d'avantage de T. Nous donnons dans cette partie une approche permettant de considérer des copules multidimensionnelles ainsi que deux méthodes ne se ramenant qu'à des copules bivariées.

5.5.1 Copules multidimensionnelles

L'écriture d'une copule en dimension $n \geq 3$ s'avère très compliquée. Différentes généralisations de copules bidimensionnelles existent pour une même famille paramétrique (voir la section 4.6.3). Elles ont toutes des avantages et des inconvénients mais leur point commun est leur complexité. Nous avons vu que les copules Archimédiennes multivariées définies par (4.20) sont une première approche. Nous redonnons la forme de ces copules pour mémoire:

$$C^n(u) = \varphi^{(-1)}(\varphi(u_1) + ... + \varphi(u_n)),$$

avec φ la fonction génératrice.

Les méthodes de décomposition de mélange de la section 5.4 peuvent s'appliquer à cette forme de copules car nous n'avons alors qu'un paramètre de copule à estimer par classe. Cependant, dans cette approche les éléments sont engendrés par la même application φ, créant ainsi un haut degré de symétrie. L'intérêt de ces copules est donc limité.

Le recours aux copules multidimensionnelles de la forme (4.22) peut résoudre le problème de symétrie. La forme de ces copules est redonnée pour mémoire:

$$C_n(u_1, ..., u_n) = \varphi_n^{-1}(\varphi_n(C_{n-1}(u_1, ..., u_{n-1})) + \varphi(u_n)),$$

avec φ la fonction génératrice.

En travaillant avec des copules de cette forme, et en nous appuyant sur les méthodes d'estimations des paramètres d'une copule Archimédienne (4.22) basées sur le maximum de vraisemblance, nous pouvons définir des méthodes multidimensionnelles de décomposition de mélange de copules. Ces méthodes travaillent selon les étapes suivantes:

– Etape 0. Initialisation d'une partition aléatoire en K classes, $P^0 = (P^0_1, ..., P^0_K)$.

- Etape 1. Estimation des proportions du mélange $(p_1^n, ..., p_K^n)$, par

$$p_k^n = \frac{card(P_k^n)}{card(\mathfrak{F})}.$$

- Etape 2. Estimation des paramètres des FDD $G_{T_j}^k$ pour chacune des classes P_k^n, $k = 1, ..., K$ (dépend de la modélisation des FDD).

- Etape 3. Estimation des paramètres $(\beta_1^n, ..., \beta_K^n)$ de copules par l'une des méthodes par maximum de vraisemblance de la section 4.7.2.

- Etape 4. Affectation des individus dans les nouvelles classes $(P_k^n)_{k=1,...,K}$:

  $P_k^n = \{\text{individus } w \text{ tq } p_k^n h_k^n(w, \beta_k^n) \geq p_m^n h_m^n(w, \beta_m^n) \; \forall \, m\}$, avec $k \leq m$ en cas d'égalité},

- Etape 5. Retour à l'Etape 1 jusqu'à convergence du critère.

Cependant, la compréhension des dépendances modélisées par les paramètres β_k n'est pas simple dans ce contexte. Pour ne pas accentuer la difficulté des équations à résoudre, nous proposons deux méthodes pour traiter deux variables différentes en même temps et/ou plus de deux T_i à la fois: la méthode par arbre binaire suggérée par E. Diday et la méthode par couplage que nous proposons immédiatement.

5.5.2 La méthode par couplage

Nous nous appuyons sur l'idée que les jointes au niveau de chaque variable peuvent être considérées comme des marginales au niveau de deux variables, puis sur l'ensemble des variables. Nous disposons d'un échantillon de n individus $W = (w_1, ..., w_n)$ décrit par deux variables Y^1 et Y^2. Déterminons deux décompositions de mélange de copules: l'une sur la variable Y^1 en K_1 classes (suivant deux T pour les FDD: T_1^1 et T_2^1), l'autre sur Y^2 en K_2 classes (suivant deux T pour les FDD: T_1^2 et T_2^2). Pour tout ω de Ω et pour chaque variable $(Y^i)_{i=1,2}$, nous disposons donc de la valeur de la fonction de répartition jointe

$$H^{Y^i}(\omega) = \sum_{k=1}^{K_i} p_k C_{\beta_k^i}(G_{T_1^i}^k(\omega_1^i, b_{T_1^i}^k), G_{T_2^i}(\omega_2^i, b_{T_2^i}^k))$$

avec

- (ω_1^i, ω_2^j) = valeurs de la fonction de répartition de la variable Y^j pour ω respectivement en T_1^i et T_2^i (i.e. $\omega_j^i = \mathbb{P}(X_j^i \leq T_j^i)$, où X_j^i est la variable aléatoire représentant l'individu ω_j pour la variable Y^i),

- β_k^i = paramètre de copule de la classe k pour la variable Y^i,

- $b_{T_j^i}^k$ = paramètre associé au FDD défini en T_j^i pour la classe k (dépend de la modélisation des FDD),

- $G_{T_j^i}^k(\omega_j^i, b_{T_j^i}^k)$ = valeur du FDD associé à la variable Y^i au point T_j^i appliqué à la valeur ω_j^i.

Nous disposons maintenant d'un couple de valeurs (H^{Y^1}, H^{Y^2}) pour chacun des n individus de Ω. Posons que les variables aléatoires H^{Y^1} et H^{Y^2} suivent respectivement les lois de distribution F_1 et F_2. Supposons que le couple (H^{Y^1}, H^{Y^2}) soit de loi jointe H. Les conditions du théorème 3 de Sklar sont vérifiées: Il existe une copule C telle que $\forall\ (x_1, x_2) \in [0,1]^2, H(x_1, x_2) = C(F_1(x_1), F_2(x_2))$. Nous pouvons par conséquent appliquer à nouveau une décomposition de mélange de lois de lois à partir des n couples obtenus. Les paramètres de copules déterminés lors des décompositions de mélange de copules sur Y^1 et Y^2 fournissent des informations sur les dépendances présentes à l'intérieur des variables (entre T_1^1 et T_2^1 et entre T_1^2 et T_2^2). Les paramètres de copules obtenues au final permettront de caractériser les dépendances qui existent alors entre les lois des deux variables Y^1 et Y^2.

Par ailleurs, cette méthode peut s'appliquer sur la même variable Y en choisissant quatre valeurs T_1, T_2, T_3 et T_4. Nous pouvons effectuer deux décompositions de mélange de copules : l'une avec les valeurs T_1 et T_2 ; la seconde avec les valeurs T_3 et T_4. Nous pouvons alors relancer la méthode sur les résultats obtenus.

5.5.3 Méthode par arbre binaire

La seconde technique (proposée par E. Diday dans [33]) consiste en une méthode d'arbre binaire. A partir de la base de distributions complète (le haut de l'arbre), nous déterminons la meilleure partition en deux classes (noeuds-fils) parmi un ensemble de familles de modèles possibles, puis la meilleure partition en deux classes des deux noeuds-fils et ainsi de suite. A chaque étape, pour chaque noeud et pour chaque variable, nous déterminons les (T_1, T_2) optimaux. Le noeud N à couper est celui dont la variable de classification et les (T_1, T_2) associés maximisent le critère de qualité Q de la division. Nous pouvons utiliser

$$Q(N) = \sum_k \sum_{\omega \in P_k} \log h_{\beta_k}(\omega).$$

La méthode par arbre binaire présente l'intérêt de pouvoir travailler avec différentes variables et de pouvoir choisir les valeurs (T_1, T_2) adaptés à chaque étape ainsi que les copules appropriées à chaque noeud (voir Figure 5.4).
Dans la Figure 5.4 la classe C_1 est l'une des classes de la décomposition sur la variable V_6 avec la famille de Frank et les valeurs optimales (T_1, T_2). Les classes C_2 et C_3 sont les deux classes de la décomposition sur la variable V_3 avec la famille de Genest-Goudi pour les valeurs optimales (T_1, T_2) sur la deuxième classe de la première décomposition. Remarquons que la même variable peut apparaitre plusieurs fois.

La méthode a néanmoins l'inconvénient d'être assez longue. En effet, pour chaque nouveau (T_1, T_2), la méthode doit calculer tous les $F(T_i(\omega))$ pour tous les individus ω de l'ensemble associé. De plus une partition en deux classes est calculée pour chaque noeud-fils. Par ailleurs, deux individus classés séparément à une étape ne pourront pas être classés ensemble par la suite. Donc dans la pratique, nous avons finalement préféré notre méthode par couplage.

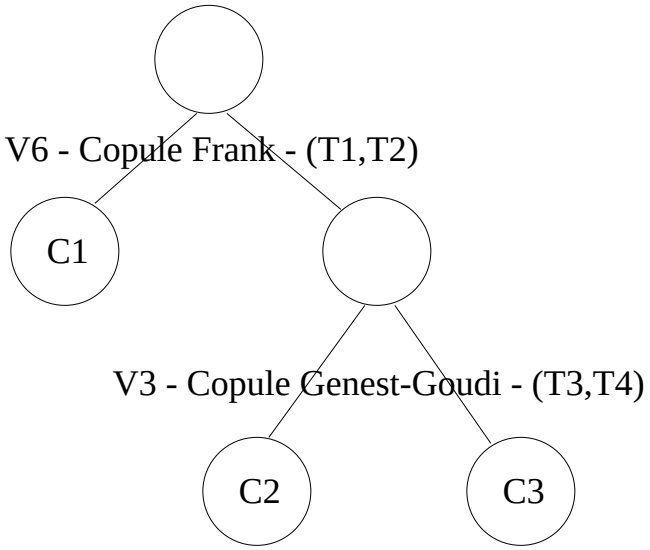


FIG. 5.4: *Arbre binaire*

5.6 Inférence

La méthode explicitée dans ce chapitre permet donc d'obtenir:

- une classification des individus
- la loi de probabilité de l'échantillon (pour une famille donnée de copules), par l'estimation des copules (et de leur paramètre) de chacune des classes de la classification
- une modélisation des dépendances entre deux points d'une variable fonction de répartition et une modélisation de la dépendance entre deux variables (dans la méthode par couplage par exemple).

Pour un nouvel individu décrit par des variables fonction de répartition, nous pouvons donc calculer sa probabilité d'apparition et déterminer sa classe d'appartenance dans la classification. La classe d'appartenance de l'individu w est définie ainsi: L'individu w de fonction de répartition F_w appartient à la classe P_k si :

$$p_k h_k(F_w(T_1), F_w(T_2), \beta_k) \geq p_m h_m(F_w(T_1), F_w(T_2), \beta_m) \ \forall \ m, \text{ avec } i < m \text{ en cas d'égalité.} \tag{5.13}$$

Si la classification a été réalisée sur des données suffisamment représentatives, les nouveaux individus peuvent être classés de manière cohérente et une nouvelle base de données peut être classée au travers du filtre des paramètres de copules et de FDD déjà obtenues.

5.7 Convergence de l'algorithme par nuées dynamiques

Proposition 11 *:*
L'algorithme de décomposition de mélange de copules par nuées dynamiques converge vers une solution localement optimale en un nombre fini d'itérations.

Démonstration :
Soit P^n la partition en K classes à l'itération n, $P^n = (P_1^n, ..., P_K^n)$
Soit L^n les différents paramètres à l'itération n, $L^n = (\alpha_1^n, ..., \alpha_K^n)$

A l'initialistion, posons :

$$\begin{cases} V_0 = (P^0, L^0) \\ et \\ U_0 = W(V_0) \end{cases}$$

et à l'itération n :

$$\begin{cases} V_n = (P^n, L^n) \\ et \\ U_n = W(V_n) \end{cases}$$

avec

$$W(P^n, L^n) = \sum_{k=1}^{K} W(P_k^n, \alpha_k^n)$$

et

$$W(P_k^n, \alpha_k^n) = \sum_{\omega \in P_k^n} \log(h_k(\omega, \alpha_k^n)).$$

- La fonction $W(P_k^n, \alpha_k^n)$ est la log-vraisemblance classifiante de la classe k,
- $P^n = f(L^{n-1})$ où f est la fonction d'affectation qui définit de nouvelles classes à partir des paramètres,
- $L^n = g(P^n)$ où g est la fonction de paramétrisation qui estime de nouveaux paramètres à partir de classes.

Montrons que la suite (U_n) converge en croissant et est stationnaire.

- Pour cela, montrons tout d'abord que $W(P^{n+1}, L^{n+1}) \geq W(P^n, L^{n+1})$.
Rappelons que la fonction d'affectation f est définie par :

$$P_k^{n+1} = \{\omega \ tq \ \ h_k(\omega, \alpha_k^{n+1}) \ \geq \ h_j(\omega, \alpha_j^{n+1}), \ \forall \ 1 \leq j \leq K \ avec \ k \leq j \ en \ cas \ d'égalité\}.$$

Par définition nous avons donc $W(P^{n+1}, L^{n+1}) \geq W(P^n, L^{n+1})$.

- Montrons maintenant que $W(P^n, L^{n+1}) \geq W(P^n, L^n)$
Par construction de la fonction de paramétrisation g, il est évident que :

$$W(P_k^n, \alpha_k^{n+1}) \geq W(P_k^n, \alpha_k^n)$$

En effet, la fonction g utilisant ici la technique du maximum de la log-vraisemblance, α_k^{n+1} vérifie

$$\max_{\alpha} \sum_{\omega \in P_k^n} \log(h_k(\omega, \alpha))$$

Par conséquent, $\forall \alpha$ (et donc pour $\alpha = \alpha_k^n$)

$$\sum_{\omega \in P_k^n} \log(h_k(\omega, \alpha_k^{n+1})) \geq \sum_{\omega \in P_k^{n+1}} \log(h_k(\omega, \alpha))$$

et donc

$$W(P_k^n, \alpha_k^{n+1}) \geq W(P_k^n, \alpha_k^n).$$

En sommant sur toutes les classes on obtient

$$W(P^n, L^{n+1}) \geq W(P^n, L^n).$$

En combinant les deux inégalités nous avons :

$$W(P^{n+1}, L^{n+1}) \geq W(P^n, L^{n+1}) \geq W(P^n, L^n)$$

$$\Longrightarrow \; U_{n+1} \geq U_n.$$

$(U_n)_{n \in \mathbb{N}}$ est croissante et ne peut prendre qu'un nombre fini de valeurs. Elle converge donc en un nombre fini d'itérations et est stationnaire.

$$\exists \; N \in \aleph \; tq \; \forall n \geq N, \; U_n = U_N.$$

Remarque:
La technique du maximum de vraisemblance ou de la log-vraisemblance pour l'estimation des paramètres de copules nécessite la connaissance des G_{T_i}. Aussi la détermination des paramètres de la loi G_{T_i} (dans le cas paramétrique) n'entre pas en compte. On suppose donc les fonctions G_{T_i} et leurs dérivées connues implicitement et induites par les classes.

5.8 Comportement asymptotique

Soit $\mathfrak{F} = \{\tilde{F}_1^{(p)}, ..., \tilde{F}_N^{(p)}\}$ un échantillon de N réalisations indépendantes et identiquement distribuées d'une variable aléatoire à valeurs fonctions de répartition. La fonction $\tilde{F}_i^{(p)}$ est en fait une estimation de la fonction de distribution réelle décrivant l'individu i, et calculée à partir de p réalisations numériques pour l'individu i. Les lois sont supposées normales $\mathcal{N}(\mu_i, \sigma_i^2)$ et les paramètres sont calculés par

$$\tilde{\mu}_i = \frac{1}{p} \sum_{j=1}^{p} x_i^j \text{ et } \tilde{\sigma}_i^2 = \frac{1}{p-1} \sum_{j=1}^{p} (x_i^j - \tilde{\mu}_i)^2. \tag{5.14}$$

Nous posons également les hypothèses suivantes :

$$\left\{ \begin{array}{l} \text{Si chaque fonction de distribution } \tilde{F}_i^{(p)} \text{ est une estimation de la fonction de} \\ \text{répartition de l'une des deux lois normales } \mathcal{N}(m_1, s_1^2) \text{ ou } \mathcal{N}(m_2, s_2^2), \end{array} \right.$$

$$\Longleftrightarrow$$

$$\left\{ \begin{array}{l} \text{S'il existe une partition en deux classes } (P_1, P_2) \text{ des } \{\tilde{F}_i^{(p)}\}_{i=1,...,N} \text{ telle que} \\ P_j = \{\tilde{F}_i^{(p)} \ / \ \tilde{F}_i^{(p)} \text{ estimation de la fonction de répartition de la loi } \mathcal{N}(m_j, s_j^2)\}, \end{array} \right.$$

et si les estimateurs des paramètres $\mu_i^{(p)}$ et $\sigma_i^{(p)}$ (respectivement de la moyenne et de l'écart-type) des lois normales de chacun des individus i, sont sans biais (c'est le cas des estimations (5.14)), alors d'après des résultats classiques de convergence d'estimateurs sans biais, on a le corollaire suivant:

Corollaire 2 *:*
$\forall i = 1, ..., N$, *quand* p *tend vers l'infini,* $\tilde{F}_i^{(p)}$ *converge uniformément vers* F_i*, la fonction de répartition de l'une des deux lois normales* $\mathcal{N}(m_1, s_1^2)$ *ou* $\mathcal{N}(m_2, s_2^2)$.

Autrement dit, quand p tend vers l'infini, toutes les fonctions de répartition $\tilde{F}_i^{(p)}$ de la base $\mathfrak{F}$ de distributions convergent vers les fonctions de distribution réelles F_i décrivant les individus.

Remarque : En associant à $\{F_1, ..., F_N\}$ la σ-algèbre engendrée par les singletons $\{F_i\}_{i=1,...,N}$, nous pouvons définir la mesure de probabilité $\mathbb{P}$ sur $(\{F_1, ..., F_N\}, \sigma(\{F_i\}_{i=1,...,N}))$, correspondant à une loi de Bernoulli de paramètre π,

$$\mathbb{P}(\{F \in \{F_1, ..., F_N\} \ / \ F \in P_1\}) = \pi = 1 - \mathbb{P}(\{F \in \{F_1, ..., F_N\} \ / \ F \in P_2\}).$$

De plus, si les FDD $(G_{T_j}^k)_{j=1,2}$ (obtenues par la base $\mathfrak{F}$) de chaque classe k sont modélisées de manière empirique, alors d'après des résultats classiques d'analyse fonctionnelle, on a:

Corollaire 3 *:*
En chaque classe $k = 1, 2$ *et pour chaque* T*, la FDD* G_T^k *de la classe* k *converge uniformément vers une fonction de distribution de distributions* G_T^{k*} *de Dirac au point* $F_{\mathcal{N}_k}(T)$*. La fonction* G_T^{k*} *est définie par*

$$\left\{ \begin{array}{ll} G_T^{k*}(x) = 0 & si \ x < F_{\mathcal{N}_k}(T) \\ G_T^{k*}(x) = 1 & si \ x \geq F_{\mathcal{N}_k}(T) \end{array} \right.$$

avec $F_{\mathcal{N}_k}$*, la fonction de répartition de la loi normale de paramètres* (m_k, σ_k^2).

Les corollaires 2 et 3 s'étendent au cas de K classes avec les hypothèses suivantes:

$$\left\{ \begin{array}{l} \text{Si chaque fonction de répartition } \tilde{F}_i^{(p)} \text{ est une estimation de la fonction de} \\ \text{distribution de l'une des } K \text{ lois normales } \mathcal{N}(m_1, s_1^2), ..., \mathcal{N}(m_K, s_K^2), \end{array} \right.$$

$$\Longleftrightarrow$$

$$\left\{ \begin{array}{l} \text{S'il existe une partition en } K \text{ classes } (P_1, ..., P_K) \text{ des } \{\tilde{F}_i^{(p)}\}_{i=1,...,N} \text{ telle que} \\ P_j = \{\tilde{F}_i^{(p)} \ / \ \tilde{F}_i^{(p)} \text{ estimation de la fonction de répartition de la loi } \mathcal{N}(m_j, s_j^2)\}, \end{array} \right.$$

et si les estimateurs des paramètres $\mu_i^{(p)}$ et $\sigma_i^{(p)}$ (respectivement de la moyenne et de l'écart-type) des lois normales de chacun des individus i, sont sans biais, alors:

Corollaire 4 *:*
$\forall \ i = 1, ..., N$, *quand* p *tend vers l'infini,* $\tilde{F}_i^{(p)}$ *converge uniformément vers l'une des* K *lois normales* $\mathcal{N}(m_1, s_1^2), ..., \mathcal{N}(m_K, s_K^2)$.

Remarque : En associant à $\{F_1, ..., F_N\}$ la σ-algèbre engendrée par les singletons $\{F_i\}_{i=1,...,N}$, nous pouvons définir la mesure de probabilité $\mathbb{P}$ sur $(\{F_1, ..., F_N\}, \sigma(\{F_i\}_{i=1,...,N}))$, correspondant à une loi multinomiale de paramètre $(\pi_1, ..., \pi_K)$

$$\mathbb{P}(\{F \in \{F_1, ..., F_N\} \ / \ F \in P_k\}) = \pi_k \text{ avec } \sum_{k=1}^{K} \pi_k = 1.$$

De plus, si les FDD $(G^k_{T_j})_{j=1,2}$ (obtenues par la base $\mathfrak{F}$) de chaque classe k sont modélisées de manière empiriquen, alors le corollaire 3 devient:

Corollaire 5 *:*
En chaque classe $k = 1, ..., K$ et pour chaque T, la FDD G^k_T de la classe k converge uniformément vers une fonction de distribution de distributions G^{k*}_T de Dirac au point $F_{\mathcal{N}_k}(T)$. La fonction G^{k*}_T est définie par

$$\begin{cases} G^{k*}_T(x) = 0 & si\ x < F_{\mathcal{N}_k}(T), \\ G^{k*}_T(x) = 1 & si\ x \geq F_{\mathcal{N}_k}(T), \end{cases}$$

avec $F_{\mathcal{N}_k}$, la fonction de répartition de la loi normale de paramètres (m_k, σ^2_k).

En s'appuyant sur ce résultat, et en conséquence directe de la propostion 8 (page 102) portant sur la copule associée à une base de distributions réduite à une fonction de répartition, nous avons :

Proposition 12 *:*
En chaque classe $k = 1, ..., K$, quelque soit le nombre n de valeurs $T_1, ..., T_n$ fixés, si la copule C_k de la classe k est définie de manière empirique, elle converge vers la copule $C^*_k = Min = \Pi$ (qui sont les mêmes dans ce cas).
De plus, $\forall\ (x_1, ..., x_n) \in \mathbb{R}^n$,

$$C^*_k(G^{k*}_{T_1}(x_1), ..., G^{k*}_{T_n}(x_n)) \in \{0, 1\}.$$

5.9 Décomposition de mélange de FDD

Nous avons vu au chapitre 3 que les fonctions de distribution de distributions sont elles-mêmes des fonctions de répartition. Au lieu de les calculer uniquement classe par classe, nous pouvons réaliser une estimation de la FDD globale en chaque T par décomposition de mélange de FDD.

Proposition 13 *:*
Si les $\{F_i\}_{i=1,...,N}$ (lois réelles des individus) sont distribuées dans les classes $(P_1, ..., P_K)$ d'une partition en K classes, suivant une loi multinomiale de paramètres $(\pi_1, ..., \pi_K)$, et si G^k_T est la FDD de la classe k en T, alors la FDD globale en T, notée G_T, est

$$G_T(x) = \sum_{k=1}^{K} \pi_k G^k_T(x). \tag{5.15}$$

Le coefficient π_k est la probabilité qu'une fonction de distribution réelle F_i soit dans la classe P_k de la partition. Il s'agit donc des mêmes π_k que dans la section 5.8.
De plus, à partir du corollaire 5 et en écrivant la FDD globale au point T comme un mélange de FDD de type (5.15), nous avons la proposition suivante:

Proposition 14 *:*
Pour chaque valeur T, la FDD globale G_T converge uniformément vers une fonction de distribution de distributions G_T^* définie par:

$$\begin{cases} G_T^*(x) = 0 & si\ x < F_{\mathcal{N}_1}(T), \\ G_T^*(x) = \sum_{j=1}^{k} \pi_j & si\ F_{\mathcal{N}_k}(T) \leq x < F_{\mathcal{N}_{k+1}}(T), \\ G_T^*(x) = 1 & si\ x \geq F_{\mathcal{N}_K}(T), \end{cases}$$

avec $F_{\mathcal{N}_j}$, la fonction de répartition de loi normale de paramètres (m_j, σ_j^2).

Dans cette proposition, il est supposé que $F_{\mathcal{N}_1}(T) < ... < F_{\mathcal{N}_K}(T)$.

Par ailleurs, nous avons vu en section 5.4 que la distribution jointe de distributions (FJDD) H_{T_1,T_2}, aux valeurs T_1 et T_2 pouvait s'écrire

$$\begin{aligned} H_{T_1,T_2}(x_1,x_2) &= \sum_{k=1}^{K} \pi_k H_{T_1,T_2}^k(x_1,x_2,\alpha_k) \\ &= \sum_{k=1}^{K} \pi_k C_{\beta_k}(G_{T_1}^k(x_1), G_{T_2}^k(x_2)). \end{aligned} \tag{5.16}$$

En posant X_1 la variable aléatoire caractérisée par G_{T_1} (la FDD globale en T_1), X_2 la variable aléatoire caractérisée par G_{T_2} (la FDD globale en T_2), la fonction de distribution jointe du couple (X_1, X_2) est donc H_{T_1,T_2}. Avec le théorème de Sklar, nous déduisons la proposition suivante:

Proposition 15 *:*
Il existe une copule C telle que $\forall\ (x_1,x_2)\ \in\ [0,1]$,

$$\begin{aligned} H_{T_1,T_2}(x_1,x_2) &= C(G_{T_1}(x_1), G_{T_2}(x_2)) \\ &= C(\sum_{k=1}^{K} \pi_k G_{T_1}^k(x_1), \sum_{k=1}^{K} \pi_k G_{T_2}^k(x_2)) \end{aligned} \tag{5.17}$$

A partir de l'équation (5.17), nous pouvons déduire qu'il existe une relation entre le mélange de copules et le mélange de FDD; cette relation portée par la copule C de la proposition 15 est :

$$\sum_{k=1}^{K} \pi_k C_{\beta_k}(G_{T_1}^k(x_1), G_{T_2}^k(x_2)) = C(\sum_{k=1}^{k} \pi_k G_{T_1}^k(x_1), \sum_{k=1}^{K} \pi_k G_{T_2}^k(x_2)). \tag{5.18}$$

5.10 Généralisation de la décomposition de mélange de densités

Nous montrons dans cette section que la décomposition de mélange de densités est un cas particulier de la décomposition de mélange de copules dans le cas d'une variable aléatoire numérique unique Z. La valeur $Z(w)$ prise par un individu w, est transformée en une fonction de distribution prenant la valeur 0 jusqu'à $Z(w)$ exclue et la valeur 1 après. De telles fonctions de répartition sont appelées "distributions de masse unitaire". Plus formellement, soient $\Omega = \{w_1, ..., w_N\}$ l'ensemble des individus et $Z(w_i) = z_i$. La fonction de distribution F_i associée à w_i est définie par $F_i(t) = \mathbb{P}(X_i < t)$ avec la variable aléatoire X_i associée à w_i telle que F_i satisfait :

$$F_i(t) = \begin{cases} 0 & si\ t < z_i \\ 1 & si\ t \geq z_i. \end{cases}$$

Pour démontrer le fait que le mélange de densités est un cas particulier du mélange de copules, regardons les propriétés d'un ensemble de distributions de masse unitaire.

5.10.1 Propriétés d'une base de distributions de masse unitaire

Si la base de distributions contient seulement de telles F_i et si F_Z est la fonction de répartition associée à la variable aléatoire Z, nous avons :

Proposition 16 (Diday, [33]) *:*
Si T_i croit avec i et que les G_{T_i} sont empiriques nous avons

1. *Si $H(x_1, ..., x_p) = C(G_{T_1}(x_1), ..., G_{T_p}(x_p))$, C est la copule Min.*
2. *Si $x_p < 1$, $Min(G_{T_1}(x_1), ..., G_{T_p}(x_p)) = G_{T_p}(x_p)$.*
3. *Pour tout $1 \leq i \leq p$, si $x_i < 1$, $G_{T_i}(x_i) = \mathbb{P}(Z > T_i) = 1 - F_Z(T_i)$*
4. *Si $x_p < 1$, $F_Z(T_p) = 1 - H(x_1, ..., x_p)$.*

Démonstration:
Pour la démonstration de la proposition 16, nous passons par le lemme:

Lemme 7 *:*
Si $\mathfrak{F}$ est un ensemble de distributions de masses unitaires et $x_i \in [0, 1[$ pour $i = 1, ..., j$, $A_j = \{F_m \in \mathfrak{F} / F_m(T_i) = 0\}$ et $B_j = \{F_m \in \mathfrak{F} / F_m(T_i) \leq x_i,\ 1 \leq i \leq j\}$ alors $A_j = B_j$ et $card(A_j) = min_{i=1,...,j} card(A_i)$.

Démonstration du lemme:
Nous avons $B_j \subseteq A_j$ car $F_m \in B_j$ implique $F_m \in \mathfrak{F}$ et $F_m(T_j) \leq x_j$ par définition de B_j. L'ensemble $\mathfrak{F}$ est un ensemble de distributions de masses unitaires et $x_j \in [0, 1[$, donc $F_m(T_j) = 0$. $F \in A_j$.
Nous avons $A_j \subseteq B_j$ car $F_m \in A_j$ implique $F_m(T_j) = 0$ qui implique $F_m(T_i) = 0$ pour $i = 1, ..., j$ car F_m est une fonction de répartition donc croissante. Donc $F_m \in B_j$ et $A_j = B_j$. Par définition de B_j, $B_j = \cap_{i=1,...,j} A_i$ et nous avons $A_j \subseteq B_j$ donc $card(A_j) = min_{i=1,...,j} card(A_i)$.

Nous pouvons démontrer la proposition 16.
1. Si $H(x_1,...,x_p) = C(G_{T_1}(x_1),...,G_{T_p}(x_p))$, C est la copule Min.
Si tous les x_i sont égaux à 1, nous avons par définition d'une fonction de distribution de distributions : $G_{T_i}(x_i) = 1$ et $H(x_1,...,x_p) = 1$. Dans ce cas, le point 1 de la proposition 16 est vraie. Supposons que quelques x_i soient inférieurs à 1, notons les $x'_1,...,x'_j$ avec $T'_1,...,T'_j$ croissants. Nous avons $H(x'_1,...x'_j) = H(x_1,...,x_p)$ car l'ensemble des fonctions de distribution inférieures à $x'_1,...,x'_j$ est le même que celui des fonctions de distribution inférieures à $x_1,...,x_p$. Nous pouvons appliquer le lemme avec $A_j = \{F_m \in \mathfrak{F} / F_m(T'_j) = 0\}$ et $B_j = \{F_m \in \mathfrak{F} / F_m(T'_i) \leq x'_i,\ 1 \leq i \leq j\}$. Or

$$\begin{aligned} G_{T'_j}(x'_j) &= \frac{card(\{F_m \in F / F_m(T'_j) \leq x'_j\})}{card(F)} \\ &= \frac{card(\{F_m \in F / F_m(T'_j) = 0\}}{card(F)} \\ &= \frac{card(A_j)}{card(F)} \end{aligned} \tag{5.19}$$

et

$$\begin{aligned} H(x'_1,...,x'_j) &= \frac{card(\{F_m \in F / F_m(T'_i) \leq x'_i,\ 1 \leq i \leq j\})}{card(F)} \\ &= \frac{card(B_j)}{card(F)}. \end{aligned} \tag{5.20}$$

Nous avons vu que $A_j = B_j$; donc $G_{T'_j}(x'_j) = H(x'_1,...x'_j)$ et $G_{T'_j}(x'_j) = H(x_1,...x_p)$. Nous savons aussi que $card(A_j) = min_{i=1,...,j}\ card(A_i)$ ce qui implique $G_{T'_j}(x'_j) = min_{i=1,...,j} G_{T'_i}(x'_i)$. De plus :

$$min_{i=1,...,j} G_{T'_i}(x'_i) = min_{i=1,...,p} G_{T_i}(x_i)$$

et donc

$$H(x_1,...,x_p) = min_{i=1,...,p} G_{T_i}(x_i) = C(G_{T_1}(x_1),...,G_{T_p}(x_p))$$

avec C la copules Min.

2. Si $x_p < 1$, $Min(G_{T_1}(x_1),...,G_{T_p}(x_p)) = G_{T_p}(x_p)$.
Nous notons $x'_1,...,x'_j$ (associés aux $T'_1,...,T'_j$) les x_i parmi $x_1,...,x_p$ qui sont inférieurs strictement à 1. Nous avons $x'_j = x_p$ et $min(G_{T_1}(x_1),...,G_{T_p}(x_p)) = G_{T_p}(x_p)$ (avec le lemme). Par la démonstration précédente nous avons :

$$min(G_{T_1}(x_1),...,G_{T_p}(x_p)) = min(G_{T'_1}(x'_1),...,G_{T'_j}(x'_j)).$$

Finalement :

$$min(G_{T_1}(x_1),...,G_{T_p}(x_p)) = G_{T_p}(x_p).$$

3. Pour tout $1 \leq i \leq p$, si $x_i < 1$, $G_{T_i}(x_i) = \mathbb{P}(Z > T_i) = 1 - F_Z(T_i)$. Par définition, $F_Z(T) = \mathbb{P}(\{Z(w) \leq T\})$ et $G_T(x) = \mathbb{P}(\{F_i \in \mathfrak{F} / F_i(T) \leq x\})$ est la proportion de distributions de masses unitaires F_i dont la valeur est $F_i(t) = 1$ pour $t > T$. $G_T(x)$ est la proportion

d'individus w tels que $Z(w) > T$.

4. Si $x_p < 1$, $F_Z(T_p) = 1 - H(x_1, ..., x_p)$. Avec 1. nous avons :

$$H(x_1, ..., x_p) = min(G_{T_1}(x_1), ..., G_{T_p}(x_p)),$$

avec 2. :

$$min(G_{T_1}(x_1), ..., G_{T_p}(x_p)) = G_{T_p}(x_p)$$

et avec 3. :

$$F_Z(T_p) = 1 - G_{T_p}(x_p).$$

5.10.2 Le mélange de densités comme cas particulier du mélange de copules

Nous introduisons les notations suivantes :

- $P = (P_1, ..., P_k)$ est une partition en K classes de l'ensemble $\Omega = \{w_1, ..., w_N\}$,
- F_Z est la fonction de répartition associée à une variable aléatoire quantitative Z décrivant les individus de l'ensemble $\Omega = \{w_1, ..., w_N\}$,
- F_{Z^k} est la fonction de répartition de la variable aléatoire Z^k définie sur les individus de la classe P_k,
- F^k est la base de distributions dont les éléments sont les distributions de masse unitaire associées à chaque valeur $Z^k(w_j)$ des individus de la classe P_k,
- G^k_T est la fonction de distribution de distributions au point T associée à la base de distributions F^k,
- $H^k_{T_1,...,T_p}$ est la fonction de distribution jointe de p de distributions associée à F^k.

Proposition 17 (Diday, [33]) *:*
Si $H_{T_1,...,T_p} = \sum_{k=1}^{K} \pi_k H^k_{T_1,...,T_p}$ *avec* $\sum_{k=1}^{K} \pi_k = 1$ *alors*

$$F_Z = \sum_{k=1}^{K} \pi_k F_{Z^k}.$$

Démonstration:
Avec le théorème de Sklar nous avons $H^k_{T_1,...,T_p}(x_1, ..., T_p) = C^k(GT_1{}^k(x_1), ..., G^k_{T_p}(x_p))$ avec C^k une sous-copule et donc $H_{T_1,...,T_p}(x_1, ..., x_p) = \sum_{k=1}^{K} \pi_k C^k(G^k_{T_1}(x_1), ..., G^k_{T_p}(x_p))$. Nous pouvons choisir $x_p < 1$ et utiliser la proposition 16.
Avec 1. nous avons : $H_{T_1,...,T_p}(x_1, ..., x_p) = \sum_{k=1}^{K} \pi_k min(G^k_{T_1}(x_1), ..., G^k_{T_p}(x_p))$.
Avec 2. nous avons : $H_{T_1,...,T_p}(x_1, ..., x_p) = \sum_{k=1}^{K} \pi_k G^k_{T_p}(x_p)$.
Avec 3. nous avons : $H_{T_1,...,T_p}(x_1, ..., x_p) = \sum_{k=1}^{K} \pi_k (1 - F_{Z^k}(T_p)) = 1 - \sum_{k=1}^{K} \pi_k F_{Z^k}(T_p)$.
Avec 4. nous avons : $F_Z(T_p) = 1 - H_{T_1,...,T_p}(x_1, ..., x_p)$ et $F_Z(T_p) = \sum_{k=1}^{K} \pi_k F_{Z^k}(T_p)$.
Le même raisonnement peut être fait pour toute suite $T_1, ..., T_p$ et finalement :

$$F_Z = \sum_{k=1}^{K} \pi_k F_{Z^k}.$$

La proposition 17 prouve que la décomposition de mélange de copules est une généralisation de la décomposition de mélange classique. Le résultat vient du fait qu'on peut induire $F_{Z^k}(T_1), ..., F_{Z^k}(T_p)$ à partir de $G^k_{T_1}(x_1), ..., G^k_{T_p}(x_p)$. De plus, en choisissant le "meilleur modèle" (celui qui adhère le mieux à $F_{Z^k}(T_1), ..., F_{Z^k}(T_p)$ pour chaque k) - Gaussien, Gamma, Poisson,... - pour chaque Z^k, nous obtenons un modèle différent pour chaque composante du mélange.

5.11 Conclusions

Nous avons proposé une extension des méthodes de décomposition de mélange de densités à des données probabilistes. Cette extension repose sur deux notions: les fonctions distribution de distributions et les copules. L'originalité de la méthode DMC est due aux données qu'elle traite (des fonctions de répartition) et à ses sorties. En effet, en plus de la classification des individus, la méthode nous permet d'obtenir une loi de données probabilistes ainsi qu'une mesure de dépendance entre les FDD d'une variable (à l'intérieur des classes) et entre les variables (lors de la méthode par couplage).

Nous pouvons étendre au cas des données fonctions de répartition la plupart des méthodes de décomposition de mélange (Nuées dynamiques, EM, SEM,...). Ces extensions sont basées sur le fait que nous travaillons sur des densités associées aux fonctions copules (elles-mêmes fonctions de répartition). Les résultats théoriques généraux portant sur la décomposition de mélange de densités restent donc valables. Nous avons également pu obtenir différents résultats théoriques basés sur la notion de distributions de distributions et nous avons pu démontrer que la décomposition de mélange de densités est un cas particulier du mélange de copules.

Par ailleurs, les méthodes de décomposition de mélange de copules que nous proposons sont en fait des méthodes applicables à l'analyse de données fonctionnelles. En effet, les données d'entrée des méthodes n'ont pas nécessairement à être des fonctions de répartition. De manière plus générale, elles peuvent être des applications numériques. Le nom de "fonction de distribution de distributions" (FDD) vient du fait que l'application climatique effectuée au chapitre 6 a été développée en parallèle avec la mise au point de la méthode. L'application ayant pour but de travailler avec pour données de base les fonctions de répartition de variables thermodynamiques pour chacun des individus, le nom de "fonctions de distributions de distributions" s'est imposé naturellement.

L'application climatique présentée au chapitre 6 n'a été réalisée que sur le développement de l'approche classification (par nuées dynamiques) pour des raisons de temps et parce que l'intérêt suscité par le mélange de copules dans la communauté de la physique atmosphérique, portait initialement sur les résultats de classification. Cette méthode apparaissait donc comme privilégiée.

Chapitre 6

Application climatique

"How can it be that mathematics, being after all a product of human thought independent of experience, is so admirably adapted to the objects of reality?"
Einstein, Albert (1879-1955)

"It is a capital mistake to theorize before one has data."
Doyle, Sir Arthur Conan (1859-1930), Scandal in Bohemia.

