Thèse

présentée par

Cosmin LAZAR

pour obtenir le titre de

DOCTEUR

de l' **Université de Reims Champagne Ardenne**

Spécialité : **Génie Informatique, Automatique et Traitement du Signal**

Méthodes non supervisées pour l'analyse des données multivariées

---

Date de soutenance : 20 Novembre 2008

Composition du jury :

Vincent WERTZ	Professeur à l'Université de Louvain, Belgique	Rapporteur
Christophe COLLET	Professeur à l'Université de Strasbourg	Rapporteur
Eric MOREAU	Professeur à l'Université de Toulon	Examinateur
Mircea CURILA	Maître de Conférences à l'Université d'Oradea, Roumanie	Examinateur
Danielle NUZILLARD	Professeur à l'Université de Reims	Directeur de thèse
Patrice BILLAUDEL	Professeur à l'Université de Reims	Co-directeur de thèse

Thèse préparée au sein du

IFTS - Institut de Formation Technique Supérieure de Charleville-Mézières

Table des matières

Liste des tableaux

Liste des figures

Introduction

Rien ne se fait sans un peu d'enthousiasme.

Voltaire

Dans de nombreux domaines de la science, les chercheurs traitent des ensembles de données multivariées. Leur analyse implique l'extraction de l'information pertinente pour la compréhension des phénomènes étudiés et les résultats d'analyse permettent d'établir les éventuelles règles de décision. Une étape essentielle dans le processus d'extraction de l'information est le regroupement des données ayant des caractéristiques similaires dans des classes ; cette étape est appelée *classification non supervisée* (ou automatique) ou *"clustering"*. La taille et la dimension de l'ensemble des données est un problème majeur dans le processus de regroupement des objets. Des techniques de réduction de dimension sont souvent utilisées conjointement avec celles de classification pour faciliter la découverte des structures d'intérêt dans l'ensemble des données. Même si de nombreux travaux ont été réalisés dans ce domaine, des problèmes comme le choix du nombre de classes, le choix de la mesure de similarité ainsi que la dimension intrinsèque d'un ensemble de données multivariées, sont des questions qui n'ont pas encore reçu de réponse satisfaisante.

Cette thèse apporte une contribution dans le domaine de l'analyse des données multivariées. A ce propos, deux problèmes ont été abordés : le premier qui n'a reçu que peu d'attention de la part des chercheurs, est lié au choix de la mesure de similarité pour la classification d'un ensemble de données quelconque. La mise en évidence du *phénomène de concentration* de la métrique euclidienne pose un doute concernant sa pertinence dans des problèmes de classification de données multivariées. Des travaux montrent que des métriques moins concentrées permettent d'obtenir un meilleur contraste entre des données multivariées ; ces métriques sont présentées comme des alternatives à la métrique euclidienne pour la mesure de similarité dans les algorithmes de classification. La première direction de recherche consiste donc à étudier le

phénomène de concentration des métriques et l'impact des métriques non euclidiennes sur les résultats de la classification ; la motivation est de trouver un moyen pour choisir la métrique optimale pour des problèmes de classification non supervisée de données multivariées.

La deuxième direction de recherche puise ses origines dans les problèmes qui surgissent lors de l'analyse des données dans des espaces de grand dimension. C'est le phénomène bien connu *de la malédiction de la dimension* ou *phénomène de Hugh* ou *phénomène de l'espace vide*. Pour comprendre ce phénomène nous rappelons l'exemple présenté dans [Ver03] : considérons une donnée multivariée comme un point dans un espace de dimension d. Un nombre fini n de données dans un espace à $d = 3$ dimensions peut présenter des structures ou des nuages de points permettant d'extraire des informations sur les données analysées. Par contre, le même nombre de données dans un espace à $d = 20$ dimensions ne présente pas de spécificité car les données apparaissent comme des points isolés dans l'espace. Ceci est dû au fait que le volume de l'espace d'analyse augmente de manière exponentielle avec sa dimension ; l'estimation de la fonction densité de probabilité d'une population dans des espaces multidimensionnels est donc quasi impossible à réaliser. Des méthodes de réduction de la dimension sont utilisées pour éviter ce phénomène. A ce propos, des méthodes de *séparation aveugle de sources* sont étudiées et mises en oeuvre dans des applications d'analyse d'images multivariées.

Une méthode de SAS issue de l'interprétation géométrique du modèle de mélange linéaire est également présentée. Cette méthode est applicable dans le cas des observations non négatives ; elle donne de bons résultats si la probabilité d'avoir des instants où chacune des sources est active et toutes les autres sources sont inactives est importante. Cette méthode est testée et comparée avec d'autres méthodes de SAS dans le contexte de séparation de sources mais aussi dans le contexte de la réduction de dimension des données multivariées.

Organisation du manuscrit

Ce travail est organisé en deux parties, chacune contenant 3 chapitres. La première partie présente de manière générale les sujets théoriques abordés dans la thèse et la deuxième partie présente la contribution de l'auteur.

Dans le premier chapitre, le domaine de la classification non supervisée des données multivariées est résumé ; les généralités sont présentées au début du chapitre ; ensuite les principales mesures de similarité et quelques méthodes de classification non supervisée sont rappelées, puis la présentation des indices de validité des classes termine ce chapitre.

Une des directions de recherche de cette thèse est l'emploi des *métriques non euclidiennes* pour la classification non supervisée des données multivariées ainsi que la recherche d'une méthode pour choisir la métrique optimale dans un problème de la classification d'un ensemble de données quelconque. A ce propos, le deuxième chapitre est dédié à l'étude du phénomène de concentration des métriques en présentant quelques résultats importants issus de la littérature.

Les méthodes de *séparation aveugle de sources* sont introduites afin d'être mises en oeuvre pour la réduction de dimension des données multivariées, dans le troisième chapitre. Les principes et les méthodes sont présentés de manière générale et ensuite quelques méthodes de séparation par analyse en composantes indépendantes et par la prise en compte de la non-négativité sont rappelées. Ceci termine la partie théorique concernant l'état de l'art et la présentation des problèmes traités.

A partir du quatrième chapitre, la contribution de l'auteur est abordée. Dans le quatrième chapitre, les travaux issus de la littérature présentés dans le chapitre 2 sont testés afin de valider l'hypothèse de la supériorité des métriques moins concentrées sur les métriques plus concentrées pour la classification non supervisée ; une discussion sur les résultats obtenus est aussi présentée à la fin de ce chapitre ainsi que notre proposition pour le choix de la métrique optimale pour la classification non supervisée.

Une méthode de SAS basée sur une interprétation géométrique du modèle de mélange linéaire est dévelopée et comparée avec d'autres méthodes de SAS dans le cinquième chapitre. Une évaluation de cette méthode ainsi que d'autres méthodes de SAS dans le contexte de la réduction de dimension de données multivariées pour la classification est réalisée et les résultats sont présentés.

Enfin, dans le dernier chapitre, les résultats obtenus précédemment ont été exploités dans le contexte de l'analyse des images multivariées et deux applications sont présentées. La première a pour but de réaliser la segmentation d'une image multivariée de microscopie. Dans une première tentative des métriques non euclidiennes sont employées et l'étape de réduction de la dimension de données a été évitée ; dans la deuxième tentative, des méthodes de réduction de la dimension ont été employées et plusieurs algorithmes de classification ont pu être mis en oeuvre. Dans la deuxième application, des séries temporelles d'images médicales ont été analysées et les méthodes de SAS prenant en compte les contraintes application ont été mises en oeuvre conjointement avec la méthode *C-moyennes*.

Partie I

Etude bibliographique

Chapitre 1

Classification non supervisée ; principes et méthodes

Notations

Notation	:	Signification
X	:	Ensemble de données multivariées ; matrice de dimension $n \times d$
n	:	Nombre de données multivariées
d	:	Dimension de données (le nombre d'attributs d'un ensemble de données)
X_{i*}, $i = 1 : d$	:	Vecteur de données
X_{*j}, $j = 1 : n$	:	Attributs d'un ensemble de données
x_i	:	Scalaire représentant une mesure d'un vecteur de données
m_j	:	Moyenne d'un attribut X_{*j}
s_j	:	Variance d'un attribut X_{*j}
D_r	:	Famille de métriques r
r	:	Exposant de la métrique
Σ	:	Matrice de covariance
α	:	Angle spectral entre deux vecteurs
R	:	Coefficient de corrélation de Pearson
C	:	Nombre de classes
G	:	Matrice des centres de classe
g_k	:	Vecteur : le centre de la classe k
u_{ij}	:	Degré d'appartenance de donnée X_{i*} à la classe j

N_j	:	Nombre de données dans la classe j
q	:	Coefficient de fuzzyfication (pour la méthode *C-moyennes floue*)
m	:	Dimension intrinsèque des données (nombre réduit d'attributs)
fdp	:	Fonction densité de probabilité
$\hat{fdp}$	:	Fonction densité de probabilité estimée
K	:	Fonction noyau utilisée pour l'estimation de la *fdp*
h	:	Largeur de la fonction noyau
$EQMI$	:	Erreur Quadratique Moyenne Intégrée
$\nabla()$	:	Fonction gradient
K_E	:	Noyau d'Epanechnikov
c_d	:	Volume de l'hypersphère unite de dimension d
S_h	:	Volume de l'hypersphère de rayon h de dimension d
M_h	:	Vecteur Mean Shift
$\hat{m}$	:	Mode de la fdp
w, ρ	:	Pour la méthode SVC : les paramètres de l'hyperplan de séparation
α_i	:	Coefficients de Lagrange
DB	:	Indice de Davies-Bouldin
C_i	:	*i-ème* classe
σ_i	:	Distance moyenne entre les objets de la classe C_i et son centre g_i
$d(g_i, g_j)$	:	Distance entre deux centres de classes
D	:	Indice Dunn
d_{min}	:	Distance minimale entre deux objets de classes différentes
d_{max}	:	Distance maximale entre deux objets de classes différentes
C_0	:	Indice C_0
l	:	Nombre des pairs d'objets dans une classe
S_{min}	:	Somme des l plus petites distances entre des paires d'objets si toutes les paires d'objets sont considérées
S_{max}	:	Somme des l plus grandes distances entre des paires d'objets si toutes les paires d'objets sont considérées
CS	:	Indice compacité-séparabilité
c_o	:	Compacité d'une partition
s_e	:	Séparabilité d'une partition

1.1 Introduction

La classification non supervisée ou *clustering* est une technique importante dans le domaine de l'analyse de données. Appliquée dans de nombreux domaines scientifiques tels que l'imagerie, la biologie, le marketing etc., elle inclut des algorithmes et des méthodes pour regrouper ou classifier des objets, selon un critère de similarité. Les objets peuvent être représentés soit par des vecteurs de mesures soit par des points dans un espace multidimensionnel. Dès le départ il est nécessaire de différencier la *classification non supervisée* et la *classification supervisée* ou *analyse discriminante*. La classification supervisée consiste à construire des règles de décision en se basant sur un ensemble de données pour lesquelles les étiquettes des classes sont connues *a priori*. Le but de la classification non supervisée est de trouver une organisation des données cohérente et valide, qui puisse mettre en évidence les vraies structures dans un ensemble de données sans aucune connaissance *a priori* sur les données traitées, [JMF99].

Dans beaucoup d'applications on ne dispose pas des connaissances *a priori* sur les données et donc, les méthodes de classification doivent faire le moins de suppositions. C'est grâce à ces restrictions que les méthodes de classification non supervisée sont particulièrement appropriées pour explorer les relations entre les données et pour offrir une vision cohérente sur leur vraie structure.

L'objectif de ce chapitre est de donner une présentation globale du domaine de la classification non supervisée afin de permettre une meilleure compréhension de ce travail. Les définitions et notations nécessaires sont d'abord exposées, ensuite quelques techniques de prétraitement de données et les problèmes majeurs liés à la classification des données multivariées sont présentés. Les mesures de similarité représentent le sujet de la troisième section. Lié à ce sujet, un point important développé dans le deuxième chapitre est constitué par l'étude du phénomène de concentration des métriques r, dont les métriques de Minkowski font partie. La quatrième section est une synthèse des méthodes de classification non supervisée : l'algorithme *C-moyennes* et son extension floue, l'algorithme *ISODATA* ainsi que les algorithmes basés sur la fonction densité de probabilité *fdp* des points, notamment *Mean-Shif*, *Parzen-Watershed* ainsi que la méthode *Suport Vector Clustering* (SVC) sont rappelés. La cinquième section est dédiée à la présentation de quelques indices pour la validation des résultats de la classification et une conclusion terminera ce chapitre.

1.2 Généralités

1.2.1 Définition et notations

Qu'est-ce qu'un *cluster*?

Dans la littérature il n'existe pas de réponse précise à cette question. Les classes ou *clusters* regroupent des objets similaires, mais la notion de similarité elle-même est ambigüe car elle peut varier d'une application à une autre. Les données peuvent présenter des structures de formes et/ou tailles différentes et donc, celles-ci peuvent être regroupées sous des hypothèses très différentes. On retient trois définitions présentées dans [Eve74] :

- *Un cluster* est un ensemble d'entités qui sont semblables, et les entités des différents clusters ne sont pas semblables.
- *Un cluster* est une agrégation des points dans l'espace d'essai tel que la distance entre deux points quelconques dans un cluster est inférieure à la distance entre n'importe quel point de ce cluster et n'importe quel point qui ne se trouve pas dans ce cluster.
- *Les clusters* peuvent être décrits en tant que régions dans un espace multidimensionnel caractérisées par une haute densité de points, séparées d'autres telles régions par une région caractérisée par une densité de points relativement faible.

Dans les deux dernières définitions les objets sont considérés comme des points dans un espace multidimensionnel.

Un cluster peut être vu comme une *source* d'objets dont la distribution dans l'espace des attributs est décrite par une densité de probabilité spécifique à cette classe.

Dans ce chapitre les termes et les notations suivantes sont utilisées [JMF99] :

- **Un objet** (ou observation, ou vecteur de données, ou point) est une donnée élémentaire utilisée dans les algorithmes. En général, une donnée élémentaire est représentée par un vecteur de d mesures :
$$X_{i*} = (x_1, \ldots, x_d) \tag{1.1}$$
- **Une mesure** d'un objet est un scalaire x_i contenu dans un vecteur de donnée X_i ;
- **Un attribut** ou *feature* est un vecteur de mesures décrivant une caractéristique générale de tous les objets ;

- **La dimension** des données d est le nombre d'attributs ou de mesures d'un vecteur de données ;
- **Un ensemble de données** est une matrice X qui contient un nombre n de données élémentaires ou objets. Souvent, il est représenté par une matrice $n \times d$;
- **Une mesure de similarité** est une fonction mathématique qui nous permet de regrouper les objets en classes. La plupart des méthodes utilisent comme mesure de similarité une des distances de Minkowski. Les mesures de similarité sont présentées dans la suite de ce chapitre ;
- **Une classe** ou *cluster* est un groupe d'objets semblables entre eux mais dissemblables par rapport aux objets se trouvant dans d'autres groupes ;
- **La classification dure** ou *hard clustering* est une technique qui attribue à chaque objet une étiquette ;
- **La classification floue** ou *fuzzy clustering* attribue à chaque objet un degré d'appartenance à chaque classe.

1.2.2 Prétraitement des données

Normalisation

Les données brutes sont rarement utilisées dans l'analyse. La préparation des données pour la classification non supervisée requiert leur normalisation ; celle-ci est effectuée en fonction de la mesure de similarité adoptée. Certaines mesures de similarité, comme par exemple la distance Euclidienne, favorisent implicitement les attributs d'un ordre d'échelle plus significatif et la contribution des attributs moins significatifs est ignorée. Quelques règles de normalisation sont présentées dans la suite.

On considère une matrice de données X^* de taille $n \times d$ qui contient n objets décrits par d attributs. Un objet est représenté par un vecteur X^*_{i*} de taille $1 \times d$ et le scalaire x^*_{ij} est une mesure d'un attribut de l'objet X^*_{i*}. Le symbole astérisque dénote les données brutes. La matrice X^* est une matrice de la forme :

$$X^* = [X^*_1 X^*_2 \dots X^*_n]^T = \begin{bmatrix} x^*_{11} & x^*_{12} & \dots & x^*_{1d} \\ x^*_{21} & x^*_{22} & \dots & x^*_{2d} \\ \vdots & & & \\ x^*_{n1} & x^*_{n2} & \dots & x^*_{nd} \end{bmatrix} \tag{1.2}$$

Chaque ligne de la matrice X^* est un objet ou une donnée et chaque point dans l'espace des données est un objet potentiel. Les données sont représentées et visualisées comme des points dans un espace multidimensionnel. Pour chaque attribut de la matrice X^*, la moyenne m_j et la variance s_j^2 s'expriment par les relations :

$$m_j = (1/n) \sum_{i=1}^{n} x_{ij}^* \tag{1.3}$$

$$s_j^2 = (1/n) \sum_{i=1}^{n} (x_{ij}^* - m_j)^2 \tag{1.4}$$

- **Normalisation par centrage** - le plus simple moyen de normaliser les données est d'extraire de chaque attribut sa moyenne, [JD88] :

$$x_{ij} = x_{ij}^* - m_j \tag{1.5}$$

  Ce type de normalisation correspond à une translation et rend les données invariantes aux déplacements des axes.

- **Normalisation par mise à l'échelle** - une deuxième méthode de normalisation consiste à redimensionner les données dans l'intervalle [0 1], par une des 2 transformations linéaires suivantes, [DAD04] :

$$X_{*j} = \frac{(X_{*j}^* - X_{*j\,min}^*)}{X_{*j\,max}^* - X_{*j\,min}^*} \tag{1.6}$$

  ou

$$X_{*j} = \frac{X_{*j}^*}{X_{*j\,max}^* - X_{*j\,min}^*} \tag{1.7}$$

  Dans [MC85] il est montré que la normalisation des données par une de ces transformations améliore le taux de classification. Dans les formules précédentes, X_{*j} et X_{*j}^* représentent les vecteurs d'attributs de l'ensemble X respectivement X^* ; $X_{*j\,max}^*$ et $X_{*j\,min}^*$ représentent les valeurs maximale et minimale observées de l'attribut X_{*j}^*.

- **Normalisation par centrage et mise à l'échelle** - le troisième type de normalisation revient à translater les données et à les redimensionner de telle sorte qu'elles soient de moyenne nulle et de variance unitaire, [JD88]. La loi de normalisation est donnée par :

$$x_{ij} = \frac{x_{ij}^* - m_j}{s_j} \tag{1.8}$$

La réduction de la dimension de données

Un prétraitement souvent utilisé avant de poursuivre la classification des données est la réduction du nombre d'attributs. Il y a plusieurs raisons d'inclure cette étape dans le processus de classification : premièrement, pour éviter le phénomène de *malédiction de la dimension*, [Bel61] connu aussi sous le nom de *phénomène de l'espace vide* ou *phénomène de Hughes* [ST83] ; deuxièmement, pour améliorer les résultats de la classification et/ou réduire le temps de calcul et finalement pour obtenir une représentation visuelle des données qui puisse servir à une meilleure inspection et ainsi à valider les résultats de la classification. Néanmoins les méthodes de réduction de dimension peuvent mettre en évidence des structures cachées dans l'ensemble des données ayant un sens réel, tout en rendant plus compréhensible l'interprétation des résultats. Il y a deux catégories de méthodes de réduction du nombre d'attributs : les méthodes de sélection d'attributs dont une présentation générale peut être retrouvée dans [DH73] et les méthodes d'extraction d'attributs.

La réduction de la dimension des données est aussi justifiée par le fait que, en réalité, les données multidimensionnelles peuvent être représentées dans un sous-espace de dimension inférieure à la dimension originale des données. Dans la littérature, la dimension du nouveau sous-espace est appelée *"dimension intrinsèque"* des données [JD88] ; ainsi, les données originales sont projetées sur les axes du nouveau sous-espace. La *dimension intrinsèque* des données est une caractéristique importante de l'ensemble des données car elle indique le nombre minimal d'attributs nécessaire pour représenter les données. Trouver la dimension intrinsèque des données est un problème qui laisse encore beaucoup de questions ouvertes. Nous allons utiliser et tester les méthodes de *séparation aveugle de sources* (SAS) comme alternative aux méthodes d'extraction d'attributs déjà existantes.

- **La sélection des attributs** - les méthodes de sélection d'attributs ne sont pas souvent utilisées dans la classification non supervisée. La sélection des attributs peut être définie comme un processus qui choisit un sous-ensemble minimal de m attributs parmi l'ensemble original de d attributs ($m < d$) de sorte que celui-ci soit réduit de manière optimale selon un certain critère d'évaluation. Le choix du sous-ensemble optimal est une procédure de recherche exhaustive. Soit d le nombre d'attributs dans l'ensemble original des données, alors le nombre de sous-ensembles candidats est $2^d - 1$. Chaque sous-ensemble candidat doit être validé selon le critère d'évaluation choisi et comparé avec le meilleur sous-ensemble

précèdent, ce qui fait que ces algorithmes sont très coûteux en temps de calcul. Par contre, ces méthodes sont plus efficaces si on dispose des étiquettes des données, c'est-à-dire pour la classification supervisée parce que le sous-ensemble optimal d'attributs sélectionnés doit être validé par une comparaison des résultats de la classification obtenus après la sélection avec les étiquettes préalablement définies.

- **L'extraction des attributs** - les méthodes d'extraction d'attributs sont totalement non supervisées, car elles sont indépendantes de la connaissance de l'étiquette des données. On peut les regrouper en méthodes de projection linéaire et en méthodes de projection non linéaires.

  En général, les méthodes de projection cherchent un sous-espace dans l'espace original, de dimension inférieure à la dimension originale d ; la nouvelle représentation des données est obtenue en projetant les données originales sur les axes du nouveau sous-espace.

  Plusieurs méthodes de projection linéaire existent dans la littérature. Parmi celles-ci, l'Analyse en Composantes Principales (ACP) [Jol02], [Smi02] ou la transformée Karhunen-Loeve est la plus connue et la plus utilisée. Récemment d'autres méthodes de projection linéaire issues du domaine de la séparation aveugle de sources (SAS), comme l'Analyse en Composantes Indépendantes (ACI) ont été utilisées pour réduire le nombre des attributs [Fod02, Hyv99b]. Dans le chapitre suivant nous étudions les méthodes de SAS dans le contexte de la réduction du nombre d'attributs ; nous accordons plus d'attention aux méthodes de séparation prenant en compte les contraintes de l'application et nous proposons une méthode géométrique pour résoudre ce problème.

1.2.3 Problèmes liés à la classification non supervisées de données multivariées

Les principaux problèmes liés à la classification non supervisée des données multivariées sont présentés dans cette partie. D'autres problèmes peuvent exister mais ceux-ci sortent de ce cadre général.

Des classes superposées

La superposition des classes dans l'espace des attributs est due au fait que des objets qui appartiennent à des classes différentes ont des attributs qui sont très similaires. Les méthodes de classification non supervisées dites *dures* ont souvent des difficultés pour classer les objets qui

se trouvent dans les régions superposées. Par contre, les méthodes floues permettent à un seul objet d'appartenir à plusieurs classes en lui attribuant des degrés d'appartenance à chacune des classes.

Des classes de forme complexe

La forme des classes est un des problèmes majeurs dans la classification non supervisée. Les méthodes basées sur les centres des classes (*e.g. C-moyennes* [For65, HW79, Mac67] et son extension floue [Bez81] ou *ISODATA*, [HB65]) donnent de bons résultats pour des classes de forme convexe, voire sphérique ou ellipsoïdale, figure 1.1 (a), (b). L'incapacité de mettre en évidence des classes de forme non convexe, figure 1.1 (d), représente le principal défaut de ces méthodes. Par contre, les méthodes basées sur l'estimation de la fonction densité de probabilité dans l'espace des attributs classifient les données tout en respectant la forme de classes.

Le nombre des classes

Le choix du nombre de classes pose probablement les plus grands défis en ce qui concerne la classification non supervisée des données. En l'absence d'informations *a priori* sur les données, les résultats de la classification sont validés par l'évaluation d'indices de validité définis sur l'ensemble de données ; ceux-ci nous offrent une information sur la cohérence de la partition faite par une certaine méthode. Plusieurs indices de validité existent dans la littérature et sont rappelés dans la suite de ce chapitre.

La taille inégale des classes

Si la population des classes est très différente, ceci peut influencer les résultats de la classification. Des situations peuvent exister où une classe importante mais de faible population n'est pas mise en évidence parce que plusieurs classes importantes en nombre d'individus dominent le résultat de la classification.