6.1 Introduction

Nous appliquons dans ce chapitre la méthode de décomposition de mélange de copules (DMC) par approche classification de type nuées dynamiques, à une base de données climatologiques. En effet, la recherche climatique mondiale est dominée par l'inquiétude provoquée par l'action de l'homme sur l'évolution du climat. Les premiers constats dressés par l' "Intergovernmental Panel on Climate Change" (IPCC) montrent que l'augmentation artificielle des concentrations de gaz à effet de serre dans l'atmosphère induit déjà et va continuer d'induire, plus rapidement que d'autres phénomènes naturels, une évolution potentiellement catastrophique du climat terrestre. Pour cette raison, les recherches actuelles se focalisent autour d'une meilleure compréhension des phénomènes dynamiques, thermodynamiques et radiatifs qui définissent le climat. La modélisation physique des lois naturelles et l'observation de la planète depuis l'espace par des satellites forment l'ossature de cette recherche.

L'avancée rapide des recherches sur la variabilité du climat et sur son évolution repose sur un couplage étroit entre la modélisation, qui tend à prendre en compte un nombre croissant de processus et leurs intéractions grâce au développement fantastique des moyens de calculs, et l'observation *in situ* ou spatiale, qui permet l'étude fine de mécanismes complexes et leur paramétrisation précise et apporte, en particulier grâce aux satellites, les données globales essentielles à l'évaluation des résultats des modèles et des hypothèses qu'ils expriment.

L'outil de base de la *modélisation* est un code de calcul, décrivant au mieux la complexité du système Terre-Océan-Atmosphère, reposant sur des équations d'évolution. De tels codes comportent plusieurs dizaines de milliers d'instructions. Malgré l'utilisation des plus gros ordinateurs, ils sont encore limités dans leurs résolutions spatiales (mailles de l'ordre de 100 km), ce qui interdit de décrire convenablement nombre de phénomènes importants de petites échelles : convection humide, nébulosité, turbulence, effets du relief, etc.

L'outil de base de *l'observation globale* et continue de la planète est le satellite et les instruments qu'il emporte, couplés à des modèles du transfert radiatif (directs et inverses). La situation actuelle, bien qu'imparfaite, permet l'observation des principales composantes du système Terre-Océan-Atmosphère grâce à des techniques de mesure des flux radiatifs, emis vers la plate-forme spatiale, dans un domaine spectral couvrant les hyperfréquences, l'infrarouge, le visible et au-delà.

6.1.1 La réanalyse de données satellitaires NOAA/TOVS de 1979 à nos jours à des fins d'étude du climat

Parmi les paramètres que l'on déduit couramment de l'observation satellitaire de type sondage vertical, les profils verticaux de température - de la surface à la stratosphère - et d'humidité sont les plus importants pour décrire l'état du milieu. D'autres paramètres tels que ceux qui décrivent la couverture nuageuse (niveau, température, répartition, propriétés radiatives), l'état de la surface (température, émissivité, présence de glace ou de neige, désert...) ou proche de la surface (température et humidité), la composition de l'atmosphère (gaz à effet de serre et/ou réactifs), sont également des paramètres clés pour la météorologie et la climatologie.
Les satellites - et plus précisément ceux de la série NOAA (National Oceanic and Atmospheric Administration) - enrichissent, année après année depuis 1979, une archive considérable d'observations globales de la Terre.
Mis au point par des groupes de travail internationaux sous l'égide, notamment de la NASA, de la NOAA, du WCRP (World Climate Research Programme) ou de la CEE (Commission des Communautés Européennes), plusieurs grands programmes d'analyse de ces données sont en cours.
L'interprétation de ces mesures en terme de variables thermodynamiques ou dynamiques des milieux observés, peut, dans certains cas, se révéler extrêmement complexe, mélant de nombreuses *disciplines*. De cet ensemble de domaines croisés sont nés des modèles et des méthodes qui ont permis, par exemple, grâce à l'implication du groupe ARA (Analyse du Rayonnement Atmosphérique) du LMD (Laboratoire de Météorologie Dynamique) dans le programme NOAA/NASA "Pathfinder", la réanalyse des observations du sondeur vertical à bord des satellites polaires opérationnels de la série TIROS-N / NOAA.

Parmi ces disciplines, l'analyse statistique de données est une discipline ancienne qui prend une dimension de plus en plus grande dans le domaine de la recherche climatique, compte tenu de la difficulté des problèmes posés. Son application est complexe parce que, d'une part, il est très difficile de réunir des échantillons réellement indépendants et, d'autre part, presque toutes les données y sont corrélées, à la fois dans le temps et dans l'espace. Cette spécificité des

applications géophysiques s'est d'ailleurs traduite par l'émergence de nombreuses techniques statistiques spécialisées.

C'est dans ce contexte, qu'a été réalisée au sein du groupe d'Analyse du Rayonnement Atmosphérique (ARA) du Laboratoire de Météorologique Dynamique (LMD) la réanalyse de près de 10 années d'observation quotidienne de la planète par les satellites NOAA (1987-1995). L'algorithme d'inversion de l'équation de transfert radiatif ("Improved Initialization Inversion", Chédin et al, [18]) développé au LMD permet entre autre d'interpréter des observations satellitaires en terme de variables thermodynamiques atmosphériques, c'est-à-dire, à partir de mesures de flux de rayonnement, de retrouver les valeurs de variables thermodynamiques de l'atmosphère (température, humidité spécifique, nuages, ozone, etc.). Les valeurs de ces variables se présentent généralement sous la forme de *profils atmosphériques*. Un profil atmosphérique (en une localisation géographique) est un ensemble de valeurs numériques associées à différentes pressions atmosphériques données, elles-mêmes associées à différentes altitudes. Pour illustrer cette notion, nous donnons un exemple de profil de température en Figure 6.1.

L'algorithme d'inversion cité plus haut, repose sur une partition de l'ensemble des profils atmosphériques représentatifs de l'atmosphère terrestre, et utilise une connaissance a priori dans une méthode d'inférence de type Bayésien. La partition utilisée doit donc regrouper les profils ayant des propriétés physiques proches à l'intérieur d'une classe et bien distinctes entre les classes.

La partition des profils, utilisée jusqu'à présent, se nomme "TIGR" (Thermodynamic Initial Guess Retrieval, Chédin et al, 1985, [18], Achard, 1991, [1], Chevallier, 1998, [19]) et est réalisée en deux étapes successives. La première étape consiste à récolter des profils parmi l'infinité de profils existants en prenant un échantillon réduit et représentatif (80 000 radio-sondages pour TIGR-2). Appelons S cette base de données initiales. La seconde étape est l'échantillonnage de S par une approche topologique basée sur un indice I qui mesure la dissimilarité entre deux situations atmosphériques. A l'étape une, une première situation atmosphérique est choisie aléatoirement et archivée dans un nouvel ensemble E. A l'étape n, une $n^{\text{ième}}$ situation atmosphérique est choisie aléatoirement et archivée dans l'ensemble E si elle est suffisamment différente (relativement à l'indice de dissimilarité I) des situations déjà sélectionnées ([19]). L'échantillon E est classé en 5 classes (deux polaires, deux tempérées, une tropicale) par une classification ascendante hiérarchique. L'approche utilisée rend les distributions sur l'espace des variables (température et humidité spécifique) plus uniformes dans E que dans S. Pour visualiser la classification en 5 classes obtenue par la méthode CAH dans [19], ainsi que pour illustrer la notion de profil présentée ci-dessus, nous traçons pour chacune des 5 classes le profil moyen de température et d'humidité en fonction de la pression atmosphérique, le profil moyen moins l'écart-type (de chaque niveau de pression), le profil moyen plus l'écart-type. Ces tracés se trouvent en Figure C.1, page 207, annexe C.

L'approche par décomposition de mélange de copules (DMC) considère l'ensemble initial S dans son entier afin d'avoir les distributions sur l'espace des variables proches des distributions réelles. En effet, la méthode de classification ascendante hiérarchique (CAH), appliquée ci-dessus sur la base de données TIGR, ne permet d'obtenir que des classifications. La méthode DMC que nous proposons d'appliquer sur des données climatiques, permet également

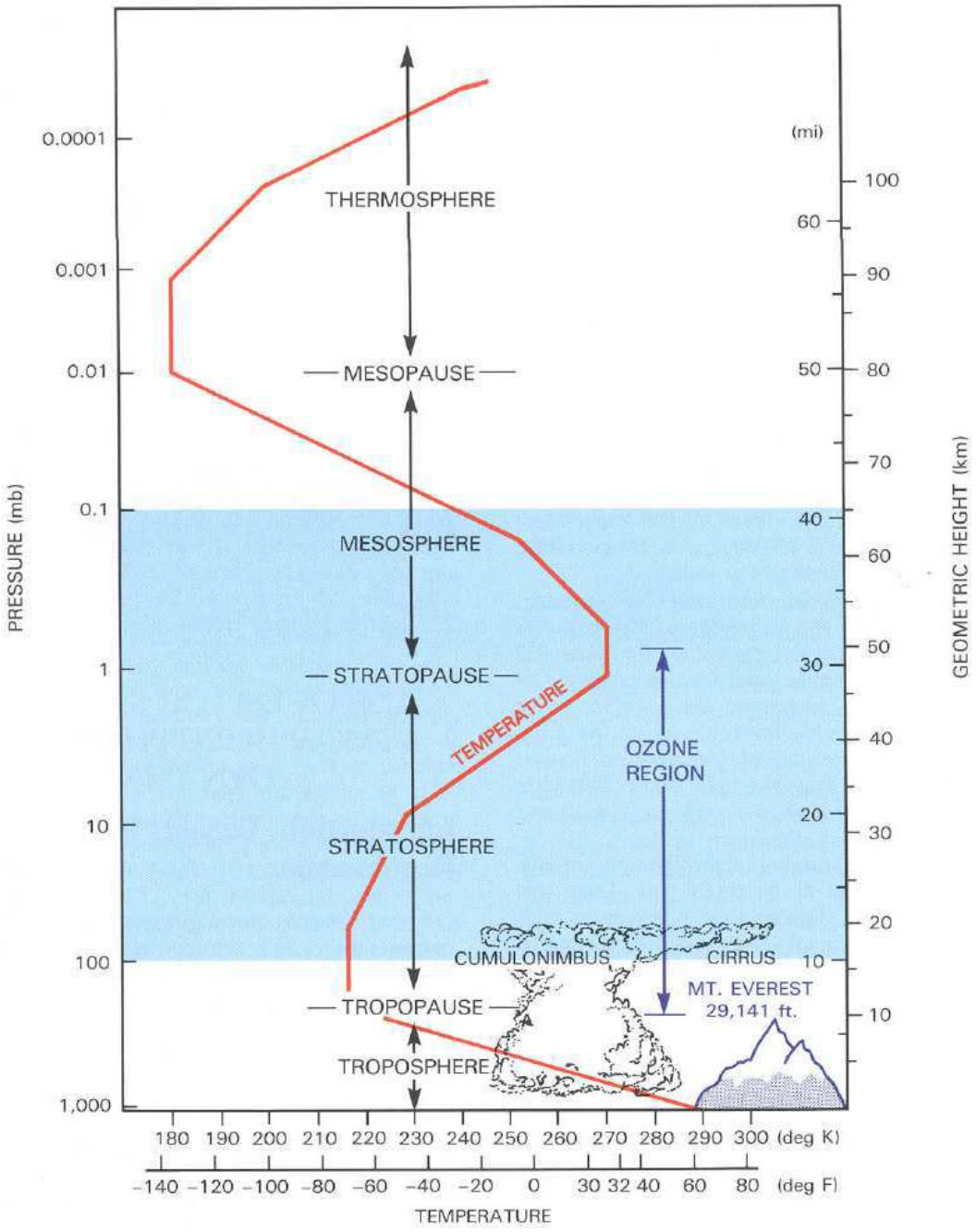


FIG. 6.1: *Exemple de profil de température*

d'obtenir des classifications en type de masse d'air, mais nous donne, en plus, pour chacune des classes, les fonctions de distribution de distributions (FDD) à des T donnés (voir section 5.3), ainsi que les paramètres de copules décrivant la relation entre ces FDD et modélisant la dépendance entre les variables thermodynamiques utilisées. L'information a posteriori apportée par la méthode DMC nous indique alors les lois de probabilité des variables considérées et les probabilités d'occurence des profils atmosphériques.

6.1.2 Les données climatiques

Pour définir une classification ayant les caractéristiques données ci-dessus, nous n'utilisons pas de données de radio-sondages. Nous disposons de données atmosphériques provenant du centre Européen ECMWF (European Center for Medium range Weather Forecasting) de Reading. Un maillage du globe terrestre est réalisé ; chaque maille correspondant à un degré de latitude et un degré de longitude. Ce maillage est étendu en altitude sur 50 niveaux appelés "coordonnées sigma". Ces 50 niveaux d'altitude varient en fonction de la pression de surface du profil atmosphérique considéré (situé en L×l). Connaissant la pression de surface PS du point de longitude L et de latitude l, la coordonnée sigma du niveau k, notée $\sigma(k)$, est donnée par $\sigma(k) = \frac{P(k)}{PS}$ où $P(k)$ est la pression atmosphérique du niveau k calculée par la formule $P(k) = \frac{1}{2}(P_h(k) + P_h(k+1))$, avec la pression de "demi niveau" notée $P_h(k)$ (pour "half level") définie par $P_h(k) = \alpha(\mathrm{k}) + \beta(\mathrm{k}) \times \mathrm{PS}$, $\alpha(k)$ et $\beta(k)$ étant fixes.
C'est-à-dire que la k^{eme} coordonnée sigma est le rapport de la pression atmosphérique au niveau k sur la pression de surface et donc les coordonnées sigma sont toujours entre 0 et 1. Pour chaque point du maillage de l'atmosphère, nous disposons des valeurs de la pression, de la température, de l'humidité, du vent, etc. Ces valeurs sont celles des prévisions ("forecast") à six heures d'intervalle, réalisées 4 fois par jour (0h, 6h, 12h, 18h) sur une période allant de décembre 1998 à décembre 1999. Nous possédons par conséquent un "quadrillage" tridimensionnel de l'atmosphère de la terre, représentant un an de son état thermodynamique complet 4 fois par jour.

Regardons les étapes préliminaires nécessaires à l'application de notre méthode ainsi que les résultats obtenus pour les variables température et humidité.

6.2 Classification par décomposition de mélange de copules - Exemple de 7 classes

A partir des données climatiques de l'ECMWF (voir section 6.1.2), nous souhaitons obtenir une classification en type de masse d'air de l'atmosphère. Pour appliquer la méthode de décomposition de mélange de copules (DMC), nous devons disposer des fonctions de répartition pour chaque profil atmosphérique et pour chaque variable physique que l'on souhaite impliquer dans la méthode. Chacune de ces fonctions est calculée de la manière suivante:

Pour chaque profil atmosphérique, et pour chaque variable physique (e.g. la température), les valeurs numériques des différents niveaux "sigma" sont remises "à plat" sur $\mathbb{R}$ et des méthodes d'estimation de densité (et donc de fonctions de répartition par intégration) sont

appliquées pour obtenir la répartition des données pour chaque profil. Dans l'application actuelle, la modélisation est effectuée par l'intégration de la densité obtenue par la méthode des noyaux de Parzen. La méthode est répétée pour chacun des profils.

Dans le but de ne pas surcharger la méthode par un trop grand nombre de données, nous ne travaillons que sur les données du 15 décembre 1998 à 0H (GMT), et nous ne prenons qu'un profil sur deux en longitude et qu'un profil sur deux en latitude, soient (360/2) longitudes $\times$ (180/2) latitudes = 16200 individus profils atmosphériques et donc autant de fonctions de distribution dans le cas d'une variable physique.

6.2.1 Classification en température

Dans un premier temps, nous souhaitons étudier la variable de température et nous fixons donc deux valeurs de température T_1 et T_2. Ces valeurs sont fixées par des méthodes visuelles de type tracé de distributions de distributions (voir section 5.3), par la visualisation de la répartition des données de température (tous niveaux confondus) avec des tracés de densités ou d'histogrammes, et par les connaissances a priori d'experts sur les valeurs de la température significatives de phénomènes physiques. Les deux valeurs de température fixées sont $T_1 = 225$ K et $T_2 = 265$ K. Différentes valeurs de T_1 et T_2 ont, de plus, été testées, prouvant que pour la variable température, la sensibilité des résultats aux 2 valeurs de T choisies n'est pas très importante. Disposant des valeurs des fonctions de répartition aux valeurs T_1 et T_2 pour chacun des profils, la méthode DMC peut être lancée.

La première classification en température comporte 7 classes. Les copules utilisées sont de la famille de Frank et les FDD sont modélisées par des lois béta. La partition est projetée sur une carte du globe avec un grossissement artificiel des pixels de manière à disposer d'un effet visuel continu. Les résultats de cette classification se trouvent Figure 6.2 et les paramètres dans le Tableau 6.1.

Classes	β	ν_1 en T_1	ν_2 en T_1	ν_1 en T_2	ν_2 en T_2
1	0.000001	6.836969	14.342546	12.208704	2.217218
2	0.300001	11.408380	69.945442	21.956680	14.064272
3	0.004093	12.180901	70.0	61.601810	70.0
4	0.000001	12.651747	70.0	56.703354	70.0
5	0.000001	13.335871	70.0	11.891472	11.261731
6	0.030567	6.040135	25.066311	8.938780	3.687328
7	0.007445	8.839353	22.021719	19.165813	2.168266

TAB. 6.1: *Paramètres de la classification en 7 classes en température*

Les classes de la partition semblent cohérentes et posséder des propriétés climatiques distinctes et réalistes. On peut retrouver par exemple une grande classe dite "tropicale" (classe 4), deux classes "polaires" correspondant à l'été dans l'hémisphère sud et l'hiver dans l'hémisphère nord (classes 1 et 7), deux classes "tempérées" (classes 2 et 5). La classe 3 fait le lien entre les zones tempérées et les zones tropicales tandis que la classe 6 fait le lien entre les zones tempérées et les zones polaires. On note que certains reliefs élevés sont identifiés (Himalaya, Cordillère des Andes) malgré l'utilisation de coordonnées σ. De plus, les classes 1,

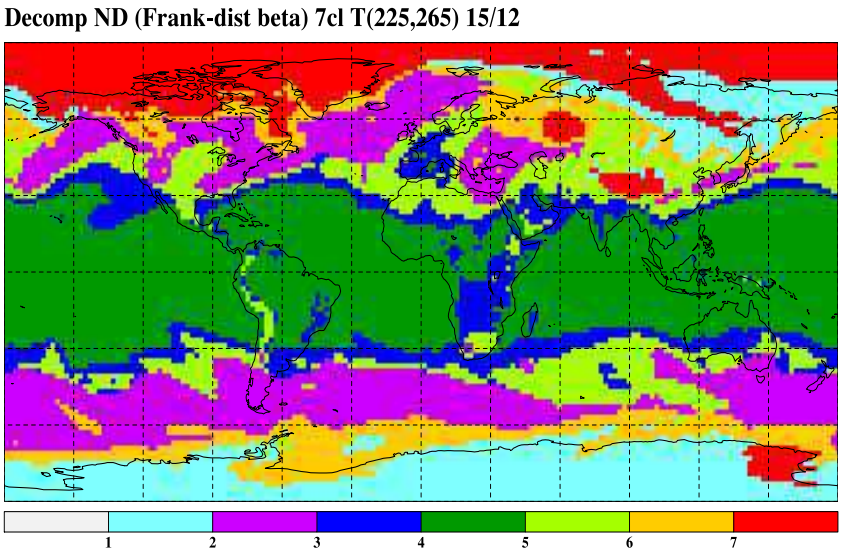


FIG. 6.2: *Classification en 7 classes par DMC sur la température ($T_1 = 225\ K$, $T_2 = 265\ K$) pour le 15 décembre 1998 à 0H*

4 et 5 ont des paramètres de copules identiques (0.000001) et signifiants que leur copule est proche de la copule *min* (voir section 4.6.2.3). Autrement dit, les fonctions de distribution à l'intérieur de chacune des classes ont tendance à évoluer parallèlement sans se couper. Une manière de décrire les classes de la partition est de visualiser le profil moyen de température (plus ou moins un ecart-type) de chaque classe aux différents niveaux sigma. Ces tracés sont placés en Figure C.2, page 208, et nous permettent de voir , par exemple, que la forme des profils moyens des classes 3 et 4 de la classification en température, est similaire à celle de la classe tropicale (classe 1) de TIGR3. Par ailleurs, une comparaison intéressante peut être faite avec le tracé de la température moyenne entre 500 et 700 hPa (Figure 6.3). Nous voyons que les classes de transitions entre les classes tempérées et tropicales de la Figure 6.2 correspondent bien à des zones de transition de la température 500-700 hPa de la Figure 6.3. Les "langues" (incursions d'air chaud dans des masses d'air plus froides ou inversement) sont bien identifiées. Le disque rouge situé à 60° N $\times$ 60° E sur la Figure 6.2, s'explique parfaitement par la Figure 6.3 de même que la forme de la classe 2 sur l'Amérique du nord. L'accord avec l'analyse synoptique de la situation est très bon (formes des incursions, position des dépressions creuses, etc.). Un autre manière de décrire les classes est de regarder la fonction de densité de probabilité pour la variable température à différents niveaux de pression de l'atmosphère. Ces densités donnent des informations sur la répartition des températures dans chaque classe et donc sur leur discrimination. Nous voyons par exemple dans les Figures 6.5 et 6.6 (correspondant aux niveaux 900 hPa et 500 hPa respectivement) que les classes sont bien discriminées (l'association classes-densités est donnée en Figure 6.4). Nous retrouvons également que les classes 1 et 7 sont les deux classes froides de la classification et que la classe 4 est la plus chaude de toutes. Nous voyons parfaitement les différences entre les classes dites de transition (ou classes plus tempérées). D'après la classification de la Figure 6.2, nous

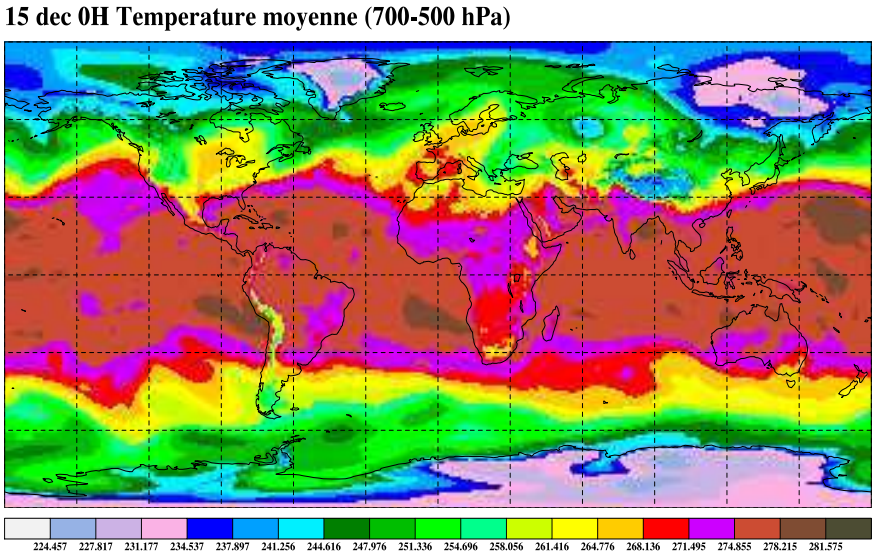


FIG. 6.3: *Température moyenne entre 500 et 7OO hPa du 15/12/98 à 0H*

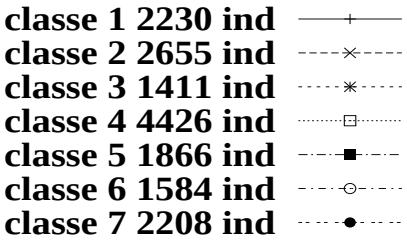


FIG. 6.4: *Association 7 classes - densités température*

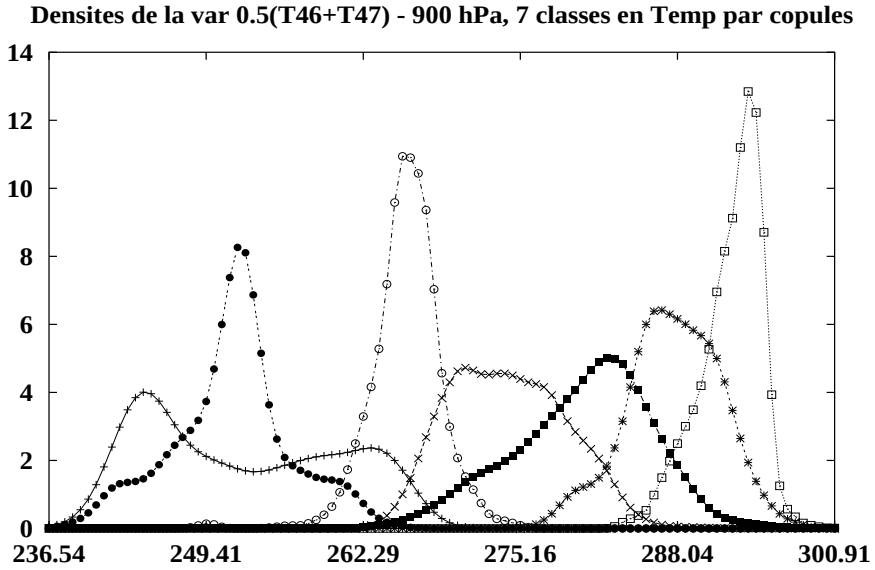


FIG. 6.5: *Densités de la température (Kelvin) par classe à 900 hPa*

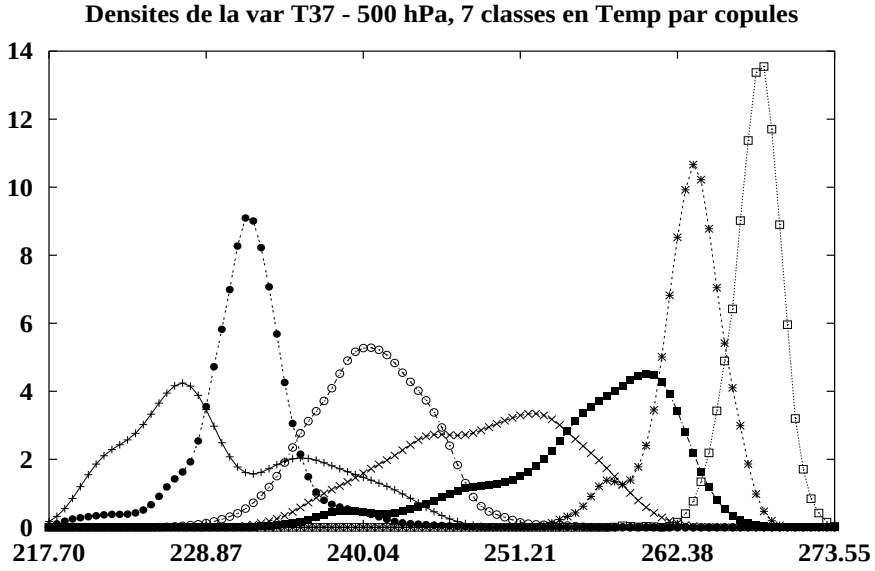


FIG. 6.6: *Densités de la température (Kelvin) par classe à 500 hPa*

pouvons trier ces classes par température moyenne croissante, de la plus proche des classes polaires (1 et 7) à la plus proche de la classe tropicale (classe 4) : classe 6, classe 2, classe 5, classe 3. Cet ordre est totalement vérifié par le calcul des densités.

Cependant, plus nous montons en altitude (i.e. plus la pression diminue), moins les classes se discriminent entre-elles : les densités empiètent davantage les unes sur les autres. C'est en particulier le cas au-dessus de la tropopause (niveau de recroissance de la température), par exemple à 70 hPa (voir Figure 6.7). La classe 4, tropicale, reste très distincte, et, cette fois,

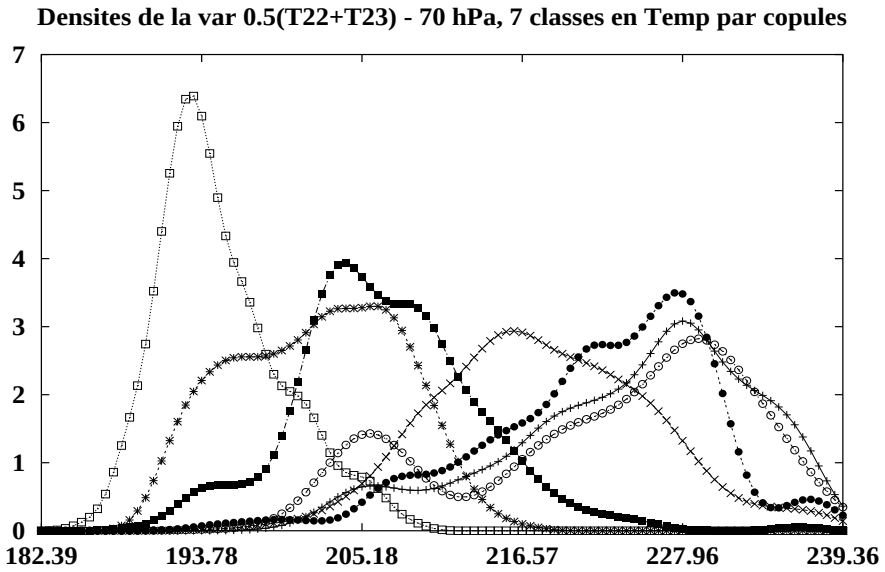


FIG. 6.7: *Densités de la température (Kelvin) par classe à 70 hPa*

devient la plus froide, comme attendu.

6.2.2 Classification en humidité

De la même manière que pour la variable température, les 16200 fonctions de répartition de la variable d'humidité sont calculées et données en entrée de la méthode DMC aux valeurs H_1 et H_2 fixés à $H_1 = 0.00003\ kg/kg$ et $H_2 = 0.006\ kg/kg$. Les copules utilisées appartiennent à la famille de Frank et les FDD sont modélisées par lois béta. La projection de la partition se trouve Figure 6.8 et les paramètres Tableau 6.2.

Le résultat visuel est plus "fouilli" que pour la température. Cet effet était prévisible du fait de la plus grande variabilité de l'humidité. Pour décrire cette partition, nous pouvons nous appuyer sur la carte du *Total Column Water Vapor* (TCWV) correspondant à la quantité totale de vapeur d'eau intégrée verticalement par profil (Figure 6.9). Les points communs entre la Figure 6.8 et la Figure 6.9 sont multiples. Les bras d'incursions d'air humide sont précisément définis et la plupart des formes sont retrouvées avec une précision étonnante. On peut remarquer que la classe tropicale précédente est scindée en deux classes (3 et 4). La classe 4 correspond aux zones les plus humides de l'analyse. De plus, toute la frontière entre la classe 3 et une masse d'air moins humide (classe 2) est bordée par la classe 5 qui correspond à des incursions d'air humide dans un milieu plus sec. Par ailleurs, la méthode

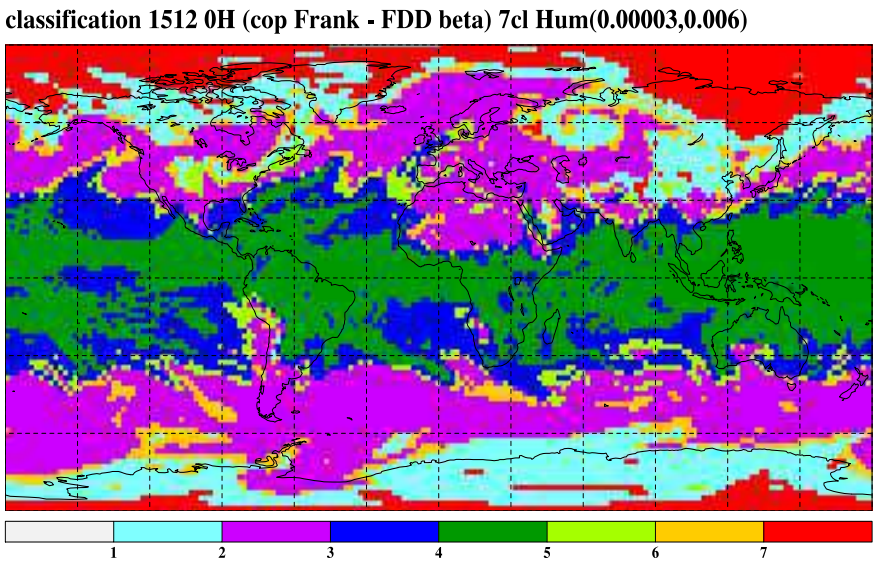


FIG. 6.8: *Classification en 7 classes par DMC sur l'humidité ($H_1 = 0.00003\ kg/kg$, $H_2 = 0.006\ kg/kg$) pour le 15 décembre 1998 à 0H*

Classes	β	ν_1 en T_1	ν_2 en T_1	ν_1 en T_2	ν_2 en T_2
1	0.200001	1.772749	36.192062	70.0	24.484861
2	0.016939	6.292977	742.659424	30.004604	13.469539
3	0.619099	0.000001	12.617126	16.619267	20.368376
4	0.445017	0.000001	12.617126	29.424589	48.242210
5	0.100001	0.000001	12.617126	6.890862	5.887025
6	0.020641	0.000001	12.617126	38.931557	14.840351
7	0.017804	2.215375	23.266142	70.0	18.847424

TAB. 6.2: *Paramètres de la classification en 7 classes en humidité*

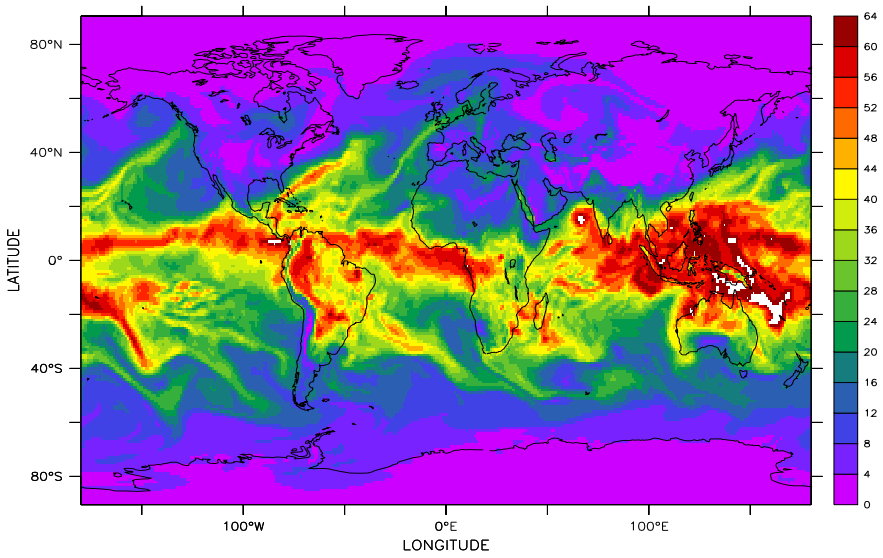


FIG. 6.9: *Total Column Water Vapor (Kg/m^2) 15/15/98 à 0H*

identifie deux classes différentes de très faible humidité (classes 1 et 7). Les profils moyens d'humidité présentés en Figure C.3 page 209 confirment cette identification. On peut voir que malgré des valeurs d'humidité du même ordre, ces deux classes ont des paramètres de copules différents (0.2 et 0.018) traduisant une dépendance différente dans le comportement de leurs fonctions de répartition. On voit également une "spirale" située à (60 ° N, 60 ° E), présente aux mêmes coordonnées sur le TCWV, et qui correspond parfaitement à la dépression centrée sur cette zone (cette particularité se retrouve sur la classification en température sous la forme d'un disque rouge, voir carte synoptique en Figures C.12 et C.13, pages 218 et 219).

De plus, nous pouvons voir que la dissymétrie induite par la différence été-hémisphère sud / hiver-hémisphère nord dans la classification en température, est moins marquée dans la classification en humidité.

Le tracé des densités de chaque classe à différents niveaux nous donne également des informations sur la manière dont les classes se discriminent. Nous voyons dans la Figure 6.11, correspondant aux densités à 900 hPa, que la classe 1, avec son "aile" de densité plus humide, correspond à des situations plus fréquentes que la classe 7, très sèche et associée à des situations de type "hiver polaire". De plus, cette figure prouve que la discrimination est très nette pour les niveaux bas de l'atmosphère. Pour les niveaux plus élevés, l'humidité diminue extrêmement rapidement, conduisant les densité à se superposer, diminuant ainsi leur pouvoir discriminant. Cependant, nous retrouvons parfaitement ce que la comparaison avec le TCWV nous laissait penser :

- les classes 1 et 7 ont les densités dont le mode statistique est le plus faible (i.e. ce sont les classes les plus sèches),

- la classe 4 est la classe la plus humide avec une densité ayant un pic vers 0.015 kg/kg,

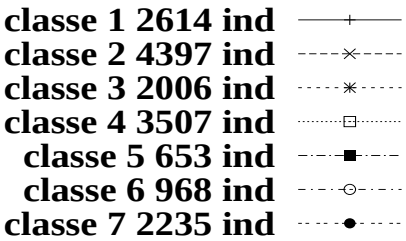


FIG. 6.10: *Association classes - densités humidité*

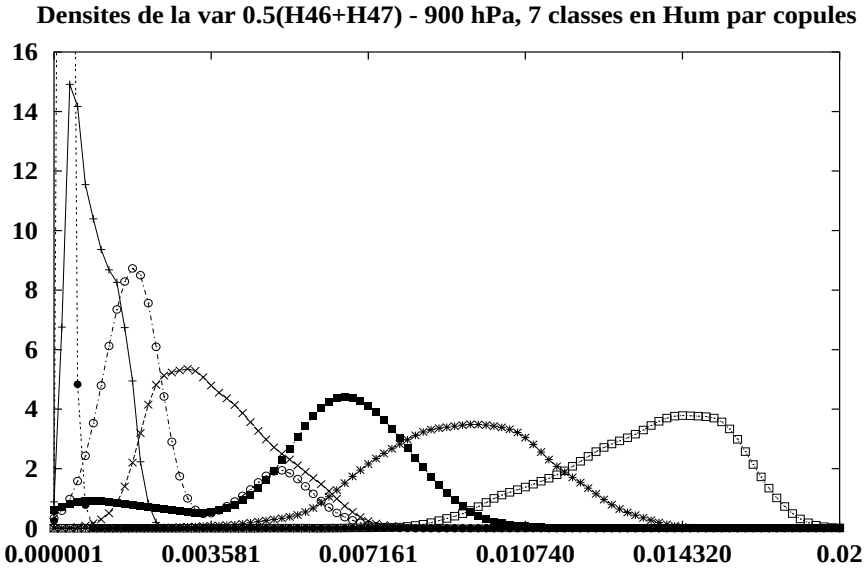


FIG. 6.11: *Densités de l'humidité spécifique (kg/kg) par classe à 900 hPa*

– les classes tempérées et de transition peuvent s'ordonner de la plus sèche à la plus humide : classe 3, classe 5, classe 2, classe 6.

Nous pouvons dire que les classifications obtenues par DMC sur la variable température puis sur la variable humidité sont très bonnes au regard de la cohérence physique des classes et de la connaissance a priori dont nous disposons. Qu'en est-il lorsqu'on souhaite une classification couplant les variables?

6.2.3 Résultat de DMC par couplage

A partir des deux classifications précédentes en température et en humidité, nous appliquons la méthode de couplage de DMC afin d'obtenir une partition qui tienne compte des deux variables physiques. La modélisation est faite par copules de Frank et par lois béta. La classification est projetée Figure 6.12 avec les paramètres dans le Tableau 6.3.

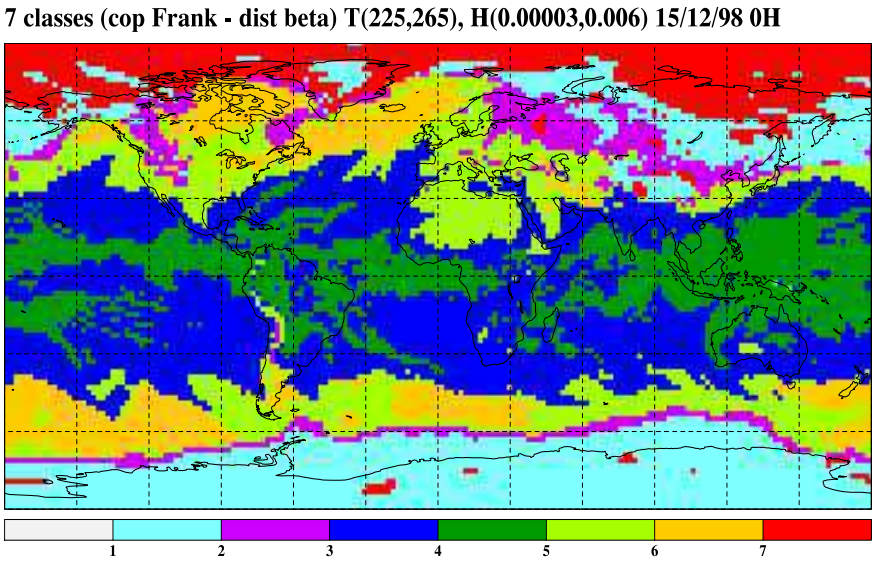


FIG. 6.12: *Classification en 7 classes par couplage de DMC sur la température et l'humidité pour le 15 décembre 1998 à 0H*

La méthode de couplage semble donner de bons résultats: mélange cohérent des deux précédentes classifications. La première remarque qu'on peut faire est que la dissymétrie été - hémisphère sud / hiver - hémisphère nord, induite par la variable température, se retrouve nettement, de même que la grande variabilité de l'humidité.

De plus, deux classes tropicales sont identifiées (classe 4 très humide et classe 3 un peu moins). La classe 4 de la Figure 6.12 correspond beaucoup mieux que la classe 4 de la figure 6.8, aux zones les plus humides de l'analyse. L'accord avec le TCWV est quasiment parfait. Les autres classes décrivent la transition des classes tropicales (chaudes et humides) aux classes polaires (froides et sèches). La "spirale" (60 ° N,60 ° E) est toujours présente et est significative de la dépression centrée sur cette zone. Les profils moyens de température et d'humidité pour chaque classe, représentés en Figures C.4 et C.5 (page 210 et page 211)

Classes	β	ν_1 en T_1	ν_2 en T_1	ν_1 en T_2	ν_2 en T_2	p_k
1	0.000001	6.712667	2.140064	5.703492	5.222391	0.197131
2	0.100001	70.0	70.0	10.42458	14.541202	0.057204
3	0.200001	18.965822	88.125916	8.056098	145.218979	0.253114
4	0.050867	19.533854	112.066284	6.489847	357.520905	0.146642
5	0.362295	12.315609	31.493969	5.033236	18.545059	0.138156
6	0.126157	0.86489	7.177879	3.316219	7.178005	0.088706
7	0.003896	23.222773	4.773149	13.366582	3.108013	0.119048

TAB. 6.3: *Paramètres de la classification en 7 classes en température et humidité*

confirment les significations données à ces classes. De manière générale, les détails, induits par la température ou par l'humidité, restent d'une grande précision : nous pouvons voir, par exemple, une zone plus froide et plus sèche que son voisinage (classe 5 et 6), dans le sud Australien, parfaitement visible sur la figure 6.9.

Si on compare les densités de probabilité de la variable température à 900 hPa, de la classification en température (Figures 6.2 et 6.5) avec celle des classes obtenues par couplage (Figures 6.12 et 6.14), ces dernières ont un pouvoir discriminant un peu inférieur.

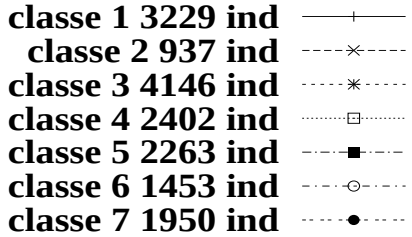


FIG. 6.13: *Association classes - densités couplage (température, humidité)*

Cet effet était prévisible du fait que la partition par couplage tient compte des deux variables (température et humidité). Les classes restent cependant relativement distinctes en température et ceci jusqu'à la tropopause ; les densités à 900 hPa (Figure 6.14) et à 300 hPa (Figure 6.15) empiètent davantage les unes sur les autres que sur la Figure 6.5 mais chaque classe ressort clairement. Au-dessus de la tropopause (i.e. dans la stratosphère), les densités sont un peu moins distinctes les unes des autres. La Figure 6.16 des densités à 70 hPa illustre cet effet. Nous pouvons voir sur cette figure que la classe 4, qui est la plus chaude entre 900 et 300 hPa, est la plus froide à 70 hPa. Ce phénomène est également présent dans la classe 3 : classe plus chaude que la plupart des classes dans les parties basses de l'atmosphère, plus froide que la plupart lorsque la pression diminue. L'inversion, attendue pour ces classes tropicales, est parfaitement mise en évidence.

Par ailleurs, de même que pour la température, les densités de probabilité de l'humidité à 900 hPa (Figure 6.17) des classes par couplage ont une discrimination inférieure à celles des classes en humidité seule (Figure 6.11). Elles restent cependant, pour la plupart, relativement

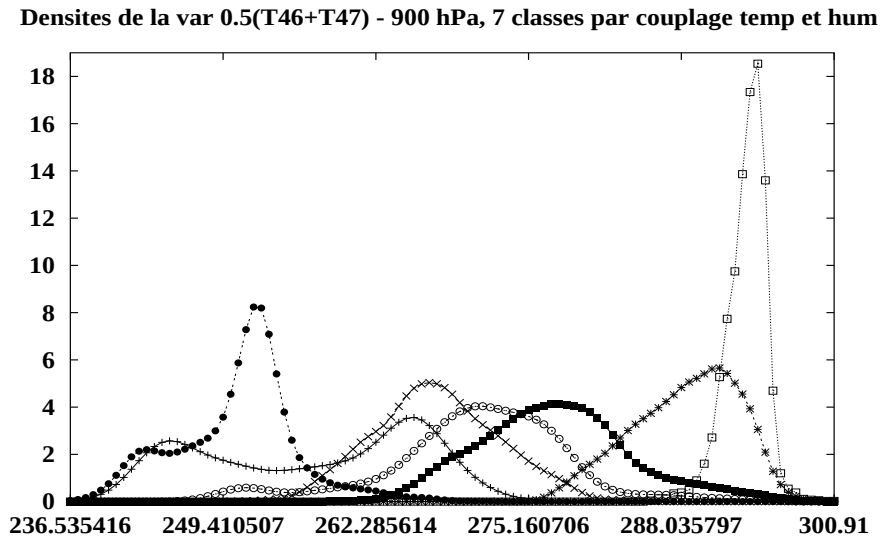


FIG. 6.14: *Densités de la température (K) par classe à 900 hPa*

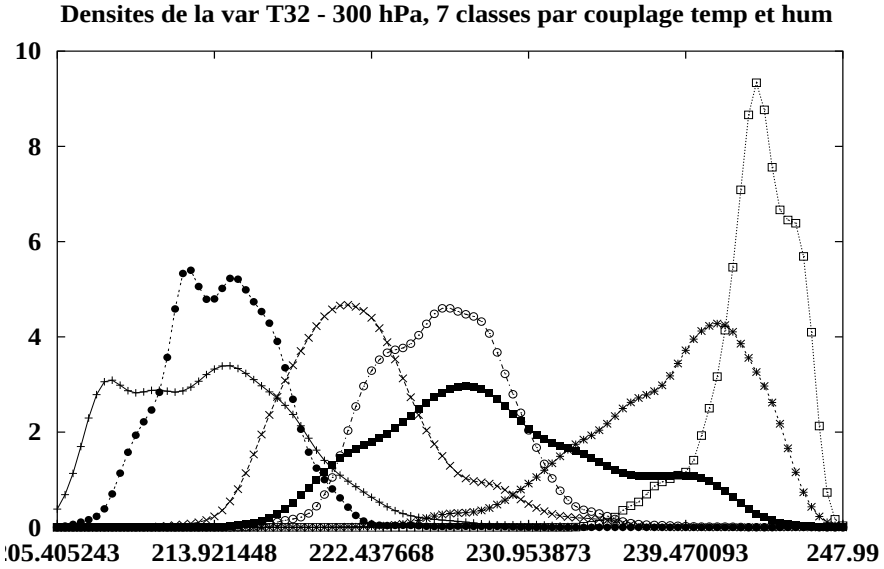


FIG. 6.15: *Densités de la température (K) par classe à 300 hPa*

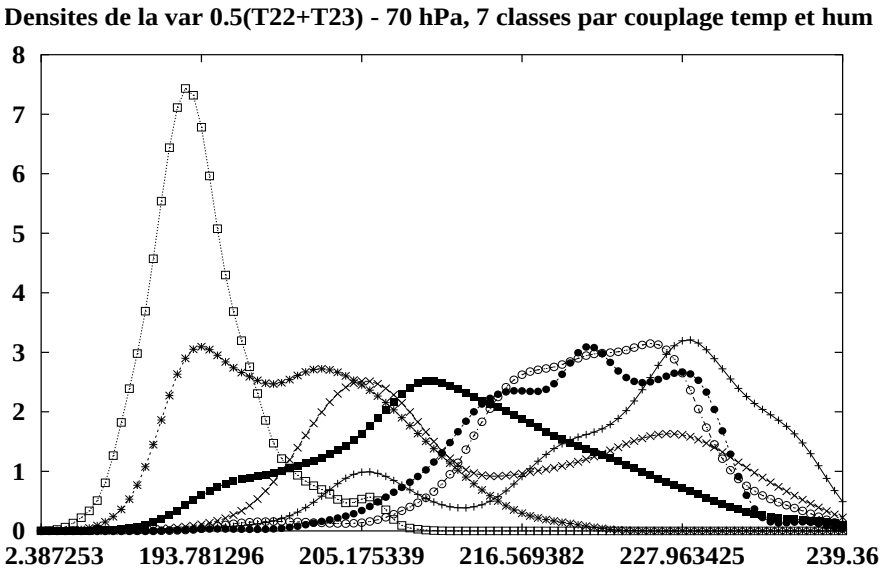


FIG. 6.16: *Densités de la température (K) par classe à 70 hPa*

distinctes, et décrivent assez bien les caractéristiques d'humidité des classes (classe 4 humide, classes 1 et 7 sèches, etc.).

Ces classes, issues de la méthode par couplage, sont aussi beaucoup plus riches puisqu'elles décrivent des comportements liant la température et la vapeur d'eau, les deux variables essentielles pour la description d'une situation synoptique. La classification ainsi obtenue est en effet en très bon accord avec celle du 15 décembre 1998.

Nous avons vu section 5.6 qu'avec les proportions du mélange, les paramètres de FDD et de copules, nous pouvons définir la classe d'appartenance d'un nouvel individu (inférence). Nous pouvons donc tenter de classifier un ensemble de profils atmosphériques correspondant à un instant différent du 15 décembre 1998 à 0H (GMT).

6.3 Inférence de DMC

Soit un nouvel individu (profil atmosphérique) w, décrit par sa fonction de distribution en température F_w^{temp} et sa fonction de distribution en humidité F_w^{hum}.

Avec les paramètres de fonctions de distribution de distributions (FDD), les paramètres de copules et les proportions de mélange obtenus par la méthode DMC sur la température et l'humidité, nous pouvons inférer les valeurs de distributions jointes de distributions pour w par :

$$H_{temp}(F_w^{temp}(T_1^{temp}), F_w(T_2^{temp})) \text{ (noté } H_{temp}(w))$$

et

$$H_{hum}(F_w^{hum}(T_1^{hum}), F_w(T_2^{hum})) \text{ (noté } H_{hum}(w)),$$

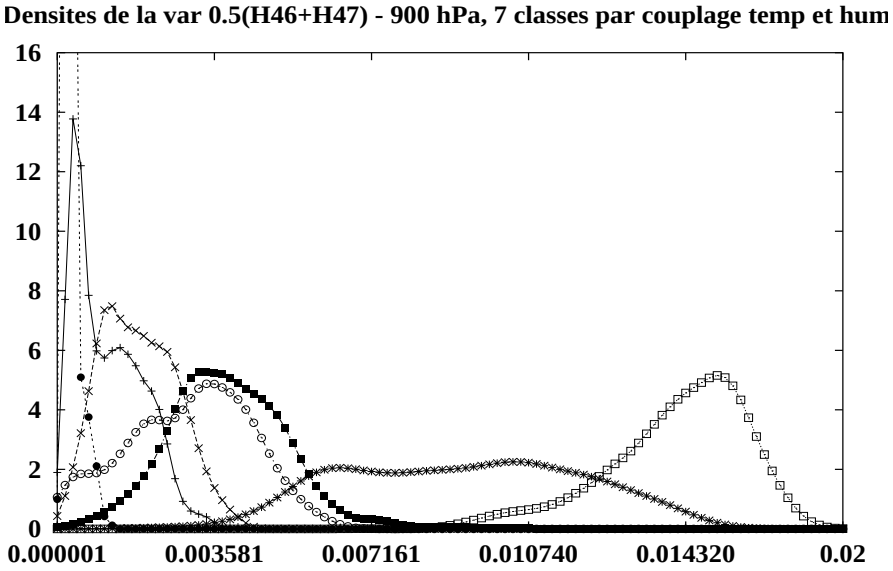


FIG. 6.17: *Densités de l'humidité spécifique (kg/kg) par classe à 900 hPa*

avec

$$\begin{aligned} & H_{temp}(F_w^{temp}(T_1^{temp}), F_w^{temp}(T_2^{temp})) \\ = & \sum_{k=1}^{K_i} p_k^{temp} C_{\beta_k^{i,temp}}(G_{T_1^i}^k(F_w^{temp}(T_1^{temp})), G_{T_2^i}(F_w^{temp}(T_2^{temp})) \end{aligned}$$

et

$$\begin{aligned} & H_{hum}(F_w^{hum}(T_1^{hum}), F_w^{hum}(T_2^{hum})) \\ = & \sum_{k=1}^{K_i} p_k^{hum} C_{\beta_k^{i,hum}}(G_{T_1^i}^k(F_w^{hum}(T_1^{hum})), G_{T_2^i}(F_w^{hum}(T_2^{hum})). \end{aligned}$$

Avec ces valeurs de fonctions de répartition jointes, et en notant h_k la densité associée à la copule de la classe k par la méthode par couplage, nous pouvons obtenir la classe à laquelle appartient l'individu w dans la classification couplée sur les variables température et humidité par : $w \in P_i$ si

$$p_i h_i(H_{temp}(w), H_{hum}(w)) \geq p_m h_m(H_{temp}(w), H_{hum}(w)) \ \forall \ m, \text{ avec } i \leq m \text{ en cas d'égalité.}$$

Cette inférence est faite par exemple pour la même journée à 12H. Les résultats sont Figure 6.18.

Nous constatons tout d'abord une évolution lente et cohérente des classes, comme cela est attendu de l'évolution de l'atmosphère sur une période de 12h. De manière plus précise, on constate, par exemple, une évolution vers l'est des classes sur l'Atlantique nord et l'Europe : classes 5 et 6, par exemple, en parfait accord avec les cartes météo (non présentées) et avec l'évolution de la position géographique des dépressions, notamment. C'est aussi le cas dans l'hémisphère sud.

Cet exemple prouve que la classification initiale est suffisamment significative de la plupart des phénomènes physiques rencontrés dans le temps. Ainsi des données ultérieures possédant ces phénomènes dans leur structure, sont classées correctement.

Nous avons également réalisé une inférence sur l'état thermodynanique de l'atmosphère du 1^{er} février 1999 à 12H, à partir des résultats obtenues sur le 15 décembre 1998 à 0H, donc

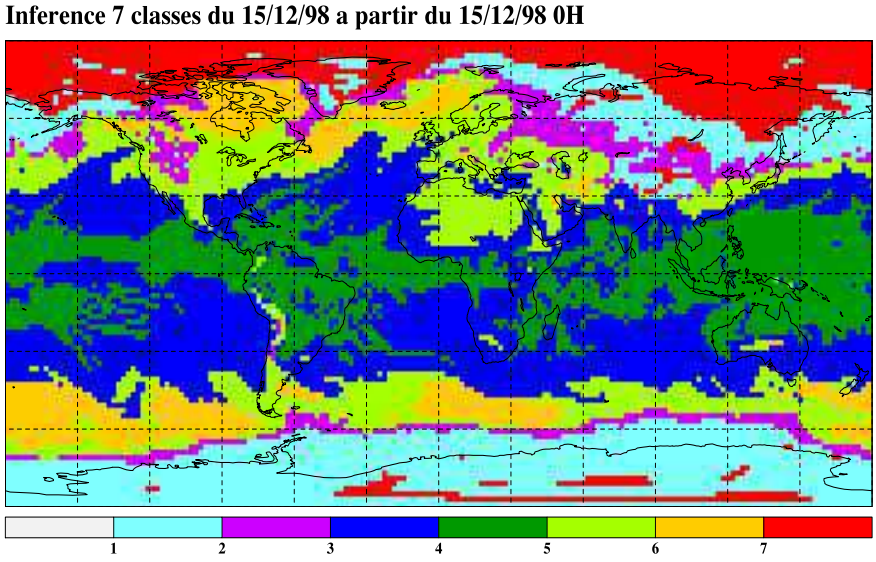


FIG. 6.18: *Inférence à 12H du couplage à 0H en température et humidité de DMC*

à beaucoup plus long terme (15 jours) que la précédente. Les résultats sont présentés sur la carte de la Figure 6.19.

Pour juger de la validité de ce résultat, nous disposons de la carte des températures moyenne entre 500 et 700 hPa (Figure 6.20) et du TCWV (Figure 6.21), pour le 1^{er} février 1999 à 12H.

L'accord avec cette réalisation a posteriori est d'une grande précision. La plupart des structures des champs de température et de vapeur d'eau sont retrouvées correctement et, le plus souvent, avec une grande exactitude. En plus d'une classification précise à un moment donné, la méthode DMC permet donc d'obtenir des informations essentielles pour inférer les classes d'appartenance de nouveaux individus de manière cohérente. Il est ainsi possible de déterminer des paramètres de copules sur un nombre limité d'exemples (i.e. un nombre limité d'états thermodynamiques de l'atmosphère) et d'appliquer l'apprentissage ainsi effectué à l'ensemble des situations atmosphériques possibles.

6.4 Résultats complémentaires par DMC - Exemple de 18 classes

Les résultats obtenus par la méthode de décomposition de mélange de copules pour une partition en 7 classes présentent l'avantage d'améliorer la décomposition classique en 5 classes (tropicale, tempérée hiver et été, polaire hiver et été) tout en préservant la relative simplicité de leur analyse et de leur interprétation. Il est toutefois intéressant de regarder ce que peut apporter une classification en un plus grand nombre de classes et d'analyser leur aptitude à apporter des informations significatives plus précises plutôt que des redondances par scission artificielle des classes d'origine. Nous avons appliqué notre algorithme à une partition en 18

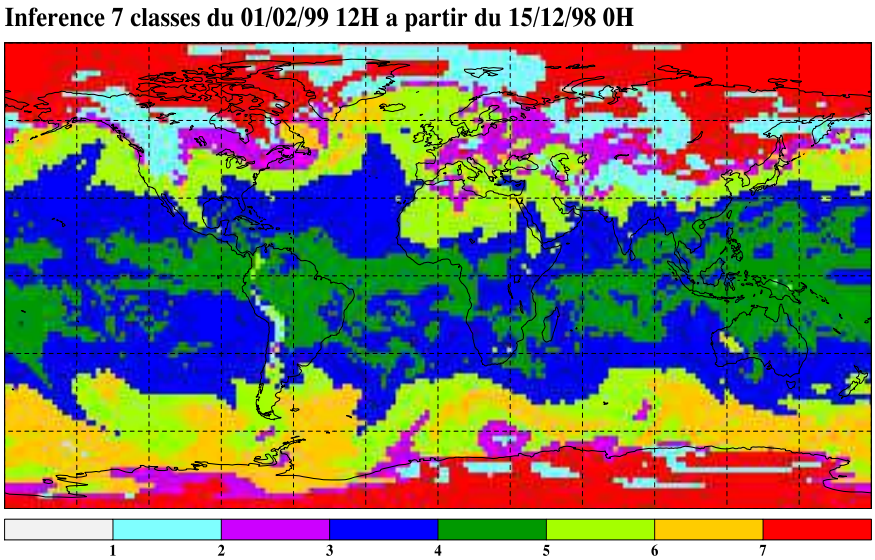


FIG. 6.19: *Inférence du 01/02/1999 à 12H du couplage du 15/12/1998 OH en température et humidité de DMC*

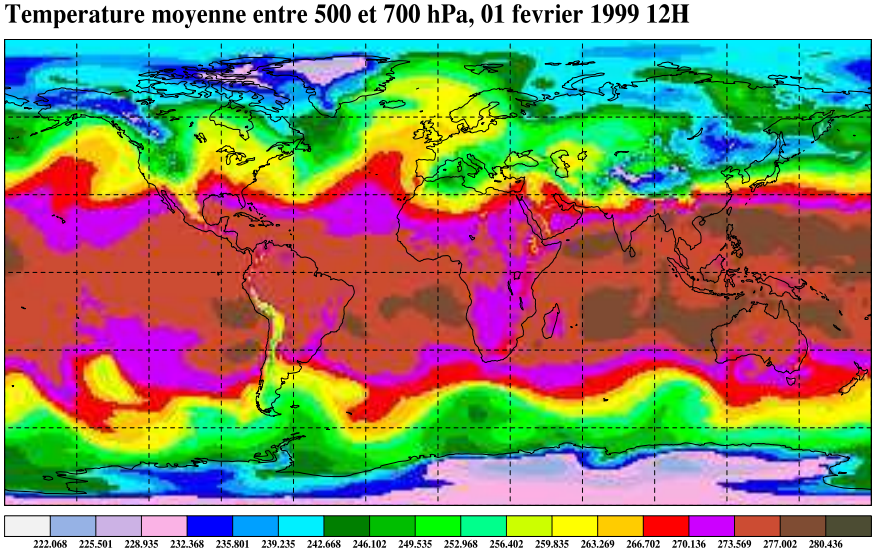


FIG. 6.20: *Température moyenne entre 500 et 700 hPa du 01/02/1999 à 12H.*

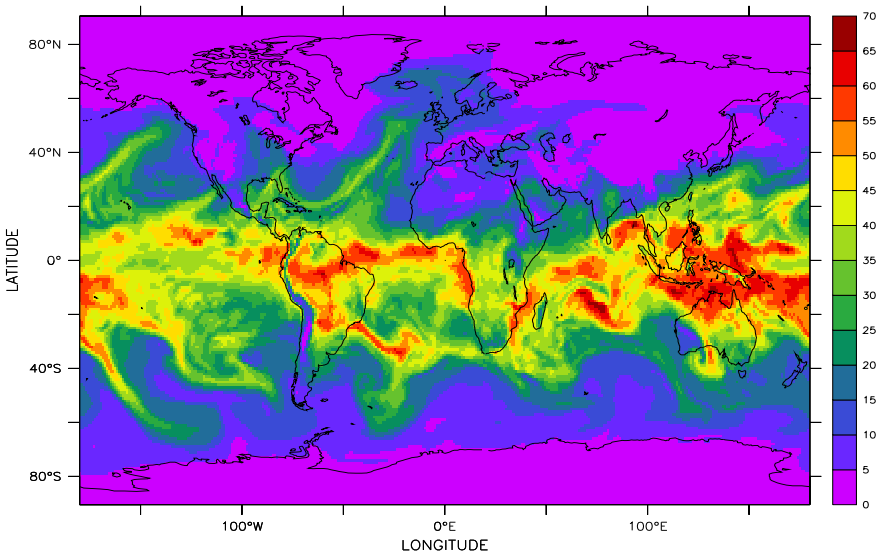


FIG. 6.21: *Total Column Water Vapor du 01/02/1999 à 12H*

classes (Figure 6.22), soit plus de deux fois le nombre de classes précédent.

Un premier jugement peut être établi par l'étude, classe par classe, des densités des variables température et vapeur d'eau à différents niveaux de l'atmosphère. Les constatations sont similaires à celles obtenues sur 7 classes (la légende des 18 classes est donnée Figure 6.23) :

- les densités en température se séparent bien en-dessous de la tropopause (voir Figure 6.24 à 900 hPa, et Figure 6.25 à 300 hPa),
- la distinction des densités en température est plus délicate au-dessus de la tropopause (voir Figure 6.26 à 70 hPa),
- les densités en humidité spécifique sont assez distinctes dans le bas de l'atmosphère (voir Figure 6.27 à 900 hPa) et se regroupent en altitude.

Le tracé des densités pour la comparaison des résultats en 7 classes et en 18 classes est particulièrement utile. En comparant les classifications en 7 classes et en 18 classes, nous pouvons par exemple constater que la classe 3 de la classification en 7 classes (Figure 6.12) semble se scinder en 3 classes dans la classification en 18 classes (Figure 6.22) : les classes 5, 6 et 11. Lorsqu'on étudie les densités de ces 3 classes (par exemple en température à 900 hPa, Figure 6.24), nous voyons qu'elles se séparent bien avec des modes distincts. Nous remarquons de plus que ces trois densités se situent les unes à côté des autres, en occupant parfaitement l'intervalle de valeurs pris par la densité de la classe 3 dans la classification en 7 classes. Ceci est visible en Figure 6.28. Ces tracés semblent indiquer que la densité de la classe 3 (dans la classification en 7 classes) se décompose principalement en un mélange de 3 densités des classes 5, 6 et 11 (de la classification en 18 classes). Cette remarque se vérifie d'ailleurs sur l'ensemble des niveaux pour la température et l'humidité spécifique (les densités aux niveaux

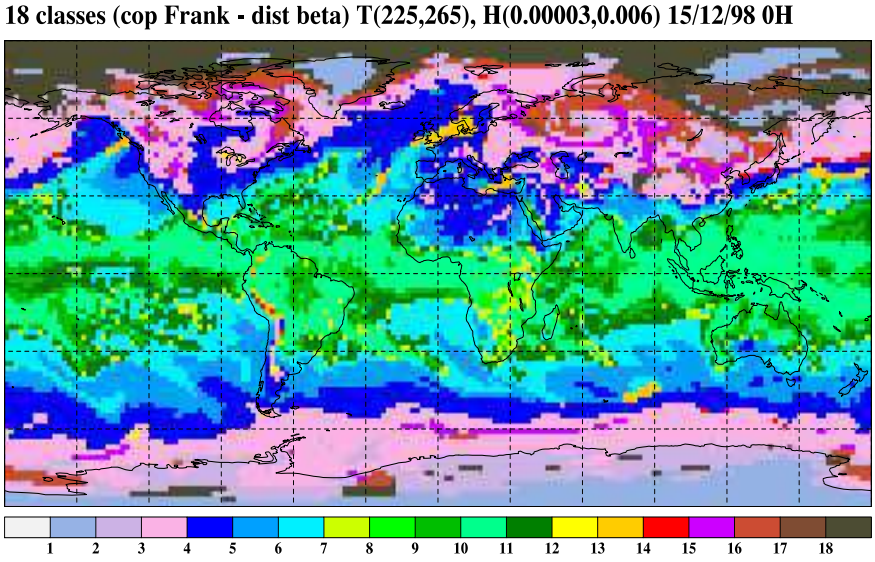


FIG. 6.22: *18 classes en temperature et humidite par DMC*

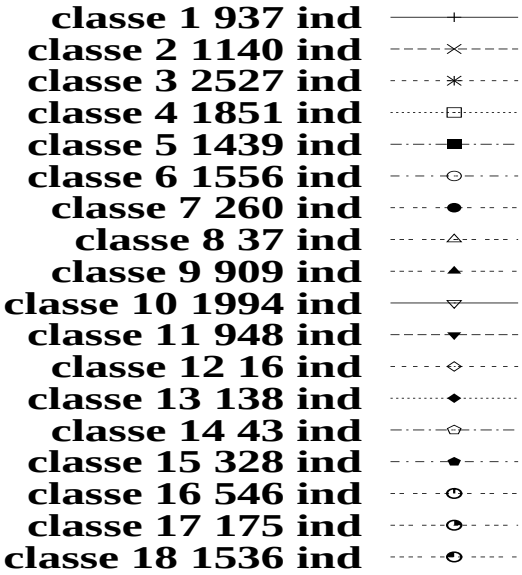


FIG. 6.23: *Association 18 classes - densités*

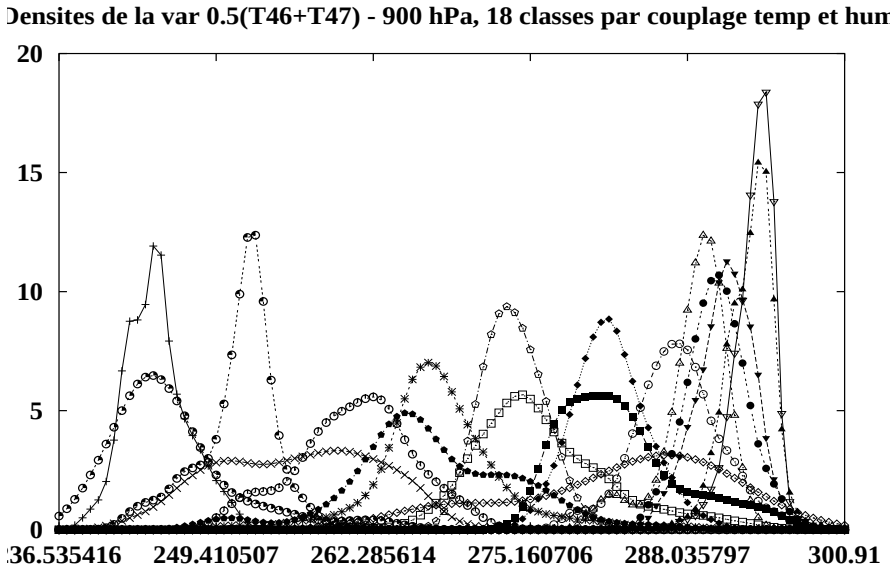


FIG. 6.24: *Densités en température à 900 hPa pour 18 classes par DMC*

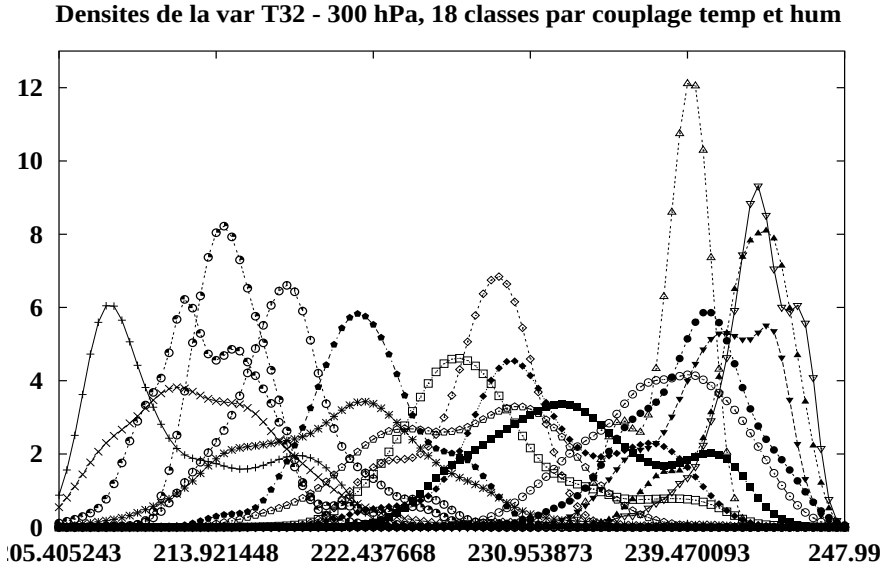


FIG. 6.25: *Densités en température à 300 hPa pour 18 classes par DMC*

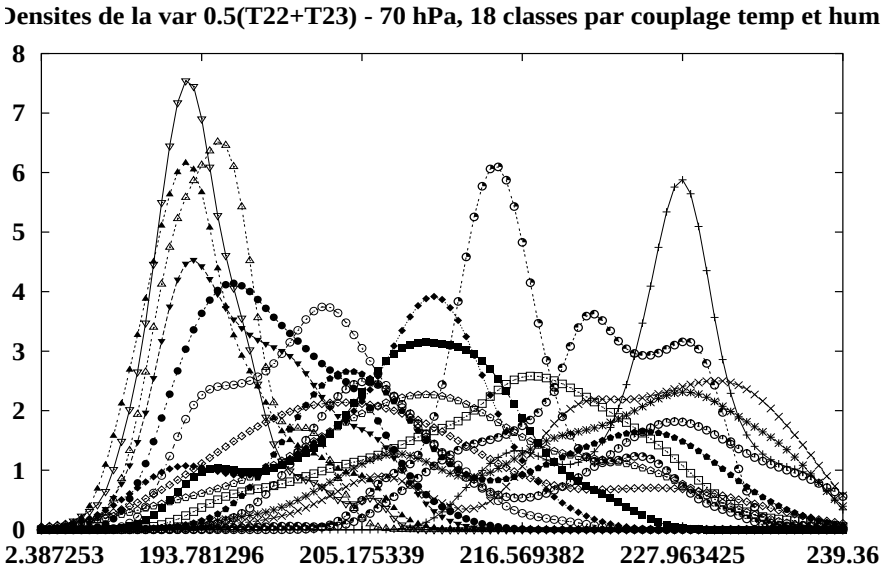


FIG. 6.26: *Densités en température à 70 hPa pour 18 classes par DMC*

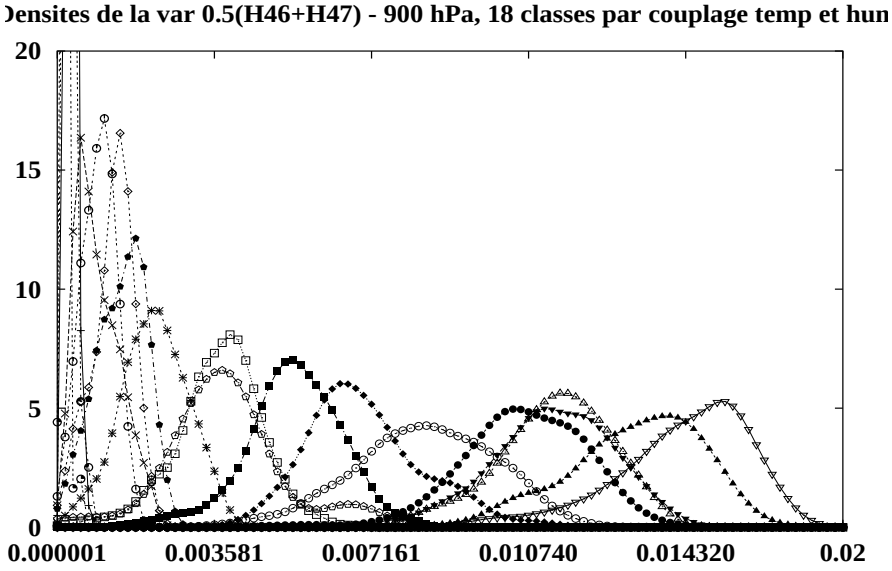


FIG. 6.27: *Densités en humidité à 900 hPa pour 18 classes par DMC*

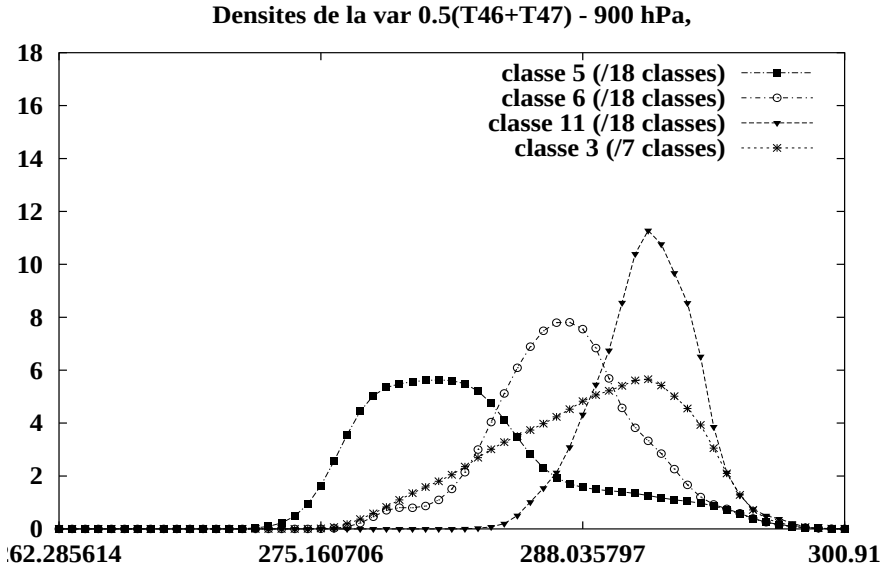


FIG. 6.28: *Comparaison des densités classe 3 (/7 classes) - classes 5, 6 et 11 (/18 classes) en température à 900 hPa*

de pression 900, 700, 500, 300, 100 et 70 hPa sont données en Figure C.6 et Figure C.7 de l'annexe C).

Une étude similaire peut être réalisée avec la classe 7 de la classification en 7 classes. En effet, cette classe semble se scinder en 2 classes : les classes 17 et 18 (de la classification en 18 classes). En comparant les densités, nous voyons que nous pouvons définir un mélange de lois avec la densité de la classe 3 (par rapport à 7 classes) s'écrivant comme la somme pondérée des densités des classes 17 et 18 (par rapport à 18 classes), avec un poids prépondérant à la classe 18.

Par ailleurs, la division de la classe 3 de la classification en 7 classes, en - au moins - 3 différentes classes (5, 6 et 11) sur les 18 nouvelles, est confirmée par la carte du TCWV du 15/12/98 à 0H. Par exemple, les incursions des masses d'air présentes à l'ouest de l'amérique du sud sur la carte de la classification couplée en 18 classes (Figure 6.22) se retrouvent parfaitement sur la carte du TCWV (Figure 6.9). Les formes des classes 5 et 6 dans cette partie du globe pour la classification en 18 classes de la Figure 6.22, correspondent exactement à celle du TCWV de la Figure 6.9, ce qui n'était pas le cas pour la classification en 7 classes, le nombre de classes n'étant pas suffisamment élevé. Le tracé des profils moyens de température et d'humidité (Figures C.9, C.10 et C.11, pages 215, 216 et 217) nous donne une description intéressante des 18 classes. Nous voyons par exemple que les classes 6 à 11 ont des profils de température très similaires de type tropical. La situation géographique de ces classes confortent cette identification. De plus, nous voyons que deux classes ayant des profils proches (à la fois en température et en humidité), se retrouvent voisines dans la classification. Par exemple, les classes 7 et 11 ont des profils très similaires et se retrouvent alors proches géographiquement.