La densité inégale de classes

La densité d'une classe en un point particulier de l'espace des attributs est donnée par le nombre de données contenues dans une unité de l'espace. Les méthodes de classification basées sur la *fdp*, rencontrent souvent des difficultés parce que les classes présentent des densités très différentes, figure 1.1 (c).

(a) (b) (c) (d)

Figure 1.1: Classes de forme sphérique (a) et elliptique (b), classes de densités différentes (c) et classes non convexes (d).

Dans la suite nous présentons quelques problèmes de classification liés plutôt à l'imagerie multivariée, mais ils peuvent apparaitre couramment dans d'autres domaines d'application.

La taille de l'ensemble de données

Les dispositifs d'acquisition d'images multivariées sont devenus de plus en plus performants et la sensibilité des capteurs a augmenté considérablement la résolution spatiale des images multivariées. Il en resulte une augmentation de la taille de l'image multivariée jusqu'à des millions de pixels. Pour beaucoup de méthodes de classification non supervisée, traiter un nombre de données si grand est un inconvénient majeur en termes de temps calcul et de mémoire.

Le nombre d'attributs

L'augmentation du nombre d'attributs améliore la résolution spectrale des images multivariées. Ceci peut être vu comme un avantage car on dispose d'un plus d'information pour l'analyse, mais le traitement mathématique devient de plus en plus compliqué. Des méthodes de réduction du nombre des attributs doivent être mises en oeuvre pour éviter les problèmes liés à la complexité des calculs mathématiques ainsi que pour éviter de prendre en compte l'information redondante.

Le bruit

L'acquisition des images est souvent accompagnée par la superposition d'un bruit sur l'information utile. Le bruit peut avoir des origines différentes : la sensibilité du capteur, des interférences ou des variations d'une autre nature. La présence du bruit favorise l'apparition de données aberrantes qui peuvent rendre les résultats très difficiles à interpréter, ou pire, donner des solutions inexactes. Des prétraitements pour supprimer le bruit sont souvent indispensables.

1.3 Mesures de similarité

Les méthodes de classification se basent sur le concept de similarité entre objets ; les mesures de similarité sont donc fondamentales pour la plupart des algorithmes. Le choix de la mesure de similarité doit être réalisé en prenant en compte toutes les informations disponibles sur l'ensemble des données car certaines d'entre elles peuvent favoriser les attributs d'un ordre d'échelle plus important et ainsi, la contribution des attributs moins significatifs sera négligée. C'est le cas de la distance euclidienne et des métriques de Minkowski d'ordre supérieur. De même, certaines mesures de similarité sont adaptées pour montrer les similarités entre objets du point de vue amplitude (les métriques de Minkowski), tandis que d'autres (le coefficient de corrélation et l'angle spectral) indiquent si les objets ont la même orientation dans l'espace des attributs. Dans la suite nous présentons les mesures de similarité les plus utilisées par les algorithmes de classification non supervisée en montrant les avantages et les inconvéniants de chacune.

1.3.1 Distance euclidienne et métriques de Minkowki

La plus simple et la plus populaire des mesures de similarité entre des données multivariées est la distance euclidienne qui est un cas particulier de la famille des métriques de Minkowski (quand l'exposant de la métrique $r = 2$). Les métriques de Minkowski sont définies par :

$$D_r(X_i, X_k) = \left(\sum_{j=1}^{d} |x_{ij} - x_{kj}|^r \right)^{1/r} \tag{1.9}$$

où $r \geq 1$. Des études récentes montrent que ces métriques sont affectées par *le phénomène de concentration* ce qui fait qu'elles ne sont pas appropriées pour estimer les similaritées entre des données multivariées. L'état de l'art lié à ce sujet est présenté dans le deuxième chapitre ; dans le quatrième chapitre nous étudions les métriques non euclidiennes dans le contexte de la classification non supervisée et nous proposons une modalité pour choisir la métrique optimale dans un problème de classification non supervisée.

1.3.2 Distance de Mahalanobis

La distance de Mahalanobis [Mah36] prend en considération la corrélation entre les données ; de plus elle n'est pas dépendante de l'échelle de données. La distance de Mahanalobis est définie par :

$$D(X_i, X_k) = (X_i - X_k)\Sigma^{-1}(X_i - X_k)^T \tag{1.10}$$

Les algorithmes à centres mobiles utilisant les métriques de Minkowski et la distance de Mahalanobis donnent de bons résultats si l'ensemble des données présente des classes compactes, isolées et de forme convexe [MJ96] (sphérique pour les métriques de Minkowski et ellipsoïdale pour la distance de Mahanalobis), figure 1.1 (a), (b), (c).

1.3.3 Angle spectral

L'angle spectral représente l'extension d-dimensionnelle de l'angle géométrique définit dans un plan. Considérons les données multivariées comme des vecteurs d-dimensionnels, alors l'angle spectral met en évidence les similarités entre la forme ou l'orientation des vecteurs. L'angle spectral entre deux vecteurs de dimension d X et Y est défini par :

$$\alpha = \arccos \frac{X^T Y}{\|X\|\|Y\|} \tag{1.11}$$

L'angle spectral est la mesure de similarité utilisée par l'algorithme supervisé de classification *Spectral Angle Mapper* [KLB+93]. Cette méthode est utilisée dans des nombreuses applications d'analyse spectrale [YGB92]. Le principal inconvénient de cette mesure de similarité est l'impossibilité de différencier les vecteurs de données qui ont des formes ou orientations similaires et des amplitudes très différentes.

1.3.4 Coefficient de corrélation de Pearson

Comme l'angle spectral, le coefficient de corrélation de Pearson est une mesure utilisée pour estimer les similarités entre objets présentant des orientations similaires dans l'espace des attributs. Il présente le même inconvénient : il est insensible aux variations d'amplitude des données. Dans [CM00] il est montré que la normalisation des données (l'extraction de la moyenne des données) améliore l'estimation des similarités entre données. Le coefficient de corrélation est ainsi défini :

$$R = \frac{(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}} \tag{1.12}$$

Ces deux mesures de similarité ne sont pas couramment utilisées dans l'analyse de données multivariées car en général, les différences entre données sont liées à l'amplitude et non à l'orientation des données dans l'espace des attributs. Pourtant elles apportent une information

complémentaire par rapport à l'information donnée par la distance euclidienne ou les autres métriques de Minkowski.

1.4 Méthodes de classification non supervisées

Les méthodes de classification non supervisée peuvent être regroupées en trois catégories ; les méthodes hiérarchiques, les méthodes à centre mobile et les méthodes basées sur la densité des objets dans leur espace de représentation. Cette taxonomie est présentée dans le tableau 1.1. Dans la littérature certains auteurs considèrent les méthodes à centre mobile et les méthodes basées sur la densité comme des méthodes par partitionnement [JMF99] tandis que d'autres considèrent les méthodes de partitionnement et les méthodes basées sur la densité comme deux catégories différentes [TWB05]. Les méthodes non supervisées de classification peuvent être *dures* ou *floues* ; les méthodes *dures* attribuent à chaque objet une seule étiquette, tandis que dans une classification *floue*, un objet peut appartenir simultanément à plusieurs classes. Les méthodes *floues* peuvent être facilement converties dans des méthodes *dures*.

Méthodes de classification non supervisée	**Algorithmes**
Méthodes hiérarchiques	L'algorithme de lien minimal, l'algorithme de lien maximal, l'algorithme de Ward
Méthodes à centre mobile	C-moyennes, C-moyennes floue, ISODATA
Méthodes à densité	Denclust, Mean-Shift, Parzen-Watershed

Tableau 1.1: Taxonomie des méthodes de classification non supervisées

1.4.1 Méthodes hiérarchiques

Les méthodes non supervisées hiérarchiques sont généralement des méthodes *dures* et consistent à trouver une organisation arborescente des classes ou un dendrogramme. La plupart de ces méthodes dérivent des algorithmes de lien minimal *single-link* [SS73], de l'algorithme de lien maximal *complete-link* [Kin67] et de la méthode de variance minimale ou méthode de Ward [War63], [Mur84].

Le principe de l'algorithme

Le principe des algorithmes hiérarchiques est résumé dans l'algorithme suivant :

Algorithme 1 Le pseudocode des algorithmes hiérarchiques

1: **Départ** : chaque objet est attribué à une seule classe
2: **Itération** : on calcule les similarités entre toutes les paires de classes i et j et les deux classes les plus similaires sont regroupées
3: **Arrêt** : l'algorithme s'arrête quand tous les objets sont regroupés dans une seule classe

Choix du nombre de classes

Le dendrogramme ainsi obtenu peut être coupé à n'importe quel niveau pour obtenir le nombre de classes désiré. Pourtant, déterminer le nombre exact de classes est très difficile. La visualisation du dendrogramme représente un moyen mais ceci est utile seulement pour un nombre réduit de données. D'autres critères de validation qui sont présentés dans la suite de ce chapitre peuvent être aussi utilisés.

Avantages et inconvénients

L'avantage des méthodes hiérarchiques est leur stabilité. Ceci est dû à deux raisons particulières [TWB05] : premièrement, l'initialisation des classes est toujours la même et deuxièmement, pour une itération quelconque, les algorithmes considèrent seulement les classes précédemment obtenues ; de cette manière, un objet appartenant à une classe ne peut pas se retrouver dans une autre classe dans les itérations suivantes. Ceci peut être vu comme un avantage mais aussi comme un inconvénient car la flexibilité de la méthode diminue. Leur principal inconvénient est lié à la taille de l'ensemble de données. A chaque itération, ces méthodes utilisent la matrice de distance interpoint ou interclasse. Ceci fait que pour des applications contenant des bases de données très grandes (*i.g.* imagerie multivariée) ces méthodes ne sont que rarement utilisées.

1.4.2 Méthodes de centre mobile

Les méthodes de centre mobile consistent à regrouper les objets en optimisant une fonction objective. En fonction de la méthode, l'optimisation est réalisée soit par minimisation soit par maximisation d'un critère objectif. Les méthodes les plus connues de cette catégories sont l'algorithme *C-moyenne* [For65], [HW79], [Mac67] avec son extension floue [Bez81] et l'algorithme *ISODATA* (*Iterative Self-Organising DATA*) [HB65], [Jen96]. Pour les méthodes *C-moyenne* et *C-moyenne floue*, la fonction à minimiser est :

$$E = \sum_{j=1}^{C} \sum_{i \in c_i} u_{ij} d(x_i, g_j) \tag{1.13}$$

où u_{ij} représente le dégré d'appartenance de l'objet x_i à la classe j et g_j est le centre de la classe j. Pour la méthode *dure*, u_{ij} prend les valeurs 0 ou 1. Pour la méthode floue, u_{ij} est remplacé par u_{ij}^q, où $u_{ij} \in (0, 1)$ et $q > 1$. Le paramètre q est appelé *coefficient de fuzzyfication* et souvent il prend la valeur 2. Pour $q = 1$ l'algorithme *C-moyenne "dur"* est obtenu.

Le principe de l'algorithme *C-moyenne* et *C-moyenne floue*

Les algorithmes *C-moyenne* et *C-moyenne floue* sont résumés par le pseudocode presenté dans l'algorithme 2 :

Algorithme 2 Le pseudocode des algorithmes *C-moyenne* et *C-moyenne floue*

1: **Départ** : choisir le nombre k de classes ; choisir aléatoirement l'ensemble initial de centre de classes $G = (g_1, g_2, \ldots, g_k)$
2: **Itération :** affecter chaque x_i à la classe c_j telle que: $d(x_i, g_j) < d(x_i, g_l)$ pour tout $l \neq j$
3: Mise à jour des centres de classes par :

- pour la méthode *dure* :

$$g_j = \frac{1}{N_j} \sum_{i=1}^{N_j} x_i \tag{1.14}$$

- pour la méthode *floue* :

$$g_j = \frac{\sum_{i=1}^{N_j} u_{ij}^q x_i}{\sum_{i=1}^{N_j} u_{ij}^m} \tag{1.15}$$

$$u_{jk} = \frac{1}{\sum_{i=1}^{C} \left(\frac{\|x_k - g_j\|^2}{\|x_k - g_i\|^2} \right)^{\frac{1}{q-1}}} \tag{1.16}$$

4: Si non stationnarité des centres de classes aller en 2
5: **Arrêt**

Choix du nombre de classes Le nombre de classes est imposé avant la classification. Ce fait constitue un des principaux inconvénients de ces méthodes. Pourtant, des indices de validité qui prennent en compte la compacité des classes et la distance interclasse peuvent être utilisés pour choisir le nombre optimal de classes. Ceux-ci sont présentés dans la section suivante.

Avantages et inconvénients L'avantage majeur de cette méthode est le temps de calcul. La complexité de l'algorithme est seulement de $n \log(n)$ où n est le nombre de données. Ceci fait que cette méthode est applicable pour des bases de données de taille très grande ; ainsi elle est bien adaptée pour des applications de l'imagerie multivariée. Ses principaux inconvénients

sont :

- le nombre des classes est défini par l'utilisateur au début de la classification ;
- les centres des classes de départ sont choisis arbitrairement. En effet le choix des centres initiaux peut conduire à des solutions totalement différentes. Ceci provient du fait que l'on recherche un minimum local. En général, ce choix est fait de façon aléatoire, ce qui ne garantit pas la pertinence de la classification finale ;
- le problème des classes de tailles inégales peut aussi influencer les résultats de la classification ; souvent, les centres des classes très petites sont attirés par les centres des classes adjacentes plus larges ;
- la forme de classes est implicitement convexe ; ces algorithmes ne sont pas adaptés pour trouver des classes de forme non convexe, figure 1.1 (d).

Le principe de l'algorithme Isodata

ISODATA reprend l'idée de *C-moyennes* mais vérifie à chaque itération certains critères d'optimisation liés à la compacité intra-classes et à la séparabilité inter-classes. Ces critères doivent permettre aux résultats de satisfaire les conditions suivantes :

- Une séparation maximale entre classes ;
- La meilleure compacité intra-classe (la distance intra-classe minimale).

Les paramètres de la méthode sont : le nombre maximal d'itérations M, l'écart-type standard intra-classe et la distance minimale entre deux classes. L'algorithme est résumé par le pseudocode suivant :

Algorithme 3 Le pseudocode de l'algorithme *ISODATA*

1: **Départ** : choisir le nombre de classes ; choisir aléatoirement l'ensemble initial de centre de classes $G = (g_1, g_2, \ldots, g_k)$ et initialiser $I = 0$
2: **Tant que** $I < M$ et non stationnarité des centres de classes
   - Affecter chaque x_i à la classe c_j telle que : $d(x_i, g_j) < d(x_i, g_l)$ pour tout $l \neq j$;
   - Mise à jour des centres de classes en utilisant l'équation 1.15 ;
   - Si écart-type de la classe trop grand on divise la classe en deux ;
   - Si distance minimale entre deux classes trop petite fusion de classes ;
   - Incrémentation de I ;
3: **Fin Tant que**
4: **Arrêt**

Choix du nombre de classes Même si le nombre de classes est initialement imposé, le nombre final de classes peut être différent car l'algorithme est capable de fusionner ou de séparer des classes si les conditions qui assurent une séparation optimale ne sont pas respectées.

Avantages et inconvénients Le principal avantage de la méthode *ISODATA* par rapport aux méthodes précédemment présentées réside dans le fait que le nombre de classes peut soit augmenter soit diminuer en fonction des valeurs des paramètres choisis. Restent tout de même plusieurs difficultés :

- Les paramètres doivent être initialisés et nécessitent donc d'avoir des connaissances statistiques *a priori* sur les différentes classes ;
- L'initialisation des classes est toujours fixée *a priori* et influe les résultats de la classification ;
- La forme de classes est implicitement convexe.

1.4.3 Méthodes basées sur l'estimation de la fonction de densité de probabilité

Un point de vue extrême dans la littérature est que l'apprentissage non supervisé est fortement lié à l'estimation de la fonction de densité de probabilité (*fdp*) des données [SPST$^+$01]. Il est évident que si on connait cette fonction on pourra résoudre tous les problèmes liés aux données en cause. Pourtant, estimer la *fdp* des données multivariées implique de nombreux défis mathématiques et on mentionne ici le *phénomène de l'espace vide* ou le *phénomène de Hughes*.

Les méthodes basées sur l'estimation de la fonction densité de probabilité constituent une troisième catégorie de méthodes non supervisée de classification. Elles sont principalement conçues pour détecter des classes de forme non convexe. Ces méthodes ont comme principe l'estimation de la densité autour de chaque objet dans l'espace de représentation. Chaque point de maximum local de la densité estimée identifie une classe et les régions de faible densité constituent les bords des classes. Le nombre des classes correspond au nombre des modes de la *fdp* estimée ; il est déterminé généralement soit par le seuillage itératif de la fonction de densité estimée [HBV01], soit par la recherche des points de maxima locaux [CM99]. La taille de la fonction noyau utilisée dans l'estimation de la *fdp* représente le principal paramètre de ces méthodes. Une fois ce paramètre estimé, le nombre de classes s'obtient automatiquement. Ces méthodes ont été proposées pour la première fois dans [FH75] et ensuite améliorées dans [Che95] ; parmi

celles-ci, les plus populaires sont *Denclust* [HK98], *CLUPOT* [CM81], *DBSCAN* [EKSX96] et *Mean Shift* [CM99]. D'autres méthodes existent comme la méthode *Parzen-Watershed* [HBV96] qui consiste à diviser l'espace des données en zones d'influence en utilisant des méthodes de morphologie mathématique et de traitement des images (la méthode *Watershed* ou *SKIZ - SKeleton by Influence Zones* [Ser82]). Dans [NL06], cette méthode est comparée à une autre issue de la théorie des vecteurs du support : la méthode *Support Vector Clustering* (SVC) [BHHSV01] qui consiste à estimer seulement le support de la *fdp*, au lieu d'estimer la *fdp* à chaque point de l'espace des données. Pour plus de détails, le lecteur est invité à consulter la référence [NL06].

L'algorithme Parzen-Watershed

Cette méthode [HBV96] est basée sur le concept des zones d'influence ; celles-ci sont déterminées en estimant la fonction de densité de probabilité (*fdp*) dans l'espace des attributs par la méthode Parzen [Par62], suivie par la division de l'espace en zones d'influence. Les données sont classifiées en fonction de la zone d'influence à laquelle elles appartiennent. La méthode est illustrée dans la figure 1.2 (b), (c), (d) pour l'ensemble des données bidimensionnelles présenté dans la figure 1.2 (a). Pour des données multivariées, l'étape de réduction de dimension est essentielle. Ensuite, les données sont représentées dans cet espace discrétisé, figure 1.2 (b). Pour un ensemble de données $2-D$ ceci peut être visualisé comme une grille rectangulaire incluant S éléments y_i, $0 \leq i \leq S$. Dans cet espace toute fonction f peut être définie par ses valeurs $f(y_i)$. Cette discrétisation de l'espace des attributs nous permet la visualisation d'une fonction comme une image de S pixels. Ainsi, les données x_k, $0 \leq k < N$ sont des éléments dans l'espace discret qui se distinguent des autres éléments par des valeurs différentes (des pixels blancs sur fond noir). L'estimation de la *fdp*, figure 1.2 (b) est réalisée par la méthode de Parzen ; la *fdp* estimée est donnée par :

$$\hat{fdp}(y_i) = \lambda \sum_{0 \leq k < N} K\left(\frac{y_i - x_k}{h}\right) \tag{1.17}$$

où K est une fonction noyau utilisée dans l'estimation, h est la largeur du noyau (ou paramètre de lissage) et λ représente le coefficient de normalisation. Souvent, le noyau utilisé est le noyau gaussien :

$$K(x) = (2\pi)^{-d/2} exp\left(-\frac{1}{2}x^T x\right) \tag{1.18}$$

La division de l'espace de représentation en zones d'influence, figure 1.2 (c) peut être réalisée selon 2 modalités. La première consiste en 2 étapes : le seuillage itératif de la *fdp* estimée

suivi par la procédure *SKIZ*. La deuxième modalité consiste à trouver directement les zones d'influence par la procédure *Watershed*. Une zone d'influence correspondra à un maximum local dans la *fdp* estimée. La dernière étape consiste à classifier les données ; celle-ci est réalisée en attribuant pour chaque donnée, l'étiquette correspondant à la zone d'influence où elle se trouve. La méthode est résumée par le pseudocode suivant :

Algorithme 4 Le pseudocode de l'algorithme *Parzen-Watershed*

1: **Départ** : réduction de dimension
2: Discrétisation de l'espace à dimension réduite et représentation des données dans cet espace
3: Estimation de la *fdp* par l'equation 1.17
4: Division de l'espace à dimension réduite en zones d'influence
5: Classification de données
6: **Arrêt**

(a) (b) (c) (d)

Figure 1.2: Illustration de la méthode Parzen-Watershed : a) représentation de données dans l'espace des attributs, b) représentation des données dans l'espace discret (image de S pixels), c) représentation de la *fdp* des données dans l'espace image, d) zones d'influence obtenues par la méthode SKIZ.

Choix du nombre de classes Le nombre de classes correspond au nombre de zones d'influence ; il dépend de la *fdp* estimée qui dépend à son tour de la taille du noyau h utilisé. Dans [HBV01] la taille du noyau est considérée comme un paramètre de la méthode ; plus les valeurs de ce paramètre sont grandes, plus lisse est la *fdp estimée*. Des valeurs trop élevées peuvent cacher des détails et ainsi, deux zones d'influence vont se regrouper en une seule. Le nombre de zones d'influence est ainsi considéré comme une fonction dépendant de la taille du noyau ; cette fonction décroit avec l'augmentation de la taille du noyau. Si les données sont bien classifiées, alors ces classes vont apparaitre pour un intervalle important du paramètre h. Si la fonction présente des plateaux, alors on suppose que ces plateaux indiquent une bonne classification.

Avantages et inconvénients Les avantages de cette méthode sont :

- Le nombre de classes dérive automatiquement de l'analyse mais il est sensible au paramètre du lisage h :
- Les classes peuvent avoir des formes complexes.

Les principaux inconvénients sont :

- Les classes de densité différente : les régions de faible densité peuvent être considérées comme du bruit ;
- Cette méthode donne de bons résultats pour des données à 2, 3 ou au maximum 4 dimensions ; pour des données multivariées, une étape de réduction de la dimension est obligatoire car la complexité de l'algorithme augmente avec la dimension ;
- Le temps de calcul dédié à l'estimation de la *fdp* est très important ;
- Le problème des classes superposées : les régions de recouvrement des classes peuvent avoir une densité plus importante que les régions voisines et donc elles peuvent former d'autres classes ou elles peuvent fusionner deux classes superposées.

L'algorithme Mean-Shift

Pour la méthode non-paramétrique *Mean-Shift* [CM99], la classification des objets est réalisée par l'estimation des *maxima* locaux de la fonction densité de probabilité associée à une distribution de points. Cette méthode est basée sur l'estimation du gradient de la *fdp* et les modes sont obtenus par la recherche des points dans l'espace des attributs pour lesquels la *fdp* estimée équation 1.17, s'annule.

$$\nabla \hat{fdp}(x_i) = \lambda \sum_{0 \leq k < N} \nabla K \left(\frac{x_i - y_k}{h} \right) = 0 \tag{1.19}$$

La fonction noyau est choisie de manière à ce que la *fdp* estimée approxime au mieux la *fdp* d'une distribution de points. Ceci peut être mesuré par l'erreur quadratique moyenne intégrée (EQMI ou MISE - Mean Integrated Square Error).

$$EQMI = E\left[\int_{\Re^d} \left(\hat{fdp}(x_i) - fdp(x_i) \right)^2 dx \right] = \int_{\Re^d} E\left[\hat{fdp}(x_i) - fdp(x_i) \right]^2 dx \tag{1.20}$$

Ce critère (équation 1.20) est minimal dans le cas de l'utilisation du noyau d'Epanechnikov :

$$K_E(x) = \begin{cases} \frac{1}{2}c_d^{-1}(d+2)(1-\|x\|^2) \ \ pour \ \ \|x\| < 1 \\ 0 \ \ sinon \end{cases} \tag{1.21}$$

où c_d est le volume de l'hypersphère unité de dimension d. En remplaçant l'equation 1.21 dans 1.17 on écrit :

$$\nabla \hat{fdp}(x) = \frac{\lambda}{c_d}\frac{d+2}{h^2}\left[\frac{1}{n_x}\sum_{x_i \in S_h(x)}(x_i - x)\right] \tag{1.22}$$

où S_h est la sphère de rayon h centrée en x contenant n_x points. Le dernier terme de l'équation 1.22 est appelé le vecteur Mean Shift et il représente une approximation du gradient de la *fdp* estimée :

$$M_h(x) = \frac{1}{n_x}\sum_{x_i \in S_h(x)}(x_i - x) \equiv \nabla \hat{fdp}(x) \tag{1.23}$$

Une présentation complète de la méthode *Mean Shift* peut être retrouvée dans [CM99] ; dans [Pet06] celle-ci est utilisée dans une application pour la segmentation des images hyperspectrales où des méthodes de réduction de dimension par projection sont employées pour l'extraction des signatures spectrales. Dans [Pet06], la méthode *Mean-Shift* est résumée par l'algorithme suivant :

Algorithme 5 Le pseudocode de l'algorithme Mean Shift

1: $\hat{m} \leftarrow x$
2: **Tant que** m est différent à chaque itération
3: Calculer $M_h(\hat{m})$
4: $\hat{m} \leftarrow \hat{m} + M_h(M)$
5: **Fin Tant que**
6: $\hat{m}$ est le mode associé à x

Choix du nombre de classes Le nombre de classes correspond au nombre de modes de la $\hat{fdp}$ et il dépend du paramètre h.

Avantages et inconvénients Par rapport à l'algorithme *Parzen-Watershed*, cette méthode a l'avantage de pouvoir classifier des données de plus grande dimension. De plus, comme toutes les méthodes basées sur l'estimation de la fdp, l'algorithme *Mean Shift* peut classifier des classes de forme non convexe, ce qui constitue un avantage par rapport aux méthodes à centre mobile. Ses points faibles sont l'estimation du paramètre de lissage h et le temps de calcul qui devient prohibitif pour des données à grande dimension.

Classification par l'estimation du support de la *fdp*

Le principe de cette méthode inspirée par une approche issue de la théorie des vecteurs du support est d'estimer le support de la *fdp* d'un ensemble de données. Estimer le support de la *fdp* consiste à estimer une fonction définie sur l'espace des attributs qui est positive dans les régions où se trouve la plupart des données et négative ailleurs. Ces régions constituent les classes. Pour un ensemble de données $X_1, \ldots, X_n \in X$, la méthode *SVC* peut être résumée en 3 étapes :

1. Transformation des données par une fonction $\Phi : X \rightarrow F$ pour laquelle le produit scalaire peut être évalué par une fonction noyau

$$K(X_i, X_j) = \Phi(X_i)\Phi(X_j) \tag{1.24}$$

comme par exemple le noyau gaussien :

$$K(X_i, X_j) = \exp^{-\frac{1}{h}\|X_i - X_j\|^2} \tag{1.25}$$

2. Dans le nouvel espace de représentation F, on cherche l'hyperplan paramétré par le couple (w, ρ) qui sépare au mieux les données par rapport à l'origine, où w représente le vecteur qui passe par l'origine de l'espace F perpendiculaire à l'hyperplan de séparation et ρ est le déplacement angulaire de l'hyperplan :

$$\min_{w \in F, \rho \in \Re} \frac{1}{2}\|w\|^2 - \rho \tag{1.26}$$

sous les contraintes $(w\Phi(x_i)) \geq \rho$

Les paramètres de l'hyperplan de séparation sont estimés par :

$$w = \sum_{i=1}^{N} \alpha_i \Phi(x_i) \tag{1.27}$$

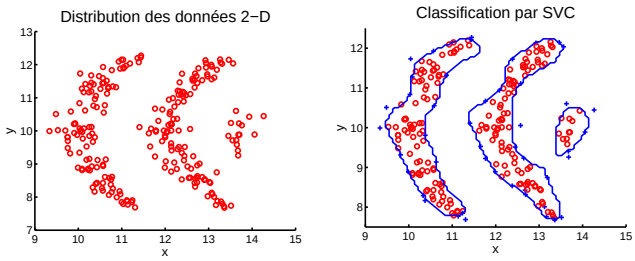
$$\rho = w\Phi(x_i) \tag{1.28}$$

Les coefficients α_i sont les coefficients de Lagrange issus de la résolution du problème d'optimisation 1.26.

3. Estimation de la fonction de décision binaire dans l'espace d'origine par :

$$f(x) = sgn(w\Phi(x) - \rho) \tag{1.29}$$

Cette fonction est positive dans les régions contenant la plupart des données et négative ailleurs. Les points qui se trouvent à l'intérieur d'un contour fermé sont mis dans la même classe. La méthode est illustrée dans la figure 1.3.



(a) (b)

Figure 1.3: Illustration de la méthode de classification basée sur l'estimation du support de la *fdp*

Choix du nombre des classes Le nombre de classes correspond au nombre de contours fermés déterminés par la fonction binaire, équation 1.29, dans l'espace de données. Comme dans le cas de la méthode *Parzen-Watershed*, il s'obtient automatiquement et il est contrôlé par la taille du noyau gaussien h.

Avantages et inconvénients Inconvénients : cette méthode est plus sensible au recouvrement des classes.

1.5 Indices de validité

La plupart des méthodes non supervisées de classification ne sont pas capables de détecter le nombre optimal de classes automatiquement. Souvent, le nombre de classes doit être défini soit de manière directe (*C-moyennes*), soit de manière indirecte (les méthodes hiérarchiques). En général, le nombre de classes est déterminé à l'aide d'un critère de validité. Les indices de validité sont des critères permettant de choisir le nombre optimal de classes pour un certain algorithme, ou de valider les résultats de la classification en les comparant avec d'autres résultats obtenus soit par un autre algorithme, soit par le même algorithme paramétré différemment. On distingue trois catégories d'indices : internes, externes et relatifs.

- *Les indices internes* confirment ou infirment les résultats de la classification en se basant seulement sur les données existantes dans l'ensemble initial ;
- *Les indices externes* se basent sur des informations *a priori* (généralement une vérité de terrain) pour la validation des résultats ;
- *Les indices relatifs* servent pour décider quel est le meilleur parmi plusieurs résultats ; les indices internes sont souvent utilisés comme des indices relatifs.

Nous rappelons quelques indices internes et nous invitons le lecteur à consulter les références [JD88, LJB06] pour plus d'information.

1.5.1 Indice de Davies-Bouldin

L'indice de Davies-Bouldin [DB79] tient compte en même temps de la compacité et de la séparabilité des classes. Si les classes sont compactes et bien séparées, alors la valeur de cet indice est faible. Si l'indice de Davies-Bouldin est estimé comme une fonction dépendante du nombre des classes, alors le nombre optimal de classes est donné par le point de minimum global de cette fonction. Cet indice favorise les classes de forme hypersphérique et il est adapté aux méthodes à centre mobile, *e.g. C-moyennes*.

$$DB = \frac{1}{C}\sum_{i=1}^{C} \max_{j=1,\cdots,C}(d_{ij}) \quad , où \quad d_{ij} = \frac{\sigma_i + \sigma_j}{d(g_i, g_j)} \tag{1.30}$$

Dans cette expression, C représente le nombre de classes, σ_i est la distance moyenne entre les objets et le centre de la classe C_i et $d(g_i, g_j)$ est la distance entre les centres de classe g_i et g_j. Ainsi, la distance d_{ij} aura une valeur faible si les classes sont compactes et bien séparées. La complexité de calcul de cet indice est faible.

1.5.2 Indice Dunn

Soit d_{min} la distance minimale entre deux objets de classe différente et d_{max} la distance maximale entre deux objets de la même classe. Alors, l'indice Dunn [Dun74] D, est défini par :

$$D = \frac{d_{min}}{d_{max}} \tag{1.31}$$

Une bonne classification est indiquée par des valeurs élevées de cet indice. Le temps de calcul de cet indice est un inconvénient majeur quand on manipule de très grands ensembles de données. Ceci, ainsi que sa faible sensibilité au bruit font que cet indice est rarement utilisé.

1.5.3 Indice C_0

L'indice C_0 mesure la compacité des classes et [HS76] est il défini par :

$$C_0 = \frac{S - S_{min}}{S_{max} - S_{min}} \tag{1.32}$$

Dans cette formule, S représente la somme des distances entre toutes les paires d'objets dans une classe. Soit l le nombre des paires d'objets dans une classe, alors S_{min} représente la somme des l plus petites distances si toutes les paires d'objets sont considérées et S_{max} représente la somme des l plus grandes distances issues de toutes les paires d'objets. Le dénominateur sert à la normalisation, $C_0 \in [0,\ 1]$. Cet indice est particulièrement bien adapté si les classes sont de taille similaire et il est de plus petit si la classe est plus compacte.

1.5.4 Indice de compacité-séparabilité

L'indice de compacité-séparabilité CS est décrit dans [Fre92] pour une partition floue et il utilise le même principe que l'indice DB : il tient compte à la fois de la compacité c_o et de la séparabilité s_e des classes. Pour une classification dure il est défini par :

$$CS = \frac{c_o}{s_e} \tag{1.33}$$

$$où\ c_o = \frac{1}{C}\sum_{i=1}^{C}\sigma_i\ et\ s_e = \min_{i \neq j}(d(g_i, g_j)). \tag{1.34}$$

1.6 Conclusion

Dans ce chapitre nous avons présenté de manière générale le domaine de la classification non supervisée des données multivariées. Parmi ces techniques, les méthodes à centre mobile et les méthodes à densité sont particulièrement appropriées à la découverte des classes dans des ensembles de données de grande taille.

Des travaux récents [AHK01, HAK99, BGRS00, FWV07, Dem94] montrent que la distance euclidienne (la mesure de similarité la plus utilisée par les méthodes à centre mobile) ainsi que toutes les métriques de Minkowski sont affectées par le *phénomène de concentration*. Ceci met en doute la pertinence de ces métriques comme mesure de similarité pour des données multivariées. Trouver une méthode pour choisir la métrique optimale dans un problème de classification non supervisée constitue la motivation pour l'étude du phénomène de concentration des métriques. L'état de l'art relatif à ce sujet est présenté dans le deuxième chapitre.

La réduction de la dimension est un autre problème majeur lié à la classification non supervisée des données de grande dimension. Son principal but est de réduire la complexité et le temps de calcul des méthodes de classification mais aussi d'extraire des facteurs pertinents qui peuvent nous aider à la compréhension des données. Pour cela, dans le troisième chapitre nous allons étudier les méthodes de *séparation aveugle de sources* (*SAS*) comme une alternative aux méthodes classiques d'extraction d'attributs (ACP - *Analyse an Composantes Principales*). La réduction de dimension nous permet d'utiliser des méthodes de classification à densité qui favorisent la mise en évidence des classes de forme complexe.

Chapitre 2

Phénomène de concentration des métriques

Notations

Notation	:	Signification
X	:	Ensemble de données multivariées ; matrice de dimension $n \times d$
n	:	Nombre de données multivariées
d	:	Dimension de données (le nombre d'attributs d'un ensemble de données)
$X_i,\ i = 1 : d$	:	Vecteur de données
x_i	:	Scalaire représentant une mesure d'un vecteur de données
L_r	:	Famille de métriques r
r	:	Exposant de la métrique
Y	:	Ensemble de données representant la différence entre toutes les paires des vecteurs de l'ensemble X
Y_{ij}	:	Vecteur de l'ensemble Y representant la différence entre les vecteurs X_i et X_j
Γ_2	:	Ensemble des normes L_2 de X
μ	:	Moyenne de Γ_2
$\frac{Dmax_d^r - Dmin_d^r}{Dmin_d^r}$	:	Fonction de contraste relatif
$\frac{Var(\|X\|_r)}{E(\|X\|_r)}$	:	Fonction variance relative

σ : Variance de Γ_2

$Dmin$: Valeur minimale observée de Γ_2

$Dmax$: Valeur maximale observée de Γ_2

M : Valeur maximale admise de Γ_2

$Var()$: Variance d'une variable aléatoire

$E()$: Espérance mathématique d'une variable aléatoire

2.1 Introduction

La distance euclidienne est le critère de similarité le plus connu et le plus utilisé par les méthodes de classification non supervisée ; cette mesure est adaptée pour calculer la distance entre des points dans des espaces à 2 ou 3 dimensions. Des travaux récents montrent que pour des données multivariées cette mesure est affectée par *le phénomène de concentration* ; cela signifie que toutes les distances entre les paires de points d'un ensemble quelconque sont presque identiques et donc la notion du *plus proche voisin* ou *de similarité* devient instable [BGRS00]. C'est à cause de ceci que la pertinence de la distance euclidienne pour résoudre des problèmes d'analyse de données multidimensionnelles et implicitement de classification non supervisée est récemment mise en question [AHK01, HAK99, BGRS00].

Pour palier à ce phénomène, une solution proposée est d'utiliser des métriques moins concentrées [AHK01]. Celles-ci augmentent le contraste entre les données et sont proposées comme alternative à la distance euclidienne pour améliorer les résultats de la classification des données multivariées. Dans [AHK01], les auteurs montrent que des valeurs inférieures à 1 de l'exposant de la norme atténuent le phénomène de concentration et ainsi, ces métriques, connues comme métriques fractionnaires, pourraient fournir de meilleurs résultats dans toute sorte d'applications traitant des données multidimensionnelles.

Contrairement à ces résultats, dans [FWV07] les auteurs montrent qu'il existe des distributions pour lesquelles les métriques d'ordre supérieur sont moins concentrées que les métriques fractionnaires. Ils montrent que les résultats obtenus en [AHK01] ne sont valables que dans le cas où les données sont uniformement distribuées, ce qui est loin d'être vrai pour des données réelles. Ils affirment que, *"chercher des métriques moins concentrées pour lutter contre la concentration est tout à fait justifié"*, mais ils ne font aucune affirmation sur l'impact de ces normes sur les résultats de la classification. Des questions restent encore ouvertes : est-ce qu'une norme moins concentrée donne des meilleurs résultats dans des problèmes de classification non supervisée

qu'une métrique plus concentrée? Si ceci est vrai, alors, trouver la métrique optimale consiste à trouver la métrique la moins concentrée pour l'ensemble des données analysées. Sinon, existe-il un moyen de choisir la métrique optimale pour résoudre le problème mentionné ?

Nous allons étudier de près ce phénomène en essayant de répondre à ces questions en ayant comme but l'amélioration des résultats des algorithmes de classification non supervisée utilisant la distance euclidienne comme mesure de similarité. Ce chapitre présente un bref état de l'art relatif au sujet de la concentration des métriques. Nous commençons en présentant les définitions de la famille de métriques r et du *phénomène de la concentration*. Ensuite nous rappelons quelques résultats théoriques sur la concentration des métriques et nous conclurons.

2.2 Métriques r

Les métriques r sont définies par la famille de métriques dépendant du paramètre r appelé *exposant* de la métrique :

$$L_r(X_i - X_k) = \left(\sum_{j=1}^{d} |x_{ij} - x_{kj}|^r \right)^{1/r} \tag{2.1}$$

où $r \in \Re_+$.

2.2.1 Métriques de Minkowki

Les métriques de Minkowski sont définies pour $r \geq 1$ où le paramètre r n'est pas forcément un nombre entier. Ces métriques satisfont les propriétés suivantes :

1. $L_r(X_i, X_i) = 0, \forall(i)$;
2. $L_r(X_i, X_k) = L_r(X_k, X_i), \forall(i, k)$;
3. $L_r(X_i, X_k) \geq 0, \forall(i, k)$;
4. $L_r(X_i, X_k) = 0$, seulement si $(i = k)$;
5. $L_r(X_i, X_k) \leq L_r(X_i, X_m) + L_r(X_m, X_k), \forall(i, m, k)$;

Parmi les métriques de Minkowski on mentionne les plus utilisées :

1. La distance euclidienne, $r = 2$ est la plus connue ; invariable aux translations et aux rotations des données dans l'espace des attributs et couramment utilisée dans des espaces

à 2 ou 3 dimensions, cette métrique donne des bons résultats si l'ensemble des données présente des classes compactes et isolées [MJ96].

2. La distance de Manhattan, $r = 1$ est plus appropriée pour mesurer la similarité entre des données multivariées ; elle est moins sensible au bruit coloré que la distance euclidienne.

Un premier inconvénient des métriques de Minkowski d'ordre supérieur $r \geq 2$ est qu'elles négligent les attributs les moins significatifs en faveur des attributs d'un ordre d'échelle plus élevé. La solution à ce problème est la normalisation des données. Selon [HAK99, BGRS00], le deuxième inconvénient majeur de ces métriques réside dans le fait qu'elles sont affectées par le phénomène de concentration qui est discuté dans la suite de ce chapitre.

2.2.2 Métriques fractionnaires

Les métriques fractionnaires sont des métriques faisant partie de la famille des métriques r, pour lesquelles l'exposant de la métrique est inférieur à 1. Dans [GP86], il est montré que seulement les propriétés 1, 2 et 4 sont nécessaires pour qu'une métrique soit utilisée comme indice de similarité dans des problèmes de classification non supervisée. Ceci fait que des valeurs inférieures à 1 du paramètre r pour lesquelles l'inégalité du triangle n'est pas réspectée (la propriété 5), peuvent être utilisées pour définir d'autres mesures de similarité. L'utilisation des métriques fractionnaires comme mesure de similarité est donc justifiée.

2.3 Phénomène de concentration des métriques

Nous introduisons le phénomène de concentration des métriques et nous présentons l'état de l'art concernant ce sujet. Dans la section suivante nous étudions l'hypothèse présentée en [AHK01] statuant que les métriques les moins concentrées donnent de meilleurs résultats dans des problèmes de classification non supervisée par rapport aux métriques plus concentrées et nous présenterons nos conclusions relatives à ce sujet.

2.3.1 Définition

La concentration des métriques est un phénomène qui peut être défini comme suit : dans des espaces multidimensionnels, toutes les distances entre les paires de points sont presque identiques. Etudier le phénomène de concentration revient à étudier la distribution de l'ensemble des distances entre toutes les paires de points d'un ensemble donné. Pour un ensemble X de n vecteurs aléatoires X_i, cela revient à construire un nouvel ensemble Y de $n(n-1)/2$ vecteurs

$Y_{ij} = X_i - X_j$ correspondant à la différence entre toutes les paires des vecteurs de X et à étudier la distribution des normes $\|Y\|_r$ de ce nouvel ensemble, [FWV07]. Le phénomène de concentration de la métrique euclidienne est très bien illustré dans [FWV07] par l'exemple suivant :

Illustration Soit X un ensemble de n vecteurs aléatoires X_i de dimension d tirés de $[0\ 1]^d$ et $\Gamma_2 = \{\|X_i\|_2\}_{i=1}^n$ l'ensemble des normes L_2 de ces vecteurs. Les valeurs de Γ_2 sont bornées : $\|X_i\|_2 \in [0\ M]$ où $M = \|1 \ldots 1\|_2 = \sqrt{d}$. Afin d'étudier le comportement de la distribution de l'ensemble Γ_2 présenté dans la figure 2.1, nous allons examiner l'évolution de ses paramètres spécifiques en fonction de la dimension des données : la moyenne μ, la variance, la valeur minimale *Dmin* et maximale *Dmax* observée et la valeur maximale admise de Γ, M.

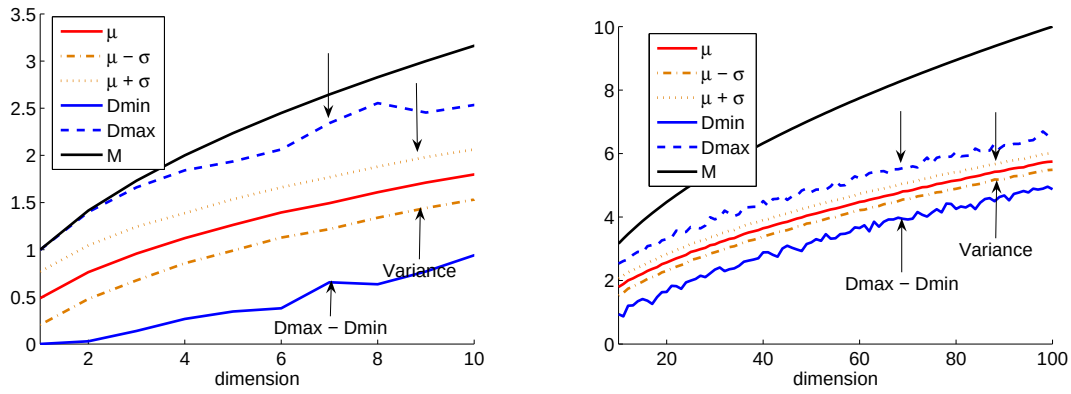


Figure 2.1: Phénomène de concentration de la métrique euclidienne.

Dans la figure 2.1 on peut observer que, au fur et à mesure que la dimension des données augmente, les normes L_2 de l'ensemble Ω vont se concentrer dans un intervalle beaucoup plus petit que l'intervalle maximal $[0\ M]$. Ceci est connu comme *le phénomène de concentration* de la métrique euclidienne.

Le phénomène de concentration peut être observé de la même manière en analysant la distribution d'autres normes du même ensemble de données X. Dans la figure 2.2, il est aussi évident que les distributions des normes L_5 et $L_{0.9}$ se concentrent dans un intervalle inférieur à l'intervalle $[0\ M]$, mais pas de la même manière ; la distribution L_5 se concentre dans un intervalle qui converge vers 0 avec la dimension tandis que la distribution $L_{0.9}$ se concentre dans un intervalle qui diverge avec la dimension. La distribution des normes euclidiennes se concentre dans un intervalle qui reste constant au fur et à mesure que la dimension des données augmente,

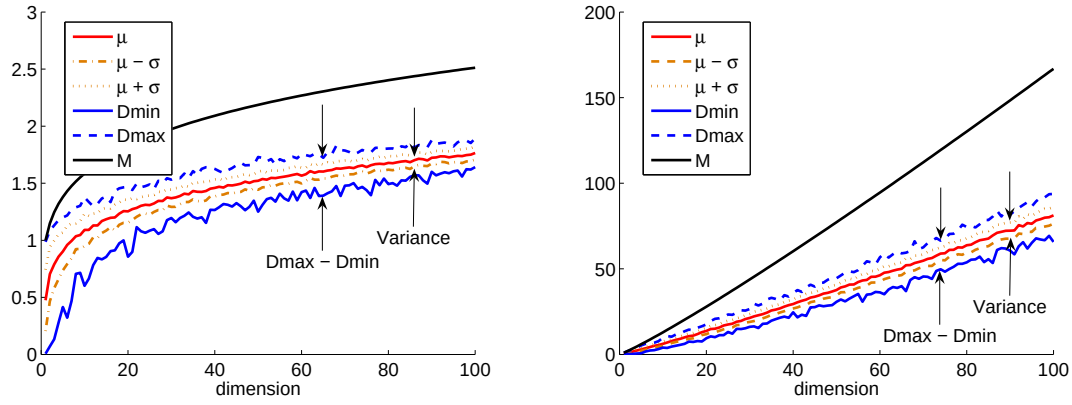


Figure 2.2: Phénomène de concentration des métriques (a) L_5 et (b) $L_{0.9}$.

figure 2.1.

2.3.2 Etat de l'art

Instabilité de la recherche du plus proche voisin dans des espaces multidimensionnels

La recherche du plus proche voisin dans des espaces multidimensionnels est un problème important pour un grand nombre d'applications traitant des données multivariées. De nombreux algorithmes de classification supervisée et non supervisée se basent sur la résolution de ce problème. Nous nous sommes intéressés à ce problème dans le contexte de la classification non supervisée.

Bayer et al. [BGRS00] ont étudié le problème de la recherche du plus proche voisin dans des espaces multidimensionnels et ils montrent que ce problème devient instable lorsque la dimension des données augmente. Ils ont expliqué cet effet par le phénomène de concentration des métriques. Soit $Dmax_d$ la distance entre un point de référence et son voisin le plus éloigné et $Dmin_d$ la distance entre le même point de référence et son voisin le plus proche dans un espace de dimension d (le point de référence est considéré comme l'origine). Alors, pour une distribution de données, sous certaines hypothèses assez générales (l'indépendance des composantes de la distribution de données), la différence entre $Dmax_d$ et $Dmin_d$ n'augmente pas avec la dimension aussi vite que $Dmin_d$. Autrement dit, le rapport $\frac{Dmax_d - Dmin_d}{Dmin_d}$ converge vers 0 avec la dimension des données pour un grand nombre de distributions aussi bien que pour un grand nombre des normes. Cela signifie que le problème de la recherche du plus proche voisin n'a pas de sens car $Dmin_d$ et $Dmax_d$ sont presque identiques. Cette affirmation est exprimée dans [BGRS00] par le théorème suivant :

Théorème 2.1 [BGRS00] *Si* $\lim_{d\to\infty} var\left(\frac{\|X_d\|_r}{E[\|X_d\|_r]}\right) = 0$, *alors* $\frac{Dmax_d^r - Dmin_d^r}{Dmin_d^r} \to_p 0$.