La carte des 18 classes (en Figure 6.22) permet de retrouver tous les éléments essentiels identifiés en 7 classes (Figure 6.12) avec un degré de précision nettement supérieur. Un effet "grain de riz" est cependant visible. Il est dû à la grande variabilité de la variable d'humidité spécifique, ainsi qu'à la résolution spatiale utilisée (un degré de latitude sur deux et un degré de longitude sur deux). Afin de montrer que cette précision accrue n'est pas artificielle, intéressons-nous à la classe 13, ici représentée par un petit nombre d'individus, notamment sur l'Atlantique nord, jusqu'à la Scandinavie, au large de la côte ouest du Canada et sur la Méditérannée. Précisément, dans ces trois cas, la carte météorologique (Figures C.12 et C.13, pages 218 et 219) fait apparaitre un système frontal, que l'on retrouve aussi au sud-est du Kamtchatka, où la classe 13 se manifeste également. Nous avons également constaté que l'analyse des vents horizontaux (composantes u et v) fait apparaitre pour la classe 13, une "aile" de densité correspondant aux plus fortes valeurs de vent, toutes classes confondues (voir Figures 6.29 et 6.30). La classe 13 de la partition en 18 classes, "noyée" dans la classe 5 de la partition

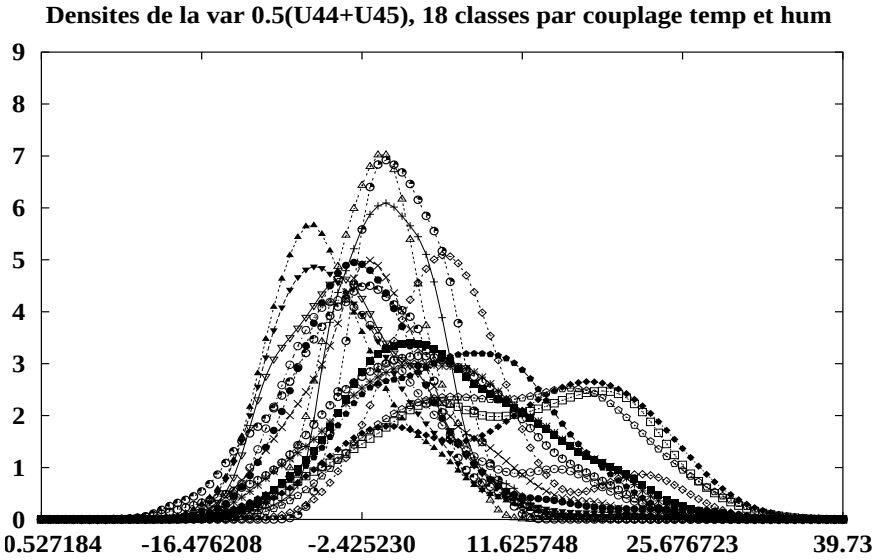


FIG. 6.29: *Densités de la variable vent (composante u) à 800 hPa pour le couplage en 18 classes*

en 7 classes, apporte bien une information précieuse pour la description d'une situation météo.

Afin de démontrer les performances de la méthode par décomposition de mélange de copules et la cohérence de ses résultats, d'autres méthodes plus "classiques" de classification sont maintenant appliquées sur les données climatiques. Nous regardons tout d'abord la classification dite "mixte", puis la méthode de référence pour la décomposition de mélange de lois, c'est-à-dire la méthode EM.

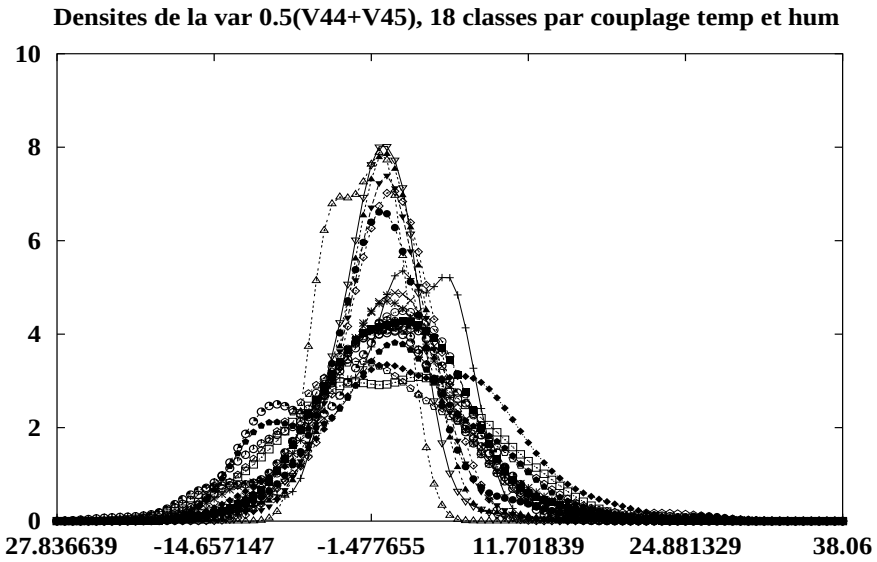


FIG. 6.30: *Densités de la variable vent (composante v) à 800 hPa pour le couplage en 18 classes*

6.5 Classification mixte

La stratégie de classification développée par Mollière (1985, [77]) et décrite par la figure 6.31 est traditionnellement bien adaptée à la classification de très grands ensembles de données en groupes homogènes. C'est une procédure assez complexe et considérée comme efficace.

Tout d'abord, un pré-traitement des données par une Analyse en Composantes Principales permet de réduire le nombre de variables de manière importante. En effet, la plupart des algorithmes de classification travaillent sur des coordonnées factorielles en abandonnant les derniers axes factoriels, généralement porteurs des composantes aléatoires (non systématiques) des données. Nous ne reviendrons pas d'avantage sur la méthode ACP bien connue par l'ensemble des communautés scientifiques.

Après un calcul d'ACP, une première étape d'agrégation autour de centres mobiles (type "Nuées Dynamiques") conduit à la construction rapide d'une partition contenant un grand nombre de petits groupes (une centaine par exemple). Ces groupes sont sensés être des morceaux de classes "réelles" que l'algorithme de partitionnement a éclatées.

Pour obtenir d'emblée une partition préalable de bonne qualité, on y intègre une procédure d'*auto-validation*. Celle-ci consiste à réaliser plusieurs partitions successives (les "partitions de base") puis à croiser les résultats obtenus. On retient comme classes finales les *groupes stables* (appelés "formes fortes") constitués par les groupes d'individus classés ensemble dans les différentes partitions.

Dans une seconde étape, on construit un arbre hiérarchique à partir des centres de ces groupes stables. Cette construction est très rapide car les éléments à agréger sont peu nom-

breux. Cette procédure est réalisée en utilisant le critère de Ward. Ce critère, basé sur la réduction minimale de variance par agrégation, est compatible avec le critère d'inertie utilisé pour la détermination des axes factoriels. L'arbre lui-même est construit avec l'algorithme rapide de recherche en chaîne des voisins réciproques, ou algorithme de Benzécri ([4]).

La troisième étape est la coupure de l'arbre. Le choix du nombre de classes peut-être effectué suivant le test du coût de Cattel. Ce test coupe l'arbre obtenu suivant la plus grande variation de l'indice de hiérarchie. Cette notion d'indice sera explicitée ulterieurement. En réalité, le choix du nombre de classes est un problème dont la solution reste très subjective. Aussi, celui-ci peut-il être déterminé après plusieurs essais, en fonction des différents résultats obtenus.

La dernière étape est la *consolidation* des classes obtenues par une procédure itérative de calcul (de type "Nuées Dynamiques"). Celle-ci conduit ainsi à une partition de qualité optimale pour le critère d'homogénéité des classes.

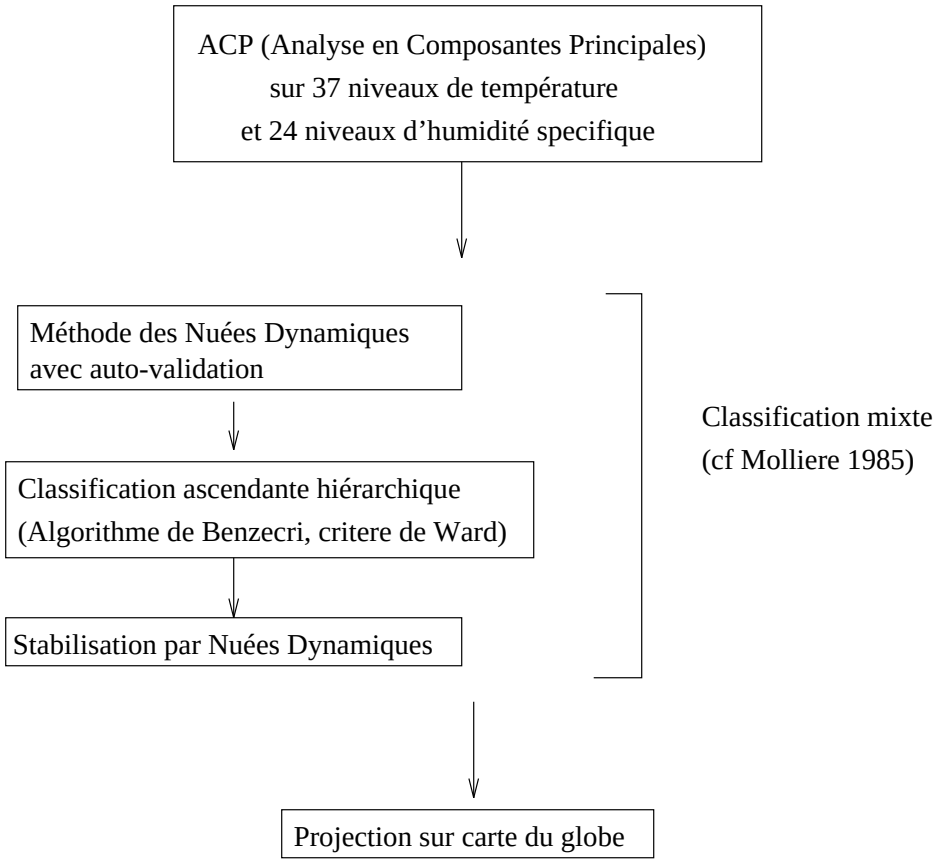


FIG. 6.31: *Stratégie de classification classique*

Nous revenons sur les deux techniques importantes évoquées ci-dessus, les Nuées Dynamiques et la classification ascendante hiérarchique des voisins réciproques avec critère de Ward.

6.5.1 La méthode des "Nuées Dynamiques" (ND)

Cette méthode bien connue (largement développée dans [35] et [31]) nécessite de fixer tout d'abord K, le nombre de clusters souhaité (une centaine dans notre cas). On tire ensuite aléatoirement K "noyaux" (i.e. K points de l'ensemble Ω à partitionner) $g_1^0, ..., g_K^0$. A partir de ces noyaux, on effectue de manière successive et itérative :

- une première étape d'affectation, $F(g_1^i, ..., g_K^i) = (P_1^i, ..., P_K^i)$. On détermine la nouvelle partition à partir des noyaux de la manière suivante

$$P_j^i = \{\text{individus } w \ ; \ d(w, g_j^i) \leq \ d(w, g_m^i) \ \forall \ m = 1, ..., K \text{ avec } j < m \text{ en cas d'égalité}\}$$

où d est une distance entre individus. C'est-à-dire que les classes sont constituées des individus les plus proches des noyaux précédents.

- une seconde étape de "représentation" de la partition, $G(P_1^i, ..., P_K^i) = (g_1^{i+1}, ..., g_K^{i+1})$ où g_j^{i+1} est le centre de gravité de la classe P_j^i et i est le numéro d'itération en cours (i=0,1,...). La fonction G est appelée la "fonction de représentation" et g_j^{i+1} est dit être le représentant" de la classe P_j^i.

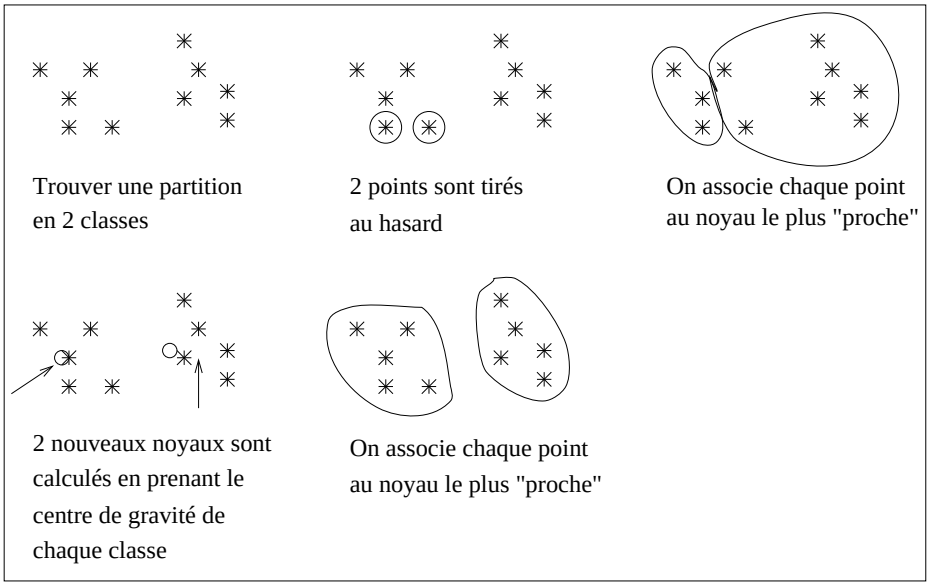


FIG. 6.32: *Exemple de la méthode des "Nuées Dynamiques"*

Le procédé résout en fait un problème d'optimisation : trouver le meilleur couple (P,L) $\in P^{(K)} \times L^{(K)}$ minimisant le critère d'adéquation W entre la partition $P = (P_1, ..., P_K)$ et sa représentation $L = (g_1, ..., g_K)$. L'ensemble $P^{(K)}$ est l'ensemble des partitions possibles en K classes de Ω et $L^{(K)}$ est l'ensemble des représentants possibles des partitions en K classes de Ω.) Le critère W s'écrit :

$$W(P, L) = \sum_{l=1}^{K} D(P_l, g_l) = \sum_{l=1}^{K} \sum_{x_i \in P_l} d_M^2(x_i, g_l)$$

où M est la métrique de la distance d. D est une "mesure d'adéquation" de la classe P_l à son représentant g_l (une petite valeur de D exprime une bonne adéquation entre P_l et g_l).

Les deux étapes sont réalisées itérativement jusqu'à ce que le critère W converge. En effet, en utilisant les fonctions d'affectation et de représentation nous définissons les suites $v_n = (P^n, L^n)$ et $u_n = W(v_n)$ qui convergent ([35]).

Remarque : La partition et la représentation obtenues ne sont que des solutions localement optimales du problème d'optimisation. Elles dépendent de la partition initiale choisie comme point de départ de l'algorithme, c'est pourquoi dans la stratégie employée par Mollière (1985), une recherche des formes fortes est introduite.

La version des Nuées Dynamiques présentée ici n'est qu'un cas particulier d'un algorithme beaucoup plus général. Les ND autorisent plusieurs sortes de noyaux: un point (e.g. centre de gravité, parangon, etc.), un ensemble de points (e.g. les plus proches du centre de gravité,etc.), une droite ou courbe de régression, des axes factoriels, etc.

6.5.2 Algorithme des voisins réciproques

Cette méthode permet d'obtenir une hiérarchie de manière ascendante. Nous ne reviendrons pas sur les classifications descendantes hiérarchiques (beaucoup moins utilisées que les ascendantes) mais rappelons ce qu'est une hiérarchie. Une hiérarchie est un *ensemble de sous-ensembles* (appelés "paliers") de l'ensemble Ω des individus à classifier. Cet ensemble de sous-ensembles contient Ω et tous les singletons (ensembles réduits aux individus seuls) et vérifie la propriété suivante : l'intersection de deux paliers est soit vide, soit identique à l'une d'entre elles. Par conséquent, dans une hiérarchie, chaque palier (non réduit à un seul individu) est la réunion d'autres paliers. Nous utilisons ici la notion de hiérarchie binaire : chaque palier est la réunion de deux paliers.

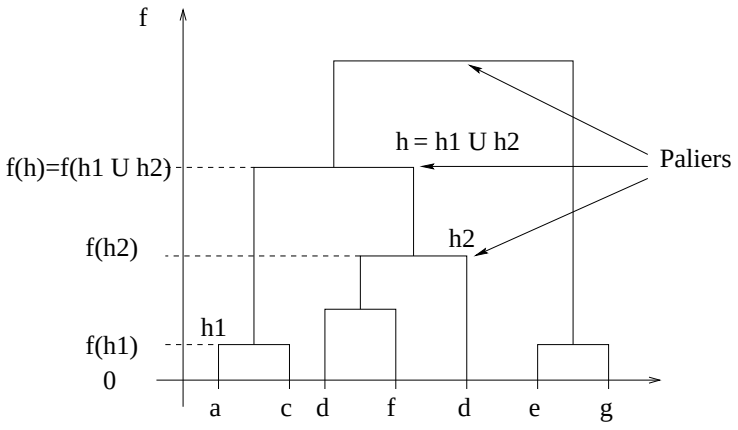


FIG. 6.33: *Hiérarchie indicée*

Soit H la hiérarchie que l'on souhaite construire à partir de l'ensemble Ω des individus. La construction de H nécessite une "mesure de ressemblance" entre groupes. Cette mesure est appelée "indice d'agrégation". C'est une application symétrique δ de $P(\Omega) \times \Omega$ dans $\mathbb{R}^+$. De

nombreux indices d'agrégation existent. Par exemple, l'indice d'agrégation du lien maximum $\delta_1(h_1, h_2) = Max_{w_i \in h_1 et w_j \in h_2} d(w_i, w_j)$ (avec h_1 et h_2 deux paliers et d une distance entre individus) ou celui du lien minimum $\delta_1(h_1, h_2) = Min_{w_i \in h_1 et w_j \in h_2} d(w_i, w_j)$ (avec h_1 et h_2 deux paliers et d une distance entre individus), etc.
Pour réaliser notre hiérarchie, nous utilisons l'indice d'augmentation d'inertie (également nommé critère de Ward) $\delta(h_1, h_2)$ ou h_1 et h_2 sont deux paliers de la hiérarchie :

$$\delta(h_1, h_2) = \frac{p(h_1)p(h_2)}{p(h_1) + p(h_2)} d^2(g(h_1), g(h_2))$$

avec $p(h_i) = \frac{card(h_i)}{card(\Omega)}$ (le poids de h_i) et g(h_i) le centre de gravité du palier h_i.

Tous ces indices d'agrégation δ permettent de définir des "indices de hiérarchie" f, application de H dans $\mathbb{R}^+$, de la manière suivante:
Soit δ un indice d'agrégation et h un palier de la hiérarchie H tel que $h = h_1 U h_2$ (deux paliers de H). Alors $f(h) = f(h_1 U h_2) = \delta(h_1, h_2)$ est un indice de hiérarchie et le couple (H,f) est appelé hiérarchie indicée.

Ayant choisi un indice d'agrégation on peut imaginer de nombreux algorithmes de construction de hiérarchie sur Ω. L'algorithme général de la classification ascendante hiérarchique (CAH) consiste à construire à l'aide de l'indice d'agrégation choisi une suite de partitions de moins en moins fines dont les classes forment la hiérarchie :

- Etape 1. Initialisation de la partition P^0 où chacune des classes ne contient qu'un seul élément (Dans notre cas, les classes initiales sont celles déterminées lors du processus d'auto-validation.),
- Etape 2. Construction d'une nouvelle partition en réunissant les deux classes de la précédente partition qui minimisent l'indice d'agrégation,
- Etape 3. Recommencer le procédé à l'étape 2 jusqu'à n'obtenir qu'une seule classe Ω.

La méthode des voisins réciproques permet d'accélérer considérablement l'exécution de l'algorithme de CAH. Cet algorithme revient à agréger tous les voisins réciproques à chaque étape de l'algorithme classique au lieu des deux éléments qui minimisent δ. Deux paliers sont voisins réciproques s'ils sont mutuellement plus proches voisins au sens de l'indice δ.

Remarque : Dans le cas de l'indice d'augmentation d'inertie (critère de Ward), la méthode des voisins réciproques est d'autant plus rapide qu'elle permet de ne stocker que la matrice des données (De Rham, 1980, [87]).

6.5.3 Résultats

La classification mixte est lancée sur toutes les données (1 degré par 1 degré) du 15 décembre 1998 à 0H. Les individus sont les profils atmosphériques décrits par 62 valeurs numériques: les 37 premières coordonnées sigma de la variable de température (jusqu'à P=10

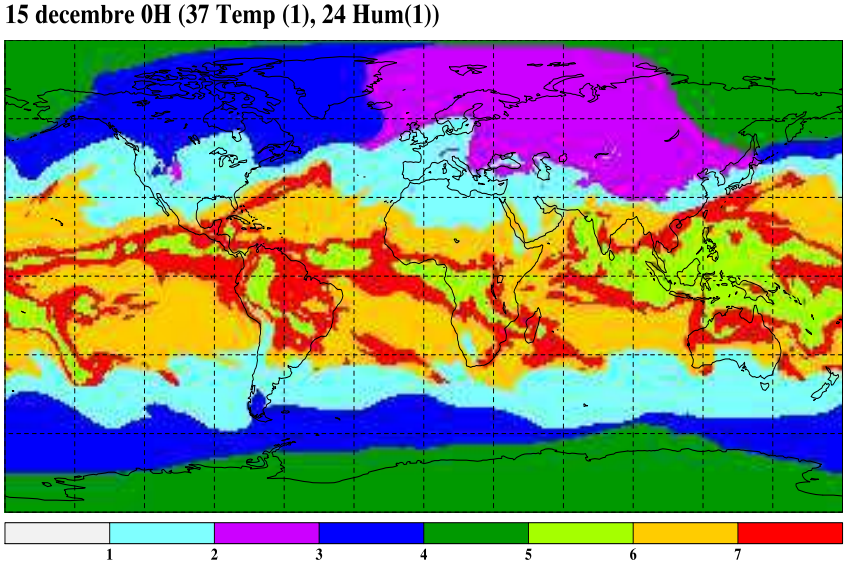


FIG. 6.34: *Classification en 7 classes sur 37 températures et 24 humidités pour le 15 décembre 1998 à 0H*

hPa si Psol=1013 hPa) et les 24 premières coordonnées sigma de la variable d'humidité spécifique (jusqu'à P=155 hPa si Psol=1013 hPa). Nous ne prenons que ces variables car les données de base sont des prévisions, et l'exactitude des prévisions, passée une certaine altitude, est contestable. Les valeurs sont centrées et normées par variable. Les données sont pondérées pour donner un poids égal à 1 à l'ensemble des données température et un poids de 1 à l'ensemble des données humidité. L'algorithme tient donc autant compte de la température que de l'humidité. La partition en 7 classes obtenue se trouve Figure 6.34.

Cette classification apparait immédiatement beaucoup plus "rustique" que celle de la Figure 6.12 : les classes de type tropical semblent être retrouvées (classes 5, 6 et 7). Cependant, nous pouvons voir sur cette classification un bras d'air chaud et humide partant du sud de la Floride vers le nord-est. Un coup d'oeil à la carte du TCWV du 15/12/98 à 0H (Figure 6.9) nous permet de voir que ce bras existe, mais qu'il en existe également un second paralelle plus à l'est qui est totalement absent. Ces deux incursions sont, par contre, bien présentes dans la classification couplée en 7 classes par la méthode par copules (Figure 6.12).

De plus, les classes autres que tropicales sont assez grossières et proches d'un comportement zonal. Par exemple, la différence entre l'été de l'hémisphère sud et l'hiver de l'hémisphère nord est nettement moins marquée. Aucun détail précis ne semble présent au-dessus de 30 ° nord et au-dessous de 30 ° sud. Nous notons également, en traçant les densités de probabilité des variables par classe, que le pouvoir discriminant de ces classes est inférieur à celui des classes de la partition par couplage de copules. Les densités de probabilité de la température et de l'humidité spécifique sont présentées pour chaque classe à différents niveaux en Figures C.14 et C.15 de l'Annexe C avec la légende en Figure C.16.

Une conclusion claire est que la partition obtenue par classification mixte sur les données numériques est inférieure à celle obtenue par DMC, au sens de la cohérence physique des

classes.

Cependant, les résultats de cette classifications peuvent nous donner des informations sur les variables les plus déterminantes (les plus "explicatives") de la partition. Pour cela, nous tentons de déterminer un nombre réduit de variables permettant de décrire les classes de la partition. Ceci est effectué par une méthode de "discrimination".

6.5.4 Variables discriminantes

Nous cherchons donc à diminuer le nombre de variables. En effet avec 61 variables (37 températures, 24 humidités) nous obtenons 99% d'inertie avec seulement 15 axes factoriels lors de l'ACP. Nous cherchons quelles sont les variables les plus discriminantes pour la partition finale, autrement dit les variables qui expliquent au mieux la partition. Une des méthodes de discrimination les plus efficaces est celle de la "segmentation par arbre binaire" (ou "arbres de décision") de type CART ([9]). Un arbre binaire est formé par des divisions successives en 2 sous-classes N_1 et N_2 (appelées "noeuds") d'un ensemble de départ (généralement l'union des classes de la partition à expliquer). Les divisions sont réalisées suivant des coupures (Y, c) où Y est une variable de coupure et c une valeur de coupure: tous les individus ω ayant une valeur inférieure (respectivement supérieure) à c en Y sont dans le noeud-fils gauche N_1 (respectivement droit N_2). Nous donnons en Figure 6.35 un exemple d'arbre binaire.

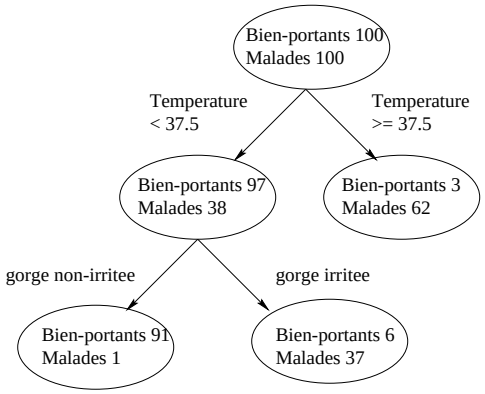


FIG. 6.35: *Exemple d'arbre binaire de décision*

La partition $P = (G_1, ..., G_K)$ étant fixée, le choix de la division d'un noeud est basé sur un critère d'impureté: l'index de Gini. Un noeud est dit pur s'il ne contient des individus que d'une même classe de la partition. L'index i de Gini mesure l'impureté d'un noeud N par rapport à la partition P par :

$$i(N) = \sum_{k \neq j} p_k p_j = 1 - \sum_{k=1}^{K} p_k^2$$

avec $p_j = \frac{card(N \cap G_j)}{card(N)}$. A chaque étape, pour chaque noeud, l'algorithme détermine la coupure (Y, c) qui maximise la variation d'impureté :

$$\Delta i = \pi_1 i(N_1) + \pi_2 i(N_2) \text{ avec } \pi_i = \frac{card(N_i)}{card(N)}.$$

La division du noeud maximisant Δi par sa coupure est effectuée.
Une étude par arbre de décision des variables les plus discriminantes (au regard de la classification obtenue) est réalisée. Quand 2 températures (ou humidités) voisines, reviennent régulièrement dans l'arbre avec la même importance, la moyenne des deux est réalisée. L'algorithme de discrimination a déterminé 11 variables discriminantes:

- 6 de température (ou moyennes de 2 températures):
  T_{50}; $\frac{1}{2}(T_{41}+T_{42})$; $\frac{1}{2}(T_{34}+T_{35})$; $\frac{1}{2}(T_{26}+T_{27})$; $\frac{1}{2}(T_{18}+T_{19})$; $\frac{1}{2}(T_{14}+T_{15})$,
- 5 d'humidité (ou moyennes):
  H_{50}; $\frac{1}{2}(H_{46}+H_{47})$; $\frac{1}{2}(H_{42}+H_{43})$; $\frac{1}{2}(H_{34}+H_{35})$; $\frac{1}{2}(H_{30}+H_{31})$.

De plus, nous souhaitons inclure la variable pression de surface (SP) des profils, afin d'enrichir la classification d'une information susceptible de différencier les situations correspondant à des reliefs élevés. A partir de ces 12 variables (6 températures, 5 humidités et SP), une classification mixte est lancée sur les profils de la journée du 15 décembre 1998 à 0H. Après plusieurs tentatives, des poids ont été données aux variables : chaque variable température et humidité spécifique a un poids de 1 et la variable de pression de surface un poids de 1.5 afin que l'information apportée par SP ressorte. La partition finale se trouve Figure 6.36. La classification est relativement proche de celle en 7 classes sur les 61 variables. La

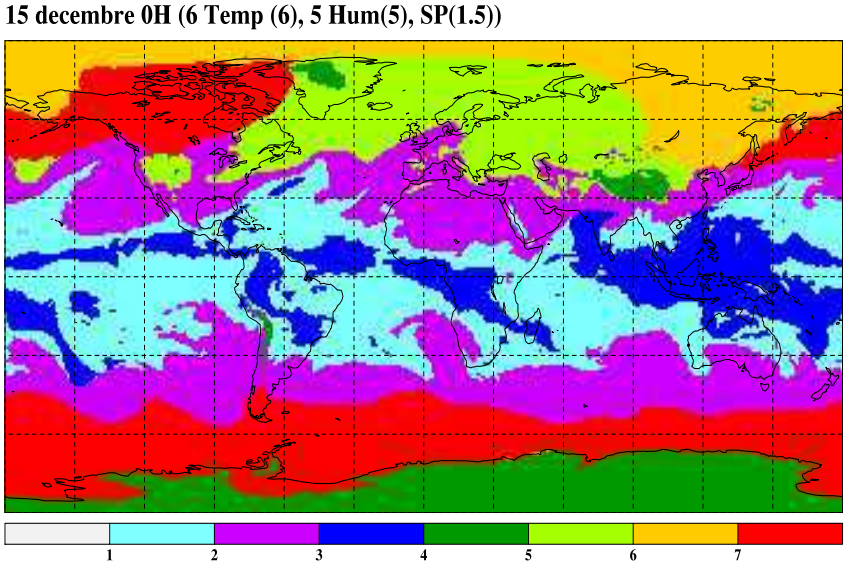


FIG. 6.36: *Classification en 7 classes (6T(6), 5H(5), SP(1.5)*

répartition des couleurs n'est pas identique (ceci est dû au choix aléatoire de la partition initiale). Nous voyons que la classe 2 (classe violette) semble avoir essentiellemnt appris l'information d'humidité moyenne présente dans les données du TCWV (Figure 6.9, couleur bleue). Cependant la classe tropicale que nous retrouvons (classe 3, chaude et humide) est nettement appauvrie par rapport aux précédentes : nous avons beaucoup moins de formes et de détails présents qu'en 7 classes par copules (Figure 6.12) ou dans le TCWV (Figure 6.9).

De plus un comportement zonal est encore fortement marqué. Cependant, la variable SP a fait ressortir les refiefs très élevés tels que l'Himalaya, les Rocheuses, la Cordillière des Andes, etc. Les défauts restent malgré tout ceux de la partition par 61 variables: certaines classes ne sont pas cohérentes (au sens physique) et les incursions ne sont pas franches.

Ces résultats pouvant être dus au fait que les données ne sont pas suffisamment nombreuses pour la méthode, 3 jours sont ajoutés au 15 décembre 1998 à 0H: le 1^{er} février 1999 à 0H, le 15 juin 1999 à 0H, le 1^{er} août à 0H. Ces trois jours permettent en plus de suivre l'évolution des classes en fonction du temps. Par ailleurs, pour enrichir la partition d'informations sur la dynamique de l'atmosphère, d'autres variables ont été introduites,

- la quantité de nuages hauts (HCC) (poids=1),
- nuages moyennement élevés (MCC) (poids=1),
- nuages bas (LCC) (poids=1),
- U1, variable de vent en abscisse moyennée sur les niveaux 30 et 31 (poids=0.5),
- U2, variable de vent en abscisse moyennée sur les niveaux 44 et 45 (poids=0.5),
- V1, variable de vent en ordonnée moyennée sur les niveaux 30 et 31 (poids=0.5),
- V2, variable de vent en ordonnée moyennée sur les niveaux 44 et 45 (poids=0.5),
- la vorticité au niveau 37 (poids=0.5)

Le nombre des variables ayant augmenté, le poids de la pression de surface est transformé et après essais, la pondération choisie est de 3. La classification sur les 20 variables nous donne la partition de la Figure 6.37 Ces résultats, bien que moins cohérents que la classification par DMC, sont intéressants car ils permettent de suivre l'évolution des classes dans le temps et de visualiser l'apport des nouvelles variables. Cependant, pour le 15/12/98, la classe chaude et humide 7 reste moins bien représentée et les incursions sont seulement ébauchées. L'aspect général reste nettement plus zonal qu'avec la méthode DMC. La comparaison du 01/02/99 avec la carte de la température moyenne entre 500 et 700 hPa du 01/02/99 (Figure 6.20) et avec le TCWV du 01/02/99 (Figure 6.21) nous conduit à des remarques similaires : la classe tropicale n'est pas trop mal représentée, mais nous distinguons cependant assez peu de details sur les grandes incursions d'air chaud et humide dans des masses d'air plus fraiches et sèches (telles qu'à l'est et à l'ouest de l'Amérique de nord). Encore une fois, le comportement des masses d'air définies par cette classification semble zonal. L'aspect général des 4 jours permet de supposer que ces conclusions sont valables sur les 2 autres journées (15/06/99 et 01/08/99).

6.5.5 Classification mixte sur valeurs de fonctions de répartition

Une combinaison intéressante entre les données probabilistes et les méthodes de classification classiques est de regarder quelle classification est obtenue lorsqu'on applique la méthode de classification mixte de Mollière aux données fonctions de répartition. Pour cela nous mettons en entrée de cette méthode, les valeurs des fonctions de distribution de température et d'humidité du 15 décembre 1998 à 0H, des 16200 profils atmosphériques décrits sections 6.1.2

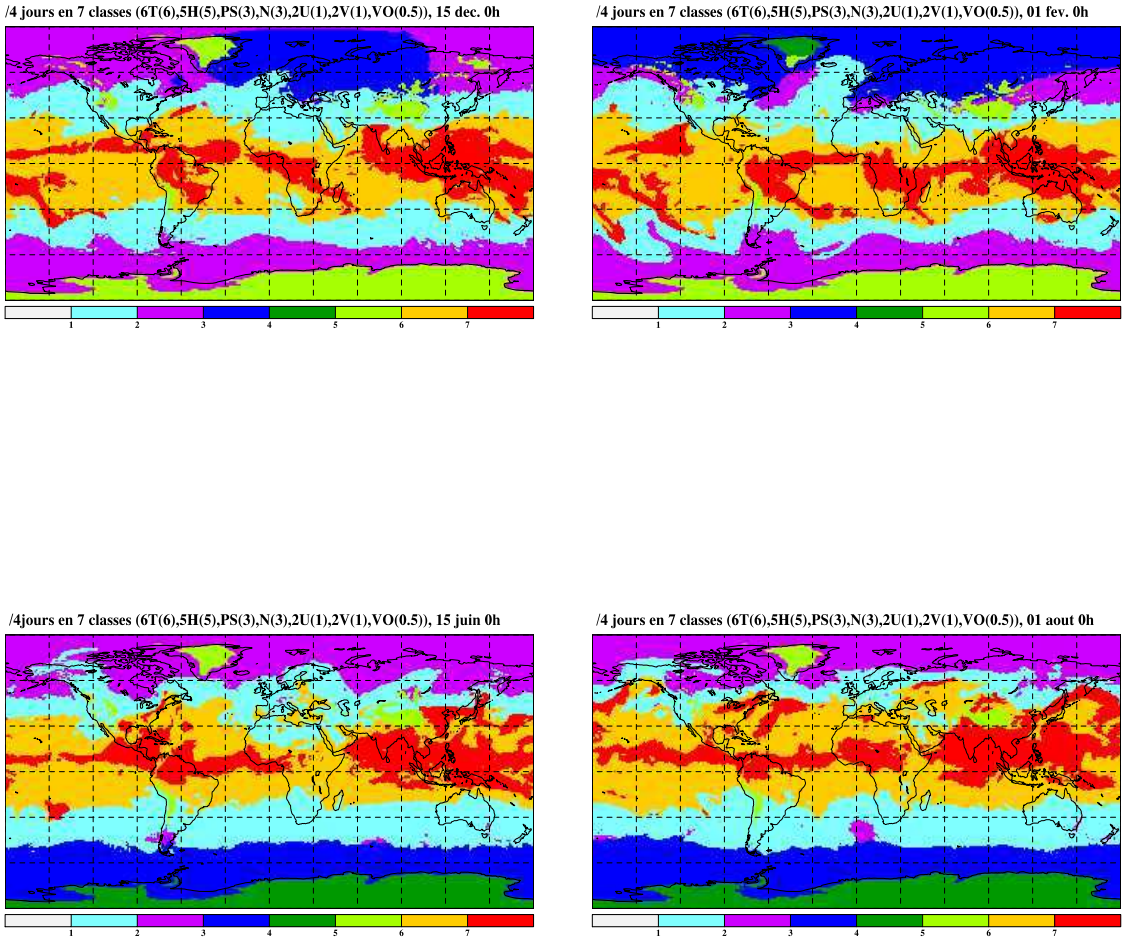


FIG. 6.37: *Classification en 7 classes (6T(6), 5H(5), SP(3), HCC(1), MCC(1), LCC(1), U1(0.5), U2(0.5), V1(0.5), V2(0.5), VO(0.5)*

et 6.2. Pour être cohérent avec l'application réalisée par la méthode de décomposition de mélange de copules (voir section 6.2), nous fixons les mêmes valeurs T_1 et T_2 en température et les mêmes valeurs H_1 et H_2 en humidité : $T_1 = 225\ K$, $T_2 = 265\ K$ et $H_1 = 0.00003\ kg/kg$, $H_2 = 0.006\ kg/kg$. Le tableau de données peut donc se résumer en 16200 lignes (correspondant aux profils atmosphériques) et 4 colonnes (correspondant aux deux valeurs de fonction de répartition de température et aux deux valeurs de fonction de répartition d'humidité). Les résultats de la classification sont projetés sur une carte en Figure 6.38. Cette partition

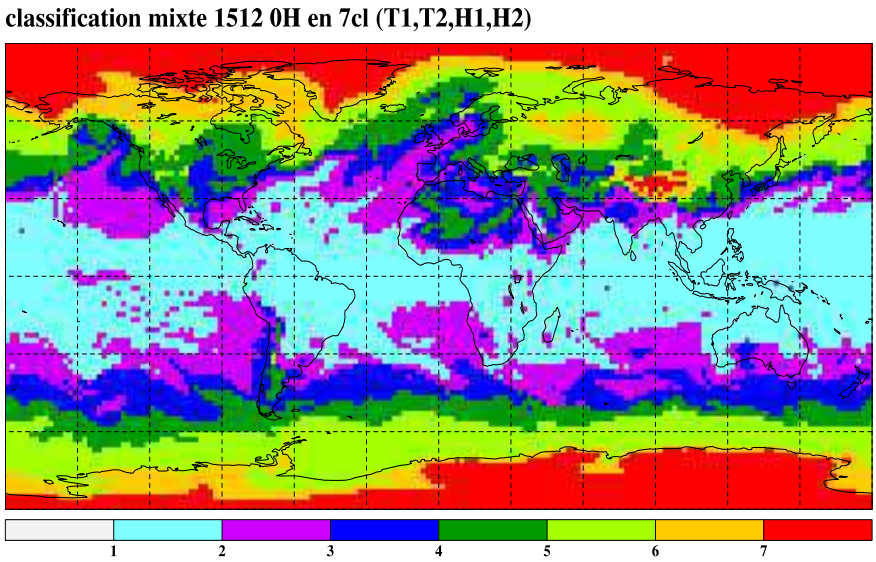


FIG. 6.38: *Classification mixte en 7 classes sur valeurs de distributions de température et d'humidité*

présente évidemment de nombreux points insatisfaisants. Par exemple, nous ne retrouvons pas pas de différence entre l'été au pôle sud et l'hiver au pôle nord. Par ailleurs, la classe tropicale (classe 1 dans la Figure 6.38, chaude et humide) n'est pas clairement définie, peu de détails sont visibles.

Malgré tout, cette figure est intéressante car elle prouve l'apport des données probabilistes. En effet, ce résultat, bien que présentant des défauts, reste cohérent. Par exemple, l'aspect est beaucoup moins zonal aux latitudes moyennes que pour les données numériques brutes (Figure 6.34) et les incursions (en température et humidité) sont dessinées avec précision (ce qui n'est pas toujours le cas avec ces mêmes données brutes). Mise à part la classe tropicale qui est moins bien définie, la classification mixte sur les valeurs de fonctions de répartition est donc de meilleure qualité que celle sur les données classiques.

6.6 Classification par EM

Nous appliquons également l'algorithme EM aux données climatiques afin de pouvoir comparer les résultats de cette méthode à ceux de la méthode DMC. Cette application,

comme la précédente, est faite de deux manières différentes:

- la première sur les données numériques fournies sous la forme de profils atmosphériques par l'ECMWF, tels que dans la méthode mixte de la section 6.5.3. L'algorithme EM prend en entrée les variables fixées en section 6.5.4.
- la seconde sur les valeurs de probabilités en température et d'humidité des profils atmosphériques. Comme dans la section 6.5.5, l'algorithme prend donc en entrée les valeurs des fonctions de répartition de température et d'humidité du 15 décembre 1998 à 0H, des 16200 profils atmosphériques décrits sections 6.1.2 et 6.2.

L'algorithme utilisé est celui implémenté par D. Peel et G.J. McLachlan dans le logiciel EMMIX. Cette algorithme travaille en estimant les paramètres de lois Normales sans restrictions sur les matrices de covariances.

6.6.1 Classification sur les données numériques brutes

Nous lançons tout d'abord EM sur le jeu de données des 16200 profils atmosphériques décrits en section 6.2 avec les 5 valeurs d'humidité spécifique fixées en section 6.5.4 et les 5 premières valeurs de température fixées en section 6.2. Nous ne prenons que 5 valeurs de température pour ne pas donner un poids plus grand à la température qu'à l'humidité. Le résultat de la classification est en Figure 6.39.
La constatation est que cette partition est de qualité inférieure à celle obtenue par couplage

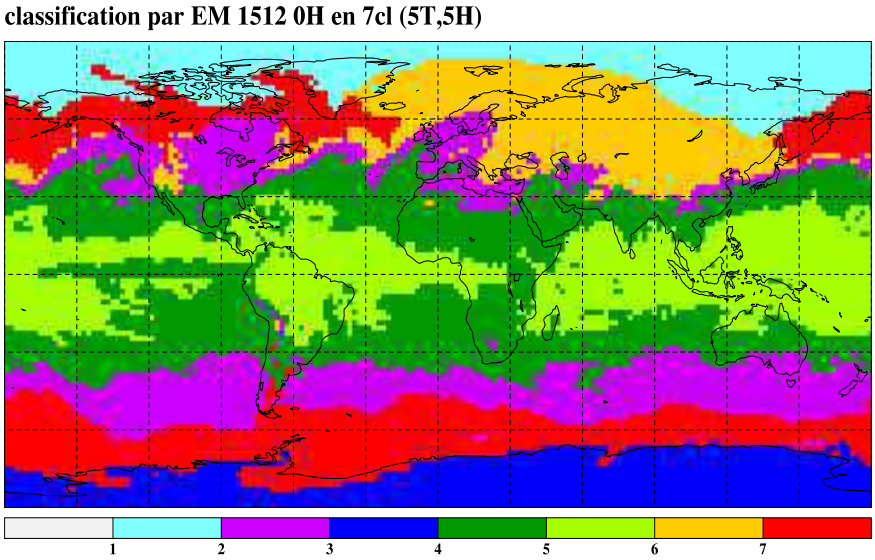


FIG. 6.39: *Classification par EM en 7 classes sur 5 températures et 5 humidités*

avec DMC. En effet, les incursions d'air sont très mal définies et ont totalement disparues dans l'hémisphère sud. De plus, la classe tropicale est grossière et le comportement zonal des classifications mixtes sur les données numériques brutes est également présent (essentiellemnt dans l'hémisphère sud).

Certains points sont tout de même cohérents: la différence entre l'été et l'hiver des deux hémisphères est présente; les incursions d'air de l'hémisphère nord, bordant la frontière entre la zone tropicale et la zone tempérée, sont réalistes.

Cependant, cette partition semble avoir perdu l'aspect dynamique. Par exemple, la dynamique de la spirale est avalée par une classe énorme (classe 6) qui se retrouve régulièrement sur les classifications de la méthode mixte avec les données numériques brutes.

6.6.2 Classification par EM sur les valeurs de fonctions de répartition

A partir des valeurs de probabilités de température et d'humidité des 16200 profils atmosphériques, l'algorithme EM de EMMIX est lancé pour 7 classes. La loi à estimer est une loi normale à 4 dimensions. Le résultat de la classification induite est projeté Figure 6.40.
Une fois encore, cette application montre l'apport des données probabilistes. La classification

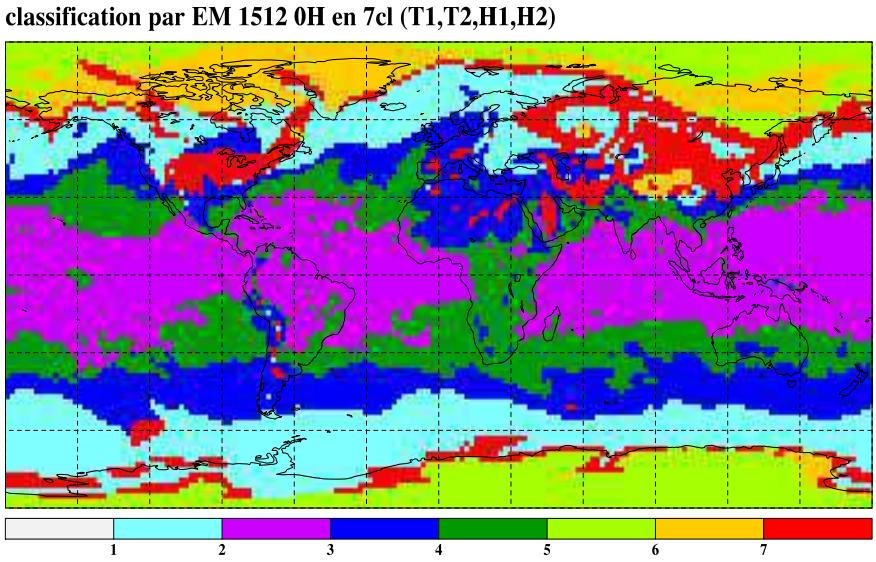


FIG. 6.40: *Classification par EM en 7 classes sur valeurs de fonctions de répartition de température et d'humidité*

obtenue par ces données semble - au premier abord - au moins tout aussi cohérente que la classification réalisée avec les données numériques brutes.

Cependant, en la regardant de plus près, nous voyons que la classe tropicale (classe 2, violette) n'est pas finement dessinée et ne possède pas de détails suffisamment précis, tels que les incursions d'air, etc. Une classe attire notre attention, la classe 7. Cette classe semble englober des profils atmosphériques tout à fait différents, groupant des profils sur des reliefs élevés (Himalaya, Alpes, etc) avec des profils océaniques, polaires ou sur des plaines américaines. Pour tenter de voir ce qu'est cette classe, nous nous intéressons aux densités de probabilité de chaque classe, à différents niveaux et pour chaque variable (la légende est en Figure 6.41). Pour la variable température, nous pouvons regarder par exemple la densité de la classe 7 (point noir) à 700 hPa (Figure 6.42) et à 300 hPa (Figure 6.43).

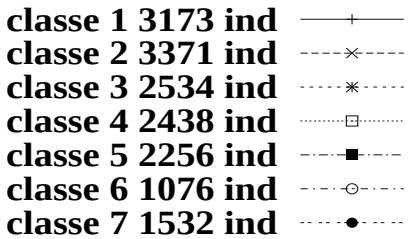


FIG. 6.41: *Association classes - densités EM sur données probabilistes*

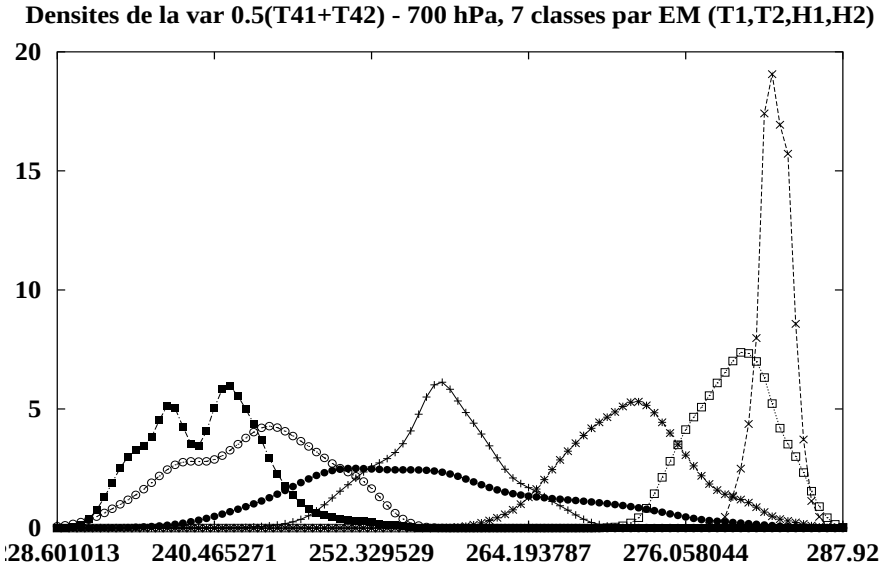


FIG. 6.42: *Densités en température à 700 hPa pour 7 classes par EM sur données probabilistes*

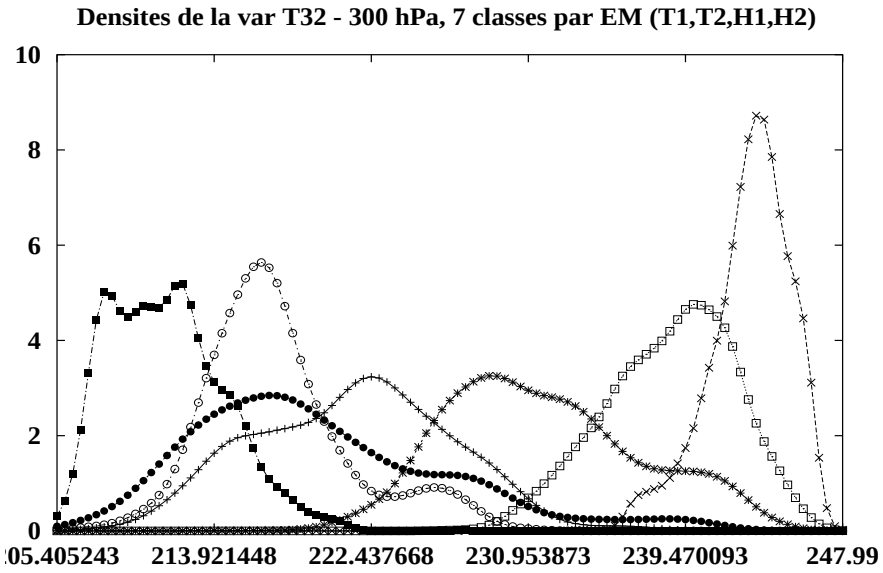


FIG. 6.43: *Densités en température à 300 hPa pour 7 classes par EM sur données probabilistes*

Ces deux tracés nous indiquent que la classe 7 contient des valeurs réparties sur l'ensemble de l'intervalle des valeurs possibles et ceci quel que soit le niveau. De plus les densités de la classe 7 semblent bimodales voir trimodales. Nous pouvons donc supposer qu'il s'agit en réalité de la réunion de 2, voire 3 classes, totalement différentes que la méthode EM n'a pu distinguer, nous donnant ainsi une classe relativement incohérente. Quant à la densité des classes en humidité (voir Figure 6.44), elles prouvent que la classe 7 est sèche. Autrement dit, la méthode EM semble avoir groupé une masse de profils d'humidité spécifique relativement similaire sans tenir compte de la température.

Encore une fois, les résultats de cette méthode sont de qualité nettement inférieure à ceux de la méthode par copules.

6.7 Conclusion

La méthode DMC semble donner de meilleurs résultats de classification que les autres méthodes testées, que celles-ci soient appliquées sur les données numériques brutes ou sur les valeurs de fonctions de répartition. En effet, la précision et la cohérence des résultats obtenus sur 7 classes par DMC, sont largement supérieures à celles des résultats sur 7 classes des autres méthodes. La partition en 18 classes par DMC offre des détails d'une qualité impressionnante avec, par exemple, la classe 13, représentante de systèmee frontaux, qui a réussi à regouper un nombre réduit, mais très cohérent, d'individus (138).

De plus, cette méthode nous permet d'obtenir une relation analytique (caractérisant la dépendance) entre les fonctions de répartition jointes (ici bivariées) et les fonctions de répartition marginales (unidimensionnelles) des variables utilisées. Cette relation est caractérisée par le paramètre de la copule Archimédienne employée. Ce paramètre nous donne également

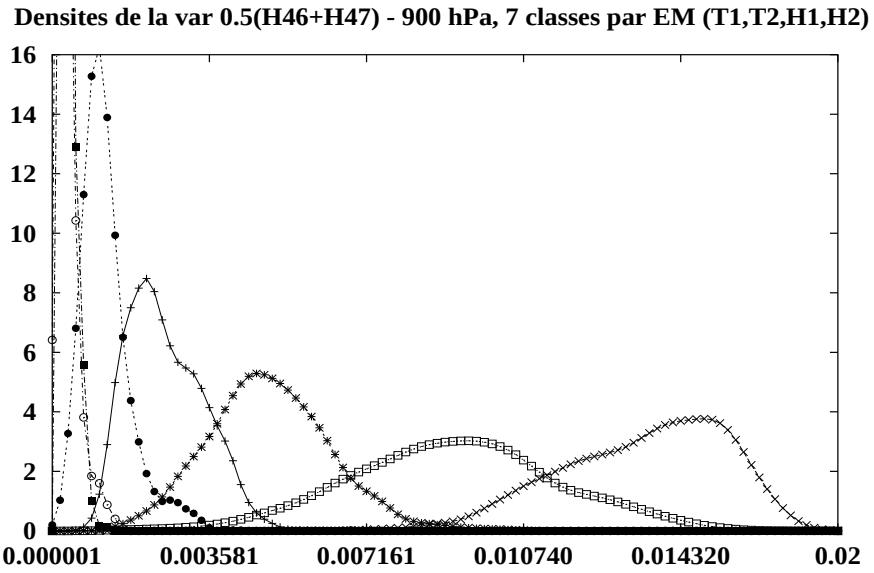


FIG. 6.44: *Densités en humidité à 900 hPa pour 7 classes par EM sur données probabilistes*

des indications sur le comportement des fonctions de répartition à l'intérieur des classes. Par exemple, pour la copule de Frank, si le paramètre est proche de 0, cela signifie que les fonctions de distribution ont tendance à évoluer parallèlement (i.e. sans se couper) entre T_1 et T_2. A l'inverse, si le paramètre tend vers l'infini, elles ont tendance à toutes se croiser entre T_1 et T_2.

Par ailleurs, les coordonnées sigma sont délicates à traiter. En effet, elles ne permettent pas de comparer les valeurs physiques d'un profil avec celles d'un autre (la pression atmosphérique et l'altitude à la surface n'étant pas identiques). Les données probabilistes permettent cette comparaison en regardant la répartition des valeurs physiques sur l'ensemble du profil. Les comparaisons entre les classifications obtenues par les données initiales et celles obtenues par les fonctions de distribution le prouvent. Quelle que soit la méthode utilisée, son application sur les données probabilistes est de qualité supérieure à celle sur les données numériques brutes.

De nombreuses perspectives peuvent alors s'ouvrir pour les applications de la méthode par copules dans un contexte climatique, dès lors que des méthodes multidimensionnelles plus simples de DMC seront développées. En effet, l'introduction de nouvelles variables (nuages, vents...) devrait permet d'accroitre le sens physique que nous pouvons donner aux classes de la partition en ajoutant de la dynamique aux données.

Chapitre 7

Perspectives

"La mesure de notre capacité intellectuelle est la capacité de nous sentir de moins en moins satisfaits de nos réponses à des problèmes sans cesse améliorés."
(C. W. Churchman)

"Quand on cherche les conditions psychologiques des progrès de la science, on arrive bientôt à cette conviction que c'est en termes d'obstacles qu'il faut poser le problème de la connaissance scientifique."
Bachelard, "La Formation de l'esprit scientifique", 1938 [chap. premier, p.13]

Nous proposons différents axes de recherche dans le but d'améliorer la méthode DMC et de l'étendre à de nouvelles applications.

Une piste de recherche est tout d'abord l'étude de l'influence des fonctions de distribution de distributions (FDD). En effet, l'estimation de la loi jointe est différente selon que les FDD sont par exemple modélisées par des lois béta, par des des lois normales tronquées ou par des estimations non-paramétriques.

De même, quelles sont les modifications apportées dans la classification et dans l'estimation de la loi pour différents choix de T_i? Les résultats sont-ils sensibles aux variations des T_i? Ceci nous amène à poser les questions:"Existe-t-il des T_i optimaux?" "Comment les trouver?"

Le nombre des T_i est également un problème complexe. En effet, plus ce nombre est grand, mieux nous pouvons décrire les variables fonctions de répartition. Cependant, un trop grand nombre alourdirait la méthode par une redondance de l'information ou une sur-information préjudiciable à la vitesse de convergence de la méthode.

Par ailleurs, nous avons développé trois approches pour considérer n T_i ($n \geq 3$), par copules multidimensionnelles, par couplage et par arbre binaire. Ces approches sont faciles de mise en oeuvre mais ont l'inconvénient d'avoir des coûts de calculs importants. Aussi le traitement du cas n-dimensionnel ($n \geq 3$) devrait être un axe de recherche fort pour l'amélioration de la méthode DMC, que ce soit d'un point de vue de précision et de compréhension

des résultats, ou d'un point de vue de vitesse de convergence de la méthode.

Dans l'application climatique détaillée chapitre 6, nous avons présenté des résultats obtenus avec la copule de Frank (section 4.6.2, page 82). Quelle(s) conséquence(s) sur les résultats aurait une autre famille de copules? Cette question entraine la suivante: Comment choisir la famille de copules la plus adaptée? De manière plus générale, l'extension des critères de choix du modèle (nombre de composantes, famille de copules) pour la méthode DMC est une piste intéressante. Les critères dans le cas classique sont nombreux. Nous pouvons citer:

- "la correction de biais de la log-vraisemblance" (Kullback-Leibler, 1951, [71]),
- "AIC" (Akaike's Information Criterion, Akaike, 1974 [2]),
- "le critère d'information par bootstrap" (également appelé EIC, Ishiguro, Sakamoto et Kitagawa, 1997, [63]),
- "NEC" (Normalized Entropy Criterion,Celeux et Soromenho, 1996, [16],Biernackie, Celeux et Govaert, 1999, [6]),
- "le critère du taux d'information minimum" (Cutler et Windham, 1993, [25])...

L'extension de ces critères pour DMC constituerait une base solide à une étude plus profonde du comportement de cette méthode. De plus, nous avons vu section 5.5.3 que l'approche par arbre binaire peut permettre de choisir un modèle de copules différent à chaque étape de coupure. Ces critères pourraient donc être appliqués sur chaque noeud de l'arbre.

La notion de densité est également très utilisée en analyse discriminante. La formulation classique de cette dernière est la suivante. Soit un échantillon $X_1, ..., X_n$ venant d'une population A et soit $Y_1, ..., Y_m$ venant d'une population B. Etant donnée une nouvelle observation Z, Z vient-il de A ou de B? Supposons que les observations $X_1, ..., X_n$ de A sont de densité de probabilité f_A et que les observations $Y_1, ..., Y_m$ de B sont de densité f_B. Une approche basée sur le maximum de vraisemblance associe Z à A si

$$f_A(Z) \geq f_B(Z).$$

Nous avons vu dans les sections 5.6 et 6.3 que cette approche s'applique parfaitement dans le cadre de travail de la méthode DMC. Cependant d'autres approches pourraient être étudiées (méthodes Bayésiennes, approche non-paramétrique...) et tous les domaines travaillant avec des méthodes dites de discrimination pourraient donc être concernés par ce type d'analyse sur des données probabilistes. Les applications sont donc riches et nombreuses.

De plus, même si les résultats théoriques de mélange de lois restent fondés, le fait de travailler sur des lois de lois implique des propositions, lemmes et corollaires nouveaux (tels que l'équation (5.18)). L'étude théorique des distributions de distributions associées aux copules doit enrichir la compréhension de la méthode (et de ses variantes), ses perspectives et sa rapidité de convergence.

Enfin, les applications de la méthode font partie des perspectives les plus importantes. Pour l'application climatique, une étape prochaine est d'introduire de nouvelles variables (nuages, vents...) permettant de mieux modéliser la dynamique de l'atmosphère. Le chapitre 6 nous a prouvé que la méthode DMC traite les données probabilistes corrélées avec une grande précision. Ce type de données est courant dans la plupart des domaines scientifiques. Les avantages de la modélisation des dépendances par copules entre FDD et entre les variables utilisées devraient rapidement être mis à profit.

Conclusion

"Le secret d'un bon discours, c'est d'avoir une bonne introduction et une bonne conclusion. Ensuite, il faut s'arranger pour que ces deux parties ne soient pas trop éloignées l'une de l'autre."
(Georges Burns)

"Cette nuit, en regardant le ciel, je suis arrivé à la conclusion qu'il y a beaucoup plus d'étoiles qu'on en a besoin."
(Quino, Mafalda)

La méthode de décomposition de mélange de copules (DMC), également appelée décomposition de mélange de distributions de distributions, que nous avons développée, est une approche originale des mélanges de lois, pour données probabilistes. A partir:

- d'un ensemble $\mathfrak{F}$ de N fonctions de répartition provenant de Ω_F (un espace de fonctions de répartition) et décrivant N individus de Ω,
- du nombre K de composantes,
- de valeurs T_1,...,T_n,
- d'un modèle pour les fonctions de distribution de distributions (paramétrique ou non) en chaque T_i,
- et d'une famille donnée de copules,

les sorties de la méthode sont:

- une classification des données,
- une modélisation de la loi des données associées à chaque classe,
- une modélisation de la loi globale des données comme somme pondérée de lois.

L'originalité de la méthode DMC repose sur deux points:
* elle permet de travailler sur des données probabilistes et de modéliser des lois de probabilité pour ces données,

* elle permet de modéliser les dépendances entre la fonction de distribution jointe et les fonctions de distribution unidimensionnelles (en différents points T d'une variable aléatoire fonction de répartition et pour deux variables de ce type dans le cas du couplage).

Le fait de travailler sur de telles données nous a permis de définir la notion de variables aléatoires de variables aléatoires pour la modélisation de la loi des données. Cette modélisation est faite au travers de la notion de distributions de distributions. Nous pouvons remarquer que le cas des données fonctions de répartition est un sous cas des données fonctionnelles. En effet, la méthode DMC marche également pour des données fonctionnelles (englobant donc les fonctions de répartition). De plus, nous avons vu section 5.10 que DMC généralise la décomposition de mélange de densités classiques en modélisant les données numériques par des distributions de masse unitaire. La méthode s'inscrit donc parfaitement dans le cadre de l'analyse de données symboliques.

Par ailleurs, Nous avons pu généraliser la plupart des méthodes de décomposition de mélange de densités (EM, SEM, etc). Ces méthodes de décomposition de mélange de copules sont en accord avec tous les résultats théoriques obtenus sur les méthodes "classiques" de décomposition de mélange de densités (section 2). En effet, l'idée principale étant de travailler sur des densités associées à des fonctions de distribution jointes de distributions modélisées par des copules, les résultats théoriques (e.g. de convergence) sont vérifiés.

L'idée principale pour travailler sur de telles densités est de choisir des valeurs $T_1, ..., T_n$ sur lesquels elles sont définies. Nous avons vu en section 5.3 que ces valeurs T ne sont pas simples à fixer. Nous avons pu proposer trois méthodes qui restent cependant à développer.

L'apport de la théorie des copules est essentielle. Le théorème de Sklar est au coeur de la méthode en décrivant la place des copules dans les relations de dépendance entre les fonctions de distribution jointes et leurs marginales unidimensionnelles. Cette dépendance est modélisée à l'intérieur de chaque classe P_i, $i = 1, ..., K$ par la formule (5.6), et pour l'ensemble des classes par les relations (5.5) et (5.18).

Les méthodes d'estimation des copules restent cependant coûteuses en calculs et les paramètres des copules Archimédiennes multidimensionnelles sont complexes à comprendre. Notre contribution à l'estimation de ces paramètres permet d'avoir une meilleure vision des dépendances exprimées. En effet, dans la copule C_{β_1,β_2} de (U, V, W), β_1 est le paramètre de la copule C_{β_1} liant U et V et β_2 est le paramètre de la copule C_{β_2} liant $C_{\beta_1}(U, V)$ et W. Cette approche est celle qui tient le plus compte de la formulation (4.22) des copules Archimédiennes n-dimensionnelles.

De plus, dans tous les exemples du chapitre 6, le traitement des données par modélisation probabiliste est meilleur que celui des données brutes. En effet, tous les résultats montrent parfaitement que, quelque soit la méthode (EM, mixte), les classifications obtenues sont plus cohérentes et plus réalistes sur les données probabilistes que sur les données brutes. Les fonctions de répartition semblent tenir d'avantage compte des phénomènes liés à la verticale

des profils étudiés.

Ceci est sans doute dû au fait que les données brutes sont exprimées pour des coordonnées sigma fixes et non des pressions atmosphériques fixes. Autrement dit, les données brutes ne sont pas nécessairement comparables d'un individu à un autre. Ceci constitue d'ailleurs l'un des apports essentiels des données probabilistes et fonctionnelles. Si nous disposons de jeu de données numériques pour des individus dont les descriptions pour les mêmes variables ne sont pas comparables, modéliser les descriptions des individus par des fonctions interpolant les données, permet la comparaison des individus. Par exemple, si nous avons un jeu de données correspondant à des mesures de la taille de différents hommes, et si ces mesures ont été effectuées à des âges différents d'un individu à un autre, elles ne sont pas comparables directement. Les méthodes d'interpolation rendent ces mesures comparables. La connaissance de l'évolution de la taille selon l'âge est une information supplémentaire qui peut être exploitée.

L'autre apport est évidemment que les données fonctionnelles et les données probabilistes (et donc les fonctions de répartition) condensent l'information sur les individus. Ce type de description rend plus facile l'analyse de bases de données gigantesques souvent intraitables directement.

Nous pouvons également remarquer que si les données brutes sont comparables d'un individu à un autre, la méthode DMC s'applique alors à ce jeu de données. Autrement dit, la méthode développée dans cette thèse est généralisable aux données numériques classiques comparables deux à deux (pour une variable, la valeur d'un individu est obtenue dans les mêmes conditions que celle d'un autre individu). Les applications peuvent donc être celles des mélanges de lois classiques, dans des buts de classification ou d'estimation de lois.

Les perspectives et conclusions que nous venons de donner s'inscrivent dans un cadre de recherche pouvant intéresser différentes disciplines désireuses d'utiliser l'analyse et la modélisation cohérente des données probabilistes. Les perspectives sont donc vastes et beaucoup reste encore à faire.

Bibliographie

[1] V. Achard. *Trois problèmes clés de l'analyse 3D de la structure thermodynamique de l'atmosphère par satellite : mesure du contenu en ozone; classification des masses d'air; modélisation hyper rapide du transfert radiatif.* Thèse de doctorat, Université Paris VII, 1991.

[2] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic control*, 19:716–723, 1974.

[3] M. Ali, N. Mikhail, and M.S. Haq. A class of bivariate distribution including the bivariate logistic given margins. *Journal of multivariate analysis*, 8:405–412, 1978.

[4] J.P. Benzécri. *L'analyse des données. Tome 1: la taxinomie. Tome 2:l'analyse des correspondances.* Dunod, Paris, 1976.

[5] J. Besag and P.J. Green. Spatial statistic and bayesian computation. *Journal of the royal statistical society*, B 55:25–37, 1993.

[6] C. Biernacki, G. Celeux, and G. Govaert. An improvement of the nec criterion for assessing the number of clusters in a mixture model. *Pattern recognition letter*, 20:267–272, 1999.

[7] J.C. Biscarat. Almost sure convergence of a class of stochastic algorithms. *Stochastic processes and their applications*, 50:83–99, 1994.

[8] H.H. Bock and E. Diday. *Analysis of symbolic data. Exploratory methos for extracting statistical information from complex data.* Springer-Verlag, Heidelberg, 2000.

[9] L. Breiman, J.H. Friedman, R.A. Ohlsen, and C.J. Stone. *Classification and regression trees.* Wadsworth, 1984.

[10] G. Celeux, D. Chauveau, and J. Diebolt. Stochastic versions of the em algorithm : an experimental study in the mixture case. *J. Statist. Comput. Simul.*, 55:287–314, 1996.

[11] G. Celeux, S. Chrétien, F. Forbes, and A. Mkhadri. A component-wise em algorithm for mixtures. Technical Report 3746, INRIA, 1999.

[12] G. Celeux and J. Diebolt. L'algorithme sem: un algorithme d'apprentissage probabiliste pour la reconnaissance de mélange de densités. *Revue de statistiques appliquées*, 34:35–51, 1986.

[13] G. Celeux and J. Diebolt. A stochastic approximation type em algorithm for the mixture problem. *Stochastics and Stochastics Reports*, 41:119–134, 1992.

[14] G. Celeux and G. Govaert. A classification em algorithm for clustering and two stochastic versions. *Computational statistics and Data analysis*, 14:315–332, 1992.

[15] G. Celeux and G. Govaert. Comparison of the mixture and the classification maximum likelihood in cluster analysis. *Journal of statist. computer*, 47:127–146, 1993.

[16] G. Celeux and G. Soromenho. An entropy criterion for assessing the number of clusters in a mixture model. *Journal of classification*, 13:195–212, 1996.

[17] K.C. Chanda. A note on the consistency and maxima of the roots of the likelihood equations. *Biometrika*, 41:56–61, 1954.

[18] A. Chédin, N. Scott, C. Wahiche, and P. Moulinier. The improved initialization inversion method : a high resolution physical method for temperature retrievals from satellites of tiros-n series. *J. Clim. Appl. Meteor.*, 24:128–143, 1985.

[19] Frédéric Chevallier. *La modélisation du transfert radiatif à des fins climatiques: une nouvelle approche fondée sur les réseaux de neurones artificiels.* Thèse de doctorat, Université de Paris 7, 1998.

[20] G. Choquet. Theory of capacities. *Ann. Inst. Fourier*, 5:131–295, 1953.

[21] L. Cohen. Probability distributions with given multivariate marginals. *Journal of mathematical physics*, 25:2402–2403, 1984.

[22] H. Cramér. *Mathematical methods of Statistics.* Princeton Univ. Press, Princeton, 1946.

[23] C.M. Cuadras. Probability distribution with given multivariate marginals and given dependance structure. *Journal of multivariate analysis*, 45:51–66, 1992.

[24] C.M. Cuadras and J. Augé. A continuous general multivariate distribution and its properties. *Comm. statist. theory methods*, A 10:339–353, 1981.

[25] A. Cutler and M.P. Windham. Information-based validity functionals for mixture analysis. In *Proceedings of the first US-Japan conference on the frontiers of statistical modeling*, 1993.

[26] G. Dall'Aglio. Sugli estremi dei momenti delle funzioni de ripartizione doppia. *Ann. Scuola Norm. Sup. Pisa*, 10:35–74, 1956.

[27] G. Dall'Aglio. Fréchet classes: the beginnings. In *Advances in probability distributions with given marginals*, pages 1–12, Rome, 1991.

[28] R.E. Davis and D.R. Walker. An upper-air synoptic climatology of the western united states. *American Meteorological Society*, 5:1449–1467, 1992.

[29] P. Deheuvels. Caractérisation complete des lois extrêmes multivariées et de la convergence des types extrêmes. *Pub. Institt. stat. Univ. Paris*, 23:1–37, 1978.

[30] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society*, 39:1–38, 1984.

[31] E. Diday. *Classification automatique des données*. Dunod informatique, Paris, 1989.

[32] E. Diday. Probabilistic, possibilist and belief objects for knowledge analysis. *Annals of operations research*, 55:227–276, 1995.

[33] E. Diday. A generalisation of the mixture decomposition problem in the symbolic data analysis framework. *Cahiers du CEREMADE*, 0112, 2001.

[34] E. Diday. Knowledge discovery from symbolic data and the sodas software. *Cahier du CEREMADE*, 105, 2001.

[35] E. Diday and al. *Optimisation en classification automatique*. INRIA, Rocquencourt, 1980.

[36] E. Diday and R. Emilion. Lattices and capacities in analysis of probabilist objects. *Ordinal an Symbolic Data Analysis*, 1995.

[37] E. Diday and R. Emilion. Treillis de galois maximaux et capacités de choquet. *CRAS Analyse Mathématique*, t 324, série 1:261–266, 1997.

[38] E. Diday, Y. Ok, and A. Schroeder. The dynamic clusters method in pattern recognition. In *Proceeding of IFIP congress*, Stockolm, 1974.

[39] E. Diday and A. Schroeder. A new approach in mixed distributions detection. *RAIRO*, 10, 1976.

[40] J. Diebolt and C.P. Robert. Estimation of finite distributions through bayesian sampling. *Journal of the royal statistical society*, 56:363–375, 1994.

[41] R. Emilion. Clustering and mixtures of stochastic processes. *C.R.A.S.*, Série I, 335:189–193, 2002.

[42] W. A. Ericson. Subjective bayesian models in sampling finite populations. *J. R. Statist. Soc. B*, 31:195–233, 1969.

[43] B. Everitt and D. Hand. *Finite mixture distributions*. Chapman and Hall, London, 1981.

[44] T. Ferguson. Prior distributions in spaces of probability measures. *Annals of Stat.*, 2:615–629, 1974.

[45] R. Féron. Sur les tableaux de corrélation dont les marges sont données, cas de l'espace à trois dimensions. *Publ. Inst. Stat. Univ. Paris*, 5:3–12, 1956.

[46] N.I. Fisher. Copulas. *Encyclopedia of statistical sciences*, 1:159–163, 1997.

[47] N.I. Fisher and P.K. Sen. *The collected works of Wassily Hoeffding*. Springer-Verlag, New-York, 1994.

[48] E. Fix and J.L. Hodges. Discriminatory analysis, nonparametric estimation: consistency properties. *Report of USAF school of aviation medicine*, 4, 1951.

[49] M.J. Frank. On the simultaneous associativity of f(x,y) and x+y-f(x,y). *Aequationes Math.*, 19:53–77, 1979.

[50] M. Fréchet. Sur les tableaux de corrélation dont les marges sont données. *Ann. Univ. Lyon*, A 9:53–77, 1951.

[51] J. Galambos. *The asymptotic theory of extrem order statistics.* John Wiley and sons, New-York, 1978.

[52] C. Genest. Frank's family of bivariate distributions. *Biometrika*, 74:549–555, 1987.

[53] C. Genest and al. De l'impossibilté de construire des lois a marges multidimensionnelles données a partir de copules. *C.R.A.S. Paris*, 1:723–726, 1995.

[54] C. Genest and K. Goudi. Une famille de lois bidimensionnelles insolite. *C.R.A.S. Paris*, 1:351–354, 1994.

[55] C. Genest and J. Mackay. Copules archimédiennes et familles de lois bidimensionnelles dont les marges sont données. *Canadian journal of statistics*, 14:145–159, 1986.

[56] C. Genest and L.P. Rivest. Statistical inference procedures for bivariate archimedean copulas. *JASA*, 88:1034–1043, 1993.

[57] I.J. Good. A bayesian significance test for multinomial distributions. *J. R. Statist. Soc. B*, 29:399–431, 1967.

[58] L.A. Goodman. Exploratory latent structure models using both identifiable and unidentifiable models. *Biometrika*, 61:215–231, 1974.

[59] Younès Hillali. *Analyse et modélisation des données probabilistes: capacités et lois multidimensionnelles.* Thèse de doctorat, Université de Paris IX Dauphine, 1998.

[60] J.P. Hobert, C.P. Robert, and D.M. Titterington. On perfect simulation for some mixtures of distributions. *Statistics and computation*, 9:287–298, 1999.

[61] W. Hoeffding. Scale invariant correlation theory. In *The collected works of Wassily Hoeffding*, pages 57–107, 1940.

[62] W. Hoeffding. Scale invariant correlation measures for discontinuous distrbutions. In *The collected works of Wassily Hoeffding*, pages 109–133, 1941.

[63] M. Ishiguro, Y. Sakamoto, and G. Kitagawa. Bootstrapping log)likelihood and eic, an extension of aic. *Annals of the institute of statistical mathematics*, 49:411–434, 1997.

[64] A.K. Jain and R.C. Dubes. *Algorithms for clustering data.* Prentice Hall Advanced Reference Series, 1988.

[65] H. Joe. Parametric families of multivariate distributions with given marginals. *Journal of multivariate analysis*, 46:262–282, 1993.

[66] H. Joe. *Multivariate models and dependence concepts.* Chapman and Hall, London, 1997.

[67] Laurence S. Kalstein, J. Scott Greene, Michael C. Nichols, and C. David Barthel. A new spatial synoptic climatological procedure. In *AMS Eight Conference on Applied Climatology*, pages 169–173, Anaheim, California, 17-22 January 1993.

[68] G. Kimeldorf and A. Sampson. Uniform representation of bivariate distributions. *Comm. statist. A - Theory methods*, 4:617–627, 1975.

[69] J.F.C. Kingman. Random discrete distributions. *J. of the Royal Statit. Society*, 1:1–22, 1975.

[70] Y. Kodratoff and E. Diday. *Induction symbolique-numérique.* Cepadues editions, 1991.

[71] S. Kullback and R.A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[72] E.L. Lehmann. *Theory of point estimation.* Wadsworth and Brooks statistic/Probability series, 1983.

[73] C.-H. Ling. Representation of associative functions. *Publ. Math. Debrecen*, 12:189–212, 1965.

[74] A.W. Marshall and I. Olkin. Families of multivariate distributions. *J. Amer. Stat. Ass*, 83:834–841, 1988.

[75] G. McLachlan and D. Peel. *Finite mixture model.* Wiley series in probability ans statistics, 2000.

[76] K. Menger. Statistical metrics. *Proc. Nat. Acad. sci. U.S.A.*, 28:535–537, 1942.

[77] J.L. Molliere. What is the real number of clusters? In *9th meeting of the German Classification Society*, 1985.

[78] A. Nataf. Détermination des distributions dont les marges sont données. *C.R.A.S. Paris*, A 255:42–43, 1962.

[79] R.B. Nelsen. Properties of one-parameter family of bivariate distribution with specified marginals. *Communications in Statistics*, A 15:3277–3285, 1986.

[80] Roger B. Nelsen. *An introduction to Copulas.* Springer Verlag, Lectures Notes in Statistics, 1998.

[81] K. Pearson. Contributions to the theory of mathematical evolution. *Philosophical transactions of the royal society of London*, A 185:71–110, 1894.

[82] C.E. Priebe. Adaptative mixtures. *Journal of the Amer. Statist. Ass.*, 89:796–806, 1994.

[83] J.O. Ramsey and B.W. Silverman. *Functional data analysis.* Springer, New-York, 1997.

[84] B.L.S. Prakasa Rao. *Nonparametric functional estimation.* Academic press, New-York, 1983.

[85] R.A. Redner. Note on the consistency of the maximum likelihood estimate for no n identifiable distributions. *Annals of statistics*, 9:225–228, 1981.

[86] R.A. Redner and H. Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM*, 26:195–239, 1984.

[87] C. De Rham. La classification hiérarchique ascendante selon la méthode des voisins réciproques. *Cahiers de l'Analyse des Données*, 5, 1980.

[88] C.P. Robert. Prior feedback: Bayesian tools for maximum likelihood estimation. *Computational statistics*, 8:279–294, 1993.

[89] C.P. Robert. Mixture of distributions; inference and estimation. *Markov vhain Monte carlo in practice. Chapman and Hall*, pages 441–464, 1996.

[90] C.P. Robert, G. Celeux, and J. Diebolt. Bayesian estimation of hidden markov chains: a stochastic implementation. *Statistics and probability letters*, 16:77–83, 1993.

[91] C.P. Robert and K.L. Mengersen. Reparametrization issues in mixture modelling and their bearing on mcmc algorithms. *Computational statistics and data analysis*, 29:325–343, 1995.

[92] A. Schroeder. Analyse d'un mélange de distributions de probabilité de même type. *Revue de statistiques appliquées*, 24:39–62, 1976.

[93] B. Schweizer. Distributions are the numbers of the futur. In *Proc. sec. Napoli meeting on the mathematics of fuzzy systems*, pages 137–149, Instituto di mathematica delle faculta di mathematica, 1985.

[94] B. Schweizer. Thirty years of copulas. In *Advances in probability distributions with given marginals*, pages 13–50, Rome, 1991.