Dans ce théorème, le rapport $\frac{Dmax_d^r - Dmin_d^r}{Dmin_d^r}$ est appelé *fonction de contraste relatif* et il est utilisé comme critère pour montrer la pertinence des métriques pour résoudre le problème de la recherche du plus proche voisin dans des espaces multidimensionnels. Ce résultat théorique est le point de départ pour d'autres travaux qui ont examiné le comportement des métriques de Minkowski et des métriques fractionnaires pour résoudre le problème mentionné dans des espaces de grande dimension.

Concentration des métriques de Minkowski

Selon [FWV07], l'effet de concentration de la métrique euclidienne a été confirmé théoriquement par Demartines [Dem94] mais son étude n'a pas été étendue aux autres métriques.

Hinneburg et al. [HAK99] ont étudié le phénomène de concentration des métriques de Minkowski lié à la recherche du plus proche voisin dans des espaces multidimensionnels. Leur principal résultat est le théorème suivant :

[chapter]

Théorème 2.2 [HAK99] *Soit X une distribution arbitraire de n points et L_r la norme de Minkowski paramétrisée par r. Alors,*

$$C_r \leq \lim_{d\to\infty} E\left[\frac{Dmax_d^r - Dmin_d^r}{d^{1/r-1/2}}\right] \leq (n-1)C_r,$$

où C_r est une constante qui dépend de la métrique L_r.

La conséquence de ce théorème est resumée dans le tableau 2.1.

Métrique	Convergence de $Dmin - Dmax$
L_1	$C_1 * \sqrt{d}$
L_2	C_2
L_r, $r \geq 3$	0

Tableau 2.1: Concentration des métriques de Minkowski.

Les résultats théoriques de Hinneburg et al. [HAK99] montrent que le *contraste absolu* défini par $Dmax_d^r - Dmin_d^r$ augmente avec le facteur $d^{1/r-1/2}$. Ceci signifie que pour la distance de Manhattan, le contraste absolu diverge vers l'infini, pour la métrique euclidienne il converge vers une constante et pour les métriques d'ordre supérieur il converge vers 0, figure 2.3. Ils concluent que pour toutes les métriques L_r avec $r \geq 3$, la recherche du plus proche voisin n'as pas de sens car la valeur maximale de la distance entre toutes les paires de points converge vers la valeur

minimale quand la dimension des données augmente. Dans [FWV07] ceci est vu comme si la métrique avait perdu ses capacitées discriminatoires entre la notion de "près" et de "loin".

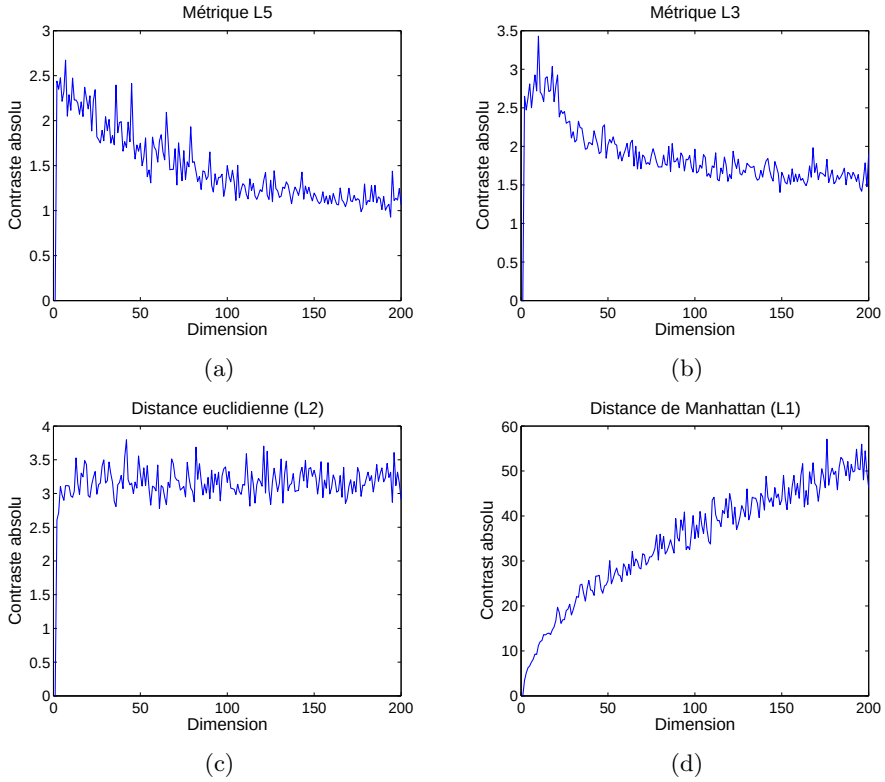


Figure 2.3: Fonction de contraste absolu en fonction de la dimension d des données pour différentes métriques de Minkowski.

Concentration des métriques fractionnaires

Les résultats précédents montrent que la recherche du plus proche voisin dans des espaces multidimensionnels dépend du paramètre r de la métrique. De même ils montrent que la métrique L_1 est supérieure en terme de contraste relatif et absolu aux métriques d'ordre supérieur. Aggarwall et al. [AHK01] étendent ces résultats aux métriques fractionnaires en montrant sur des données synthétiques que ces métriques sont supérieures en terme de contraste absolu et relatif aux métriques de Minkowski et qu'elles améliorent aussi les résultats de la classification obtenus par des algorithmes tels que *C-moyenne*.

Théorème 2.3 [AHK01] *Soit X une distribution uniforme de n points et L_r la famille des normes paramétrisées par $r > 0$. Alors :*

$$\left(\frac{C}{(r+1)^{1/r}}\right)\sqrt{\left(\frac{1}{2r+1}\right)} \leq \lim_{d\to\infty} E\left[\frac{Dmax_d^r - Dmin_d^r}{d^{1/r-1/2}}\right] \leq \left(\frac{C(n-1)}{(r+1)^{1/r}}\right)\sqrt{\left(\frac{1}{2r+1}\right)}.$$

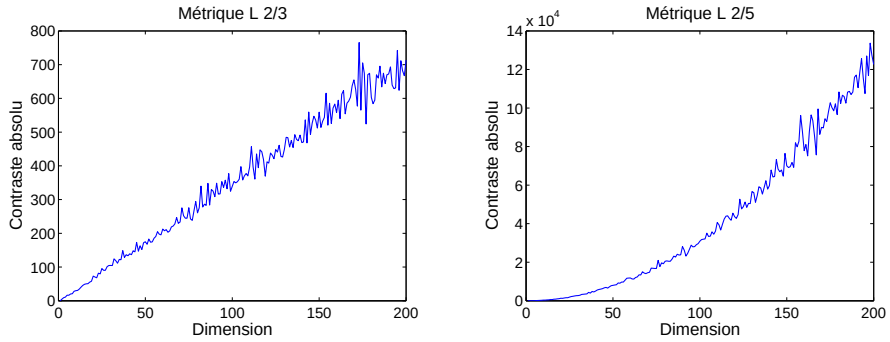
où C est une constante qui ne dépend pas de la métrique L_r.

Théorème 2.4 [AHK01] *Soit F une distribution uniforme de n points et L_r la famille des normes paramétrées par $r > 0$. Alors :*

$$C_r\sqrt{\left(\frac{1}{2r+1}\right)} \leq \lim_{d\to\infty} E\left[\frac{Dmax_d^r - Dmin_d^r}{Dmin_d^r}\right] \leq C_r(n-1)\sqrt{\left(\frac{1}{2r+1}\right)}.$$

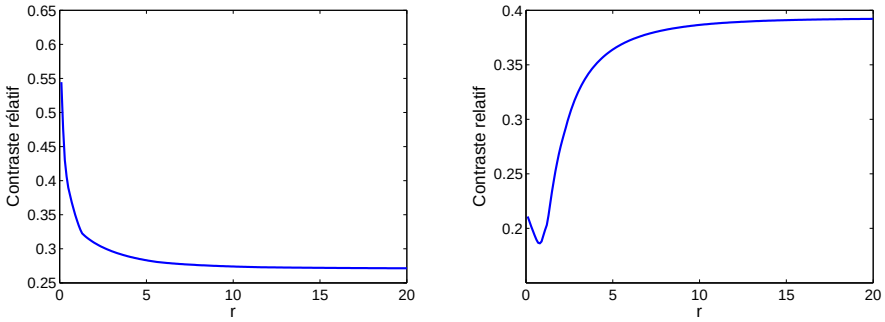
où C_r est une constante qui dépend de la métrique L_r.

Ces deux résultats représentent une extension des résultats de Bayer et al. [BGRS00] et de Hinneburg et al. [HAK99] aux métriques fractionnaires.



(a) (b)

Figure 2.4: Fonction de contraste absolu en fonction de la dimension d des données pour différentes métriques fractionnaires.



(a) (b)

Figure 2.5: Fonction de contraste relatif pour différentes valeurs du paramètre de la métrique r. Les données sont tirées aléatoirement d'une distribution uniforme (a) et d'une distribution normale (b). La dimension des données est $d = 20$ et le nombre de données $n = 1000$.

En étudiant la *fonction de contraste relatif*, figure 2.5 (a) ainsi que le *contraste absolu*, figures 2.4 pour différentes métriques, sur des données tirées aléatoirement d'une distribution uniforme, ils montrent que les métriques fractionnaires offrent un contraste plus élevé que les

métriques d'ordre supérieur, figure 2.3. C'est la raison pour laquelle ils considèrent les métriques fractionnaires comme plus appropriées pour résoudre le problème de classification dans des espaces multidimensionnels.

Mais ces résultats sont valables seulement pour des données uniformes, figure 2.5 (a) ; pour d'autres distributions, la supériorité des métriques fractionaires sur les métriques d'ordre supérieur en terme de contraste absolu ou relatif n'est toujours pas valable comme cela s'observe dans la figure 2.5 (b) où il est montré que, pour des données tirées d'une distribution gaussienne, le contraste relatif des normes d'ordre supérieur dépasse celui estimé par les normes fractionnaires.

Les résultats de Aggarwall et al. [AHK01] ont deux points faibles rendant ceux-ci non applicables sur n'importe quel ensemble de données. D'abord les résultats dépendent de la taille de l'ensemble des données ce qui peut nous faire croire que le phénomène de concentration n'est qu'un phénomène particulier qui se manifeste à cause d'un nombre trop petit de données par rapport à la dimension des données, ou comme une conséquence du phénomène de l'espace vide. Ensuite, la distribution des données est supposée uniforme. Or, dans des problèmes réels de classification, la distribution des données est loin de l'être. Ces deux problèmes ont été abordés dans [FWV07] et sont discutés dans la suite.

La concentration - une propriété intrinsèque de toutes les normes

François et al. [FWV07] s'attaquent à ces deux problèmes ; ils affirment que le phénomène de la concentration est une propriété intrinsèque de toutes les normes en montrant l'indépendance de la concentration de la taille de l'ensemble de données. Leur principal résultat est constitué par le theorème suivant :

Théorème 2.5 [FWV07] *Soit X une distribution quelconque, $X_i = (x_1 \dots x_d)$ un vecteur aléatoire dont les composantes sont indépendamment et identiquement distribuées (i.i.d.) tirées de X et L_r la famille de normes paramétrées par $r > 0$. Alors,*

$$\lim_{d\to\infty} \frac{Var(\|X\|_r)}{E(\|X\|_r)} = 0.$$

Dans ce theorème, le raport $\frac{Var(\|X\|_r)}{E(\|X\|_r)}$ est appelé *variance relative* et a la même signification que le *contraste relatif*. Les auteurs montrent ainsi que toutes les normes paramétrées par r sont affectées par le phénomène de concentration, mais pas toutes de la même manière. Ce théorème est valable seulement si les données sont uniformément distribuées. Pourtant ils montrent que ce théorème reste toujours valable si les données sont normalisées. Cette affirmation peut être

vue comme le fait que, la normalisation des données rend la distribution plus uniforme que la distribution des données brutes.

Un autre résultat de François et al. [FWV07] qui étend le résultat d'Aggarwall et al. [AHK01] est la proposition suivante :

Proposition 1 *La variance relative (ou le contraste relatif) est 1) une fonction strictement décroissante avec le paramètre r seulement si les variables sont distribuées selon une distribution uniforme dans l'intervalle* [0 1]*, et 2) il existe des distributions non uniformes pour lesquelles cela n'est pas vrai. Les normes fractionnaires ne sont pas toujours moins concentrées que les normes d'ordre supérieur.*

Ces résultats montrent que la concentration des normes dépend de la distribution de données et du paramètre de la norme r. Ils concluent que le paramètre r peut être ajusté pour lutter contre le phénomène de concentration mais ils ne font aucune affirmation sur l'impact des normes moins concentrées concernant les résultats des algorithmes de classification non supervisée.

Les résultats obtenus dans [AHK01] ont constitué le point de départ pour d'autres travaux essayant d'utiliser les métriques fractionnaires pour améliorer les résultats de la classification non supervisée des données réelles. Dans [DAD04] et [DAD07] les auteurs ont évalué les performances de plusieurs algorithmes non supervisés (*C-moyennes*, *Neural Gas*, *Growing Neural Gas* et *Self Organising Map*) sur plusieurs ensembles de données réelles en utilisant différentes métriques et ils concluent que les améliorations apportées par les métriques fractionnaires sur les résultats de la classification dépendent fortement de la distribution des données. De même ils montrent que la normalisation des données a comme effet invariable une amélioration des résultats de la classification et réduit aussi l'influence imprédictible des normes. Ils affirment que *les données doivent toujours être normalisées avant la classification et ainsi la métrique euclidienne peut toujours être utilisée sauf s'il y a des raisons particulières (e.g. la présence du bruit coloré) pour justifier l'utilisation d'une autre norme.*

Les métriques fractionnaires ont été appliquées dans le contexte de la recherche des images à base de contenu [HR05]. Les résultats montrent que même si la valeur optimale de la métrique dépend de l'ensemble des données, une métrique fractionnaire est plus efficace que les métriques de Manhattan et euclidienne pour cette application.

François et al. [FWV05] donnent une autre justification de l'emploi des métriques fractionnaires comme mesures de similarité dans le contexte de la classification non supervisée. Ils montrent que celles-ci sont moins sensibles à la présence du bruit non gaussien que les métriques de Minkowski en mettant en évidence leur supériorité dans un problème de classification de

données affectées par un bruit coloré.

2.4 Conclusion

Quelques remarques très pertinentes liées à ce sujet sont présentées dans [FWV07]. Nous finissons ce chapitre en les présentant comme conclusion. Elles représentent le point de départ du 4-ème chapitre où l'hypothèse de la supériorité des métriques moins concentrées pour la classification des données multivariées est testée ; le but est de chercher des indices pour le choix de la métrique optimale.

- le phénomène de concentration est une propriété intrinsèque de toutes les normes paramétrées par $r > 0$ [FWV07] ;
- le contraste relatif ou la variance relative sont des fonctions qui mesurent la concentration des métriques ;
- les métriques fractionnaires sont moins concentrées que les métriques d'ordre supérieur seulement si les données sont uniformement distribuées [AHK01] ; pour d'autres distributions, cela n'est toujours valable [FWV07] ;
- il existe des distributions de données pour lesquelles les métriques d'ordre supérieur sont moins concentrées que les métriques fractionnaires ;
- on peut lutter contre le phénomène de la concentration en ajustant la valeur de r [AHK01] ;
- il n'est pas encore montré (à notre connaissance) qu'une norme moins concentrée donne de meilleurs résultats qu'une métrique plus concentrée dans des problèmes de classification non supervisée ; ceci va être étudié dans le chapitre suivant ;
- il n'existe pas de méthode pour choisir la métrique optimale dans les problèmes de classification non supervisée.

Chapitre 3

Séparation aveugle de sources en vue de la réduction de dimension

Notations

Notation	:	Signification
X	:	Matrice d'observations
X_i	:	Vecteur d'observations
S	:	Matrice des sources
A	:	Matrice des mélange
E	:	Matrice du bruit additif
$x_{i,k}$	:	Valeur du vecteur d'observation X_i à l'instant k
$s_{j,k}$	:	Valeur du vecteur d'observation S_j à l'instant k
m	:	Nombre de vecteurs d'observation
p	:	Nombre de sources
n	:	Nombre d'instants
$a_{i,j}$	:	Coefficient de mélange
$p(X)$	:	Probabilité conjointe de X
$p(x_i)$	:	Probabilité marginale x_i
$E[]$	:	Fonction espérance mathématique
f, g	:	Deux fonctions quelconque
KL	:	Divergence Kullback-Leibler
$H(X)$	:	Entropie différentielle conjointe de X

$H(x_j)$	:	Entropie différentielle marginale de x_i
V	:	Ensemble de variable aléatoire gaussienne
v_i	:	Variable aléatoire gaussienne
σ_i	:	Variance d'une variable aléatoire gaussienne
Σ_i	:	Matrice de covariance de V
$J(x_i)$	:	Néguentropie d'une variable aléatoire
$P_x(u)$	:	Première fonction caractéristique d'une variable aléatoire
$\log P_x(u)$	:	Deuxième fonction caractéristique d'une variable aléatoire
k_i	:	Cumulant d'ordre i d'une variable aléatoire
F	:	Fonction de contraste
P	:	Matrice de permutation
D	:	Matrice diagonale
W	:	Matrice de séparation
U	:	Ensemble de variables aléatoires uniformes
u_i	:	Variable aléatoire uniforme
L	:	Log-vraissemblence
$N = Q(M)$	:	Matrice cumulante associée à une matrice M
W	:	Estimation de la matrice de melange A
H	:	Estimation de la matrice des sources S

3.1 Introduction

Dans de nombreux domaines de la science, de plus en plus d'applications conduisent à l'émergence de données multivariées. Le traitement de ce type particulier de données implique l'utilisation de techniques de fouille de données dont une catégorie importante est représentée par les méthodes de classification non supervisée. Si d'une part, l'aspect multivarié des données peut être vu comme un avantage dans le sens où chaque attribut décrivant les données amène un plus d'information, d'autre part les données multivariées contiennent beaucoup d'information redondante ainsi que du bruit. Ceci fait que l'information pertinente est souvent noyée parmi les attributs révélant un aspect sans intérêt pour l'utilisateur. Par conséquent, l'information redondante peut nuire à la découverte de structures intéressantes dans l'ensemble des données. *Le phénomène de la malédiction de la dimension* connu aussi sous le non de *phénomène de l'espace vide* [ST83] est un autre aspect qui doit être pris en compte par les techniques d'analyse

de données multivariées. La réduction de dimension est une approche utilisée pour contrer les effets indésirables de ces deux phénomènes. On peut résumer les principaux objectifs de la réduction de dimension par :

- réduction de l'information redondante,
- éviter le phénomène de l'espace vide,
- réduction de la complexité et du temps de calcul des algorithmes de classification,
- identification des facteurs pertinents dans l'ensemble des données (*e.g.* des signatures spectrales ou temporelles qui peuvent identifier des composées chimiques ou des régions d'intérêt dans des images multispectrales ou séries temporelles),
- visualisation des données multivariées.

Généralement deux types d'approche sont utilisés pour résoudre ce problème : la *sélection* et *l'extraction des attributs*.

Les méthodes de sélection d'attributs sont mieux adaptées dans le cadre supervisé d'analyse de données multivariées. Néanmoins, le fait d'éliminer complètement des variables peut être considéré comme un inconvénient majeur car on risque de perdre d'information pertinente. Nous nous sommes donc orientés vers les méthodes d'extraction d'attributs, même si ces méthodes imposent un effort pour interpréter et comprendre la nouvelle représentation des données.

Les méthodes d'extraction d'attributs utilisent toute l'information contenue dans l'ensemble de données pour obtenir une nouvelle représentation dans un espace de plus petite dimension. Ces techniques peuvent être linéaires ou non-linéaires ; ce travail est réalisé autour des méthodes linéaires d'extraction d'attributs. Le principe de base de ces techniques est de projeter les données originales dans un espace de dimension plus petite. Une ancienne approche linéaire de réduction de dimension est l'analyse en composantes principales (ACP) [Jol02, Smi02] qui consiste à projeter les points de l'espace original de représentation sur les axes orthogonaux qui maximisent la variance des observations. Des méthodes issues du domaine de traitement du signal, *e.g.* l'analyse en composantes indépendantes (ACI) [Com94, CJ07] proposent de rechercher des directions de projection qui ne sont pas forcément orthogonales. Ces méthodes remplacent la contrainte de décorrélation par celle d'indépendance statistique. Certains auteurs considèrent l'ACI comme une généralisation de l'ACP car la contrainte d'indépendance est plus générale et plus puissante que celle de décorrélation. A l'origine, l'ACI a été proposée pour

résoudre le problème de *séparation aveugle de sources* (SAS). Bien que de nombreuses autres méthodes d'extraction d'attributs existent dans la littérature (les lecteurs sont invités à consulter à ce propos la référence suivante [Fod02]), la suite de ce chapitre est dédiée aux méthodes de SAS afin de les tester et de les utiliser dans le contexte de la réduction de dimension de données multivariées.

3.2 Séparation aveugle de sources : principes et méthodes

3.2.1 Introduction à la séparation aveugle de sources

La séparation aveugle de sources est un problème fondamental dans le domaine du traitement du signal ; la résolution de ce problème implique la prise en compte de contraintes assez variées sur les sources : l'orthogonalité, l'indépendance, la non-négativité etc. Les techniques de SAS reposent sur l'hypothèse que le signal observé est un mélange de plusieurs signaux appelés *signaux sources* ; le but de la SAS est de retrouver les signaux sources à partir de plusieurs *observations* en tenant compte de toute l'information disponible sur les signaux observés et sur le processus de mélange. Même si à l'origine cette technique a été utilisée dans le domaine du traitement du signal et de la parole, elle se retrouve dans de nombreux autres domaines : dans l'imagerie multivariée (microscopie, télédétection) où les pixels sont considérés comme des mélanges de spectres spécifiques de différentes composantes pures présentes dans la substance, en communications numériques, en sonar et radar où les signaux provenant de plusieurs émetteurs ou réflecteurs interfèrent au niveau des antennes de réception. Dans ce chapitre nous présentons les principales méthodes de SAS et nous mettons en évidence les points forts de chacune des techniques afin de les utiliser pour la réduction de dimension de données multivariées appliquée à l'imagerie multivariée.

3.2.2 Principe de la séparation de sources

La *séparation aveugle de sources* (SAS) inclut un ensemble de techniques qui, à partir de plusieurs observations du même phénomène physique, permettent d'obtenir des répliques proportionnelles aux signaux sources ainsi que la contribution de chaque source aux observations. Dans [Mou05], le problème de SAS est divisé en deux sous-problèmes : (1) l'identification du mélange et (2) la reconstitution des sources. Sans aucune information "*a priori*" sur les signaux sources et sur le processus de mélange, ce problème admet une infinité de solutions. Des hypothèses supplémentaires doivent être introduites afin d'obtenir une solution unique et adéquate à ce

problème.

Un point de vue souvent adopté pour résoudre le problème de SAS est celui de la décomposition des signaux observés sur une base de signaux élémentaires permettant d'éliminer la redondance d'information entre les différentes observations. Plusieurs mesures pour estimer la redondance peuvent être définies introduisant ainsi des contraintes sur les composantes recherchées. La contrainte d'orthogonalité aboutit à l'Analyse en Composantes Principales (ACP), tandis que la contrainte d'indépendance statistique des sources aboutit à l'Analyse en Composantes Indépendantes (ACI). D'autres hypothèses prennent en compte des contraintes spécifiées par l'application : on peut retrouver des hypothèses sur la distribution des sources ou des vecteurs de mélange, leur structure temporelle ainsi que d'autres contraintes telles que *la parcimonie* ou *la non-négativité*.

Modèle du mélange

Un point essentiel dans la séparation de sources est le choix d'un modèle de mélange décrivant la relation entre les sources et les observations. Celui-ci peut être linéaire ou non linéaire, convolutif ou instantané, variant ou invariant dans le temps. Le modèle le plus utilisé est le modèle linéaire instantané invariant dans le temps car c'est un modèle simple dont les applications sont nombreuses.

Ce modèle suppose qu'à chaque instant k, $k = 1 : n$ les m observations $\left\{x_{(i,k)}\right\}_{i=1}^{m}$ sont des mélanges linéaires instantanés de p sources $\left\{s_{(j,k)}\right\}_{j=1}^{p}$:

$$x_{(i,k)} = \sum_{j=1}^{p} a_{ij} s_{(j,k)} + e_{(i,k)}, \text{pour } i = 1, \cdots, m \tag{3.1}$$

où $a_{ij} \in \mathbb{R}$ pour $i \in 1, \ldots, m$ et $j \in 1, \ldots, p$ sont les coefficients de mélange.

Sous la forme matricielle, ce modèle de mélange s'exprime par :

$$X = AS + E \tag{3.2}$$

où $X \in \mathbb{R}^{m \times n}$ est la matrice des observations, $A \in \mathbb{R}^{m \times p}$ est la matrice de mélange, $S \in \mathbb{R}^{p \times n}$ est la matrice des signaux source et $E \in \mathbb{R}^{m \times n}$ est la matrice du bruit additif.

Indéterminations

Sans information *a priori* sur le mélange ou sur les sources, une identification complète des matrices A et S est impossible. Pourtant, même si des contraintes sur les sources ou sur le

mélange sont prises en compte, deux indéterminations persistent : (1) *l'indétermination d'échelle* c'est-a-dire que les sources estimées sont des répliques proportionnelles à un facteur d'échelle des sources originales, et (2) *l'indétermination d'ordre*, signifiant que l'indice associé à chaque source est arbitraire car les sources ne seront connues qu'à une permutation près des vraies sources.

3.2.3 Méthodes de séparation aveugle de sources

La résolution du problème de SAS nécessite de formuler des hypothèses supplémentaires sur la solution recherchée, hypothèses fondées sur la prise en compte des informations sur le mélange et sur les sources. En fonction de ces hypothèses, les méthodes de SAS peuvent être regroupées en trois catégories : des méthodes basées sur l'hypothèse d'orthogonalité, des méthodes basées sur l'hypothèse de l'indépendance statistique et des méthodes basées sur la prise en compte de la non-négativité des sources. Une catégorie à part de méthodes de SAS, basée sur l'interprétation géométrique du modèle de mélange linéaire est représentée par les méthodes géométriques. Une taxonomie des méthodes de SAS pour le modèle de mélange linéaire instantané peut être celle présentée dans le tableau 3.1.

Méthodes de séparation	**Algorithmes**	**Contraintes**
Séparation par décorrelation	ACP, SOBI etc.	Décorrelation des sources
Séparation par ACI	FastICA, JADE, InfoMax, FOBI etc.	Indépéndance des sources
Séparation par la prise en compte de la non-négativité	NMF, Sparse NMF, ALS etc.	Non-négativité des sources et de la matrice de mélange
Séparation par approches géométriques	SAS géométrique	Les *fdp* des sources sont bornées

Tableau 3.1: Présentation des méthodes de SAS pour le modèle de mélange linéaire instantané.

3.3 Séparation par analyse en composantes indépendantes

Le principe des méthodes de séparation par ACI est d'appliquer une transformation aux signaux observés afin d'obtenir des signaux statistiquement indépendants. Dans [Com94] il est montré que l'utilisation de l'indépendance statistique comme hypothèse de séparation ne garantit pas l'unicité de la solution sauf dans le cas sur-déterminé ($m \geq p$) et qu'au plus une source a une

distribution gaussienne. Dans cette section les mesures d'indépendance statistique ainsi que quelques algorithmes résultant de l'utilisation de ces mesures sont présentés.

3.3.1 Mesures de l'indépendance statistique

Avant de présenter les principales mesures utilisées pour la séparation de sources par ACI nous introduisons la définition de l'indépendance statistique des variables aléatoires.

Définition - indépendance statistique des variables aléatoires

Soit $X = \{x_j\}_{j=1}^{n}$ un ensemble de n variables aléatoires. Les variables x_j sont dites statistiquement mutuellement indépendantes si et seulement si :

$$p(x_1, x_2, \cdots, x_n) = \prod_{j=1}^{n} p(x_j) \tag{3.3}$$

Propriétés des variables aléatoires statistiquement indépendantes :

1. si deux variables aléatoires x_1 et x_2 sont statistiquement indépendantes, alors:

$$E\left[f(x_1)g(x_2)\right] = E\left[f(x_1)\right] E\left[g(x_2)\right], \tag{3.4}$$

   $\forall f, g$.

2. l'indépendance statistique implique la décorrélation, mais l'inverse n'est pas toujours vrai, sauf si les variables ont une distribution gaussienne.

Dans la suite nous présentons les principales mesures d'indépendance statistique des variables aléatoires.

Divergence de Kullback-Leibler

La divergence de Kullback-Leibler [Kul59] permet d'estimer l'indépendance mutuelle de variables aléatoires en mesurant la distance entre leur densité de probabilité. Soit $X = [x_1, \ldots, x_n]$ un vecteur de n variables aléatoires, alors la divergence de Kullback-Leibler est définie par :

$$KL\left(p(X), \prod_{j=1}^{n} p(x_i)\right) \triangleq \int_{\Re^n} p(X) \log \frac{p(X)}{\prod_{j=1}^{n} p(x_j)} \tag{3.5}$$

Propriétés :

1. la divergence KL est non-négative et n'est nulle que lorsque les variables x_j sont statistiquement indépendantes ;

2. la divergence KL est invariante par permutation ou par changement d'échelle. Par conséquent, l'annulation ou la minimisation de cette fonction est un critère pertinent pour séparer les sources.

Information mutuelle

L'indépendance des variables aléatoires peut être egalement estimée en utilisant l'information mutuelle exprimée par :

$$I(X) = \sum_{j=1}^{n} H(x_j) - H(X) \tag{3.6}$$

où:

$$H(X) = -\int_{\Re^n} p(X) \log p(X) dX = -E[\log p(X)] \tag{3.7}$$

$$H(x_j) = -\int_{\Re} p(x_j) \log p(x_j) dx = -E[\log p(x_j)] \tag{3.8}$$

sont respectivement, les entropies conjointe et marginale de X et x_j. L'information mutuelle est liée à la divergence de Kulback-Leibler. La séparation des sources est réalisée par la minimisation de ce critère.

Néguentropie

La néguentropie est définie comme étant une mesure de l'éloignement entre la distribution d'une variable aléatoire x_j et la distribution d'une variable gaussienne. Soit $V = v_1, \ldots, v_n$ un ensemble de variables aléatoires tirées d'une distribution gaussienne pour lesquelles on peut écrire :

$$H(v_j) = \frac{1}{2}\left(\log \sigma_j^2 + \log 2\pi + 1\right) \tag{3.9}$$

$$H(V) = \frac{1}{2}\left(\log \det \Sigma + p(1 + \log 2\pi)\right) \tag{3.10}$$

où σ_j^2 et Σ représentent, respectivement, la variance et la covariance de v_j et V. Si les variables x_i et v_j ont la même variance, alors la néguentropie est donnée par :

$$J(x_j) = H(v_j) - H(x_j) \tag{3.11}$$

Cette mesure est toujours positive (l'entropie d'une variable gaussienne est maximale pour une variance donnée) et n'est nulle que lorsque la variable x_j est gaussienne. Dans [Hyv99a, HO00,

HK01] il est montré que la maximisation de la néguentropie correspond à la recherche de composantes non-gaussiennes, ce qui, d'après le théorème de la limite centrale tend à rechercher des composantes indépendantes.

Statistiques d'ordre supérieur

Les statistiques d'ordre supérieur sont aussi employées pour évaluer l'indépendance statistique de variables aléatoires non-gaussiennes [DLM07, PM01, CRMP07]. Afin d'introduire les statistiques d'ordre supérieur d'une variable aléatoire, nous allons définir ses fonctions caractéristiques. Pour une variable aléatoire x, la première fonction caractéristique est décrite par la transformée de Fourier de la densité de probabilité $p(x)$:

$$P_x(u) = E(\exp(jux)) = \int_{-\infty}^{+\infty} p_x(x) \exp(jux) dx \tag{3.12}$$

$$= \int_{-\infty}^{+\infty} p_x(x) \left(1 + jux + \frac{j^2u^2x^2}{2!} + \frac{j^3u^3x^3}{3!} + \frac{j^4u^4x^4}{4!} + \dots\right) dx \tag{3.13}$$

$$= 1 + juE(x) + \frac{j^2u^2}{2!}E(x^2) + \frac{j^3u^3}{3!}E(x^3) + \frac{j^4u^4}{4!}E(x^4) + \dots \tag{3.14}$$

où $E(x^k)$ représente le moment d'ordre k de la variable x. La deuxième fonction caractéristique d'une variable aléatoire est décrite par le logarithme de la première fonction caractéristique :

$$\log P_x(u) = k_1u + k_2\frac{u^2}{2!} + k_3\frac{u^3}{3!} + k_4\frac{u^4}{4!} + \dots \tag{3.15}$$

où les coefficients k_i s'appellent les cumulants de la distribution de la variable x. Pour une variable aléatoire de moyenne nulle, les trois premiers cumulants sont définis par :

$$k_1 = E(x) \tag{3.16}$$

$$k_2 = E(x^2) \tag{3.17}$$

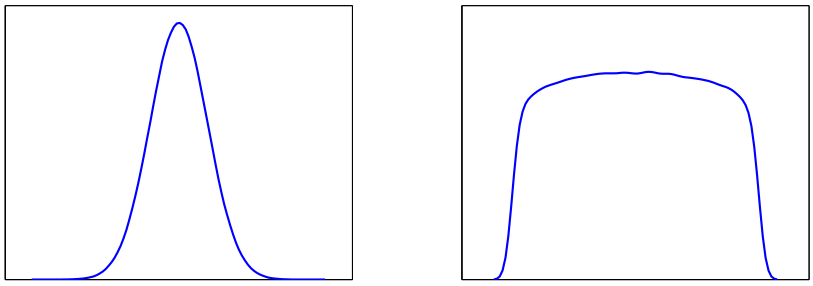
$$k_3 = E(x^3) \tag{3.18}$$

et ils représentent *la moyenne*, *la variance* et *l'assymétrie (ou skewness)* de la distribution d'une variable aléatoire. Le cumulant d'ordre 4 ou le Kurtosis est un cas particulier :

$$k_4 = E(x^4) - 3(E(x^2))^2 \tag{3.19}$$

Une observation souvent utilisé pour obtenir l'indépendance statistique des variables aléatoires est le fait que deux variables aléatoires non gaussiennes sont indépendantes. Chercher l'indépen-

dance statistique revient donc à chercher des variables non gaussiennes. Le kurtosis est considéré comme une mesure de la non-gaussianité de la distribution d'une variable aléatoire x. Pour une distribution gaussienne $k_4 = 0$, pour une distribution sur-gaussienne $k_4 > 0$, figure 3.1 (a) et pour une distribution sous-gaussienne $k_4 < 0$, figure 3.1 (b). Dans la littérature, les cumulants sont utilisés soit d'une façon directe pour construire des mesures d'indépendance [PM01], soit comme des outils d'approximation d'autres mesures d'indépendance telle que la néguentropie [Com94, Car99, HO97].



(a) (b)

Figure 3.1: a) *Fdp* d'une distribution sur-gaussienne, $k_4 > 0$ et b) *fdp* d'une distribution sous-gaussienne, $k_4 < 0$.

3.3.2 Méthodes de séparation par ACI

Généralement, les méthodes de séparation par ACI sont fondées sur la minimisation d'une fonction de contraste. La notion de fonction de contraste est introduite pour la première fois dans le domaine de la séparation de sources par Comon [Com94]. Il définit une fonction de contraste de la manière suivante : une fonction F est dite fonction de contraste pour la matrice de vecteurs aléatoires X si elle vérifie les conditions suivantes :

1. $F(PX) = F(X)$, $\forall P$, matrice de permutation,
2. $F(DX) = F(X)$, $\forall D$, matrice diagonale,
3. si les composantes de X sont indépendantes alors :
   - $F(MX) \leq F(X)$, $\forall M$ matrice inversible
   - $F(MX) = F(X) \Leftrightarrow M = DP$.

En exploitant la richesse des mesures d'indépendance statistique des variables aléatoires, plusieurs fonctions de contraste peuvent être dérivées ce qui conduit à de nombreux critères de séparation. Dans la suite, nous présentons quelques critères de séparation en montrant le lien entre eux (lorsqu'il y en a un).

ACI par minimisation de l'information mutuelle

L'information mutuelle équation (3.6), est une mesure de la dépendance des variables aléatoires issue de la théorie de l'information. Ce critère est utilisé comme fonction de contraste pour trouver les composantes indépendantes dans une matrice d'observations.

Lien avec la néguentropie

Une propriété importante de l'information mutuelle est que, pour une transformation linéaire inversible $s = Wx$ on peut écrire :

$$I(s_1, s_2, ..., s_n) = \sum_i H(s_i) - H(x) - log|detW| \tag{3.20}$$

Si s_i sont décorrelées et de variance unitaire, il est possible d'écrire $E\left\{ss^T\right\} = WE\left\{xx^T\right\}W^T = I$, impliquant :

$$\det I = 1 = \left(\det WE\left\{xx^T\right\}W^T\right) = (\det W)\left(\det E\left\{xx^T\right\}\right)\left(\det W^T\right) \tag{3.21}$$

et ceci implique que $\det W = \frac{1}{\sqrt{det(Exx^T)}}$. La néguentropie d'une variable s est définie par :

$$J(s) = H(y_{gauss}) - H(s) \tag{3.22}$$

Dans [Hyv99a] il est montré que la minimisation de l'information mutuelle correspond à la maximisation de la néguentropie. Une approximation de la néguentropie par des cumulants d'ordre supérieur est le critère à maximiser utilisé par la première version de l'algorithme FastICA [Hyv99a]. Des versions ultérieures de cet algorithme utilisent des estimateurs plus robustes de la néquentropie. Pour plus d'information nous invitons le lecteur à consulter les références [Hyv99a, HO00].

ACI par InfoMax

L'ACI par le principe InfoMax a été développée par A. J. Bell et T.J. Sejnowski dans [BS95] ; cette méthode se base sur l'observation que la minimisation de l'information mutuelle $I(u)$ entre

un vecteur u et ses composantes uniformes $u_1, \ldots, u_n$ est équivalente à la maximisation de l'entropie de u, (l'entropie d'une variable uniforme étant nulle) :

$$I(u) = \sum_{i=1}^{n} H(u_i) - H(u) = -H(u) \tag{3.23}$$

En d'autres termes, le principe de l'ACI par InfoMax consiste à trouver une matrice de séparation W de sorte que les sources soient les plus uniformes possible.

ACI par maximum de vraisemblance

Une approche très répandue pour résoudre le problème de séparation de sources par l'ACI est obtenue par la maximisation de la vraisemblance. Cette méthode est proposée par [Pha96]. Soit $W = (w_1, \ldots, w_n)^T = A^{-1}$, la log-vraisemblance est donnée par :

$$L = \sum_{i=1}^{n} \sum_{j=1}^{m} \log f_j(w_j^T x(i)) + nlog|\det W| \tag{3.24}$$

où f_i représentent les densités de probabilité des sources (supposées connues).

Lien avec l'information mutuelle Afin de mettre en évidence le lien avec l'information mutuelle, considérons l'espérance mathématique de la log-vraisemblance :

$$\frac{1}{n} E\{L\} = \sum_{i=1}^{m} E\left\{\log f_i(w_i^T x)\right\} + \log|\det W| \tag{3.25}$$

Si $f_i(w_i^T x)$ sont les vraies densités de sources, on peut écrire :

$$E\left\{\log f_i(w_i^T x)\right\} = -\sum_i H(w_i^T x) \tag{3.26}$$

et donc on peut écrire la log-vraisemblance :

$$\frac{1}{n} E\{L\} = \log|\det W| - \sum_i H(w_i^T x) \tag{3.27}$$

La maximisation de la log-vraisemblance est donc équivalente à la minimisation de l'information mutuelle. L'inconvénient de cette méthode est que les densités des sources doivent être estimées correctement. Si l'information sur la nature des composantes indépendantes n'est pas correcte cette méthode donne des résultats complètement erronés. Ce problème n'apparaît pas si une mesure de la non-gaussianité des sources est utilisée.

ACI par statistiques d'ordre supérieur

Ces méthodes utilisent comme mesure d'indépendance les statistiques d'ordre supérieur, plus précisément, le kurtosis d'une variable aléatoire. Dans [Car89], Cardoso a introduit la notion de matrice cumulante $N = Q_x(M)$ associée à une matrice M de dimension $n \times n$:

$$N = Q_x(M) \Longleftrightarrow \left\{ N_{ij} = \sum_{k=1}^{n} \sum_{l=1}^{n} Q_{ij}^{kl} M_{lk} | 1 \leqslant i, j \leqslant n \right\} \tag{3.28}$$

où :

$$Q_x \triangleq \left\{ Q_{ij}^{kl} = Cum[x_i, x_j^*, x_k, x_l^*] | 1 \leq i, j, k, l \leq n \right\} \tag{3.29}$$

Nous allons mentionner deux algorithmes utilisant la matrice cumulante pour la séparation de sources : l'algorithme FOBI (Forth Order Blind Identification) [Car89, Car92] dont le principe est de diagonaliser une seule matrice cumulante $Q(M)$ pour obtenir une matrice de rotation unitaire assurant l'indépendance des sources estimées et l'algorithme JADE (Joint Approximate Diagonalisation of Eigen-matrices) [CS93] dont le principe est la diagonalisation conjointe de plusieurs matrices cumulantes. Nous rappelons l'algorithme JADE qui peut être résumé en quatre étapes :

Algorithme 6 L'algorithme JADE

1: Calculer la matrice de blanchiment B permettant d'obtenir une matrice d'observation X blanche,
2: Calculer le tenseur cumulant d'ordre quatre de la matrice d'observations blanchies $Z = \hat{B}X$; calculer les matrice propres les plus significatives du tenseur cumulant,
3: Diagonalisation jointe de ces matrices pour obtenir la matrice de rotation unitaire $\hat{U}$ assurant l'indépendance de sources,
4: Estimation de la matrice de mélange $\hat{A} = \hat{B}^T \hat{U}$.

Pour plus d'informations nous vous invitons de consulter les références [CS93].

ACI par approche bayesienne

L'approche bayesienne a été introduite pour la première fois dans la séparation de sources par [Rob98, Knu98, MD99]. Le schéma général de l'approche bayesienne est présentée dans [CJ07] et il peut être résumé par des étapes suivantes :

Algorithme 7 L'approche bayesienne pour l'ACI

1: Décrire le modèle de mélange et en déduire la loi $p(X|A,S)$, appelée la vraisemblance des inconnues,
2: Attribuer des lois *a priori* à toutes les inconnues du problème (aux sources $S \rightarrow p(S)$ et à la matrice de séparation $A \rightarrow p(A)$),
3: En utilisant la règle de Bayes, déduire la loi *a posteriori* $p(A,S|X)$,

$$p(A,S|X) = \frac{p(X|A,S)p(A)p(S)}{p(X)} \propto p(X|A,S)p(A)p(S) \tag{3.30}$$

où $p(X) = \int p(X|A,S)p(A)p(S)dS$
4: Utiliser cette loi *a posteriori* pour définir une solution ou un ensemble de solutions pour le problème de séparation.

La résolution du problème de séparation de sources par l'approche bayesienne peut être abordée par trois directions [CJ07] :

1. Estimation jointe des sources S et de la matrice de mélange A,

2. Estimation de la matrice de mélange A,

3. Estimation des sources S.

Pour plus de détails, nous invitons les lecteurs à consulter les références [CJ07].

3.4 Séparation par la prise en compte de la non-négativité

La non-négativité des sources et/ou des coefficients de mélange est une contrainte imposée par des applications traitant des phénomènes physiques réels décrits par des grandeurs physiques non-négatives telles que la température, la longueur d'onde, le champ gravitationnel, la masse etc. Les méthodes de séparation par décorrélation ou par ACI présentées ne sont pas adaptées pour prendre en compte cette contrainte. La solution donnée par celles-ci présente souvent des coefficients négatifs ce qui rend impossible l'interprétation des résultats de la séparation. De plus, la solution n'est unique que dans le cas où au plus une source est gaussienne. Les méthodes de séparation par la prise en compte de la non-négativité s'affranchissent de ces inconvénients, sauf celui de l'unicité de la solution. Il est à noter que la plupart de ces méthodes de séparation ne font aucune supposition sur les statistiques des sources et/ou de la matrice de mélange, seul la contrainte de non-négativité étant prise en compte. Même si des méthodes algébriques existent pour résoudre le problème de séparation sous la contrainte de non-négativité, la plupart sont

fondées sur le critère des moindres carrés sous la contrainte de non-négativité. Dans la suite nous présentons quelques méthodes.

3.4.1 Méthode PMF

La méthode PMF (Positive Matrix Factorization) [PT94, Paa97] est la toute première à s'adresser au problème de factorisation en matrices non-négatives, bien que ce travail soit rarement cité par des auteurs. Le but de la méthode PMF était de réaliser une analyse factorielle des données représentant des observations de l'environnement ; ceci implique de trouver un nombre minimal de *causes* (facteurs) qui puissent expliquer l'ensemble d'observations. L'idée de base est que dans des circonstances réelles, un facteur est présent (et donc il a une contribution positive) ou absent (et sa contribution est nulle). La contrainte de non-négativité des facteurs a donc du sens. Soit X la matrice d'observations, W la matrice de facteurs et H leur contribution, la méthode d'analyse factorielle proposée par Paatero [PT94] consiste à minimiser la fonction objective suivante :

$$f(W,H) = \|X - WH\|^2 \tag{3.31}$$

sous les contraintes $W \geq 0$ et $H \geq 0$. Dans sa version originale, l'algorithme ALS (Alternative Least Square) a été proposé pour l'optimisation. Le pseudocode de cet algorithme est présenté dans la suite :

Algorithme 8 Le pseudocode de l'algorithme ALS

Input : A
$W = rand(m,k)$
for i = 1 : maxiter **do**
  (LS) Resoudre $W^TWH = W^TA$ pour trouver H
  (NONNEG) Fixer toutes les composantes négatives de H à 0
  (LS) Resoudre $HH^TW^T = HA^T$ pour trouver W
  (NONNEG) Fixer toutes les composantes négatives de W à 0
end for

En ce qui concerne la convergence de l'algorithme, Paatero [Paa99] affirme que celle-ci est assurée par le fait que les étapes 4 et 6 de l'algorithme sont des problèmes convexes ; pourtant, la solution n'est pas unique. En fonction de son implémentation, cet algorithme peut être très rapide. Des améliorations de l'algorithme sont proposées par Paatero dans [Paa99].

3.4.2 Méthode NMF

L'algorithme NMF (Non-negative Matrix Factorization) proposé par Lee et Seung [LS99] est conçu pour fournir une représentation *"en parties"* d'une matrice d'observation X. Si la matrice d'observation est non-négative, l'algorithme NMF consiste à trouver les matrices W et H non-négatives qui approximent au mieux la matrice d'observation :

$$X \approx WH \tag{3.32}$$

Dans [LS00] sont proposées deux formulations de l'algorithme NMF :

1. Minimiser l'erreur quadratique entre la matrice d'observations X et l'estimation de la matrice des observations $\hat{X} = WH$ par rapport à W et H sous la contrainte de non-négativité,

$$\min_{W\geq 0, H\geq 0} \|X - WH\|^2 \tag{3.33}$$

   Les règles de mise à jour qui assurent la convergence de l'erreur quadratique vers son minimum sont :

$$H^+_{a\mu} \leftarrow H_{a\mu} \frac{(W^T X)_{a\mu}}{(W^T WH)_{a\mu}} \tag{3.34}$$

$$W^+_{ia} \leftarrow W_{ia} \frac{(XH^T)_{ia}}{(WHH^T)_{ia}} \tag{3.35}$$

2. Minimiser la divergence Kulback-Leibner entre la matrice d'observations X et son estimation $\hat{X} = WH$ par rapport à W et H sous la contrainte de non-négativité.

$$\min_{W\geq 0, H\geq 0} \left(X_{ij} \log \frac{X_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij} \right) \ \forall i,\, j \tag{3.36}$$

   Les règles de mise à jour qui assurent la convergence de la divergence Kulback-Leibler vers son minimum sont :

$$H^+_{a\mu} \leftarrow H_{a\mu} \frac{\sum_i W_{ia} X_{i\mu}/(WH)_{i\mu}}{\sum_k W_{ka}} \tag{3.37}$$

$$W^+_{ia} \leftarrow W_{ia} \frac{\sum_\mu H_{a\mu} X_{i\mu}/(WH)_{i\mu}}{\sum_\nu H_{a\nu}} \tag{3.38}$$

où les deux indices représentent la ligne et la colonne des matrices employées dans les notations. La méthode NMF est simple à implémenter. Pourtant, l'inconvénient majeur de cette méthode est la non-unicité de la solution ainsi que la vitesse de convergence. Le pseudocode de cet algorithme est présenté dans la suite :

Algorithme 9 Le pseudocode de l'algorithme NMF

```
W = rand(m, k)
H = rand(m, k)
for i = 1 : maxiter do
    (Mise à jour H) Eq. 3.34 ou 3.37
    (Mise à jour W) Eq. 3.35 ou 3.38
end for
```

3.4.3 Méthode de factorisation sous des contraintes auxiliaires

Le problème de factorisation en matrices non-négatives peut être étendu pour inclure des contraintes supplémentaires sur les matrice W et H. Ceci permet d'imposer les connaissances *a priori*, ou d'obtenir des solutions avec une certaine particularité.