[95] B. Schweizer and A. Sklar. *Probabilistic Metric Spaces.* Elsevier North-Holland, New-York, 1983.

[96] A. Scott and M. Symons. Clustering methods based on likelihood ratio criteria. *Biometrics*, 27, 1971.

[97] I. Shlezinger. An algorithm for solving the self organization problem. *Cybernetics*, 2, 1968.

[98] B.W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London, 1986.

[99] A. Sklar. Fonction de répartition à n dimensions et leurs marges. *Inst. Statist. Univ. Paris Pub.*, 8:229–231, 1959.

[100] A. Sklar. Random variables, distribution functions and copulas - a personal look backward and forward. In *Distributions with fixed marginals*, pages 1–14, Institute of mathematical statistics, Hayward, 1996.

[101] A.F.M. Smith and U.E. Makov. A quasi-bayes sequential procedure for mixtures. *Journal of the royal statistical society*, B 40:106–112, 1978.

[102] A.F.M. Smith and G.O. Roberts. Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. *Journal of the royal statistical society*, B 55:3–23, 1993.

[103] A. Stuart and K. Ord. *Kendall's advanced theory of statistics, volume 1 : Distribution Theory*. Edward Arnold, 1994.

[104] A. Stuart, K. Ord, and S. Arnold. *Kendall's advanced theory of statistics, volume 2a : Classical Inference and the linear model*. Edward Arnold, 1999.

[105] M. Symons. Clustering criteria and multivariate normal mixtures. *Biometrics*, 37, 1981.

[106] M.A. Tanner and W.H. Wong. The calculation of posterior distribution by data augmentation (with discussion). *Journal of the american statistical association*, 82:528–550, 1987.

[107] P. Tassy and S. Legait. *Théorie des probabilités en vue des applications statistiques*. technip, 1990.

[108] M. Vrac and E. Diday. Description symbolique de classes. *Cahiers du CEREMADE*, 2001.

[109] M. Vrac, E. Diday, A. Chédin, and P. Naveau. Mélange de distributions de distributions. In *SFC'2001 8^{mes} Rencontres de la Société Francophone de Classification*, Université des Antilles et de Guyane, Guadeloupe, 17-21 décembre 2001.

[110] M. Vrac, E. Diday, S. Winsberg, and M.M. Liman. A top down binary tree method for symbolic class description. In *IPMU 2002*, Annecy, 2002.

[111] M. Vrac, E.Diday, and A. Chédin. Mixture decomposition of copulas and application to climatology. In *IPMU 2002*, Université de Savoie, Annecy, 2002.

[112] M. Vrac, M.M. Limam, S. Winsberg, and E. Diday. Symbolic class description. In *IFCS 2002*, Crakovie, Pologne, 2002.

[113] A. Wald. Note on the consistency of the maximum likelihood estimate. *Annals of mathematical statistics*, 20:595–601, 1949.

[114] G.C.G. Wei and M.A. Tanner. A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the american statistical association*, 85:699–704, 1990.

[115] D.V. Widder. *The Laplace transform*. Princeton University press, Princeton, 1941.

Annexe A

Article : "Symbolic Class Descriptions"

Cet article a été publié dans le proceeding de la conférence internationale "IFCS 2002" de Crakovie en juillet 2002. Le but est la description d'une classe d'individus décrits par des histogrammes à modalités nominales en tenant compte à la fois de l'homogénéité des sous-classes caractérisant la classe à décrire, et de sa discrimination par rapport à une partition a priori donnée. L'article aborde donc une méthode supervisée et non-supervisée.

Symbolic Class Descriptions

Mathieu Vrac[*, ***], Edwin Diday[*],
Suzanne Winsberg [**], Mohamed Mehdi Limam [*]

[*] LISE-CEREMADE, Université Paris IX Dauphine,
Place du Maréchal de Lattre-de-Tassigny, 75775 Paris

[**] IRCAM, 1 Place Igor Stravinsky, Paris 75004, France

[***] ARA, Laboratoire de météorologie dynamique,
Ecole Polytechnique, 91128 Palaiseau Cedex

Résumé

Our aim is to describe a partition of a class by a conjunction of characteristic properties. We use a stepwise top-down binary tree method. At each step we select the best variable and its optimal splitting to optimize simultaneously a discrimination criterion given by a prior partition and a homogeneity criterion. Moreover, this method deals with a data table in which each cell contains a histogram of nominal categories but the method can be extended or reduced to other types of data. The method is illustrated on both simulated data and real data.

A.1 Introduction

Classification methods are often designed to split a population of statistical individuals to obtain a partition into L classes. Generally, a partition is designed to optimize an intra-class homogeneity criterion as in classical clustering, or equivalently an inter-class criterion, as in classical decision or regression trees. In practice, when the aim is class description, it may be desireable to consider both types of criteria simultaneously. Here, our aim is to produce a description of a class which induces a partition of the class, both satisfying an intra-class homogeneity criterion and a discrimination criterion with respect to a prior partition. So, our approach has both unsupervised and supervised aspects. For example, the class to describe, C, could be young people of ages between 15 and 25, and the discriminatory categorical variable or prior partition could be smokers and nonsmokers. We want to obtain a description of C which induces a homogeneous partition of C which is well discriminated for the prior partition. The context here differs from that considered in Huygen's theorem, in which the intra-class and inter-class inertias are based on the same initial set. Here, we calculate the homogeneity criterion for the class we want to describe, but the discrimination criterion is based on the prior partition.

Our approach is based on divisive top-down methods, which successively divide the population into two classes, until a suitable stopping rule prevents further divisions. We use a

monothetic approach such that each split is carried out using only one variable, as it provides a clearer interpretation of the obtained clusters. Divisive methods of this type are often referred to as tree structured classifiers with acronyms such as CART and ID3 (see Breiman et al.(1984), Quinlan (1986)).

Not only does our paper combine the two approaches: supervised and nonsupervised learning, to obtain a description induced by the synthesis of the two methods, which is in itself an innovation, but it can deal with histogram data. We call histogram data, data in which the entries of the data table are weighted categorical or ordinal variables. These data are inherently richer, possesing potentially more information than the data previously considered in the classical algorithms mentionned above. This type of data is encountered when dealing with more complex, aggregated statistical units found when analyzing very large data sets. It may be more interesting to deal with aggregated units such as towns rather than with the individual inhabitants of the towns. Then the resulting data set, after the aggregation will most likely contain symbolic data rather than classical data values. By symbolic data we mean that rather than having a specific single value for an observed variable, an observed value for an aggregated statistical unit may be multivalued. For example, as in the case under consideration, the observed value may be a multivalued weighted categorical variable. For a detailed description of symbolic data analysis see Bock and Diday (2000). Naturally, classical data are a special case of the histogram type of data considered here. This procedure thus works for classical numerical or nominal data. It can also be applied when dealing with other types of symbolic data such as interval data. Others have developped divisive algorithms for data types encountered when dealing with symbolic data, considering either a homogeneity criterion or a discrimination criterion based on an a priori partition, but not both simultaneously. Chavent (1997) has proposed a method for unsupervised learning, while Périnel (1999), and Gettler-Summa (1999) have proposed ones for supervised learning. This method is an extension of that proposed by Vrac and Diday (2001).

First we describe our method, including some practical details necessary to implement it. For example we define a cutting or split for weighted categorical variables and we define a cutting value for this type of data. Then we outline the approach used to combine the two criteria. We present some examples of simulated data of histogram type to test the behavior of our new method. Finally we illustrate the algorithm with a real example dealing with unemployment data.

A.2 The Method

Four inputs are required for this method: 1) the data, consisting of n statistical units, each described by K histogram variables; 2) the prior partition into classes; 3) the class, C, the user aims to describe; and 4) a coefficient which gives more or less importance to the discriminatory power of the prior partition or to the homogeneity of the description of the given class, C. Alternatively, instead of specifying this last coefficient, the user may choose to determine an optimum value of this coefficient, using this algorithm.

The method uses a monothetic hierarchical descending approach working by division of a set into two nodes, that is sons. At each step l (l nodes corresponding to a partition into l classes), one of the nodes (or leaves) of the tree is cut into two nodes in order to optimize a quality criterion Q for the constructed partition into $l+1$ classes. The division of a node N into two nodes N_1 and N_2 is done by "cutting", where y is called the cutting variable and c the cutting value. We denote as N_1 and N_2, respectively, the left and right node of N.

The algorithm always generates two kinds of output. The first is a graphical representation, in which the class to describe, C, is represented by a binary tree. The final leaves are the clusters constituting the class and each branch represents a cutting (y, c). The second is a description: each final leaf is described by the conjunction of the cutting values from the top of the tree to this final leaf. The class, C, is then described by a disjunction of these conjunctions. If the user wishes to choose an optimal value of α using our data driven method, a graphical representation enabling this choice is also generated as output.

Let $H(N)$ and $h(N_1; N_2)$ be respectively the homogeneity criterion of a node N and of a couple of nodes $(N_1; N_2)$. then we define $\Delta H(N) = H(N) - h(N_1; N_2)$. Similarly we define $\Delta D(N) = D(N) - d(N_1; N_2)$ for the discrimination criterion. The quality Q of a node N (respectively q of a couple of nodes $(N_1; N_2)$) is the weighted sum of the two criteria, namely $Q(N) = \alpha H(N) + \beta D(N)$ (respectively $q(N_1; N_2) = \alpha h(N_1; N_2) + \beta d(N_1; N_2)$) where $\alpha + \beta = 1$. So the quality variation induced by the splitting of N into $(N_1; N_2)$ is $\Delta Q(N) = Q(N) - q(N_1; N_2)$. We maximize $\Delta Q(N)$. Note that since we are optimizing two criteria the criteria must be normalized. The user can modulate the values of α and β so as to weight the importance that he gives to each criterion.

To determine the cutting $(y; c)$ and the node to cut: first, for each node N select the cutting variable and its cutting value minimizing $q(N_1; N_2)$; second, select and split the node N which maximizes the difference between the quality before the cutting and the quality after the cutting, $max \Delta Q(N) = max[\alpha \Delta H(N) + \beta D(N)]$.

We recall that we are working with multivalued weighted categorical variables (histograms). So we must define what constitutes a cutting for this type of data and what constitutes a cutting value. The main idea is that the cutting value of a histogram variable is defined on the value of the frequency of just one category, or on the value of the sum of the frequencies of several categories.

To illustrate consider the following example: we have n statistical units in the class N (take $n = 3$), and consider variable Y_k, (say the variable color with categories red(r), blue(b), green(g), yellow(y)). Say, $unitA$ has values $r = 0.2, b = 0.1, g = 0.2, y = 0.5$; $unitB$ has values $r = 0.5, b = 0.2, g = 0.1, y = 0.2$; and $unitC$ has values $r = 0.1, b = 0.4, g = 0.1, y = 0.4$. For this variable Y_k we first order the units in increasing order of the frequency of just one category, eg red. We obtain $C(r = 0.1) < A(r = 0.2) < B(r = 0.5)$. So we can determine $n - 1 = 2$ cutting values by taking the mean of two different consecutive values (here cutting value $1 = (0.1 + 0.2)/2 = 0.15$ and cutting value $2 = (0.2 + 0.5)/2 = 0.35$). The-

refore we can also determine $n-1$ partitions into two classes (here partition $1 = \{N_1 = \{unitC\}; N_2 = \{unitA; unitB\}\}$ and partition $2 = \{N_1 = \{unitC; unitA\}; N_2 = \{unitB\}\}$ and so we have $n-1$ quality criterion values $q(N_1; N_2)$. We do it in turn for each single category (just red, just blue, just green, just yellow). Then we sort units in increasing order of the sum of frequencies of two categories. For example, with "(red + blue)" we obtain $unitA(r+b=0.3) < unitC(r+b=0.5) < unitB(r+b=0.7)$. We thus get $n-1$ new cutting values, ($n-1$ new partitions into two classes and $n-1$ new quality criterion values $q(N_1; N_2)$. In general if a histogram allows m categories we can look at the sorting on the sum of at most $m/2$ categories for even m and on the sum of at most $[m/2] = (m-1)/2$ categories for odd m. Indeed, in our little example we can see that the partitions obtained with "red + blue" are the same as those obtained with "green + yellow". We remark that if a multivalued weighted categorical variable Y has m categories, we have 2^{m-1} ways to sort the units in increasing order. Indeed, $2^{(m-1)}-1$ is the number of partitions in two non-empty classes from a set with m categories. Moreover, for each way of sorting we can have $(n-1)$ partitions of the units, so for a variable with m categories we have at most $(2^{(m-1)}-1)(n-1)$ partitions of the units.

The clustering or homogeneity criterion we use is an inertia criterion. This criterion is used in Chavent (1997). The inertia of a class N is

$$H(N) = \sum_{w_i \in N} \sum_{w_j \in N} \frac{p_i p_j}{2\mu} \delta^2(w_i, w_j),$$

and

$$h(N_1, N_2) = H(N_1) + H(N_2);$$

where p_i = the weight of individual w_i, and $\mu = \sum_{w_i \in N} p_i$ = the weight of class N, and δ is a distance between individuals. For histograms with weighted categorical variables, we can imagine many distances. We choose δ, defined as,

$$\delta(w_i; w_j) = \sum_{k=1}^{K} \sum_{m=1}^{mod[k]} |y_k^m(w_i) - y_k^m(w_j)|^2,$$

where $y_k^m(w)$ is the value of the category m of the variable k for the individual w, $mod[k]$ is the number of categories of the variable k, and K is the number of variables.

This distance must be normalized. We normalize it to fall in the interval [0,1]. So δ must be divided by K to make it fall in the interval [0,1].

Let us turn to the discrimination criterion. The discrimination criterion we choose is an impurity criterion, Gini's index. Gini's index, which we denote as D, was introduced by Breiman et al (1984) and measures the impurity of a node N with respect to the prior partition $G_1, G_2, ..., G_J$ by

$$D(N) = \sum_{l \neq j} p_l p_j = 1 - \sum_{j=1,\ldots,J} p_j^2,$$

with $p_j = n_j/n$, $n_j = card(N \cap G_j)$ and $n = card(N)$ in the classical case. In our case n_j = the number of individuals from G_j such that their characteristics verify the current description of the node N. To normalize $D(N)$ we multiply it by $J/(J-1)$; where J is the number of prior classes; it then lies in the interval [0,1].

Let us now discuss the robustness of the results. In many situations we obtain trees yielding unstable predicting models. Then it is necessary to prune the tree by removing the less significant branches. Consider a fixed value of α. The main idea is to estimate the inertia and discrimination rate for every subtree. The inertia and discrimination rate R associated with tree A is $R(A) = \sum_{t \in A} \frac{n_t}{n} Q(t)$ where n_t is the number of individuals in terminal node t and n is the total number of individuals. The optimal tree is the subtree minimizing this rate. We use a bootstrap method to estimate these rates. Here, pruning consists of selecting the best subtree from all subtrees obtained by removing branches from the main or starting tree. The tree with lowest value of R is the "best" subtree. Starting from the set of individuals we construct the main tree A_{max}. For each subtree A_h we estimate R using the bootstrap, so we have B samples, say 100, from the initial set of individuals. Then for each bootstrap sample we calculate R for each subtree A_h, and for each subtree A_h we calculate the mean $\overline{R}(A_h)$ of all the samples. Finally we choose the best subtree A_h^*, that is the one that has the minimum $\overline{R(A_h)}$.

The user may choose to optimize the value of the coefficient α. To do so, one must fix the number of terminal nodes. The influence of the coefficient α can be determinant both in the construction of the tree and in its prediction qualities. The variation of α (or of β since $\alpha + \beta = 1$) from 0 to 1 increases the importance of the homogeneity and decreases the importance of discrimination. This variation influences splitting and consequently results in different terminal nodes. We need to find the inertia of the terminal nodes and the rate of misclassification as we vary α. Then we can determine the value of α which optimizes both the inertia and the rate of misclassification ie the homogeneity and discrimination simultaneously. If on the contrary the user fixes the value of $\alpha = 0$, considering only the discrimination criterion, and in addition the data are classical, the algorithm functions just as CART. So CART is a special case of this algorithm.

For each terminal node t of the tree T associated with class c_s we can calculate the corresponding misclassification rate $R(s/t) = \sum_{r=1}^{L} P(r/t)$ where $r \neq s$ and $P(r/t) = \frac{n_r(t)}{n_t}$ is the proportion of the individuals of the node t allocated to the class c_s but belonging to the class c_r. The misclassification MR of the tree T is the sum over all terminal nodes ie $MR(A) = \sum_{t \in A} \frac{n_i}{n} R(s/t) = \sum_{t \in A} \sum_{r=1}^{L} \frac{n_r(t)}{n}$, where $r \neq s$. For each terminal node of the tree T we can calculate the corresponding inertia, $H(t)$ and we can calculate the total inertia by summing over all the terminal nodes. So, $H(t) = \frac{1}{2n|t|} \sum_{w_i \in t} \sum_{w_j \in t} \delta(W_i, W_j)$ with $|t| = card(t)$, and the total inertia of T, $I(A) = \sum_{t \in T} H(t)$.

The idea is to build for each value of α several trees from many samples and then to calculate the inertia and misclassification rate for each tree. Starting from our initial set of n individuals we extract B bootstrap samples of size n (by randomly sampling n individuals with replacement). For each sample and for each value of α between 0 and 1, we build a tree and calculate our two parameters (inertia and misclassification rate). Varying α from 0 to 1 (say with a stepsize of 0.1) gives us 11 couples of values of inertia and misclassification rate corresponding to the mean values of these parameters for the B bootstrap samples.

In order to visualize the variation of the two parameters, we display a curve showing the inertia and a curve showing the misclassification rate as a function of α. The optimal value of α is the one which minimizes the sum of the two parameters.

A.3 Examples

First we consider three sets of simulated data. These simulated data examples are presented to give a clear picture under controlled conditions, of how the algorithm works and how it permits an optimal choice of α. We also consider a real data set. The first example consists of 90 individuals with a prior partition. The class to describe, $C = C1 \cup C2$. Each individual is described by two variables of histogram type. This first simulation is constructed so as to make the class to describe have two subclasses, $C1$ and $C2$ with optimal homogeneity. Then we added a prior partition which perfectly distinguishes C from the rest of the population. In this very special case the inertia and the misclassification rate should not vary with the choice of α. Any value of α from 0 to 1 should give a result with the same inertia and misclassification rate. In fact for this example when we graphically display the results we obtain for these two parameters as a function of α we obtain a constant value.

The second simulated example is a modification of the first example. We modify the data from example 1 such that we deteriorate only the discrimination by changing the value of the discriminatory variable for some individuals, while keeping the homogeneity of the class to describe identical to that in example 1. This change should make it necessary to increase the importance of discrimination and thus the optimum value of β should be close to one (that is, α close to zero). In fact our results show that the inertia remains constant as a function of α while the misclassification rate increases as α increases from 0 to 1 indicating a choice of $\alpha = 0$ as expected.

In the third simulated example we modify the data from example one by deteriorating both the discrimination and the inertia. We find as expected that the optimal level of α depends upon the degree of deterioration of these two factors. For example, the optimal value of α decreases as the number of individuals whose discriminatory variable is changed increases.

Finally the fourth example deals with real unemployment data from 35 towns (districts) in Great Britain. The aim is to explain the factors which discriminate towns with high unemployment from those with low unemployment. But we also wish to have good descriptors of the resultant clusters due to their homogeneity. Because we have aggregated data for the

inhabitants of each town, we are not dealing with classical data with a single value for each variable for each statistical unit, here the town. The class to describe is the 35 towns, and the prior partition is low versus high employment. Here, each variable for each town is a histogram. There are $K = 4$ variables. The first is age with 6 categories: 0-4 years; 5-14 years; 15-24 years; 25-44 years; 45-64 years; greater than or equal to 65. The second is racial origin with 4 categories: Whites; Blacks; Asians; Others. The third is type of dwelling with 4 categories: owner occupied; public sector accomation; private sector accomadation; other. The fourth is social class with 4 categories: household head in social class 1 or social class 2; household head in social class 3; household head in social class 4 or social class 5; other. The discriminatory variable is unemployment rate for men and women. For these districts the rate varies between 0% and 18%. Two prior classes are defined: class1 for unemployment rate $\leq 9\%$ and class2 for unemployment rate $> 9\%$.

We stopped the algorithm with four terminal nodes of description. We obtain four symbolic descriptions of each node. An example of such a description is: [the proportion of people in social classes 1 and 2 is less than 33.9%], and [the proportion of people between 45 and 64 years of age is less than 21.8%]. When we use only a homogeneity criterion, that is we fix $\alpha = 1$, ($\beta = 0$), each description gathers homogenous groups of towns. The total inertia of the terminal nodes is minimized and equals 0.135. However, the misclassification rate is high, equal to 20.4%. Next we use only a discrimination criterion, (that is we fix $\alpha = 0, \beta = 1$). We choose an intial partition with P1 = towns where the unemployment rate is high and P2 = towns where the unemployment rate is low. We have the same set of towns to describe. In this case we have a misclassification rate of 6.25% considerably reduced from 20.4%. However the inertia is equal to 0.26 which is higher than above. So we have good discrimination but inferior homogeneity. Finally, we use our method and choose a value of α based on the data which optimizes both the inertia and the misclassification rate simultaneously. The inertia decreases when we increase α. But, there is a gradient from $0.4 \leq \alpha \leq 0.6$ showing that the inertia decreases sharply in this region. The misclassification rate increases when we increase α. However, there is a gradient between $0.5 \leq \alpha \leq 0.7$; showing that the rate increases sharply between these values. However, at $\alpha = 0.6$ the rate of misclassification is only slightly increased over that for $\alpha = 0$, which is the best rate. If we choose $\alpha = 0.6$ the inertia is 0.233 and the misclassification rate is 6.52%. So we have an almost optimal misclassification rate and a better class description, than that which we obtain when considering only a discrimination criterion; and we have a much better misclassification rate than that which we obtain when considering only a homogeneity criterion.

A.4 Conclusion

In this paper we present a new approach to get a description of a set. This method applicable to histogram data is new for the classical case as well. The main idea is to mix a homogeneity criterion and a discrimination criterion to describe a set according to an initial partition. The set to describe can be a class from a prior partition, the whole population or any class from the population. Having chosen this class, the interest of the method is that the user can choose the weights α and $\beta = 1 - \alpha$ he/she wants to put on the homogeneity and

discrimination criteria respectively, depending on the importance of these criteria to reach a desired goal. Alternatively, the user can optimize both criteria simultaneously, choosing a data driven value of α. We show on a real data set that a data driven choice can yield an almost optimal discrimination, while improving homogeneity, leading to improved class description. One of the future evolutions of this algorithm will be the treatment of other types of symbolic data, such as symbolic data dealing with taxonomies and rules.

References

BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A., and STONE, C.J. (1984): *Classification and Regression Trees*. Wadsworth, Belmont, California.

CHAVENT, M. (1997): *Analyse de Données Symboliques, Une Méthode Divisive de Classification*. Thèse de Doctorat, Université Paris IX Dauphine.

DIDAY, E. (1999): Symbolic Data Analysis and the SODAS Project: Purpose, History, and Perspective In: H.H. Bock, and E. Diday (Eds.): *Analysis of Symbolic Data*. Springer, Heidelberg, 1–23.

GETTLER-SUMMA, M. (1999): *MGS in SODAS : Marking and Generalization by Symbolic Objects in the Symbolic Official Data Analysis Software*. Cahiers du CEREMADE, Paris, France.

PÉRINEL, E. (1999): Construire un arbre de discrimination binaire à partir de données imprécises. *Revue de Statistique Appliquée*, *47*, 5–30.

QUINLAN, J.R. (1986): Induction of Decision Trees. *Machine Learning*, *1*, 81–106.

VRAC, M. and DIDAY, E. (2001): *Description Symbolique de Classes*, Cahiers de CEREMADE, Paris,France.

Annexe B

Article : "Mixture decomposition of copulas and application to clymatology"

Cet article a été publié dans le proceeding de la conférence internationale "Information Processing and Management of Uncertainty 2002", Annecy, juillet 2002.

Mixture decomposition of copulas and application to climatology

Mathieu Vrac
Address
vrac@lmd.polytechnique.fr

Edwin Diday
Address
diday@ceremade.dauphine.fr

Alain Chédin
Address
chedin@lmd.polytechnique.fr

Abstract

The goal of this work is to cluster a data set described by probability distribution functions in opposition to "classical" data set described by numerical or categorical values. The proposed method extends the decomposition of densities mixture and allows to take into account dependencies existing between variables and dependencies existing inside a variable between different points of the cumulative distribution. These dependencies are modeled with multidimensional functions called copulas (or "dependencies functions"). Copulas link every unidimensional cumulative distribution (called "margin") to the multidimensional cumulative distribution. First, we will link the searched partition to the decomposition of densities mixture. Secondly, We will look at the useful notions of distribution of distributions and copulas. Finally we will apply the clustering method to a climate data set.

Keywords: mixture model, distributions of distributions, copulas

1 Introduction

The classical method in mixture densities consists in estimating a probability density function from a given sample in $\Re^p$, considering that the reached function f is a finite mixture of K densities:

$$f(x_1, ..., x_p) = \sum_{l=1}^{K} p_l f(x_1, ..., x_p, \alpha_l) \qquad (1)$$

with $\forall l = 1, ..., K,\ 0 < p_l < 1$ and $\sum_{l=1}^{K} p_l = 1$, $f(., \alpha)$ is a density function with parameter α belonging to R^d (d is the number of coordinates of α) and p_l is the probability that one element of the sample get the density $f(., \alpha_l)$.

This problem has been investigated by many authors with two different approaches. The most widespread consists in seeing an estimation problem of parameters $(p_l, \alpha_l)_{l=1,...,K}$ ("estimation approach") [Everitt, Hand, 1981]. Classical methods of estimation are maximum likelihood technics. Generally, optimization algorithms of likelihood are EM (Estimation, Maximization) [Dempster, Laird, Rubin,1977], [Redner, Walker, 1984], [Shlezinger, 1968].
The second approach ("clustering approach") considers a partition $P = (P_1, ..., P_K)$, where each cluster P_l is assimilated to a sample with law $f(., a_l)$, K is supposed given ([Diday, Schroeder, 1976], [Scott et Symons, 1971], [Symons, 1981]). Used algorithms are "nuées dynamiques" methods. In this paper, we are interested in the generalization of the clustering approach : mixture of laws when data is probability distribution functions.
Our approach is in the framework of the analysis of symbolic data. In a symbolic data table, a cell can contain a distribution function (Schwe zer (1984) says that "distribution are the number of the future"). For more details on symbolic data analysis see Diday (1998), Diday, Bock (2000). The idea is to operate with dis-

tribution functions as numerical values. Here, a variable Y^j is considered to be a random variable from a set of units Ω in an infinite set of distributions F^j. After a short presentation of useful notions, we will explicit the proposed method and apply it to a climate data set.

2 Mixture decomposition applied to probability distributions

2.1 Input and output

We have a data table of n rows and p columns where the i^{th} row is associated to the unit (or "individual") $\omega_i \in \Omega$ and each column is defined by a variable $Y^j \in \{Y^1, ..., Y^p\}$. Each cell (i, j) of this table contains a distribution $Y^i(\omega_i) \in F^j$.

	Variable 1	.....	Variable p
Ind 1		.....	
.....	.....	.....	.....
Ind N		.....	

FIG. 1 – *data table of distributions*

Let's the sample note $F=(F_1, ..., F_n)$. F is called the "distributions base".
Each $F_i = (F_i^1, ..., F_i^p)$ corresponds to individual i and each F_i^j to the distribution of the individual i for the variable j.
We have a double objective: to decompose the probability law (and therefore to get a classification) and to model the dependencies existing inside a feature and between the variables. In output, we hope to get:

⋆ a partition of the sample

⋆ the parameters of copulas and PDD corresponding to the clusters and proportions $(p_i)_{i=1,...,K}$.

To reach this goal, we will use copulas and distributions of distributions notions.

2.2 Distributions of distribution and copulas

To work with distribution, we need distributions of distributions notion developped by E. Diday [2001].

Definition : *Let* F$= (F_1, ..., F_n)$ *be a set of distributions. A "Point of Distribution of Distributions" (PDD) associated to* F*, at the point* T*, is defined by* $G_T(x) = Pr(\{F_i \in F \ / F_i(T) \leq x\}) \forall\, x \in \overline{\Re}$.

For instance, if $x = 0.5$, $G_T(x)$ is the percentage of distributions taking a smaller value than 0.5 at the point T (see Figure 2). Let's define a q-dimensional PDD.

Definition : *A "Point of Joint Distributions of Distributions" (PJDD) associated to* F *at the point* $(T_1, ..., T_q)$ *is defined by*

$$H(x_1, ..., x_q) = Pr(\{F_i \in F / F_i(T_1) \leq x_1\} \wedge ...$$

$$... \wedge \{F_i \in F / F_i(T_q) \leq x_q\})$$

From these definitions, we can see that G_T

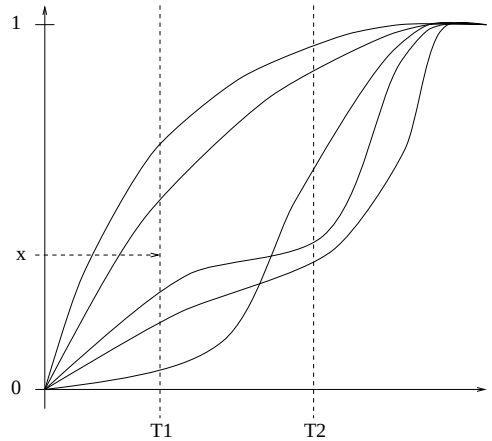


FIG. 2 – *Points of Distributions of distributions for* (T_1, T_2)

is a probability distribution function [Did 01] and that H is a q-dimensional joint distribution with margins $G_{T_1}, ..., G_{T_n}$.

2.3 To model dependence between different PDD with copulas

A function copula (Schweitzer et Sklar ,1983) links multidimensional probability distributions to the margins of a multidimensional random variable.

The most important theorem in copula's theory is the Sklar's theorem:
Let H a joint distribution function with margins F and G. Then there exists a copula C such that for all x,y in $\overline{R}$,

$$H(x,y) = C(F(x), G(y)). \qquad (2)$$

If F and G are continuous, then C is unique; otherwise C is uniquely determined on $RanF \times RanG$. Conversely, if F and G are distribution functions and C is a copula, the function H defined by [2] is a joint distribution function with margins F and G.

Modelisation of dependencies of PDD is based on the following proposition [Did 01], coming from the Sklar's theorem.

Proposition : *It exists a q-copula C such that for all* $(x_1, ..., x_q)$ *belonging to* $\overline{R}^q$,
$H(x_1, ..., x_q) = C(G_{T_1}(x_1), ..., G_{T_q}(x_q))$.
Moreover, C is uniquely determined on $Ran(G_{T_1}) \times \ ... \ \times \ Ran(G_{T_q})$

As H is a joint distribution, we can put

$$H(x_1, ..., x_q) = \sum_{k=1}^{K} p_k H(x_1, ..., x_q, \alpha_k)$$

with for all $k = 1, ..., K$, $0 < p_k < 1 \sum_{k=1}^{K} p_k = 1$, where $H(..., \alpha)$ is a distribution with parameter α belonging to $\Re^d$ and p_k is the probability to get the law $H(..., \alpha_k)$.
From the previous proposition, we can write

$$C(G_{T_1}(x_1), ..., G_{T_q}(x_q))$$
$$= \sum_{k=1}^{K} p_k C(G_{T_1}(x_1, b_1^k), ..., G_{T_q}(x_q, b_q^k), \beta_k) \qquad (3)$$

where β_k is the parameter of the copula corresponding to the cluster k. $G_{T_i}(., b_i^k)$ is the PDD with parameter b_i^k, at the point T_i in the cluster k.

Let's $h(x_1, ..., x_q)$ note the probability density function corresponding to H:
$h(x_1, ..., x_q) = \frac{\partial^q H}{\partial x_1 ... \partial x_q}(x_1, ..., x_q)$.

$$h(x_1, ..., x_q) = (\prod_{i=1}^{q} \frac{dG_{T_i}}{dx}(x_i)) \times$$
$$\frac{\partial^q C}{\partial u_1 ... \partial u_q}(G_{T_1}(x_1), ..., G_{T_q}(x_q)). \qquad (4)$$

From the equations [1] and [4], if we work with densities (and no distributions), we get

$$h(x_1, ..., x_q) = \sum_{k=1}^{K} p_k (\prod_{i=1}^{q} \frac{dG_{T_i}}{dx}(x_i, b_i^k)) \times$$
$$\frac{\partial^q C}{\partial u_1 ... \partial u_q}(G_{T_1}(x_1, b_1^k), ..., G_{T_q}(x_q, b_q^k), \beta_k).$$

3 Clustering algorithm

The developped method is an extension to distributions of the "nuées dynamiques" method (Diday, Ok, Schroeder, 1974) for densities mixture. The main idea is to estimate, at each step, the parameters of copulas which describe the best the clusters of the current partition, according to a given quality criterion. First we have to fix a model of copula. Either we consider non-parametrical copulas, or we use a family of parametrical copulas. In this case, many families can be investigated. Here, we will just use the Frank's family. Moreover, a criterion to mesure adequation between a partition P and a set of copulas $(C_{\beta_i})_{i=1,...,K}$ is needed. We will work with the classifier log-likelihood

$$lvc(P, \beta) = \sum_{k=1}^{K} \sum_{w \in P_k} log(h(w, \beta_k)).$$

Other criteria of fit could give good results.
After the initialization of a partition, the algorithm is defined in two successive and iterative steps:
Step 1. Estimation of the parameters $(\beta_1, ..., \beta_K)$ maximizing the given criterion.
Step 2. Distribution of the units in the new classes $(P_i)_{i=1,...,K}$:
$P_i = \{\omega \ tq \ p_i h(\omega, \beta_i) \geq p_m h(\omega, \beta_m) \ \forall \ m\}$,

with $i \leq m$ in case of equality. If we choose a non-parametrical copula this algorithm can be applied in the same way (see 3.2).

3.1 Estimation of mixture ratios and parameters of copulas

3.1.1 mixture ratios

Estimation of $(p_i)_{i=1,...,K}$ is classical. We use $p_i = \frac{cardP_i}{card(F)}$ as in the algorithm of the mobil centres. Some other estimations can be found in Celeux, Govaert, 1993.

3.1.2 Parameters of copulas

The parameters β of the copulas to determine must maximize $lvc(P, \beta)$. The parameters β maximise

$$\sum_{w=(x_1,...,x_q)\in P_k} log[(\prod_{i=1}^{q} \frac{dG_{T_i}}{dx}(x_i, b_i^k)) \times$$

$$\frac{\partial^q C}{\partial u_1 ... \partial u_q}(G_{T_1}(x_1, b_1^k), ..., G_{T_q}(x_q, b_q^k), \beta_k)].$$

These equations can be complexe even if we work with a copula in two dimensions. The Frank's copula in two dimension is:

$$C_\beta(u, v) = \frac{\log(1 + \frac{(\beta^u - 1)(\beta^v - 1)}{(\beta - 1)})}{\log(\beta)},$$

with β strictely positive and $\beta \neq 1$ and $\forall$ u,v $\in$ [0,1]. It has the following properties:

$$\lim_{\beta \to 0} C_\beta(u, v) = min(u, v),$$

$$\lim_{\beta \to 1} C_\beta(u, v) = uv \text{ (independence copula)}$$

$$\text{and } \lim_{\beta \to \infty} C_\beta(u, v) = max(u + v - 1, 0).$$

This Archimedean copula ([NEL 98]) is generated by the function $\phi_\beta(t) = -\log(\frac{1-\beta^t}{1-\beta})$ and we have

$$\frac{\partial^2 C_\beta}{\partial u \partial v}(u, v) = \frac{(\beta - 1)\log(\beta)\beta^{u+v}}{[(\beta - 1) + (\beta^u - 1)(\beta^v - 1)]^2},$$

$0 \leq u, v \leq 1$. The parameters β will be determined by the resolution of the log-likelihood equations with a numerical estimation.

If we are more interested in the classification than in the parameters of copulas, the non-parametrical estimation of copulas can be an answer. The second derivative of C according to u and v is the two dimensional density for the PDD (in two dimensions, C is the distribution function of the PDD $(G_i^1(x^1), G_i^2(x^2))$ $\forall(x^1, x^2)$). We use the classical methods to estimate densities .

3.2 To model and estimate the PDD

Estimation of the PDD corresponds to estimate a distribution function. Methods of determination are numerous. In this paper, we are interested in three methods.
The first one is the simplest, the empirical frequency.

$$G_T(x) = \frac{card(\{F_i \in F \ / F_i(T) \leq x\})}{card(F)}$$

It gives only some values for $G_T(x)$.

The second one is the "troncated" Parzen method. From a sample $(X_1, ..., X_n)$, the density estimation can be

$$\hat{f}(x) = \frac{1}{c_n} \frac{1}{nh} \sum_{i=1}^{n} Ke(\frac{x - X_i}{h}),$$

with c_n such that $\int \hat{f} = 1$. Ke is the kernel (density function), h is the window width ([Sil 86]), automically estimated with the Mean Integrated Square Error (MISE) formulae $h = 1.06\sigma N^{-1/5}$ with σ the standard deviation from the sample.

The last model is a beta law (Dirichlet's law in one dimension). The density function is

$$f_{a,b}(x) = \frac{x^{a-1}(1 - x)^{b-1}}{\int_0^1 y^{a-1}(1 - y)^{b-1} dy},$$

with $a > 0$ and $b > 0$ the two parameters for the beta law. The essential particularity is to go from [0,1] to [0,1].

4 Classification of multidimensional distributions

A multidimensional copula is complex. Numerous kinds exist, they have avantages and

desavantages, the commun point is the complexity ([Hil 98]). We propose two methods to treat two distributions at the same time and/or more than two T_i.

4.1 Coupling

The first one is the coupling. From Ω a sample of n individuals described by two distributions variables Y^1 and Y^2, we determine two copulas mixture decomposition: one on Y^1 on K_1 clusters (with two thresholds T_1^1 and T_2^1), the other on Y^2 on K_2 clusters (with 2 thresholds T_1^2 and T_2^2). The value of the joint distribution is known for all $\omega \in \Omega$ and for all $(Y^i)_{i=1,2}$

$$H^{Y^i}(\omega) = \sum_{k=1}^{K_i} p_k C_{\beta_k^i}(G^k_{T_1^i}(\omega_1^i, b^k_{T_1^i}), G_{T_2^i}(\omega_2^i, b^k_{T_2^i}))$$

with (ω_1^i, ω_2^i) = values of the distribution of the variable Y^i for ω in T_1^i and T_2^i respectively ($\omega_j^i = Pr(X_j^i \leq T_j^i)$ and X_j^i is the random variable corresponding to the individual ω_j and the variable Y^i).
β_k^i = parameter of the copula from cluster k and variable Y^i
$b^k_{T_j^i}$ = parameter associated to the PDD defined in T_j^i from cluster k
$G^k_{T_j^i}(\omega_j^i, b^k_{T_j^i})$ = value of the PDD associated to Y^i at T_j^i in ω_j^i. We have a couple of values of distributions (H^{Y^1}, H^{Y^2}) for each individual ω. We put $(H^{Y^1}(\omega_i))_{i=1,\ldots,n}$ are from distribution law F_1, $(H^{Y^2}(\omega_i))_{i=1,\ldots,n}$ are from distribution law F_2 and $(H^{Y^1}(\omega_i), H^{Y^2}(\omega_i))_{i=1,\ldots n}$ are from law H. Hypothesis from Sklar's theorem are cheked. There exists a copula C such that for all $(x_1, x_2) \in [0,1]^2$, $H(x_1, x_2) = C(F_1(x_1), F_2(x_2)$. We can apply a new copulas mixture from the n couples. The parameters of copulas from the decompositions on Y^1 and Y^2 give information on dependencies inside the variables (between T_1^1 and T_2^1 and between T_1^2 and T_2^2). The parameters from coupling give informations on the dependencies between Y^1 and Y^2.