Factorisation par moindres carrés pénalisés

Un exemple des contraintes supplémentaires est représenté par les termes de pénalisation inclus dans la fonction de coût 3.4.1 :

$$f(W,H) = \|X - WH\|^2 + \alpha J_1(W) + \beta J_2(H) \tag{3.39}$$

où $J_1(W)$ et $J_2(H)$ sont introduits afin de prendre en compte les contraintes imposées par l'application et les paramètres α et β règlent le compromis entre l'erreur d'approximation et les contraintes.

Factorisation sous la contrainte de parcimonie

La contrainte de parcimonie sur les matrices W et/ou H peut être également imposée. La notion de parcimonie fait référence à une représentation où seulement une partie des attributs est utilisée pour décrire les vecteurs de données [Hoy02, Hoy04]. Plusieurs mesures sont proposées pour mesurer la parcimonie des variables aléatoires. Parmi celles-ci nous rappelons les normes l^p proposées par [KC03] et la mesure proposée par Hoyer en [Hoy04] :

$$sparseness(x) = \frac{\sqrt{n} - \frac{\|x\|_1}{\|x\|_2}}{\sqrt{n} - 1} \tag{3.40}$$

où $\|x\|_1$ et $\|x\|_2$ représentent les normes L_1 et L_2 d'une variable aleatoire x. La contrainte de parcimonie est ainsi incluse dans la fonction de coût sous la forme de termes de pénalité comme on peut le voir dans l'équation 3.39. Les nouvelles règles de mise à jour sont données par :

$$W_{ia}^{+} \leftarrow W_{ia} \frac{(VH^T)_{ia}}{(WHH^T)_{ia} + \alpha \frac{\partial J_1(W)}{\partial w_{ij}}} \tag{3.41}$$

$$H^{+}_{a\mu} \leftarrow H_{a\mu}\frac{(W^T X)_{a\mu}}{(W^T W H)_{a\mu} + +\beta\frac{\partial J_2(H)}{\partial h_{ij}}} \tag{3.42}$$

3.5 Séparation par des approches géométriques

Les méthodes géométriques représentent des outils très intéressants pour résoudre le problème de séparation aveugle de sources. L'attractivité de ces méthodes réside dans leur description visuelle ainsi que dans la facilité de leur mise en oeuvre. L'idée de base des approches géométriques est le fait que le modèle de mélange linéaire n'est rien d'autre qu'une transformation géométrique des données de l'espace de sources vers l'espace des observations.

La première approche géométrique pour résoudre le problème de séparation de sources a été proposée par Puntonet en [Pun95, PPJ^{+}95, PMJ95]. L'algorithme proposé consiste à trouver les pentes des arêtes du parallélogramme renfermant les observations qui représentent les paramètres de la matrice de démélange. Considérons le cas d'un mélange de 2 sources, l'algorithme peut être résumé en deux étapes :

Algorithme 10 L'algorithme géométrique de base

1: Calculer la nouvelle origine des observations par la relation suivante :

$$O^{'} = (x_1(t_0), x_2(t_0)) \tag{3.43}$$

où $t_0 = argmax_t(x_1^2(t) + x_2^2(t))$

2: Estimer les pentes des arrêtes du parallélogramme et les points se trouvant sur les arrêtes :

$$r_{min} = \min_t \left(\frac{x_2(t) - x_2(t_0)}{x_1(t) - x_1(t_0)}\right) \rightarrow X_{min}(x_1(t_{min}), x_2(t_{min})) \tag{3.44}$$

$$r_{max} = \max_t \left(\frac{x_2(t) - x_2(t_0)}{x_1(t) - x_1(t_0)}\right) \rightarrow X_{max}(x_1(t_{max}), x_2(t_{max})) \tag{3.45}$$

où $t_{min} = argmin_t \left(\frac{x_2(t)-x_2(t_0)}{x_1(t)-x_1(t_0)}\right)$ et $t_{max} = argmax_t \left(\frac{x_2(t)-x_2(t_0)}{x_1(t)-x_1(t_0)}\right)$

La matrice de démélange est $W^{-1} = \begin{pmatrix} x_1(t_{min}) & x_1(t_{max}) \\ x_2(t_{min}) & x_2(t_{max}) \end{pmatrix}$.

L'avantage de cette méthode est sa simplicité. Pourtant elle n'est applicable que si les sources sont tirées d'une *fdp* bornée et que le nombre maximal de sources est égal à 2 ; néanmoins cette méthode est sensible au bruit. En [PP98, MPO01], les auteurs ont adapté cette méthode au cas où, les observations représentent le mélange linéaire de plusieurs sources. La qualité de la séparation dépend de la probabilité d'avoir des points près des arrêtes de l'hyperpolyèdre dans l'espace d'observations.

3.6 Conclusions

Dans ce chapitre nous avons présenté quelques méthodes de séparation de sources, l'objectif étant d'étudier quels sont les avantages apportés par les différentes approches de séparation dans le contexte de la réduction de données multivariées. Nous avons mis en évidence quelques points importants : tout d'abord, les approches étudiées (la séparation par ACI, la prise en compte de la non-négativité et les approches géométriques) amènent deux avantages majeurs dans le cadre de l'analyse exploratoire des données multivariées. Le premier est la réduction implicite de la dimension et le deuxième est la mise en évidence des structures cohérentes dans l'ensemble des données. Ceci, peut s'avérer très utile dans différentes applications. Quelques commentaires peuvent être écrits sur les méthodes de séparation de sources étudiées avant de clore ce chapitre :

- Les méthodes de séparation par ACI utilisent une contrainte très générale afin de réaliser la séparation et sont les plus efficaces parmi les algorithmes de séparation. Pourtant, l'apparition de coefficients négatifs représente un inconvénient important car beaucoup d'applications imposent des contraintes de non-négativité sur les sources et/ou la matrice de mélange et les résultats obtenus avec ces méthodes ne sont pas interprétables ;

- Les méthodes de séparation par la prise en compte de la non-négativité enlèvent cet inconvénient mais se montrent moins efficaces en terme de séparation ; pourtant, le fait de prendre en compte des contraintes de l'application (là où la contrainte de non-négativité s'impose) et de pouvoir mettre en évidence des facteurs ayant un sens physique réel les rend plus attractives que les méthodes d'ACI pour ce type d'applications ;

- Les méthodes de séparation géométriques n'imposent pas de contraintes ni sur les statistiques des sources ni sur leur signe. Elles sont simples, tiennent compte des contraintes imposées par l'application mais sont moins efficaces que les méthodes d'ACI en terme de séparation. Pourtant elles offrent une solution unique ce qui représente un avantage par rapport aux méthodes NMF. Ces méthodes sont sensibles au bruit et conditionnées par des sources bornées.

Partie II

Contributions et expérimentations

Chapitre 4

Métriques non-euclidiennes pour la classification non supervisée

4.1 Motivation

Tout au long de ce chapitre, nous invitons le lecteur de prendre en compte les notations définies au début du deuxième chapitre.

Le critère de similarité utilisé par la plupart des algorithmes de classification de données est la distance euclidienne. Même si ce critère fournit de bons résultats dans des espaces à deux ou trois dimension, son utilisation dans des espaces de grand dimensions est problèmatique. En effet, il accorde une importance plus élevée aux données dont l'ordre d'échelle est plus grand et ainsi, la contribution des attributs d'un ordre d'échelle moins important n'est pas prise en compte. Ce problème peut être résolu en normalisant les attributs ; ainsi la distance euclidienne peut toujours être utilisée comme mesure de similarité. Le deuxième problème est lié au *phénomène de concentration* ; il traduit l'incapacité de la distance euclidienne à distinguer les voisins le plus proche et le plus eloigné d'un point de référence quelconque. La solution proposée pour résoudre ce problème est d'utiliser des métriques moins concentrées. Celles-ci montrent un meilleur contraste entre les données et elles sont recommandées pour améliorer les résultats des différentes applications traitant des données multivariées.

Dans cette partie nous proposons d'étudier les métriques non euclidiennes dans le contexte de la classification non supervisée des données multivariées pour trouver une solution pour choisir la métrique optimale. Nous allons étudier si les métriques moins concentrées améliorent les résultats de la classification des algorithmes utilisant la distance euclidienne comme mesure de similarité (*e.g. C-moyenne*) et nous montrerons que, en fonction de la distribution des données, la concentration des métriques peut être vue comme un inconvénient mais aussi comme un

avantage. Nous montrerons également que les fonctions *contraste relatif* et *variance relative* ne nous offrent pas d'indices sur la supériorité des métriques moins concentrées pour résoudre des problèmes de classification non supervisée. L'étude des résultats obtenus sur des données synthétiques ainsi que sur des bases de données réelles montrent que la distance interclasse peut être utilisée comme indice pour déterminer la métrique optimale si les classes sont gaussiennes. Nous proposons d'utiliser les indices de validité *Davies-Bouldin* et *compacité-séparabilité* pour résoudre le problème du choix de la métrique optimale.

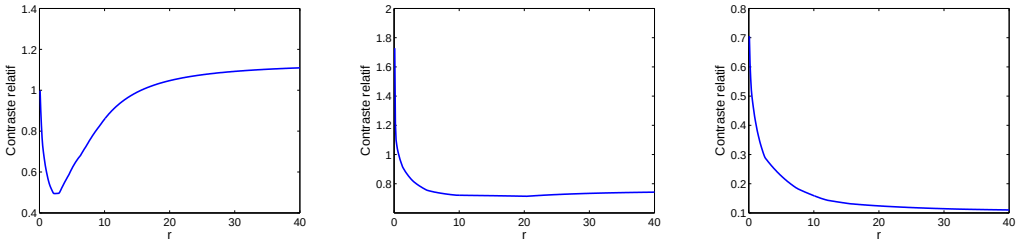
4.2 Choisir la métrique optimale

Dans [AHK01], les métriques fractionnaires sont étudiées et proposées pour éviter le phénomène de concentration des métriques de Minkowski dans des espaces multidimensionnels. Pour des données tirées aléatoirement d'une distribution uniforme, celles-ci sont moins concentrées que toutes les autres métriques de Minkowski, en montrant un meilleur contraste entre données. Cet argument est utilisé par les auteurs pour affirmer que les métriques fractionnaires sont plus appropriées pour résoudre des problèmes comme la recherche du plus proche voisin ou la classification non supervisée dans des espaces multidimensionnels. Il est montré aussi que pour d'autres distributions, les métriques d'ordre supérieur sont moins concentrées que les métriques fractionnaires, [FWV07]. Dans la suite nous allons etudier si les métriques moins concentrées (fractionnaires ou d'ordre supérieur) donnent de meilleurs résultats dans des applications de classification non supervisée de données multivariées et ainsi vérifier si la fonction de contraste relatif introduite par [BGRS00] est un critère pertinent pour déterminer la métrique optimale. Autrement dit, nous allons étudier si la valeur maximale de la fonction de contraste relatif définie par

$$CR_r = \frac{Dmax_d^r - Dmin_d^r}{Dmin_d^r} \tag{4.1}$$

indique la métrique la plus discriminante.

Dans [FWV07] il est montré que la normalisation des données dans l'intervalle [0 1] a comme effet une uniformisation des données ; les résultats obtenus par [AHK01] et [FWV07] (c'est-à-dire la supériorité des normes fractionnaires en terme de contraste relatif ou variance relative) peuvent être donc appliqués sur n'importe quelle distribution de données à condition qu'elle soit normalisée. En comparant la fonction de contraste relatif d'une distribution bimodale de dimension 20 normalisée dans l'intervalle [0 1] figure 4.1 (b), avec celle d'une distribution uniforme de même dimension figure 4.1 (c), on observe une forte similarité : dans les deux cas,



(a) (b) (c)

Figure 4.1: Contraste relatif d'un ensemble de données tirées de deux distributions gaussiennes de dimension 20 : a) données brutes, b) données normalisées, c) données ayant la même dimension, tirées d'une distribution uniforme.

les métriques fractionnaires dépassent invariablement en terme de contraste relatif les métriques d'ordre supérieur. Par contre, la fonction contraste relatif des données brutes figure 4.1 (a), présente des valeurs plus élévées pour les métriques d'ordre supérieur. Quelques conclusions présentées dans [FWV07] sont rappelées :

- la concentration des métriques varie avec la distribution des données ; il existe des distributions pour lesquelles les métriques d'ordre supérieur offrent un meilleur contraste relatif que les métriques fractionnaires ainsi que des distributions pour lesquelles les métriques fractionnaires sont moins concentrées que les métriques d'ordre supérieur ;
- en normalisant n'importe quel ensemble de données dans l'intervalle [0 1], la nouvelle distribution des données tend vers une distribution uniforme et donc, les métriques fractionnaires vont montrer un contraste relatif plus important que les métriques d'ordre supérieur.

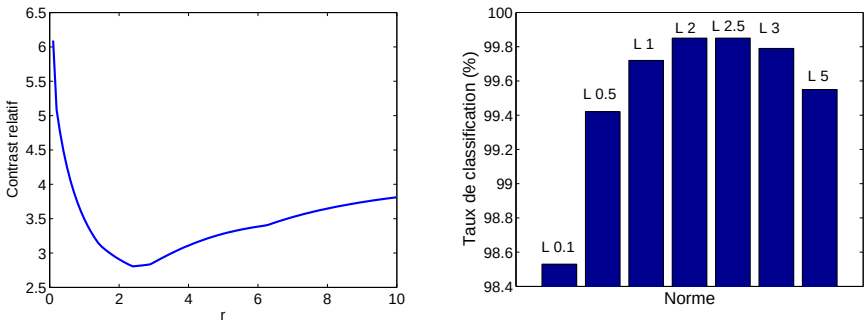
4.3 Expérimentation et évaluation

Dans la suite nous allons poursuivre quelques simulations pour tester si les métriques moins concentrées sont supérieures aux métriques plus concentrées dans le contexte de la classification non supervisée ; pour différentes bases de données multivariées pour lesquelles l'attribut indiquant les classes est connu, nous allons tracer la fonction de contraste relatif en fonction de l'exposant de la norme. Ensuite nous allons classifier les données par un algorithme non supervisé ; en l'occurance la méthode *C-moyenne*. Le taux de classification est estimé en comparant l'attribut indiquant les classes avec les résultats de la classification. Le dernier pas consiste à vérifier s'il existe une corrélation entre la fonction de contraste relatif et les résultats de la classification ; les métriques indiquant un contraste relatif supérieur doivent offrir de meilleurs résultats de

classification que les métriques indiquant un contraste relatif moins important. Les simulations sont réalisées sur des données synthétiques et réelles.

4.3.1 Etude sur des données synthétiques

Dans la première expérience, on considère un ensemble de données tirées de deux distributions gaussiennes, $\mu_1 = 1$, $\mu_2 = 3$, $\sigma_1 = \sigma_2 = 1.5$, correspondant à 2 classes. Les données comportent $d = 15$ attributs (chaque attribut représente un tirage aléatoire de deux gaussiennes ayant la moyenne et l'écart type μ_1, σ_1 respectivement μ_2, σ_2) et la taille de l'ensemble de données est $n = 1000$. La fonction de contraste relatif est présentée dans la figure 4.2 a. Le taux de classification est présenté dans la figure 4.2 b.



(a) (b)

Figure 4.2: a) Fonction de contraste relatif pour des données synthétiques de dimension 15, en fonction de l'exposant r de la métrique - les données sont tirées de deux distributions gaussiennes et b) taux de classification de l'algorithme *C-moyennes* pour différentes normes.

La fonction de contraste relatif montre que les métriques fractionnaires sont moins concentrées que les métriques d'ordre supérieur ; selon [AHK01] elles doivent être plus discriminantes. Pourtant, les résultats de la classification contredisent les résultats théoriques obtenus en [AHK01] ; pour cet ensemble de données, la métrique euclidienne qui présente un contraste relatif moins important dépasse en terme de taux de classification toutes les autres métriques. Les meilleurs taux de classification sont obtenus pour la métrique euclidienne et pour la métrique $L_{2.5}$.

4.3.2 Etude sur des bases de données réeles

La même expérience a été répétée sur des données réelles. Les bases de données utilisées pour les tests sont mises à disposition par l'Université de Californie à Irvine [UCI]. Elles font partie d'un

ensemble très large de jeux de données multivariées utilisés par la communauté d'apprentissage artificiel pour tester leurs méthodes. Pour nos simulations nous avons retenu 5 jeux de données différents :

Iris : Cette base de données est très utilisée dans le domaine de la reconnaissane de formes. Elle contient 3 classes de 50 objets chacune, chaque classe représentant un type de plante Iris. Les deux premières classes sont linéairement séparables et la dernière classe n'est pas linéairement séparable des deux autres. Chaque objet est représenté par 4 attributs.

WBC : La base de données WBC a été obtenue dans les hôpitaux de l'Université de Wisconsin par le Dr. William H. Wolberg. Les données ont été recueillies pendant une période de 2 ans. Les objets, en nombre de 699 sont décrits par 10 attributs et sont groupés en 2 classes représentant des tumeurs malignes ou bénignes. Les classes sont linéairement séparables.

WDBC : Les données de cette base ont été recueillies à partir d'images numérisées d'un prélèvement par biopsie d'une masse éventuellement cancéreuse. Elles décrivent les caractéristiques de noyaux de cellule présents dans l'image. Les objets sont répartis en deux classes selon qu'il s'agit de tumeurs malignes (212 exemples) ou bénignes (357 exemples) et le nombre d'attributs est $d = 32$. On notera qu'il s'agit d'un problème relativemement simple : les classes sont linéairement séparables et l'état de l'art fait mention d'une précision supérieure à 97% en classement.

Ionosphere : Cette base de données contient des données radar recueillies à l'aide d'un système à 16 antennes de haute fréquence. Les objets représentent des électrons libres dans la ionosphère ; ils sont décrits à l'aide de 34 attributs et ils sont groupés en deux classes qui ne sont pas linéairement séparables. Le nombre de données est $n = 351$.

Wine : Cette base de données recense les résultats d'une analyse chimique de différents vins produits dans une même région d'Italie à partir de différents cépages. La concentration de 13 constituants est indiquée pour chacun des 178 vins analysés qui se répartissent ainsi : 59 dans la classe 1, 71 dans la classe 2 et 48 dans la classe 3.

Dans les simulations nous avons utilisé les données brutes ainsi que les données normalisées dans l'intervalle [0 1]. Pour chacun des cas, la fonction de contraste relatif est estimée en balayant le paramètre de la norme r dans l'intervalle [0 20]. Ensuite, l'algorithme *C-moyenne* a été utilisé pour la classification en considérant plusieurs normes L_r comme mesure de similarité. Pour chacune des bases de données de test, le nombre de classes est connu. Nous disposons aussi de l'attribut *étiquette* ; cet attribut n'est pas utilisé dans la classification mais il nous servira

pour valider les résultats.

Selon [AHK01], des valeurs élevées pour la fonction de contraste relatif indiquant une norme moins concentrée, doivent identifier le paramètre r de la norme la plus discriminante. Nous allons comparer les résultats de classification pour plusieurs normes plus ou moins concentrées pour confirmer ou pour infirmer si la concentration des normes peut nous fournir des indices pour choisir la métrique optimale dans un problème de classification.

Présentation des résultats

Les résultats obtenus sur les bases de données utilisées ne nous offrent aucun indice sur la pertinence de la fonction de contraste relatif pour le choix de la métrique optimale dans des problèmes de classification non supervisée.

Données brutes Pour les données Iris et WBC, le contraste relatif, figure 4.3 (a) et (b) indique que les métriques fractionnaires sont plus discriminantes, contrairement aux résultats de classification, tableau 4.1 indiquant les meilleures performances de classification pour les métriques d'ordre supérieur. Pour les données Wine, le contraste relatif figure 4.3 (e), indique les métriques d'ordre supérieur comme étant les plus appropriées pour la classification, mais les résultats de la classification montrent la supériorité des métriques fractionnaires. La seule cohérence entre la fonction de contraste relatif et les résultats de la classification est obtenue pour les bases de données WDBC et Ionosphere.

	Iris	WBC	WDBC	Ionosphere	Wine
$L_{0.3}$	84.66%	92.24%	83.35%	61.82%	85.95%
$L_{0.5}$	84.66%	93.26%	84.35%	60.96%	75.28%
$L_{0.7}$	84.66%	94.14%	84.88%	60.68%	70.78%
L_1	85.33%	94.43%	85.23%	60.96%	70.22%
$L_{1.5}$	88.66%	94.72%	85.41%	60.11%	70.22%
L_2	88.66%	95.9%	85.41%	60.39%	70.22%
L_3	88.66%	96.77%	85.41%	60,11%	70,22%
L_4	88.66%	96.92%	85.41%	59.82%	70.22%
L_5	88.66%	96.92%	85.41%	60.68%	70.22%

Tableau 4.1: Performances de l'algorithme *C-moyennes* sur les bases de données Iris, WBC, WDBC, Ionosphere et Wine pour différentes métriques.

Données normalisées Pour les données normalisées, la fonction de *contraste relatif* indique dans tous les cas un contraste plus important pour les métriques fractionnaires figure 4.4 ;

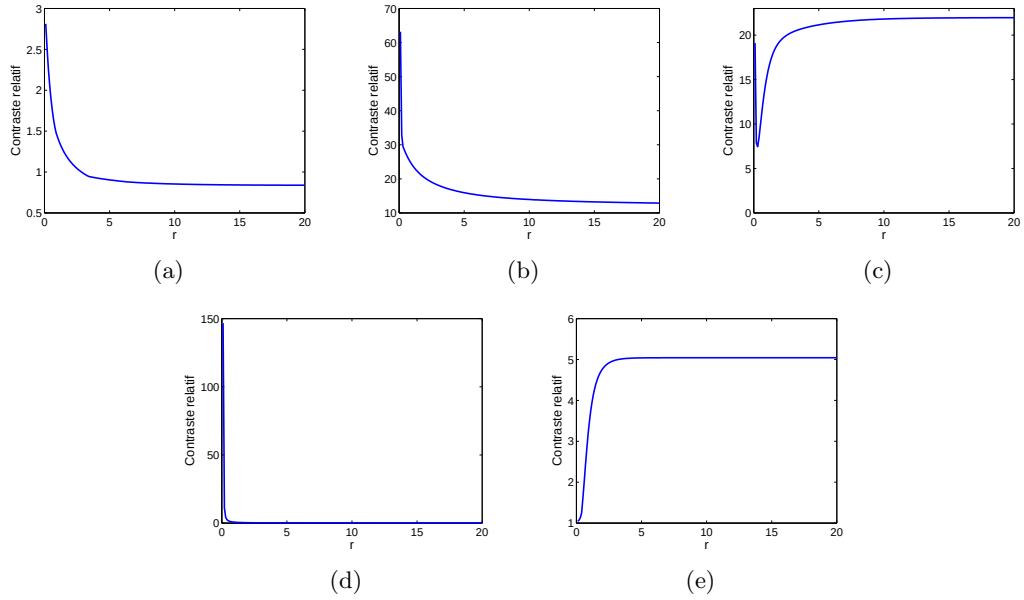


Figure 4.3: Contraste relatif pour les bases de données réelles : a) Iris, b) WBC, c) WDBC, d) Ionosphere, e) Wine en fonction du paramètre de la métrique.

ceci est dû à notre avis, au fait que la normalisation des données dans l'intervalle [0 1] rend la distribution des données plus uniforme et donc, selon [AHK01], les métriques fractionnaires sont moins concentrées que les métriques d'ordre supérieur. Par contre, les résultats de la classification tableau 4.2 ne confirment pas l'hypothèse que si une métrique est moins concentrée, elle est plus appropriée pour résoudre les problèmes de classification non supervisée.

	Iris	WBC	WDBC	Ionosphere	Wine
$L_{0.3}$	86.59%	93.17%	89.80%	61.82%	92.69%
$L_{0.5}$	88%	93.26%	90.51%	61.53%	94.94%
$L_{0.7}$	87.53%	94.14%	91.56%	60.68%	94.94%
L_1	87.79%	94.43%	92.26%	60.96%	96.62%
$L_{1.5}$	88.66%	94.72%	92.45%	60.11%	96.06%
L_2	88.39%	95.9%	92.79%	60.11%	94.94%
L_3	88.66%	96.77%	92.26%	63,07%	92,13%
L_4	88.56%	96.92%	91.74%	65.56%	92.13%
L_5	87.33%	96.92%	91.74%	66.85%	89.32%

Tableau 4.2: Performances de l'algorithme *C-moyennes* sur les bases de données Iris, WBC, WDBC, Ionosphere et Wine, normalisées. Plusieurs métriques sont utilisées dans la classification.

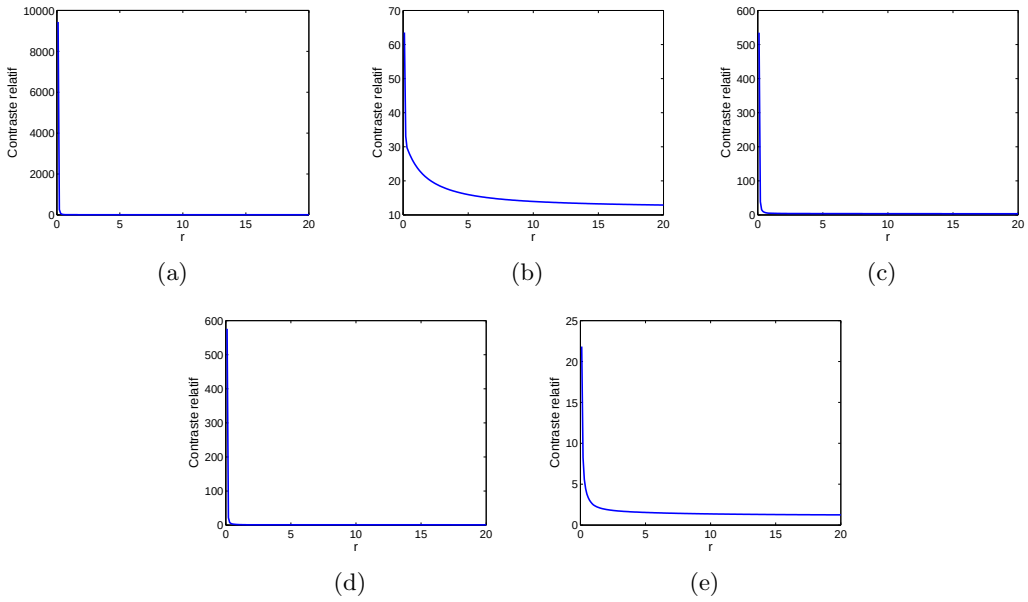


Figure 4.4: Contraste relatif pour les bases de données réelles normalisées : a) iris, b) WBC, c) WDBC, d) ionosphere, e) wine en fonction du paramètre de la métrique.

4.3.3 Discussion sur les résultats

Nous étudions ces résultats en essayant de comprendre pourquoi les métriques montrant un contraste plus important entre données sont souvent moins discriminantes pour la classification.

Pour un ensemble quelconque de données, une métrique moins concentrée implique par définition une distribution des normes plus large qu'une métrique plus concentrée ; la figure 4.5 montre cette propriété sur un ensemble de données de dimension 15 tirées d'une distribution uniforme. Les métriques moins concentrées montrent donc un contraste plus important et donc, selon [AHK01] une meilleure discrimination entre les données. Mais cela n'est pas souvent le cas, comme nous l'avons vu pour les résultats précédents.

Pour expliquer ce fait, nous envisageons un exemple très simple. Considérons le cas d'un ensemble de données qui contient deux classes tirées de deux distributions gaussiennes de dimension d et considérons $\{\|X_i\|\}_r$ l'ensemble des normes L_r de ces données. Conformément au phénomène de concentration, une norme moins concentrée va présenter une distribution plus large comme on peut voir, figure 4.5 ; moins la norme est concentrée, plus sa distribution est large.

Cela peut avoir comme effet un recouvrement entre les distributions des normes des deux classes car la distribution globale des normes a tendance à s'uniformiser. Cet effet est présenté

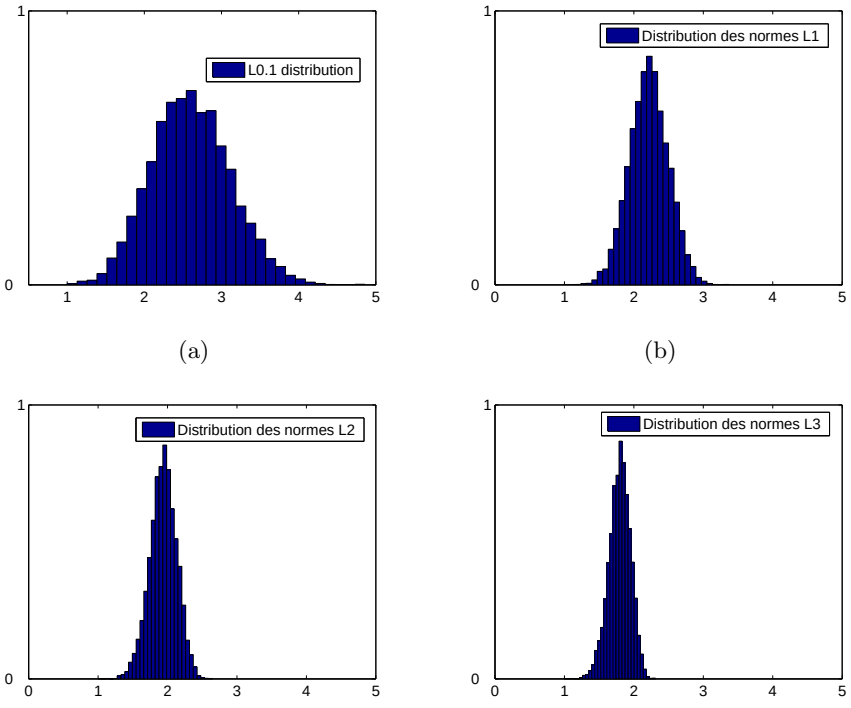


Figure 4.5: Distribution des différentes métriques d'un ensemble de données uniformes

dans la figure 4.6 ; les deux modes correspondant aux classes sont plus visibles quand les normes L_1 et L_2 sont utilisées. Par contre, la métrique $L_{0.1}$ qui est moins concentrée, a comme effet un recouvrement plus important des distributions des deux classes. Pourtant, cette inspection visuelle ne suffit pas pour montrer laquelle est la métrique la plus discriminante. Pour montrer cela, nous mesurons la distance entre les distributions des normes des deux classes ; pour une distribution bimodale univariée contenant 2 classes gaussiennes, la distance entre les deux classes peut être estimée par :

$$|Dmin^2 - Dmax^1| \tag{4.2}$$

où $Dmin^2$ et $Dmax^1$ représentent respectivement les valeurs minimales et maximales de l'ensemble des normes de la classe 2 et de la classe 1 et $\mu_2 > \mu_1$.

Pour une distribution des normes r d'un ensemble de données de dimension d contenant 2 classes gaussiennes, la distance relative entre les deux classes devient :

$$\frac{|Dmin_r^2 - Dmax_r^1|}{mean(\|X\|_r)} \tag{4.3}$$

où X est l'ensemble des données et les exposants 1 et 2 indiquent la classe. Plus la distance

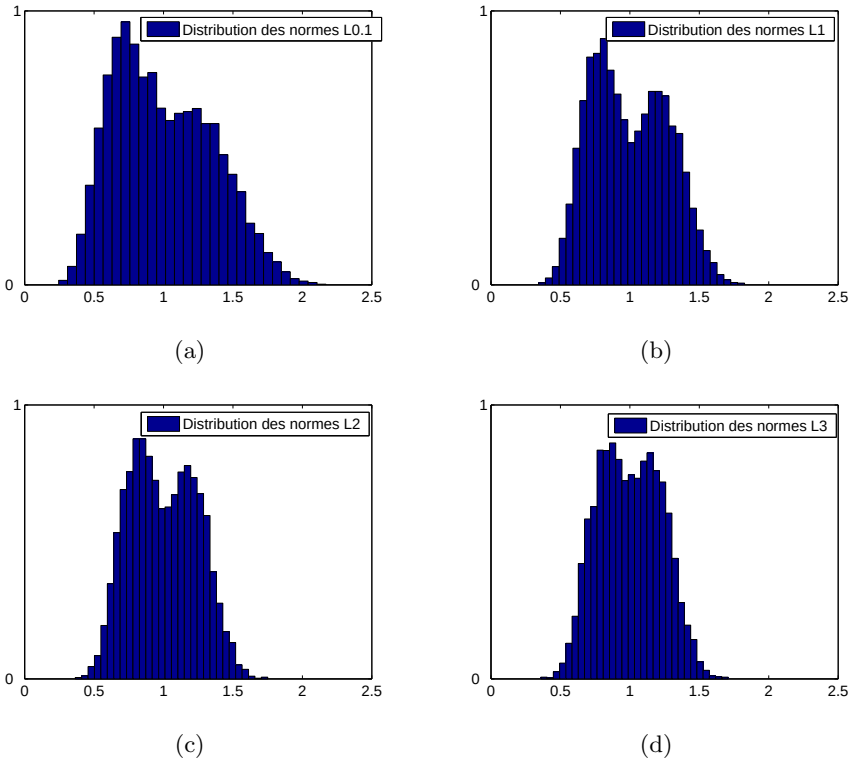
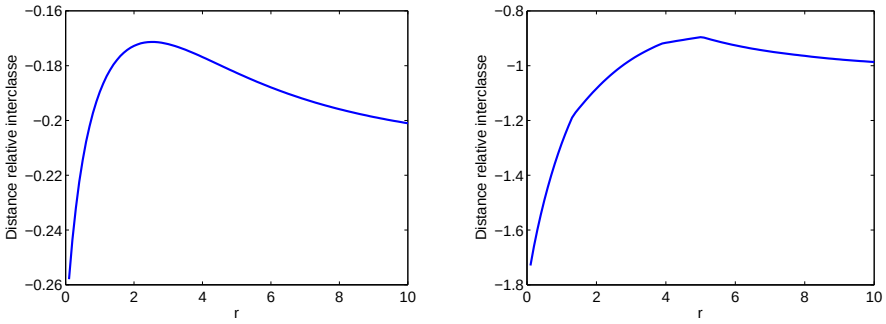


Figure 4.6: Distribution des différentes métriques d'un ensemble de données tirées de deux distributions gaussiennes

relative est grande, mieux les classes sont séparées.

Dans la suite, nous allons tester si cette fonction dépendant de l'exposant de la métrique peut nous aider à trouver la métrique optimale. Pour l'exemple présenté ci-dessous nous avons calculé la distance interclasse relative en balayant le paramètre r dans l'intervalle [0.1 10] figure 4.7 a. La valeur maximale de cette fonction doit nous indiquer la métrique optimale. La valeur maximale de la distance relative est obtenue pour $r = 2.5$; pour cette métrique ainsi que pour la métrique euclidienne nous obtenons le meilleur taux de classification de l'algorithme *C-moyennes*, figure 4.2 (b) ; la valeur minimale de cette fonction indique la métrique pour laquelle nous obtenons le plus faible taux de classification.

Des tests ont été réalisés sur les bases de données réelles utilisées précédemment. Parmi celles-ci, nous avons choisi la base de données WBC présentant des classes gaussiennes. Pour cette base de données, les résultats obtenus sur les données synthétiques sont confirmés. La fonction *distance relative* indique que la norme L_5 est la plus discriminante figure 4.7 b, ce qui est confirmé par les résultats de la classification tableau 4.2 où les meilleurs taux de classification



(a) (b)

Figure 4.7: Distance relative en fonction de l'exposant de la métrique r : a) entre deux classes gaussiennes, b) pour la base de données WBC.

sont obtenus pour les normes L_4 et L_5. Ces résultats ont été obtenus sur les données normalisées. Pour les autres bases de données la distance relative n'indique pas la métrique optimale car les classes ne sont pas gaussiennes.

La distance relative, équation 4.3.3 peut être un moyen de déterminer la métrique optimale dans un problème de classification où les classes sont gaussiennes. Pour estimer la métrique optimale dans un problème de classification non supervisée, le seul moyen est de considérer l'exposant de la métrique comme un paramètre de l'algorithme de classification. Au lieu d'utiliser la distance relative pour le choix de la métrique, nous proposons d'utiliser un des indices de validité des classes présentés dans le chapitre 1 prenant en compte la distance interclasse. Dans le chapitre 6 nous présentons une application de segmentation des images multivariées par classification des pixels, où le choix de la métrique a été réalisé en utilisant les indice Davies-Bouldin et compacité-séparabilité.

4.4 Conclusion

Dans ce chapitre les métriques non euclidiennes sont étudiées dans le contexte de la classification non supervisée. Le point de départ de cette démarche est constitué par quelques résultats récents qui montrent que dans des espaces multidimensionnels, la distance euclidienne est affectée par le phénomène de concentration et ainsi, la recherche du plus proche voisin (indispensable dans la classification non supervisée) devient un problème instable. Des normes moins concentrées sont proposées pour éviter la concentration, ainsi que des fonctions permettant de calculer et de comparer la concentration des différentes normes : *le contraste relatif* et *la variance relative*. Ces

métriques montrent un contraste plus élévé entre les données et sont proposées pour améliorer les résultats des algorithmes de classification utilisant la distance euclidienne comme mesure de similarité. Si ceci est vrai, les fonctions de *contraste relatif* ou *variance relative* peuvent être utilisées comme critère pour choisir la métrique optimale, car la valeur maximale de ces fonctions indique la métrique la plus discriminante.

L'hypothèse de la supériorité des métriques moins concentrées pour la classification de données multivariées est testée sur des données synthétiques ainsi que sur des données réelles et nos résultats l'infirment ; la métrique optimale dans un problème de classification non supervisée dépend fortement de la distribution de données. Nous avons vu que le phénomène de concentration est à la fois un avantage ou un inconvénient en fonction de l'application. Le choix de la métrique optimale peut seulement se faire en considérant l'exposant de la métrique comme un paramètre de l'algorithme de classification, mais ceci a comme inconvénient une augmentation du temps de calcul.

L'étude des résultats des simulations nous permet d'observer que la distance relative interclasse peut nous servir pour choisir la métrique optimale dans un problème de classification si les classes sont de forme gaussienne. Pour des applications où des méthodes non supervisées sont demandées, nous proposons d'utiliser les indices DB ou CS (qui sont adaptés pour des classes gaussiennes) car ceux-ci prennent en compte la distance interclasse.

Chapitre 5

Méthodes de SAS pour la réduction de la dimension des données multivariées

5.1 Motivation

Dans ce chapitre nous utilisons les notations définies au début du troisième chapitre.

La dimension des données est un des problèmes majeurs des approches traitant des données multivariées. Souvent, des méthodes de réduction de dimension, groupées en méthodes de sélection d'attributs et méthodes d'extraction d'attributs sont utilisées avant l'analyse. Parmi les méthodes d'extraction d'attributs, l'ACP, l'Analyse Factorielle sont des méthodes souvent citées dans la littérature. Les méthodes de SAS constituent une alternative très puissante à ces méthodes ; pour celles-ci, l'avantage de la réduction de dimension est complété par la mise en évidence des structures cohérentes (*e.g.* des données qui identifient la présence d'une espèce minérale, ou des composées chimiques etc.) dans l'ensemble de données. Ces structures ont un un sens physique réel et permettent une meilleure interprétation et compréhension des données multivariées.

Dans ce chapitre nous ramenons le problème de réduction de la dimension par extraction des attributs à un problème de SAS. Nous présentons une nouvelle approche pour résoudre le problème de SAS des mélanges linéaires et ainsi trouver le sous-espace optimal de représentation des données multivariées. Cette méthode est basée sur une interprétation géométrique du modèle de mélange linéaire, idée qui apparaît pour la première fois dans [Pun95]. La méthode proposée est très simple et elle est applicable pour l'extraction de sources non-négatives à partir de mélanges dont les coefficients sont non-négatifs. Une évaluation des méthodes de SAS comme des méthodes de réduction de dimension pour la classification non supervisée est aussi présentée.

5.2 Approche proposée

5.2.1 Préliminaires

Nous allons définir le problème de réduction de la dimension comme le problème de la recherche du sous-espace optimal pour la représentation d'un ensemble de données multivariées. Le sous-espace optimal doit décrire au mieux l'ensemble original des données, avec le minimum de dimensions. Le principe des méthodes de réduction de dimension par extraction d'attributs est de projeter les données sur les axes du nouveau sous-espace de représentation. Dans ce chapitre, le choix de la dimension du nouveau sous-espace (ou la dimension intrinsèque des données) n'est pas discuté.

La SAS consiste à trouver les signaux originaux ainsi que la matrice de mélange à partir d'un ensemble de mesures appelés *observations*. Si la matrice de mélange est connue, alors la matrice des sources peut être estimée par la projection des observations sur les vecteurs ligne de la matrice inverse ou pseudoinverse de la matrice de mélange. Considérons que les observations sont le résultat d'un mélange linéaire entre la matrice des sources et la matrice de mélange, alors nous pouvons décrire le problème par le modèle de mélange linéaire où seule la matrice X est connue :

$$X^{m\times n} = A^{m\times p} S^{p\times n} \tag{5.1}$$

Dans le cadre de la SAS les vecteurs lignes de la matrice A représentent des vecteurs de mélange tandis que les vecteurs lignes de la matrices S sont des sources. Les matrices A et S peuvent être obtenues sous différentes contraintes, comme nous l'avons vu dans le troisième chapitre. Dans le cadre du problème de réduction de dimension par extraction d'attributs, les vecteurs colonne de la matrice A représentent les directions dans l'espace original des données qui décrivent le nouveau sous-espace de représentation, tandis que la matrice S représente les projections des données X sur les vecteurs ligne de la matrice inverse (ou pseudoinverse si la matrice A n'est pas carrée) de A, autrement dit, la représentation de X dans le nouvel espace.

La méthode proposée est inspirée par l'observation que le modèle de mélange linéaire a comme effet une transformation géométrique des *sources* dans l'espace de représentation. L'idée est de trouver les extrémités de la matrice d'observation X dans l'espace de représentation. Ceux-ci représentent les points extrêmes de la matrice de sources S après le mélange linéaire (ou après la transformation géométrique). Supposons que les points extrêmes des sources sont connus, alors les vecteurs de mélange sont facilement trouvés et donc les sources.

5.2.2 Représentation géométrique du modèle de mélange linéaire

Pour des raisons de simplicité et de visualisation, nous réduisons cette discussion au cas où le nombre de sources est égal à 2 ou 3. Considérons d'abord le cas de 2 sources uniformes *indépendamment et identiquement distribuées* (*i.i.d.*), non négatives, de distribution bornée [0 1], figure 5.1 (a). Dans le plan (s_1, s_2), chaque instant i est représenté par un point (s_1^i, s_2^i) et tous les points dans l'espace de sources sont contenus dans le cône dont les génératrices sont définies par l'origine (le point $(0, 0)$) et les points $(0, 1)$ respectivement $(1, 0)$. La particularité de ces deux vecteurs est que l'angle défini par ces derniers a la valeur maximale. On mentionne que tous les points de la forme $(0, s_2^i)$ et $(s_1^j, 0)$ se retrouvent sur les génératrices du cône et présentent cette particularité. On va appeler ces points *les points extrêmes* de la distribution des points sources.

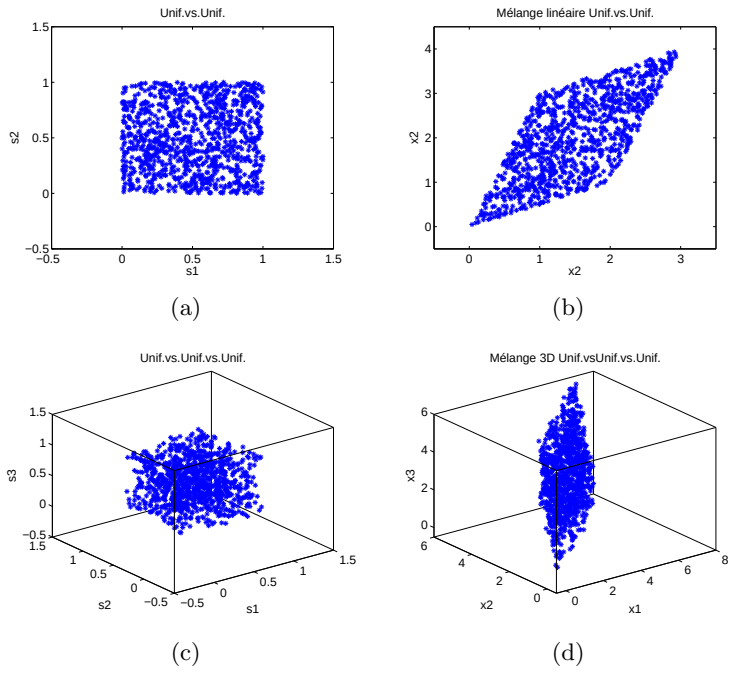


Figure 5.1: Représentation des sources uniformes et des observations (mélange linéaire) : a) et b) 2 sources, c) et d) 3 sources.

De manière évidente, l'effet du modèle de mélange linéaire sur les sources est une transformation géométrique dans l'espace de représentation, comme on peut le voir dans la figure 5.1 (b). Cette transformation géométrique fait que dans l'espace des observations X, les données restent toujours à l'intérieur des génératrices d'un autre cône. Comme la transformation géométrique du mélange linéaire s'applique sur tous les points (s_1^i, s_2^j), elle s'applique implicitement sur les

points extrêmes définissant les génératrices du cône initial ; les points extrêmes vont toujours garder le plus grand angle parmi toutes les autres paires de points. Ces points se retrouvent sur les génératrices du nouveau cône qui enferme toutes les données X_i. Dans la suite de ce chapitre nous allons montrer que sous certaines hypothèses ces points représentent les vecteurs de mélange multipliés par un facteur d'échelle.

Les figures 5.1 (c) et (d) illustrent la géométrie du modèle de mélange linéaire pour trois sources uniformes : (c) la représentation des sources, (d) la représentation des observations (les sources après le mélange linéaire avec une matrice de mélange $A^{3\times3}$). La figure 5.2 montre la représentation géométrique du modèle de mélange linéaire pour des sources tirées d'une distribution bi-gaussienne.

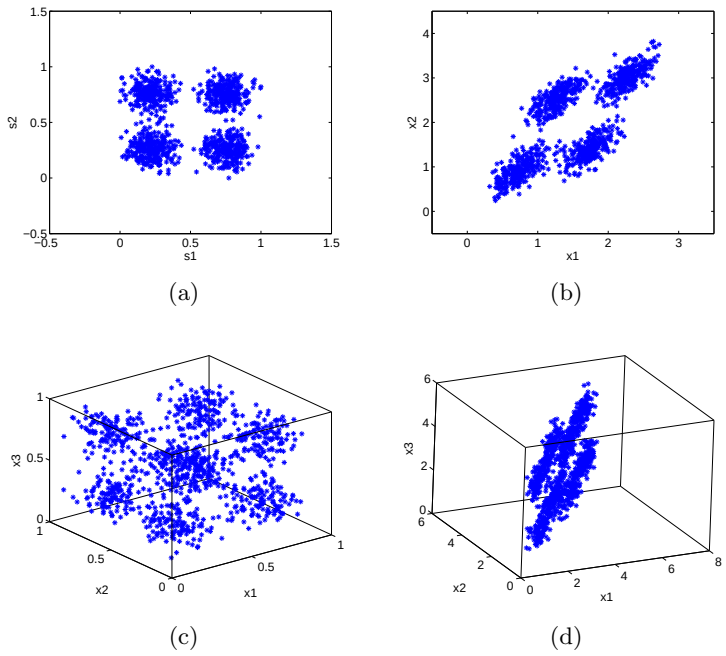


Figure 5.2: Représentation des sources tirées d'une distribution bimodale et des observations (mélange linéaire) : a) et b) 2 sources, c) et d) 3 sources.

5.2.3 Etude analytique

Soit $X \in \mathbb{R}^{2\times n}$ une matrice d'observations représentant le mélange linéaire entre une matrice de sources $S \in \mathbb{R}^{2\times n}$ et une matrice de mélange $A^{2\times2}$. On suppose que :

1. les vecteurs ligne s_1 et s_2 de la matrice S sont tirés d'une distribution quelconque à support positif $s_1 \geq 0$ et $s_2 \geq 0$,

2. il existe un instant i pour lequel la première source est active et la seconde source est inactive $(s_1^i \neq 0)\&(s_2^i = 0)$,

3. il existe un autre instant j pour lequel la première source est inactive et la seconde source est active $(s_1^j = 0)\&(s_2^j \neq 0)$.