4.2 Binary tree

The second method is a step down binary tree method. From the distributions base, the best partition in two classes is computed and the best partition in two classes from the new clusters and so on. For each node and each variable, optimal (T_1, T_2) is determined. The node to cut N has the cutting variable and the associated (T_1, T_2) which maximize the quality criterion Q,
$Q(N) = \sum_k \sum_{\omega \in P_k} \log(h_{\beta_k}(\omega))$.
The binary tree method can work with different variables, (T_1, T_2) and copulas for each node. The method can be longer than coupling. For each (T_1, T_2), $F(T_i(\omega))$ is estimated for every individual ω of the node. Moreover, two individuals in two different classes cannot be grouped in the following.

5 Application to a climatic data base

Climate study is an important axis of the world research. In short-term weather forcasts ([KAL 93], [DAV 92]) or in long-term dynamic evolution of climate, statistics and data analysis play an essential part. For example, in the algorithm of inversion of the equation of radiative transfert (allows to interpret satellital observations in thermodynamic atmospheric variables [ACH 91]), a partition of atmospheric profiles is used to determinate an initial solution near the real solution. The used partition must group profiles with similar physic properties inside a cluster and distinct physic properties between the clusters.
The method we will apply allows to know the probability laws of the variables and the probabilities of occurency of the atmospheric profiles.

We have atmospheric data from the European Center for Meteorological Weather Forecasting (ECMWF) from Reading. We realize a wire of the earth, each mesh corresponds to a latitude degree and a longitude degree. The wire is extended in altitude to 50 levels called "sigma coordinates". For each point of the wire, we have pression values, temperature, specific humidity, wind, etc. These values are forcasts every 6 hours, that is 4 times a day (0 a.m., 6 a.m., 12 a.m., 6 p.m.) from 1998 december to 1999 december. We have a 3-

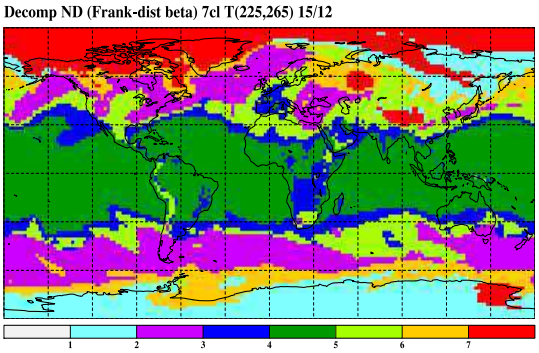


FIG. 3 – *classification on distributions of temperature (Frank's copula, PDD beta law)*

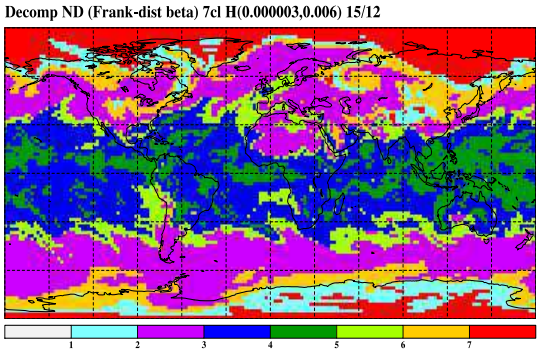


FIG. 4 – *classification distributions of humidity (Frank's copula, PDD law beta)*

dimensional wire of the earth representing a year of its thermodynamic state. From these data we want to study the classification and estimations for the temperature variable for the atmospheric profiles of 1998 december the 15^{th} at 0 a.m.. Temperature data are numerical values for each profile and each sigma coordinate. Estimation of the probability distribution function is realized. Each temperature profile has a distribution. Two thresholds T_1 and T_2 are fixed by a prior knowledge for the estimation of the PDD $(G_{T_i}(x))_{i=1,2}$. $T_1 = 225K$ *and* $T_2 = 265K$ (K=Kelvin degrees).

5.1 Results

We classify the profiles of temperature from 1998 december the 15^{th} at 0 a.m. in 7 clusters from the Frank's copula and with a PDD modeled with a beta law with parameters ν_1 and ν_2. We have taken one profile on 4 (that is one on two on latitude and one on two on longitude), the distributions base contains more than 16.000 distributions. The classification in Figure 3 is obtained for 2 iterations. Each pixel is bigger than in real to get a continuous visual effect. Parameters of copulas and PDD are in Table 1. The clusters of the partition seem to be coherent and get realistic and distinct climatic properties. We can see a big class called "tropical class" (cluster 4), two "polar" classes, winter in north pole (cluster 1) and summer in south pole (cluster 7), two "temperate" classes (clusters 2 and 5). The cluster 3 links moderate zones and tropical zones and cluster 6 links polar zones and moderate zones. The High reliefs are well-identified (Himalaya, Andies). Moreover the clusters 1, 4 and 5 have similar parameters of copulas (0.000001) near zero, meaning that their copulas are near the Min copula $C(u, v) = min(u, v)$. Distributions inside the clusters have tendency to evolve in a parallel way without cutting. Generally the fit with the synoptic analisys of the situation is good (shape of hot or cold air incursions, deression, etc.).

A similar classification in 7 classes has been realized on distributions of humidity. The obtained parameters *beta* are in table 1 and the classification in Figure 4.

The previous tropical class is cut in two classes (3 and 4). The cluster 4 contains the most humid zones. Moreover, the boudary between the cluster 3 and the cluster 2 (less humid air mass) is edged with the cluster 2 (humid air incursions in a drier air mass). The method identifies 2 clusters with a weak humidity (1 and 7). The two clusters have similar parameters of copulas (0.0078 and 0.001) meaning a similar dependency of the behaviour of their distributions. Moreover, a spiral at $60^\circ N, 60^\circ E$ fixes the position of a depression (it is a red disc for the classification on temperature).

From the classifications in 7 classes on distributions of temperature and humidity a coupling has been realized. The results of the partition in 7 classes are in table 1 and Figure 5.

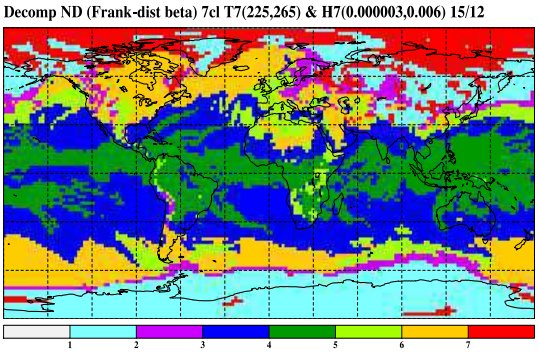


FIG. 5 – *classification on distributions of temperature and humidity (Frank's copula, PDD law beta)*

Classes	Temp	Hum	Temp+Hum
Classe 1	0.000001	2.1	0.0078
Classe 2	0.3	0.045	0.1
Classe 3	0.004	0.000001	0.1
Classe 4	0.000001	30	0.022
Classe 5	0.000001	0.009	0.041
Classe 6	0.0308	0.14	0.1
Classe 7	0.007	0.1	0.001

Table 1. *Parameters β at convergence (temperature, humidity and temperature+humidity)*

The coupling gives good results, coherent mixture of the two previous classifications. We see winter in north pole (cluster 7) and summer in south pole (cluster 1) coming from temperature with variations from humidity. Moreover, two tropical classes are identified (cluster 4 and 3). The cluster 4 gives better the humid zones than the cluster 4 in classification on humidity Figure 4. The other clusters are transitions from tropical classes (hot and humid) to polar classes (dry and cold). The spiral (60 ° N,60 ° E) is present.

6 Conclusion

The used approach of decomposition of mixture generalizes the classical one in using a superior abstraction level [Did 01]. The first results on complex data give a realistic classification in climatology and are encourageant for the next variables. The interests in climatology and meteorology are modeling of dependencies between variables and modeling of probabilty distribution functions. Moreover, the parameters of copulas have been estimated by maximum likelihood, Kendall's tau or Spearman's rho could be used [Nel 98]. PDD are modeled with beta law but can be modeled by a classical decomposition of mixture. The choice of the thresholds T_i is a delicate choice. The tries to automatically get an optimal T_i are not successful. The multidimensional case must be a strong axis of research. The complexity of the formulae over two dimensions is such that cleverness of simplification is needed. An extension of EM and SEM is in progres and other algorithms can be generalized.

Acknowledgements

The autors would like to thank all the ARA group from the LMD and particularly R. Armante, N. A. Scott and S. Serrar for their help, their support and their friendship.

Références

[1] V. Achard. *Trois problemes clés de l'analyse 3D de la structure thermodynamique de l'atmosphere par satellite : mesure du contenu en ozone; classification des masses d'air; modélisation hyper rapide du transfert radiatif.* Thèse de doctorat, Université Paris VII, 1991.

[2] R.E. Davis and D.R. Walker. An upper-air synoptic climatology of the western united states. *American Meteorological Society*, 5:1449–1467, 1992.

[3] Edwin Diday. *Classification automatique des données.* Dunod informatique, Paris, 1989.

[4] Edwin Diday. A generalisation of the mixture decomposition problem in the symbolic data analysis framework. *Cahiers du CEREMADE*, (0112), 2001.

[5] Younès Hillali. *Analyse et modélisation des données probabilistes: capacités et lois multidimensionnelles.* Thèse de doctorat, Université de Paris IX Dauphine, 1998.

[6] Laurence S. Kalstein, J. Scott Greene, Michael C. Nichols, and C. David Barthel. A new spatial synotic climatological procedure. In *AMS EIght Conference on Applied Climatology*, pages 169–173, Anaheim, California, 17-22 January 1993.

[7] Roger B. Nelsen. *An introduction to Copulas*. Springer Verlag, Lectures Notes in Statistics, 1998.

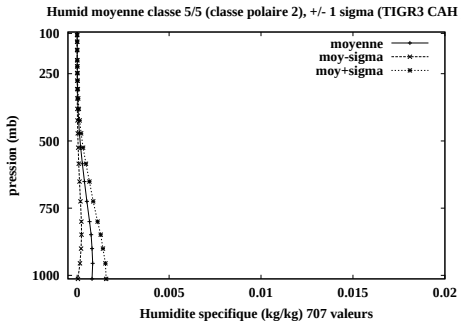
[8] B. Schweizer and A. Sklar. *Probabilistic Metric Spaces*. Elsevier North-Holland, New-York, 1983.

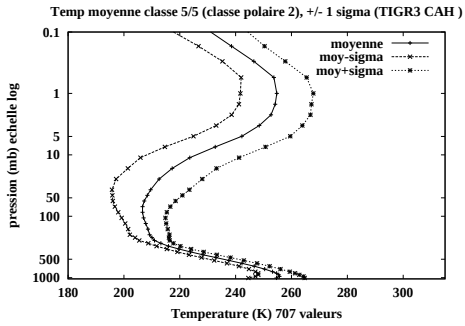
[9] B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.

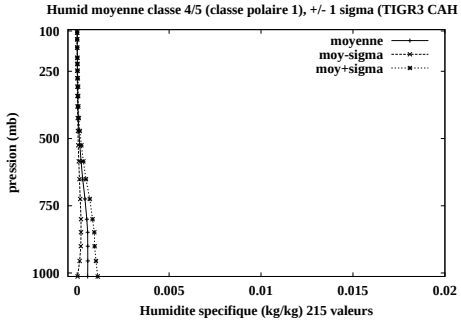
[10] Mathieu Vrac, Edwin Diday, Alain Chédin, and Philippe Naveau. Mélange de distributions de distributions. In *SFC'2001 8^{mes} Rencontres de la Société Francophone de Classification*, Université des Antilles et de Guyane, Guadeloupe, 17-21 décembre 2001.

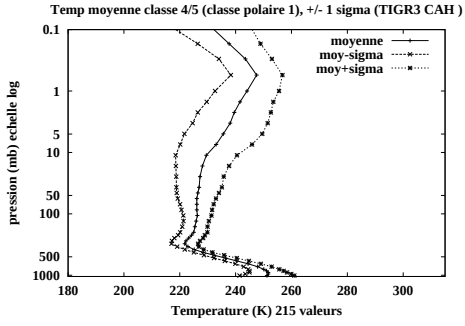
Annexe C

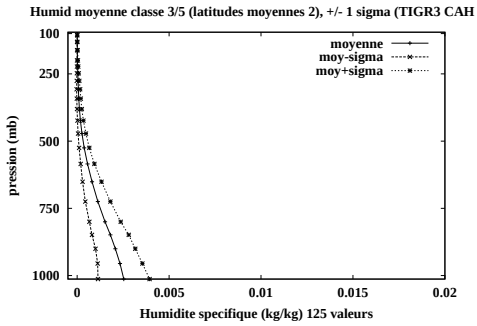
Cartes et graphiques

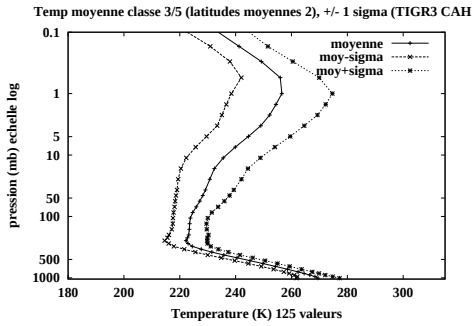


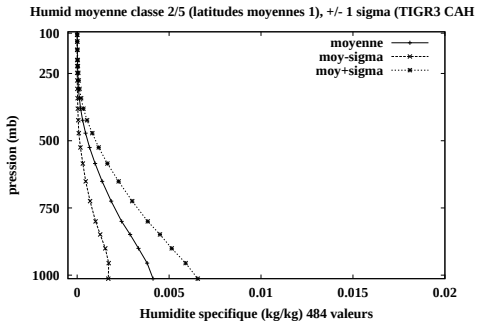


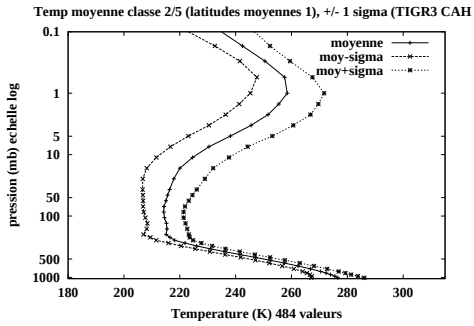


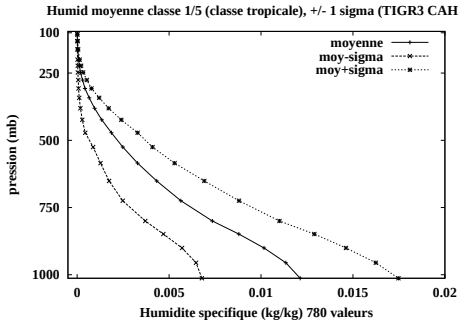












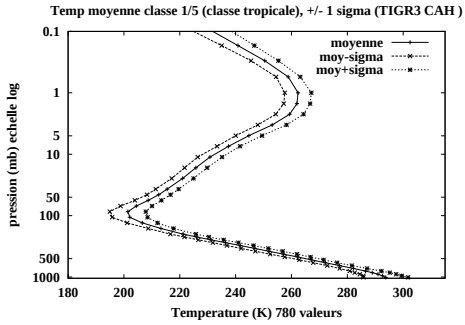
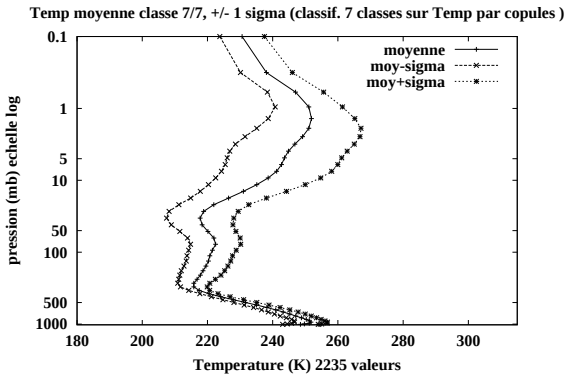
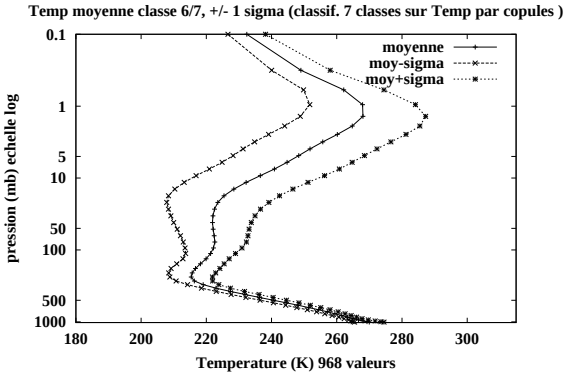
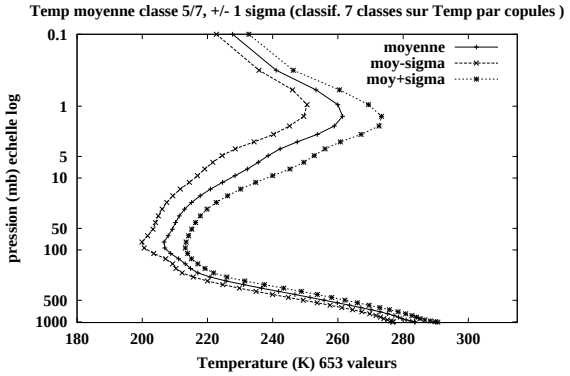
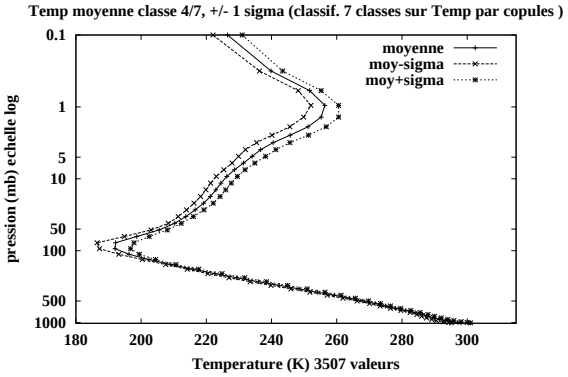


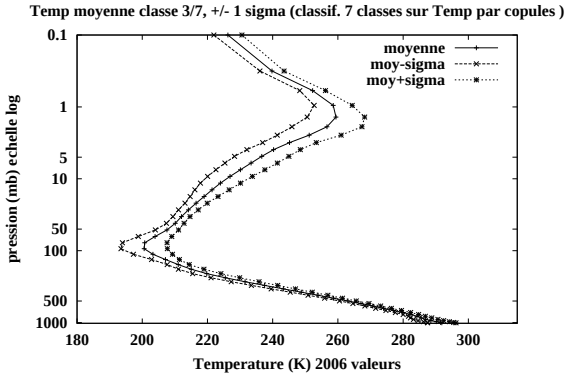
FIG. C.1: *Profils moyens de TIGR3 ± un écart-type (Température et Humidité)*

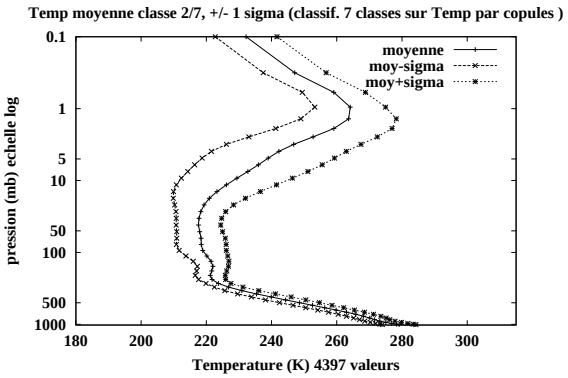












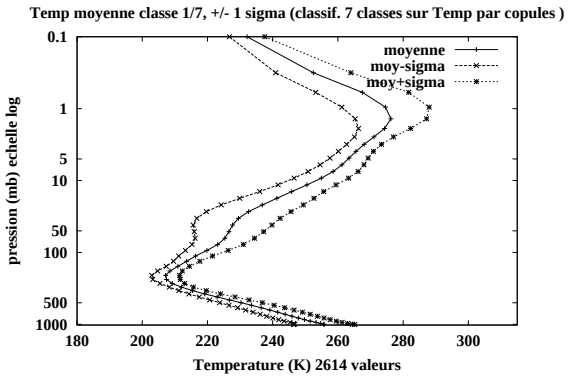


FIG. C.2: *profils de température pour la classification en température par copules*

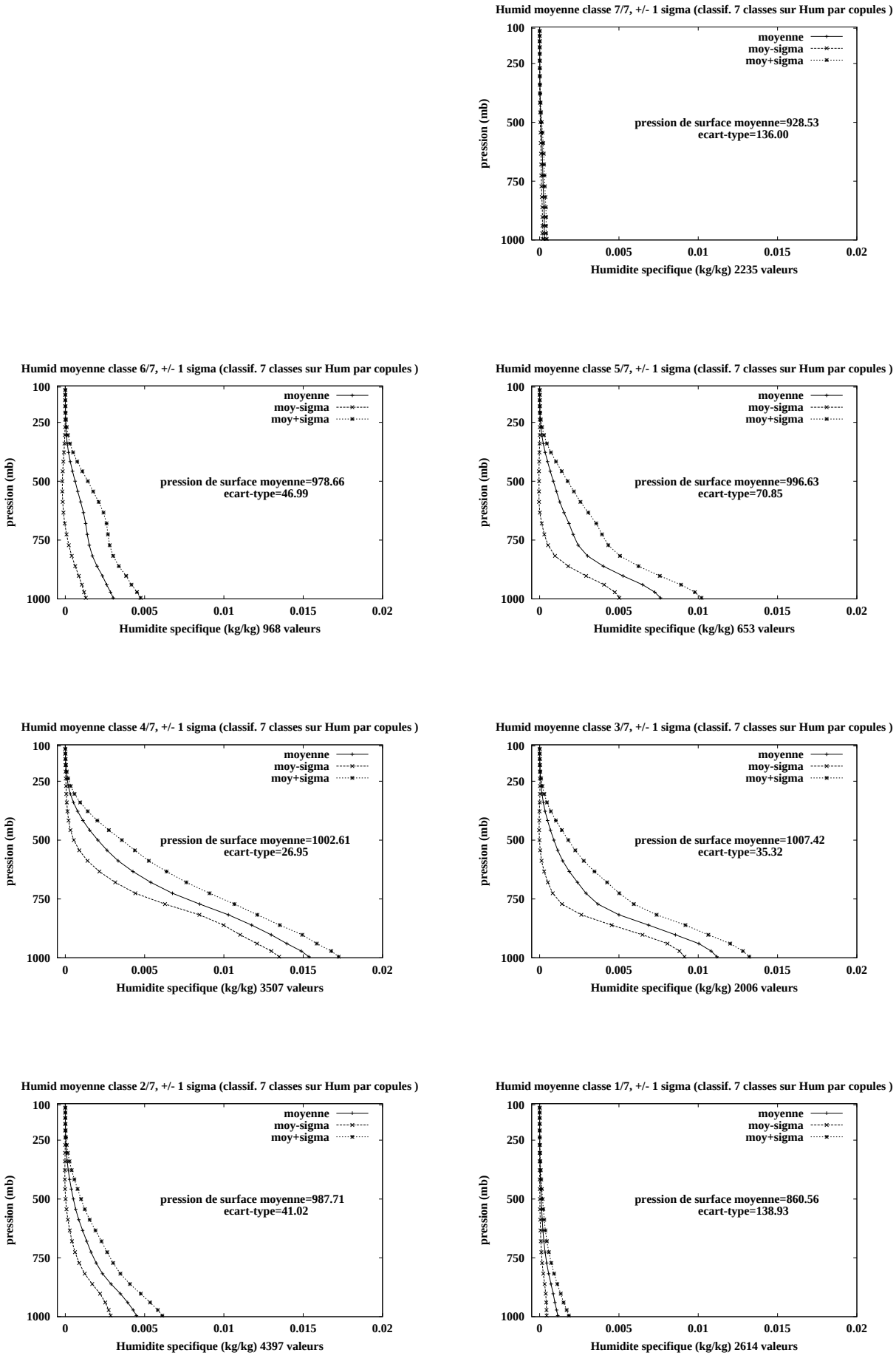


FIG. C.3: *profils d'humidité pour la classification en humidité par copules*

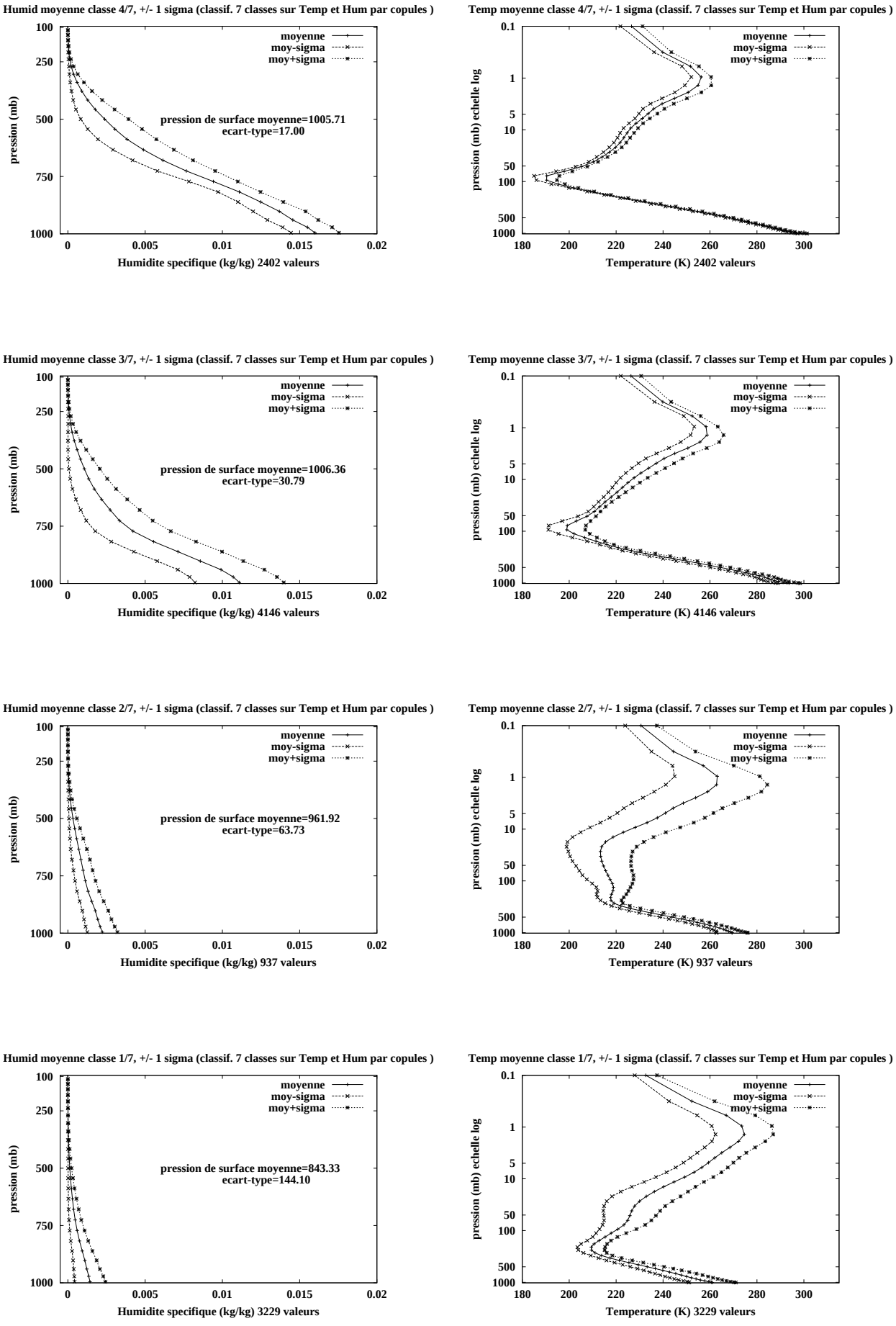


FIG. C.4: *profils de température et d'humidité des classes 1 à 4 pour le couplage en 7 classes par DMC*

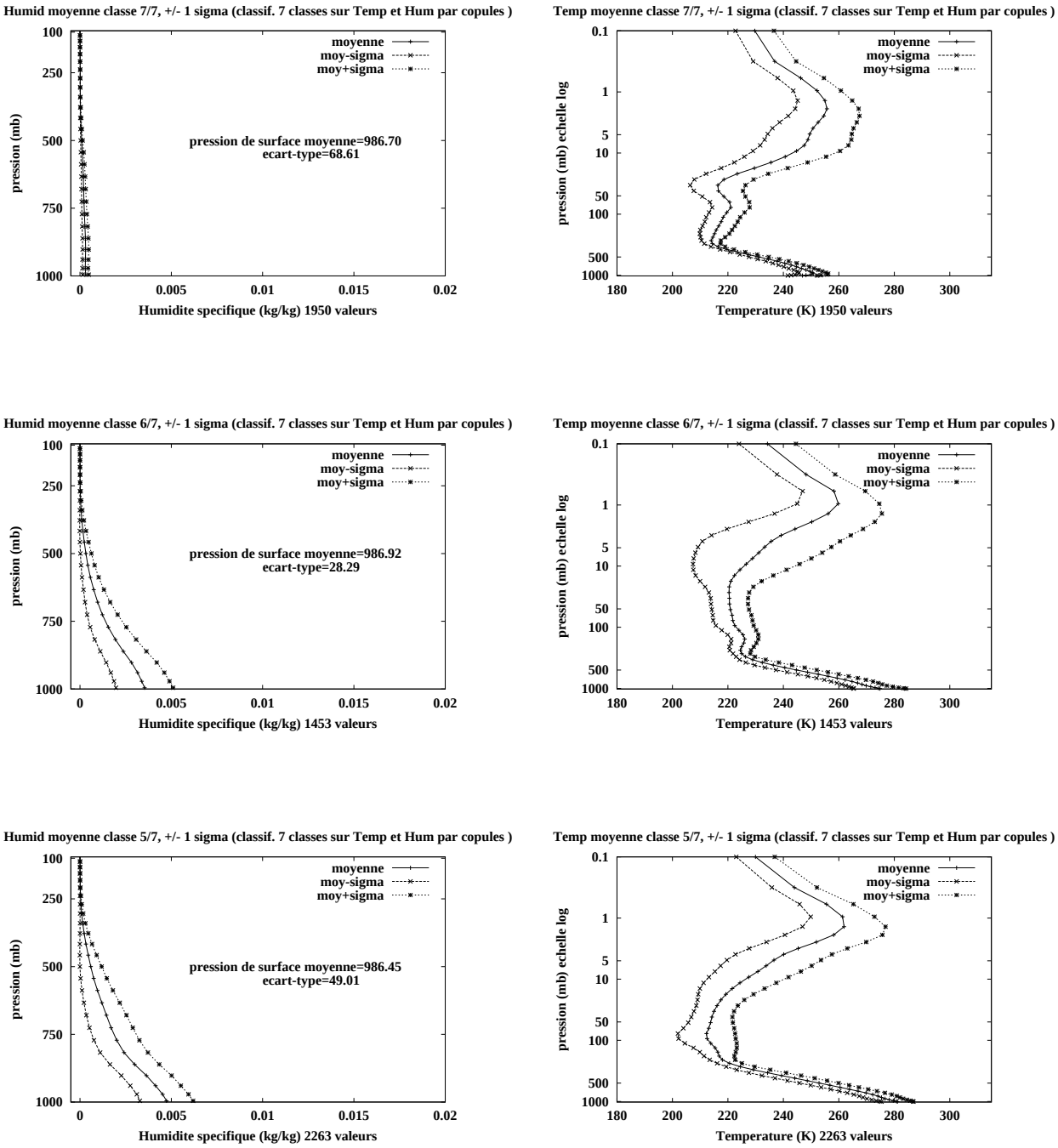


FIG. C.5: *profils de température et d'humidité des classes 5 à 7 pour le couplage en 7 classes par DMC*

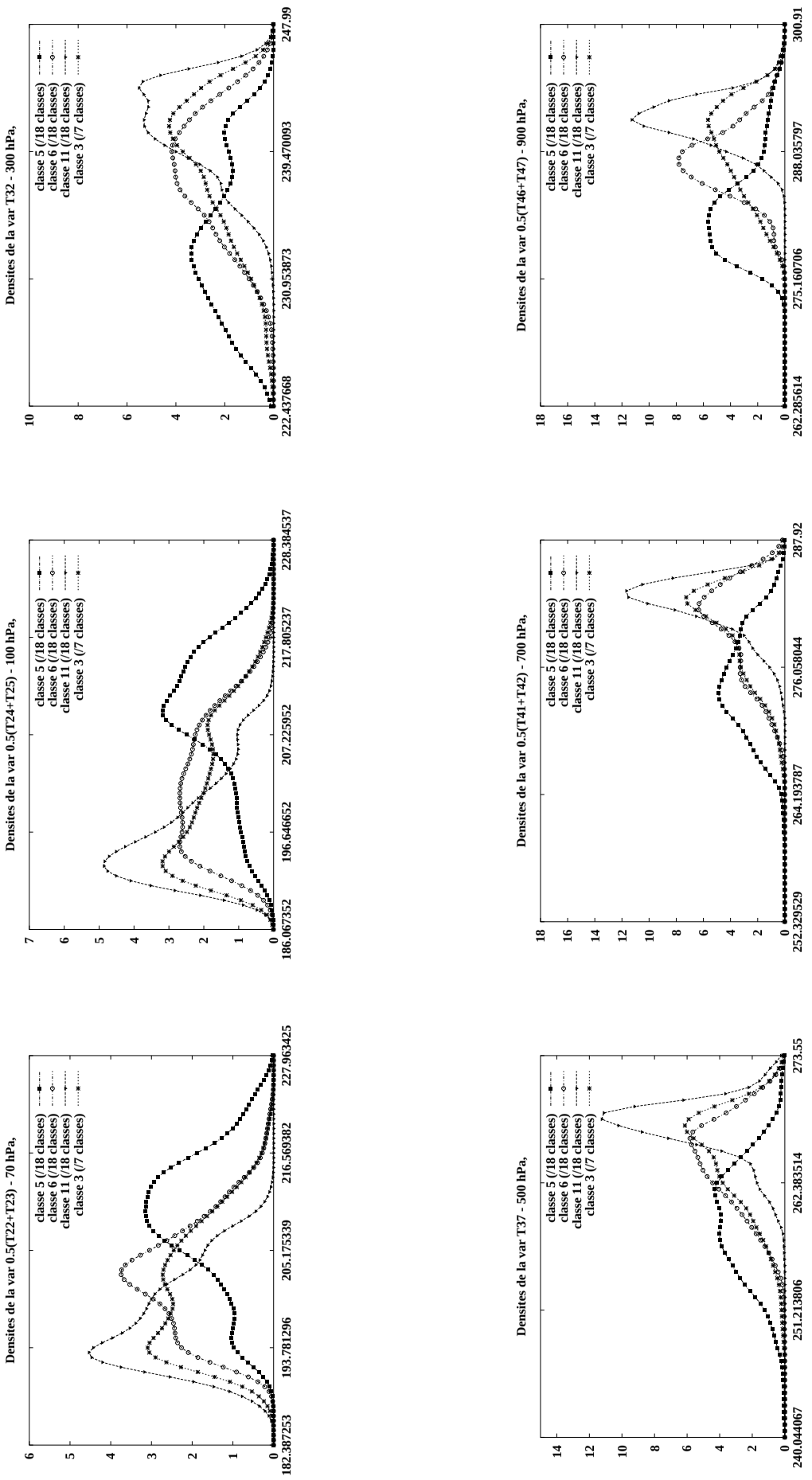


FIG. C.6: *Comparaison des densités en température - classe 3 (/7 classes) / classes 5, 6 et 11 (/18 classes)*

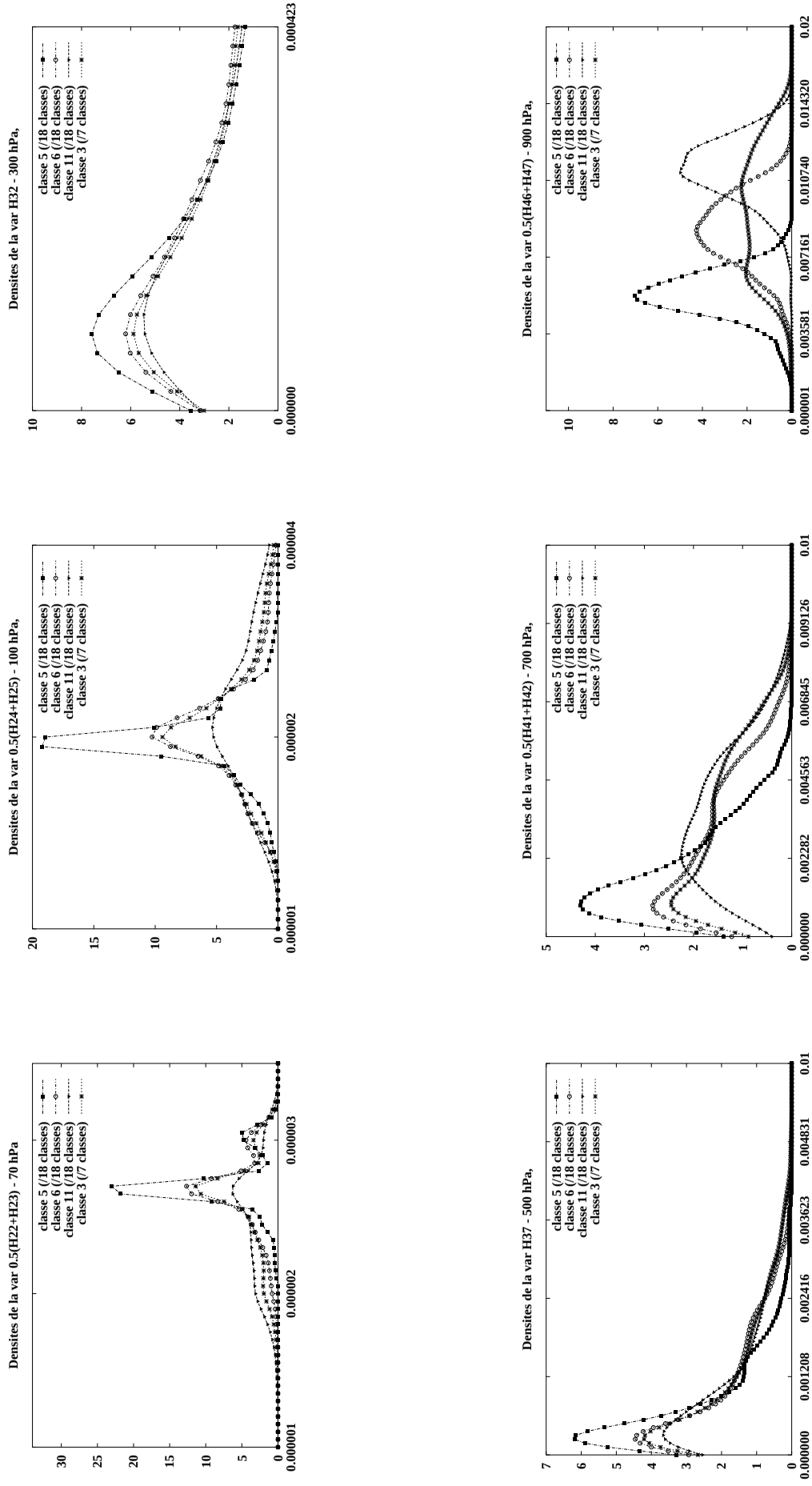


FIG. C.7: *Comparaison des densités en humidité - classe 3 (/7 classes) / classes 5, 6 et 11 (/18 classes)*

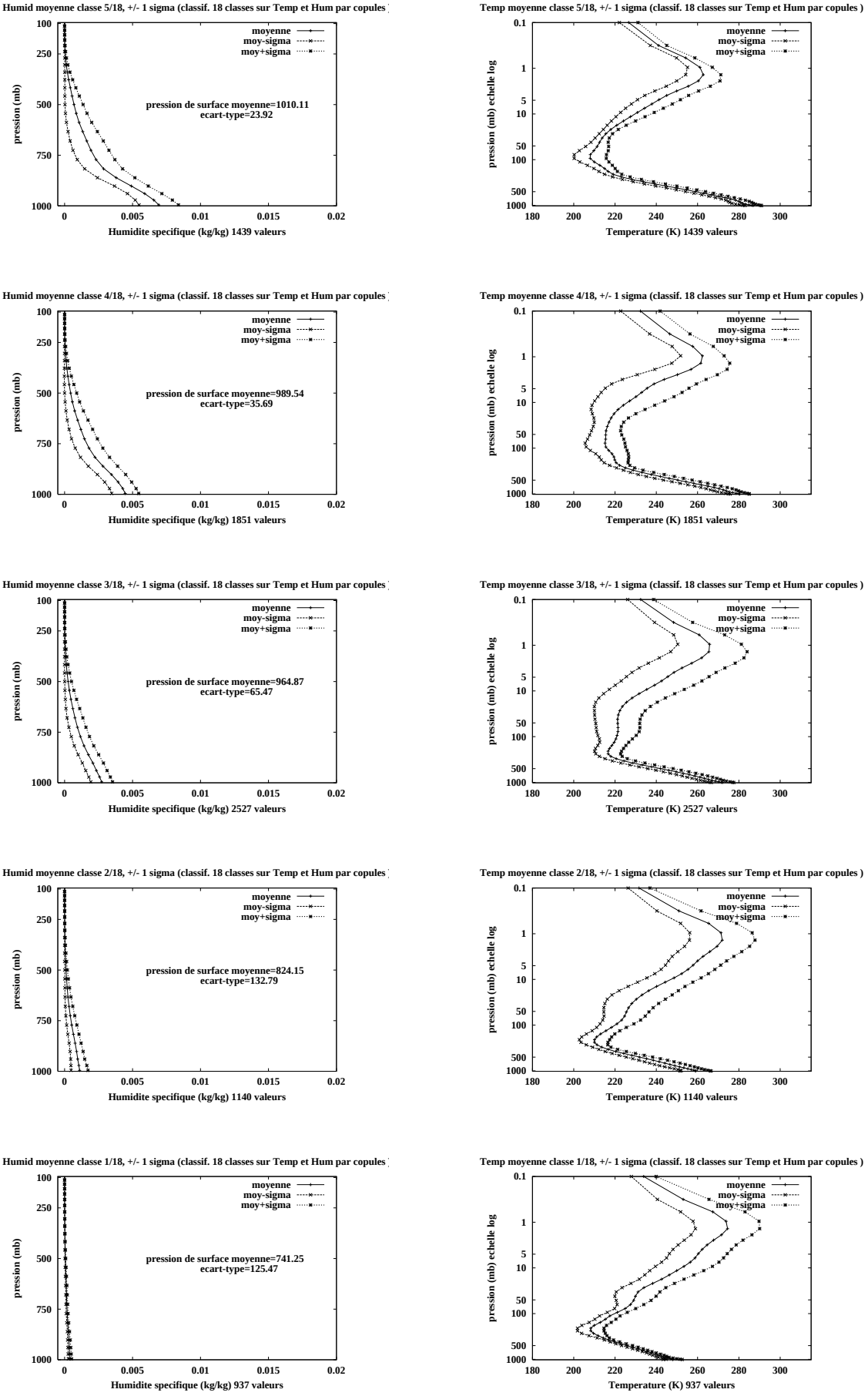


FIG. C.8: *profils de température et d'humidité des classes 1 à 5 pour 18 classes DMC*

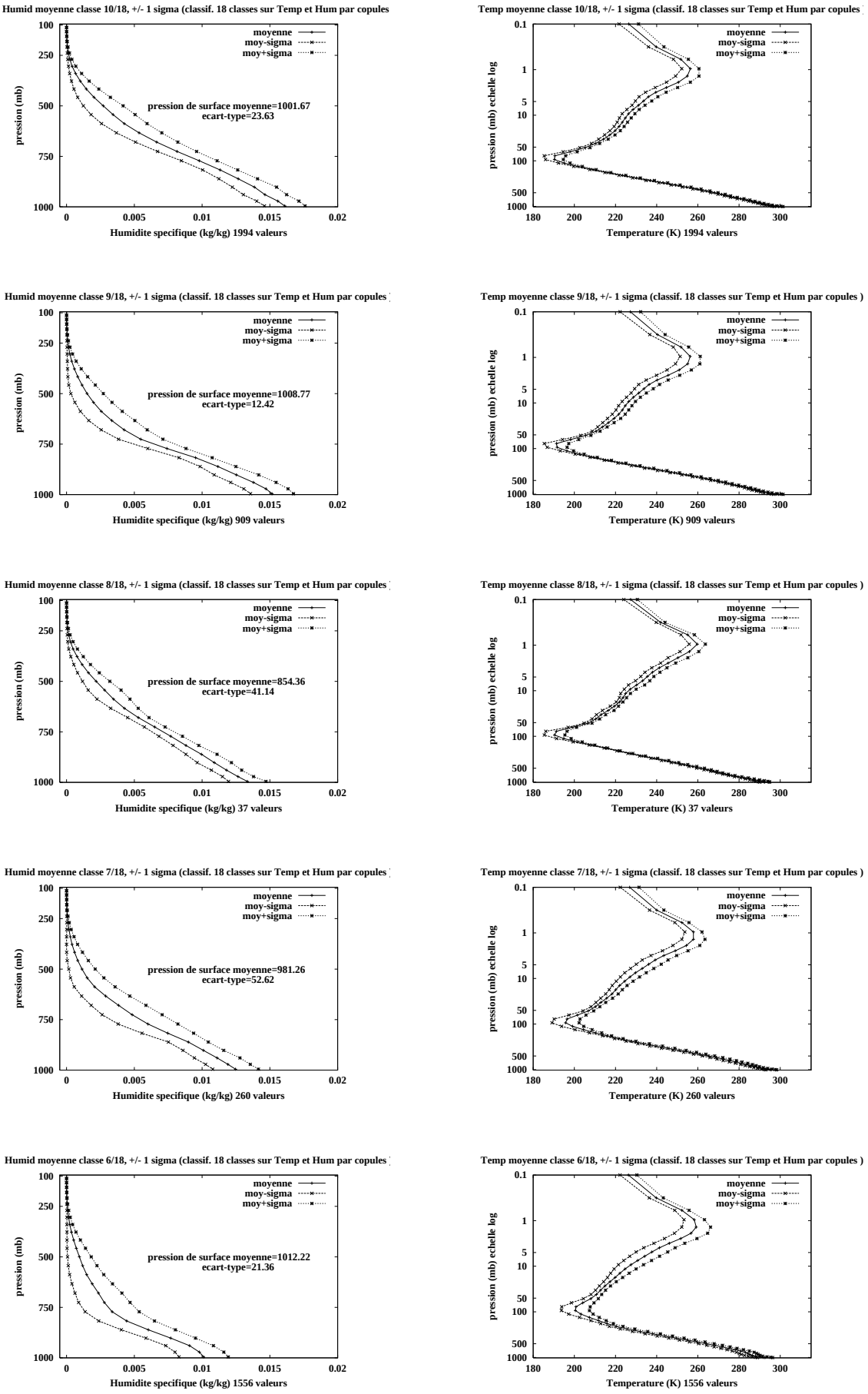


FIG. C.9: *profils de température et d'humidité des classes 6 à 10 pour 18 classes DMC*

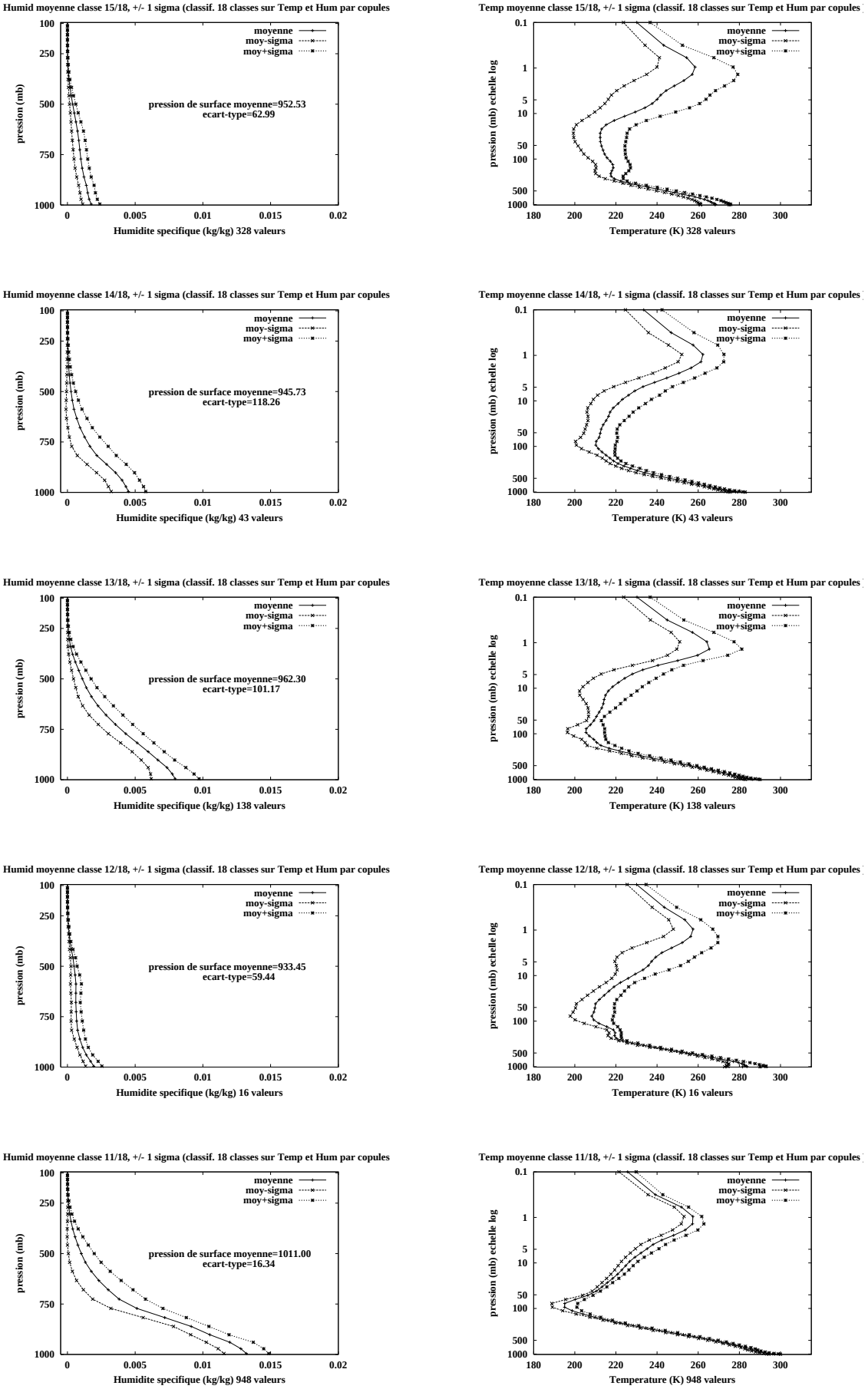


FIG. C.10: *profils de température et d'humidité des classes 11 à 15 pour 18 classes DMC*

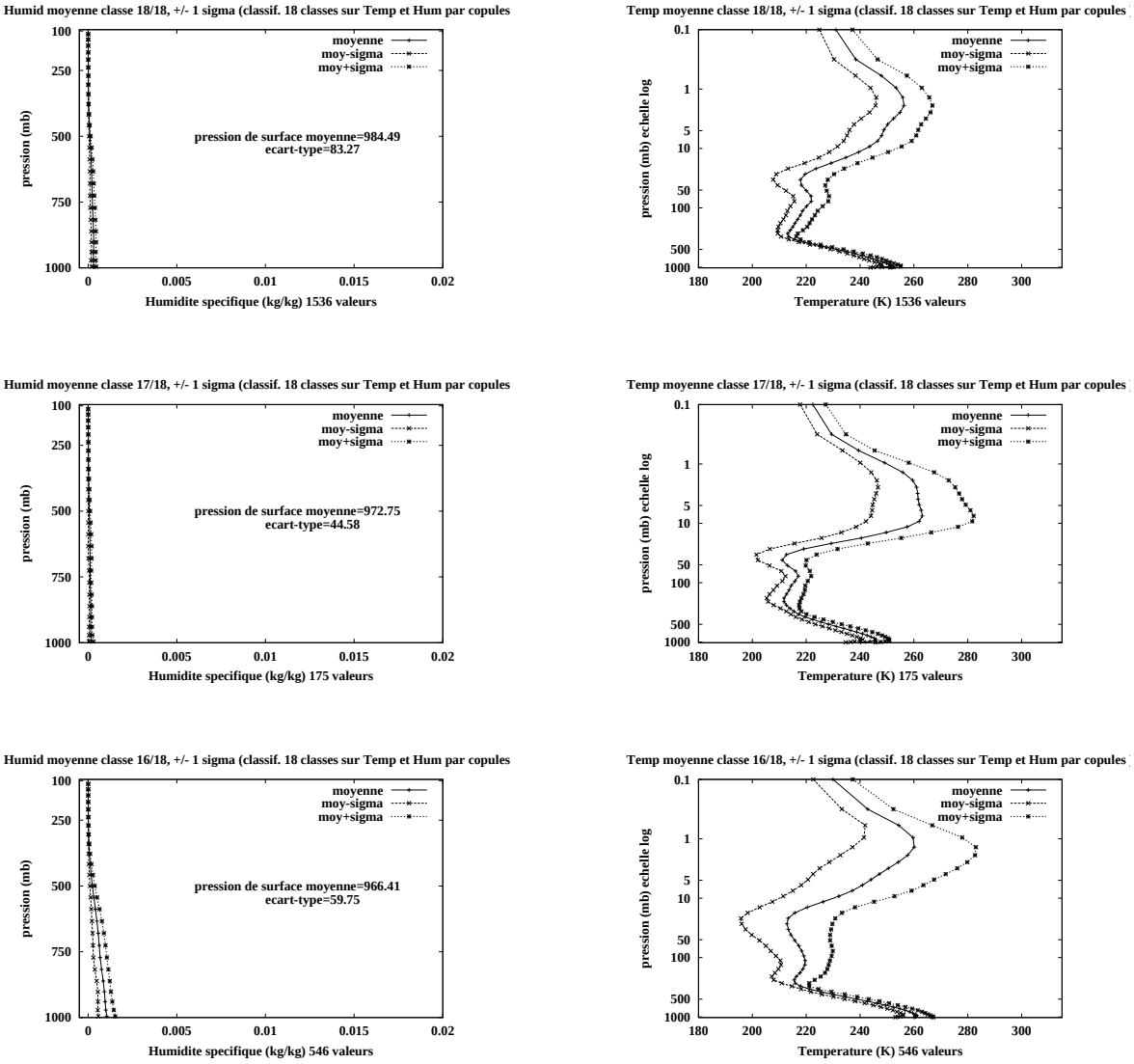


FIG. C.11: *profils de température et d'humidité des classes 16 à 18 pour 18 classes DMC*

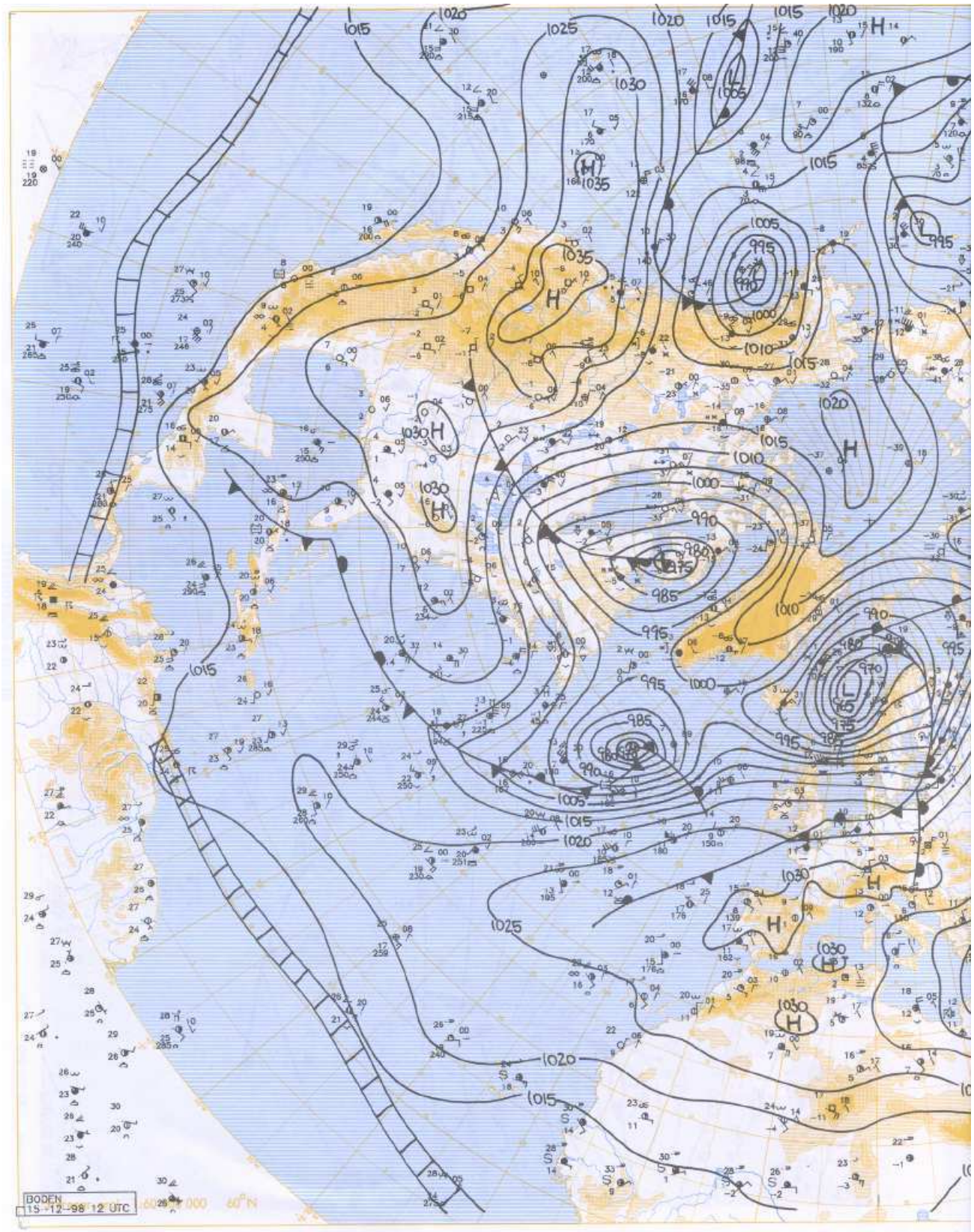


FIG. C.12: *Carte synoptique (partie 1) du 15/12/98 à 12H*

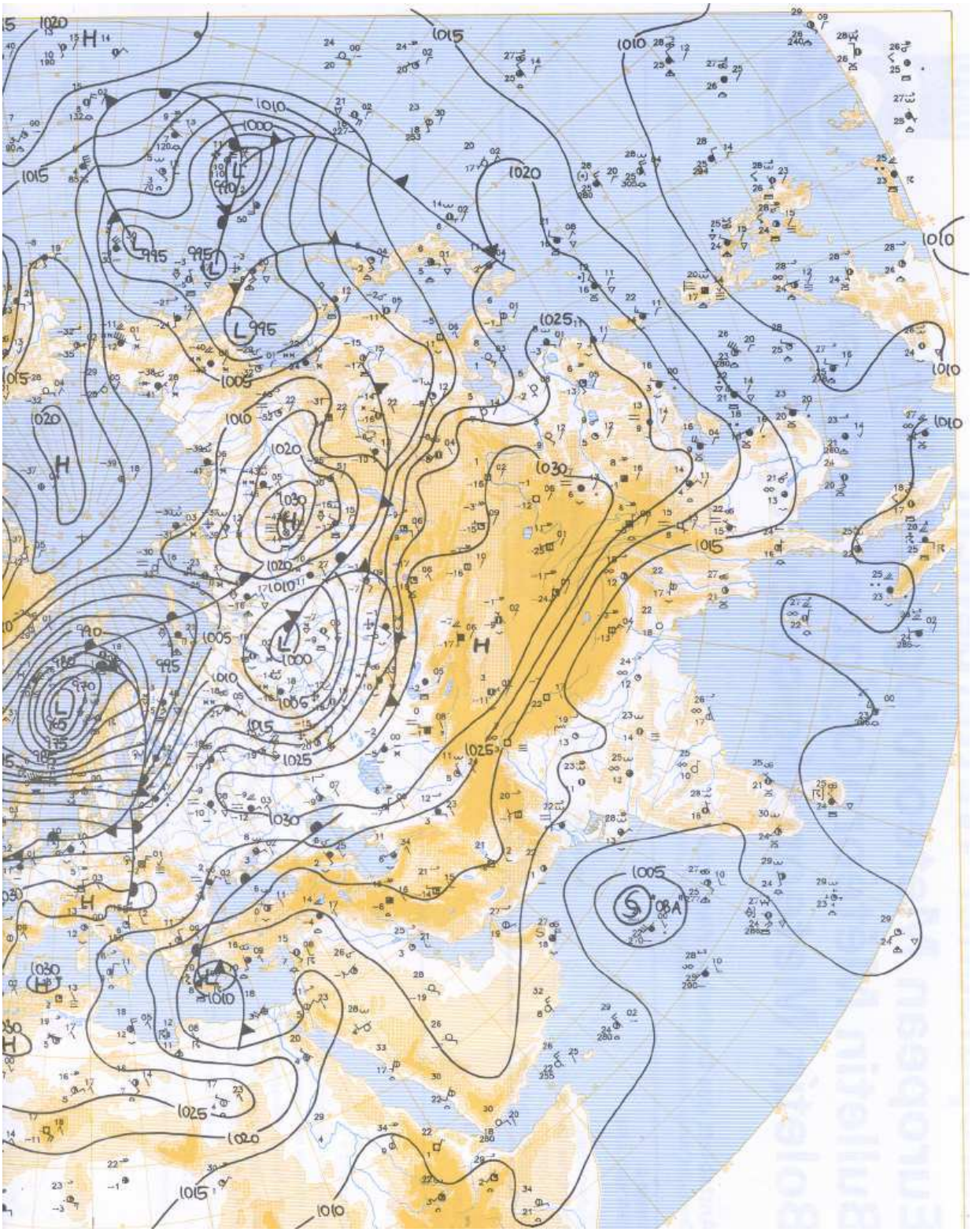


FIG. C.13: *Carte synoptique (partie 2) du 15/12/98 à 12H*

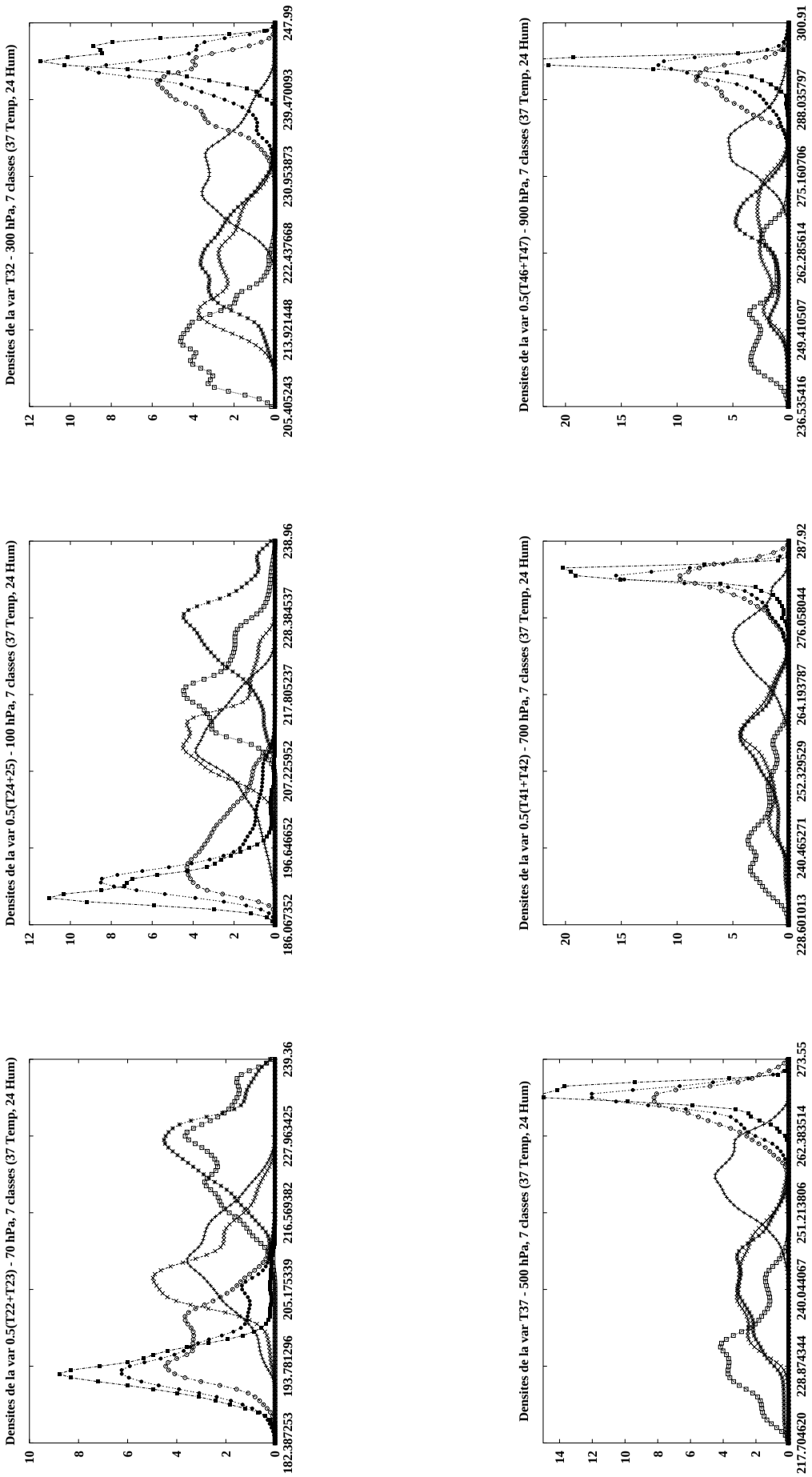


FIG. C.14: *densités de température pour la classification avec 37T, 24H*

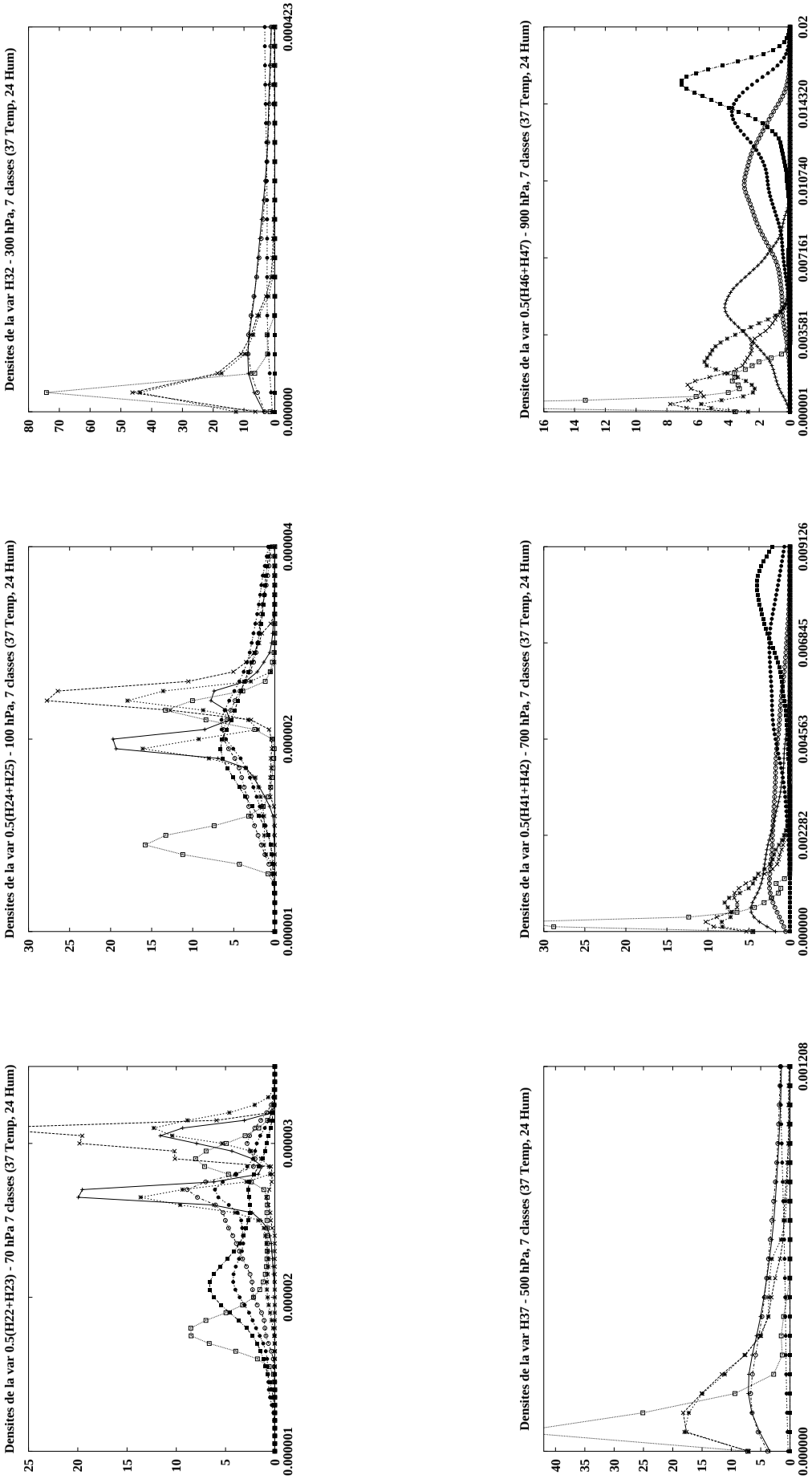


FIG. C.15: *densités d'humidité pour la classification avec 37T, 24H*

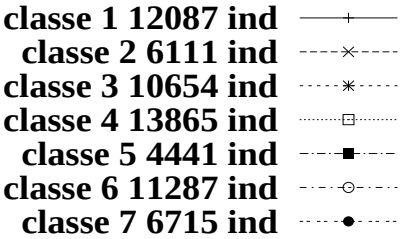


FIG. C.16: *Association classes - densités (37T,24H)*

Table des figures

1.1 Schéma du cadre de l'analyse symbolique 26
1.2 Tableau de données fonctions de répartition 28

2.1 Mélange de deux densités gaussiennes 33

3.1 Fonction de distribution de distributions pour 2 valeurs de T 49
3.2 Fonction empirique de distribution de distributions 50
3.3 Exemple d'histogramme (h=5 K) 55
3.4 Exemple d'un estimateur naif 56
3.5 Exemple d'un estimateur par noyaux (h=2.85) 57
3.6 Exemple d'un estimateur par noyaux (h=0.25) 57
3.7 Exemple d'un estimateur du plus proche voisin (k=100) 59
3.8 Exemple d'un estimateur du plus proche voisin (k=900) 59
3.9 Exemple d'un estimateur par noyaux variables 61
3.10 Exemple de surface de distributions de distributions 63
3.11 Exemple de surface de densités de distributions, associée à la surface de distribution de distributions 64

4.1 Graphes de copules 72

5.1 Tableau de données fonctions de distribution 98
5.2 Position des points 103
5.3 Exemple de surface de densités de distributions d'humidité 105
5.4 Arbre binaire 116

6.1 Exemple de profil de température 130
6.2 Classification en 7 classes par DMC sur la température ($T_1 = 225\ K$, $T_2 = 265\ K$) pour le 15 décembre 1998 à 0H 133
6.3 Température moyenne entre 500 et 7OO hPa du 15/12/98 à 0H 134
6.4 Association 7 classes - densités température 134
6.5 Densités de la température (Kelvin) par classe à 900 hPa 135
6.6 Densités de la température (Kelvin) par classe à 500 hPa 135
6.7 Densités de la température (Kelvin) par classe à 70 hPa 136
6.8 Classification en 7 classes par DMC sur l'humidité ($H_1 = 0.00003\ kg/kg$, $H_2 = 0.006\ kg/kg$) pour le 15 décembre 1998 à 0H 137
6.9 Total Column Water Vapor (Kg/m^2) 15/15/98 à 0H 138

6.10 Association classes - densités humidité 139
6.11 Densités de l'humidité spécifique (kg/kg) par classe à 900 hPa 139
6.12 Classification en 7 classes par couplage de DMC sur la température et l'humidité pour le 15 décembre 1998 à 0H 140
6.13 Association classes - densités couplage (température, humidité) 141
6.14 Densités de la température (K) par classe à 900 hPa 142
6.15 Densités de la température (K) par classe à 300 hPa 142
6.16 Densités de la température (K) par classe à 70 hPa 143
6.17 Densités de l'humidité spécifique (kg/kg) par classe à 900 hPa 144
6.18 Inférence à 12H du couplage à 0H en température et humidité de DMC . . . 145
6.19 Inférence du 01/02/1999 à 12H du couplage du 15/12/1998 OH en température et humidité de DMC 146
6.20 Température moyenne entre 500 et 700 hPa du 01/02/1999 à 12H. 146
6.21 Total Column Water Vapor du 01/02/1999 à 12H 147
6.22 18 classes en temperature et humidite par DMC 148
6.23 Association 18 classes - densités 148
6.24 Densités en température à 900 hPa pour 18 classes par DMC 149
6.25 Densités en température à 300 hPa pour 18 classes par DMC 149
6.26 Densités en température à 70 hPa pour 18 classes par DMC 150
6.27 Densités en humidité à 900 hPa pour 18 classes par DMC 150
6.28 Comparaison des densités classe 3 (/7 classes) - classes 5, 6 et 11 (/18 classes) en température à 900 hPa 151
6.29 Densités de la variable vent (composante u) à 800 hPa pour le couplage en 18 classes 152
6.30 Densités de la variable vent (composante v) à 800 hPa pour le couplage en 18 classes 153
6.31 Stratégie de classification classique 154
6.32 Exemple de la méthode des "Nuées Dynamiques" 155
6.33 Hiérarchie indicée 156
6.34 Classification en 7 classes sur 37 températures et 24 humidités pour le 15 décembre 1998 à 0H 158
6.35 Exemple d'arbre binaire de décision 159
6.36 Classification en 7 classes (6T(6), 5H(5), SP(1.5) 160
6.37 Classification en 7 classes (6T(6), 5H(5), SP(3), HCC(1), MCC(1), LCC(1), U1(0.5), U2(0.5), V1(0.5), V2(0.5), VO(0.5) 162
6.38 Classification mixte en 7 classes sur valeurs de distributions de température et d'humidité 163
6.39 Classification par EM en 7 classes sur 5 températures et 5 humidités 164
6.40 Classification par EM en 7 classes sur valeurs de fonctions de répartition de température et d'humidité 165
6.41 Association classes - densités EM sur données probabilistes 166
6.42 Densités en température à 700 hPa pour 7 classes par EM sur données probabilistes 166

6.43 Densités en température à 300 hPa pour 7 classes par EM sur données probabilistes 167
6.44 Densités en humidité à 900 hPa pour 7 classes par EM sur données probabilistes 168

C.1 Profils moyens de TIGR3 ± un écart-type (Température et Humidité) 207
C.2 profils de température pour la classification en température par copules . . . 208
C.3 profils d'humidité pour la classification en humidité par copules 209
C.4 profils de température et d'humidité des classes 1 à 4 pour le couplage en 7 classes par DMC 210
C.5 profils de température et d'humidité des classes 5 à 7 pour le couplage en 7 classes par DMC 211
C.6 Comparaison des densités en température - classe 3 (/7 classes) / classes 5, 6 et 11 (/18 classes) 212
C.7 Comparaison des densités en humidité - classe 3 (/7 classes) / classes 5, 6 et 11 (/18 classes) 213
C.8 profils de température et d'humidité des classes 1 à 5 pour 18 classes DMC . 214
C.9 profils de température et d'humidité des classes 6 à 10 pour 18 classes DMC . 215
C.10 profils de température et d'humidité des classes 11 à 15 pour 18 classes DMC 216
C.11 profils de température et d'humidité des classes 16 à 18 pour 18 classes DMC 217
C.12 Carte synoptique (partie 1) du 15/12/98 à 12H 218
C.13 Carte synoptique (partie 2) du 15/12/98 à 12H 219
C.14 densités de température pour la classification avec 37T, 24H 220
C.15 densités d'humidité pour la classification avec 37T, 24H 221
C.16 Association classes - densités (37T,24H) 222

Liste des tableaux

1.1 Tableau de données classiques 20
1.2 Exemple de données pour objet symbolique booléen 23
1.3 Tableau de données symboliques 25
1.4 Généralisation de C 26

6.1 Paramètres de la classification en 7 classes en température 132
6.2 Paramètres de la classification en 7 classes en humidité 137
6.3 Paramètres de la classification en 7 classes en température et humidité 141

Résumé

Nous étendons les méthodes de décomposition de mélange de densités de probabilité au cas des données "fonctions de répartition", permettant ainsi de classifier ces fonctions et de modéliser une loi pour ces données fonctionnelles particulières. Cette loi est donnée par la notion de "fonctions de distribution de distributions" (FDD), basée sur la définition d'une fonction de répartition pour des variables aléatoires à valeurs dans un espace probabiliste. Les extensions sont effectuées en associant les FDD aux fonctions "copules" par le théorème de Sklar. Les copules "couplent" les fonctions de répartition à n dimensions (jointes) et à 1-dimension (marginales) d'un n-uplet de variables aléatoires. Nous regardons principalement une classe de copules paramétriques, les copules Archimédiennes, et proposons trois nouvelles méthodes d'estimation des paramètres dans le cas de copules multivariées : par coefficients de corrélation de Kendall, de Spearman, et par maximisation de la vraisemblance. L'association des FDD et des copules caractérise l'évolution des données fonctionnelles (i.e. la forme de ces fonctions) entre différents points à l'intérieur des classes pour chaque variable, et donne une mesure de dépendance entre les variables utilisées. Les méthodes sont tout d'abord développées pour une variable, puis divers généralisations sont proposées pour n dimensions. Certains points théoriques sont ensuite discutés, tels que la convergence de l'algorithme et le fait que la méthode par copules est une généralisation du cas classique. Une application de la méthode "approche classification" par copules est réalisée sur des données climatiques de l'atmosphère terrestre. Le but est la classification de "profils" atmosphériques et l'estimation de la loi sous-jacente des données. Les résultats sont comparés avec ceux de méthodes "classiques", prouvant ainsi les performances nettement supérieures de la méthode par décomposition de mélange de copules (DMC) et l'intérêt de l'utilisation des données probabilistes.

Abstract

We extend the mixture decomposition of densities methods to the case of data "distribution functions", allowing to classify these functions and to model a probability law for this particular functional data. This law is given by the notion of "distribution of distribution functions" (FDD in French), based on the definition of a distribution function for random variables with values in a probabilistic space. The extensions are realised in associating the FDD to "copulas" functions according to the Sklar's theorem. Copulas joint the multivariate distribution functions (joint) with the univariate distribution functions (margins) for a vector of n random variables. We essentially look at one class of parametric copulas, the Archimedian copulas and we propose three new methods for the estimation of parameters in the multivariate copulas case : with Kendall's rank correlation coefficients, Spearman's coefficient and in maximising the likelihood. The association of the copulas with the FDD, characterises the evolution of the functional data (i.e. the shape of these distribution functions) between different points of the functions inside the classes for each variable, and gives a measure of dependency between the used variables. The methods are first developed for one variable, then several generalisations are proposed for n dimensions. Some theoretical points are discussed, such as the convergence of the algorithm and the fact that the method with copulas is a generalisation of the classical case. An application of the "classification approach" method by copulas is realised on a climate database of the terrestrial atmosphere. The goal is to classify atmospheric "profiles" and to estimate the probability law of these data. The results are compared with those of "classical" methods, showing the performances of our method by mixture decomposition of copulas, and the interest of using probabilistic data.