Ces vecteurs sont des points extrêmes de la distribution des points sources et représentent les génératrices du cône qui renferme tous les points sources. Si les conditions 1, 2 et 3 sont satisfaites alors on peut énoncer la proposition suivante :

Proposition 1 1 *Les points extrêmes de la matrice d'observations X représentent les vecteurs colonne de la matrice de mélange A multipliés par un facteur d'echelle. Ils peuvent être retrouvés comme étant la paire des points (X_k, X_l), $X_{k,l} \in X$ pour laquelle l'angle :*

$$\alpha(X_k, X_l) = \arccos \frac{X_k^T X_l}{\|X_k\|\|X_l\|}$$

est maximal parmi toutes les autres paires de points de l'ensemble X.

Démonstration

Notre démonstration se base sur deux observations:

1. l'effet du modèle de mélange linéaire sur les données source est une transformation géométrique dans l'espace de représentation,

2. une transformation géométrique d'un ensemble de données a le même effet que si on l'effectuait sur l'ensemble des données normalisées. Autrement dit, les vecteurs de l'ensemble $X = AS$ et de l'ensemble $X_{norm} = AS_{norm}$ ont la même direction. Cette observation est illustrée dans la figure 5.3 dans le cas de 2 sources et dans la figure 5.4 dans le cas de 3 sources.

Soit la matrice de mélange A (2×2)

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \tag{5.2}$$

la matrice de sources

$$S = \begin{pmatrix} s_1^1 & \dots & 0 & \dots & s_1^i & \dots & s_1^n \\ s_2^1 & \dots & s_2^j & \dots & 0 & \dots & s_2^n \end{pmatrix} \tag{5.3}$$

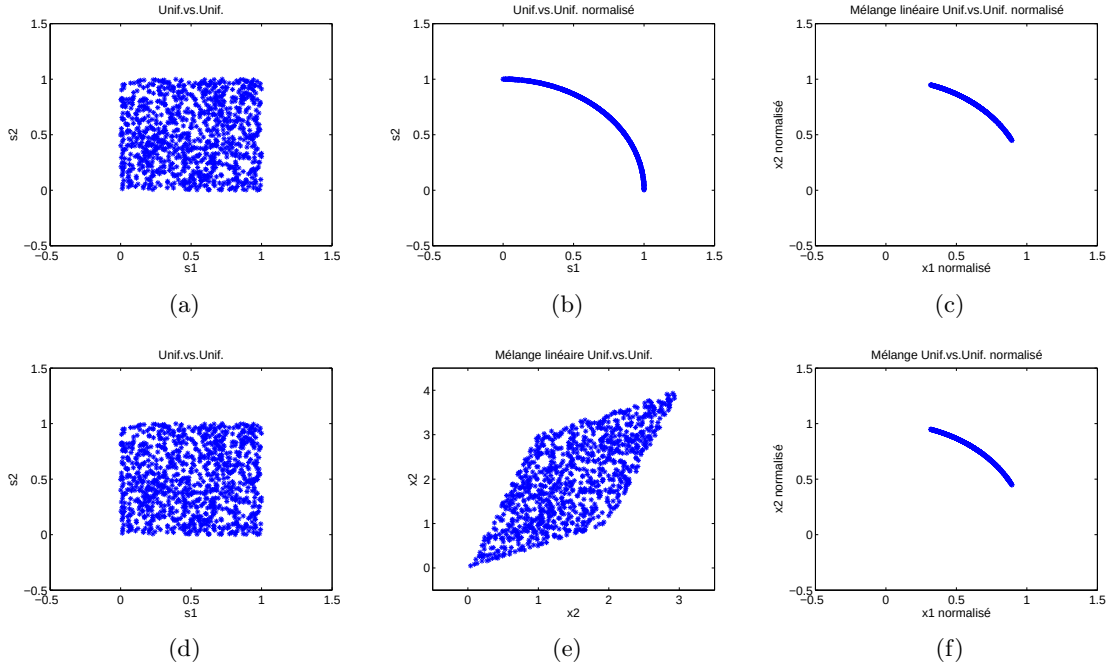


Figure 5.3: Première ligne : a) Représentation des deux sources uniformes, b) représentation des deux sources uniformes - chaque donnée à l'instant i est normalisé, c) représentation du mélange des deux sources après la normalisation de chaque donnée à l'instant i. Deuxième ligne : d) représentation des deux sources uniformes, e) représentation du mélange de deux sources uniformes, f) représentation du mélange de deux sources après la normalisation de chaque donnée à l'instant i.

alors la matrice des observations $X = AS$ est donnée par :

$$X = \begin{pmatrix} x_1^1 & \dots & a_{12}s_2^j & \dots & a_{11}s_1^i & \dots & x_1^n \\ x_2^1 & \dots & a_{22}s_2^j & \dots & a_{21}s_1^i & \dots & x_2^n \end{pmatrix} \tag{5.4}$$

La matrice des sources normalisées est donnée par :

$$S_{norm} = \begin{pmatrix} s_{1norm}^1 & \dots & 0 & \dots & 1 & \dots & s_{1norm}^n \\ s_{2norm}^1 & \dots & 1 & \dots & 0 & \dots & s_{2norm}^n \end{pmatrix} \tag{5.5}$$

et la matrice X_{norm} est :

$$X_{norm} = \begin{pmatrix} x_{1norm}^1 & \dots & a_{12} & \dots & a_{11} & \dots & x_{1norm}^n \\ x_{2norm}^1 & \dots & a_{22} & \dots & a_{21} & \dots & x_{2norm}^n \end{pmatrix} \tag{5.6}$$

Dans la matrice X, équation 5.4, les *i-ème* et *j-ème* colonnes représentent les vecteurs colonnes de la matrice de mélange, multipliés par un facteur d'echelle. Dans la matrice X_{norm}, équation 5.6, les *i-ème* et *j-ème* colonnes représentent les vecteurs colonne de la matrice de mélange. Les vecteurs de mélange peuvent être ainsi déterminés comme étant les points extrêmes de la

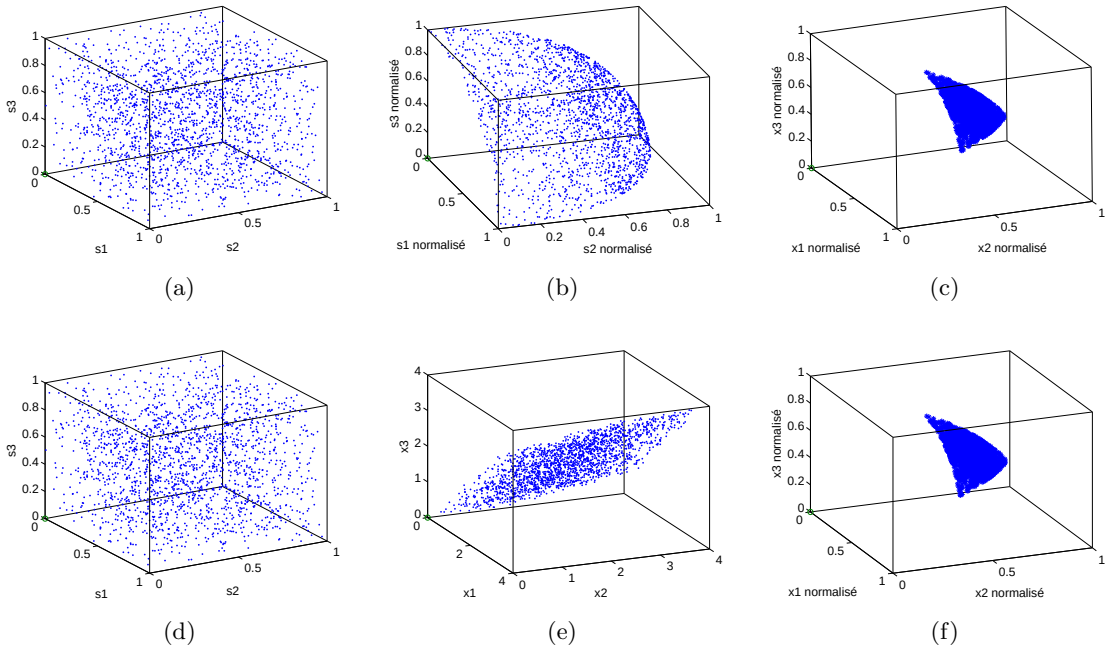


Figure 5.4: Première ligne : a) Représentation des trois sources uniformes, b) représentation des trois sources où chaque donnée à l'instant i est normalisé, c) représentation du mélange des trois sources après la normalisation de chaque donnée à l'instant i. Deuxième ligne : d) représentation des trois sources uniformes, e) représentation du mélange de trois sources uniformes, f) représentation du mélange de trois sources après la normalisation de chaque donnée à l'instant i.

matrice X_{norm}. Ceux-ci représentent la paire de points (X_{inorm}, X_{jnorm}) pour lesquels l'angle entre eux défini par $\arccos(X_{i\,norm}^T X_{j\,norm})$ est maximal parmi toutes les autres paires de points de l'ensemble X_{norm}. En remplaçant $X_{i\,norm}$ et $X_{j\,norm}$ par $\dfrac{X_i}{\|X_i\|}$ et par $\dfrac{X_j}{\|X_j\|}$, la proposition est démontrée.

Nous allons généraliser la proposition 1 pour p sources.

Soit $X \in \mathbb{R}^{m\times n}$ une matrice d'observations représentant le mélange linéaire entre une matrice de sources $S \in \mathbb{R}^{p\times n}$ et une matrice de mélange $A^{(m\times p)}$. On suppose que :

1. les vecteurs ligne s_i de la matrice S sont tirés d'une distribution quelconque à support positif $s_i \geq 0$,

2. pour chaque source i il existe un instant k où elle est active et toutes les autres sources sont inactives $(s_i^k \neq 0)\&(s_j^k = 0)$, $j = 1 : p$, $j \neq i$,

Si ces conditions sont satisfaites alors on peut énoncer la proposition suivante :

Proposition 2 1 *Les points extrêmes de la matrice d'observations X représentent les vecteurs colonne de la matrice de mélange A multipliés par un facteur d'echelle. Ils peuvent être retrouvés comme étant le p-tuplet des points $X_p \in X$ pour lequel la somme des angles entre toutes paires de points définies sur le p-tuplet X_p est maximale parmi toutes les autres p-tuplets de l'ensemble X.*

Démonstration

La démonstration de cette proposition suit le même principe que dans la démonstration de la première proposition.

5.2.4 Algorithme proposé

Pour une matrice d'observations X obtenue lors d'un mélange linéaire des sources *i.i.d.*, l'algorithme peut être résumé en 2 étapes :

Algorithme 11 Le pseudocode de l'algorithme PExSAS (**S**éparation **A**veugle de **S**ources par la rechèrche des **P**oints **Ex**trèmes)

1: Projection des observations X_i sur la sphère unité :

$$X_i = \frac{X_i}{\|X_i\|} \tag{5.7}$$

Cette normalisation fait que tous les points se trouvent sur le cercle unité ou dans une région sur la surface de la sphère (ou hypersphère) unité, comme on peut voir dans la figure 5.3 (f) dans le cas de 2 sources ou dans la figure 5.4 (f) dans le cas de 3 sources.

2: Sur la surface de la sphère (ou hypersphère) unité, on cherche le *p-tuplet* qui maximise la fonction :

$$\sum \arccos(x_i^T x_j) \tag{5.8}$$

où $i, j = 1 : p$, $i \neq j$. Ces points représentent les extrémités de la distribution des points sur la sphère unité, figure 5.5 et constitueront les vecteurs colonne de la matrice de mélange.

Stratégie pour la recherche des points extrêmes d'une distribution de points

- initialisation du point de départ par le barycentre de la distribution des points,
- le premier point extrême est le point le plus éloigné du barycentre,
- le deuxième point extrême est le point le plus éloigné du premier point extrême,
- le k-ème point extrême est le point $x_j \in X$ pour lequel,

$$x_j = \max_j \sum \arccos(x_i^T x_j) \tag{5.9}$$

où $i = 1 : k - 1$.

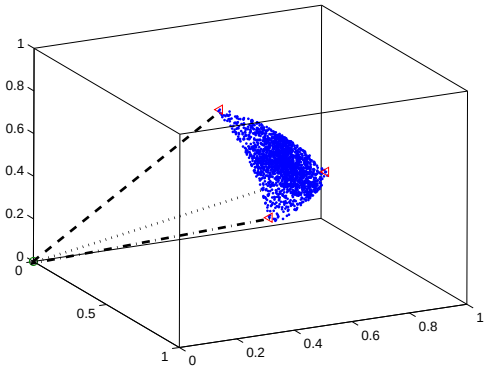


Figure 5.5: Matrice des observations X : les points extrêmes de la matrice des observations constituent les vecteurs colonne de la matrice de mélange.

5.3 Evaluation et comparaison avec quelques méthodes usuelles

Dans cette section les performances de l'algorithme proposé sont évaluées dans le contexte de la séparation aveugle de sources et les résultats sont comparés avec celles obtenus avec quelques méthodes usuelles de séparation de sources : FastICA [Hyv99a], JADE [CS93], NMF [LS99] et NMF ALS [Paa97]. L'algorithme est testé dans le cas de 2, 3 et 4 sources tirées aléatoirement à partir de plusieurs distributions. L'influence du niveau de bruit sur les performances de l'algorithme est aussi discutée.

5.3.1 Mise au point des simulations

Simulation des sources Les sources ont été simulées par tirage aléatoire à partir des différentes distributions présentées dans la figure 5.6. La distribution de chaque source est non-négative $[0\ \infty]$. Les distributions choisies pour les simulations sont souvent utilisées pour tester des algorithmes de séparation. Nous nous sommes particulièrement intéressés à la séparation des sources non négatives pour pouvoir utiliser ces techniques dans le cadre de l'analyse des images multivariées. L'organisation temporelle ou spatiale des sources n'est pas prise en compte dans ce travail. C'est la raison pour laquelle des algorithmes utilisant ce critère ne sont pas discutés.

Coefficients de mélange Les coefficients de la matrice de mélange sont simulés arbitrairement à partir d'une distribution uniforme à support positif.

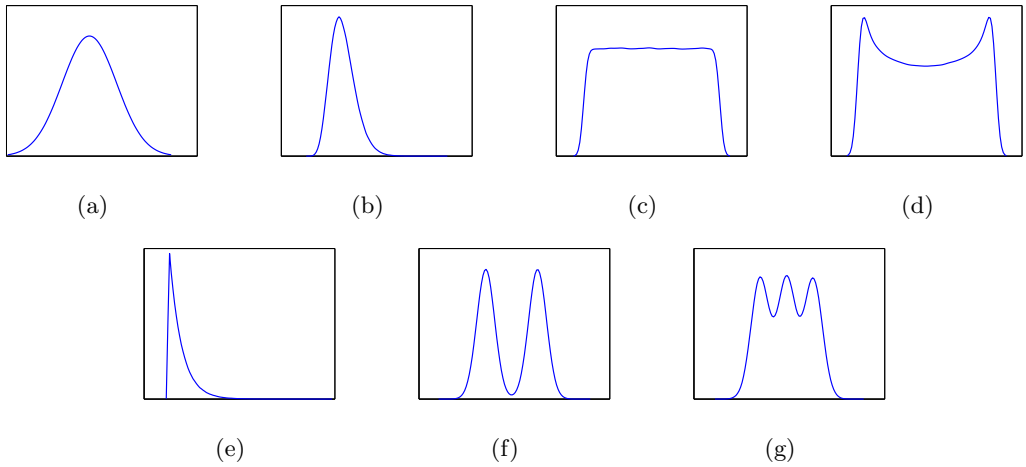


Figure 5.6: *Fdp* des sources : a) Gaussienne, b) Gamma, c), Uniforme, d) Beta, e) Exponentielle, f) mélange de 2 Gaussiennes, g) mélange de 3 Gaussiennes.

Matrice des observations La matrice des observations X est construite à partir des signaux sources simulés et de la matrice de mélange, selon le modèle de mélange linéaire avec bruit aditif, blanc gaussien *i.i.d.* et spatialement décorrélé.

Indices de performances Afin de comparer les résultats, nous utilisons comme indice de performance l'erreur d'Amari, défini par [ACY96] :

$$d(A,W) = \frac{1}{2p(p-1)} \sum_{i=1}^{p} \left(\frac{\sum_{j=1}^{p} |b_{ij}|}{\max_j |b_{ij}|} - 1 \right) + \frac{1}{2p(p-1)} \sum_{j=1}^{p} \left(\frac{\sum_{i=1}^{p} |b_{ij}|}{\max_i |b_{ij}|} - 1 \right) \tag{5.10}$$

où $b_{ij} = (AW^{-1})_{ij}$, A est la matrice de mélange et W est la matrice de mélange estimée. Cet indice mesure les performances de la séparation et vaut zéro pour une estimation parfaite de la matrice de mélange. On peut également utiliser comme indice de performance l'*Erreur Quadratique Moyenne* (EQM) d'estimation pour chaque source j, définie par :

$$EQM_j = \sum_{k=1}^{n} (s_{jk} - \hat{s}_{jk})^2 \tag{5.11}$$

ou le *Rapport Signal à Distorsion* (RSD) sur chaque source, défini par :

$$RSD_j(dB) = 10 \log_{10} \left(\frac{\sum_{k=1}^{n} s_{jk}^2}{\sum_{k=1}^{n} (s_{jk} - \hat{s}_{jk})^2} \right) \tag{5.12}$$

Niveau de bruit Le niveau de bruit dans la matrice d'observations X, mesuré en terme de rapport signal-bruit (RSB), est exprimé par :

$$RSB(dB) = 10 \log_{10} \left(\frac{\sum_{k=1}^{n} x_k^2}{\sum_{k=1}^{n} e_k^2} \right) \tag{5.13}$$

Les résultats de séparation (les vecteurs de mélange) dans le cas de 3 sources exponentielles sont présentés dans la suite. Pour la première experience nous avons utilisé une matrice de mélange $A^{3\times 10}$; les vecteurs colonne de la matrice A sont présentés dans les figures 5.7, 5.8 et 5.9 en rouge continu. La matrice des sources est de dimension 3×1000 et donc la matrice d'observations est de dimension 10×1000.
Un récapitulatif des résultats pour 2, 3 et 4 sources tirées à partir des distributions présentées dans la figure 5.6 est présenté à la fin de cette section.

5.3.2 Séparation par ACI

Dans ces simulations, la forme des vecteurs de mélange a été choisie afin de permettre leur identification visuelle. Les performances de séparation des algorithmes FastICA et JADE sont bonnes mais leur principal inconvénient est l'indétermination de signe et l'apparition des coefficients négatifs, comme on peut voir dans la figure 5.7.

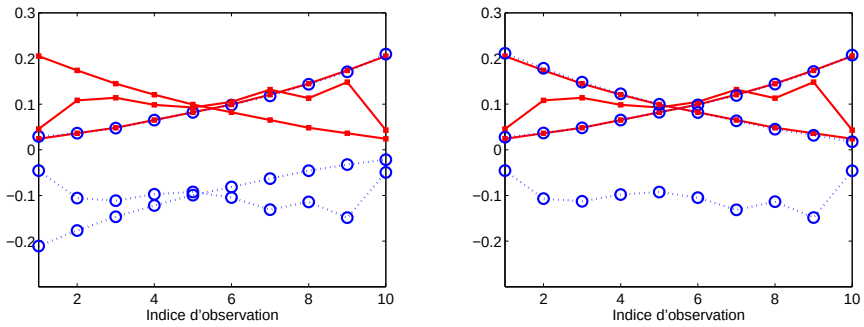
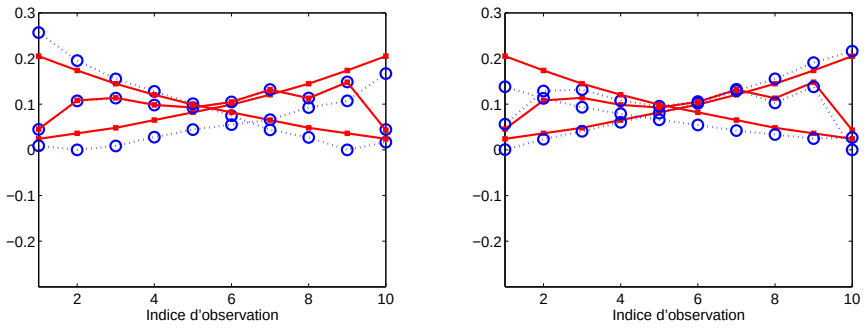


Figure 5.7: Résultats de la séparation en utilisant : a) FastICA et b) JADE : coefficients de mélange originaux (continu) et estimés (discontinu).

5.3.3 Séparation par la prise en compte de la non-négativité

Les méthodes basées sur la prise en compte de la non-négativité éliminent l'indétermination de signe des méthodes par ACI mais les performances de séparation sont inférieures, figure 5.8. La solution obtenue par ces méthodes n'est pas unique ce qui constitue un inconvénient majeur de ces méthodes. Pourtant, dans certaines applications les contraintes des lois physiques imposent que les résultats soient positifs. C'est le cas des applications en imagerie, spectroscopie, microscopie etc.



(a) (b)

Figure 5.8: Résultats de la séparation en utilisant : a) NMF-ALS et b) NMF : coefficients de mélange originaux (continu) et estimés (discontinu).

5.3.4 Séparation par l'algorithme PExSAS

La méthode proposée basée sur la recherche des points extrêmes d'une distribution fournit une solution unique au problème de séparation de sources non-négatives, figure 5.9. L'indétermination de signe est aussi éliminée car les vecteurs de mélange sont des points dans l'ensemble des observations X. Pourtant, cette méthode ne garantit pas la non-négativité des sources, mais celle-ci peut être obtenue par une seuillage des sources estimées.

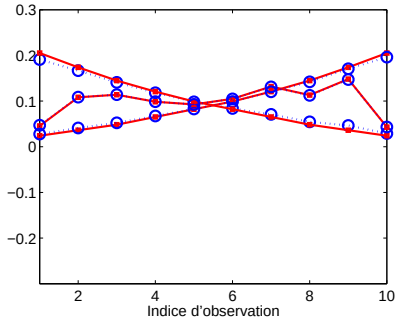


Figure 5.9: Résultats de la séparation en utilisant PExSAS : coefficients de mélange originaux (continu) et estimés (discontinu).

5.3.5 Récapitulatif des résultats

Deux sources

Dans le cas de deux sources, nous avons testé l'algorithme sur toutes les combinaisons possibles des distributions présentées dans la figure 5.6. Comme nous avons vu précédemment, l'algorithme proposé fournit de très bons résultats dans le cas où la probabilité d'avoir une

des sources active et toutes les autres inactives est importante. C'est le cas des sources parcimonieuses avec des distributions proches d'une distribution exponentielle. Pourtant, dans le cas de deux sources non-négatives, l'algorithme proposé dépasse tous les autres algorithmes de séparation en terme de performance, comme on peut voir dans le tableau 5.1.

	Fdp des sources	FastICA	JADE	NMF ALS	NMF(Lee-Seung)	PExSAS
1	a+a	0.2003	0.1772	0.0115	0.1586	**0.0045**
2	b+b	0.1077	0.0439	0.0039	0.0939	**0.0010**
3	c+c	0.0293	0.0190	0.0066	0.1021	**0.0020**
4	d+d	0.0221	0.0148	0.0026	0.0878	**0.0003**
5	e+e	0.1602	0.0396	0.0037	0.0996	**0.0009**
6	f+f	0.0168	0.0145	0.0104	0.1226	**0.0035**
7	g+g	0.0225	0.0160	0.0125	0.1186	**0.0044**
8	a+b	0.0876	0.0951	0.0223	0.1147	**0.0117**
9	a+c	0.0674	**0.0410**	0.1775	0.3244	0.0959
10	a+d	0.0499	**0.0375**	0.1975	0.3242	0.0955
11	a+e	0.0664	**0.0553**	0.0840	0.2369	0.0560
12	a+f	0.0288	0.0182	0.0272	0.1698	**0.0132**
13	a+g	0.0393	0.0238	0.0261	0.2016	**0.0127**
14	b+c	0.1501	0.1564	0.0182	0.2920	0**.0037**
15	b+d	0.1460	0.1666	0.0188	0.2900	**0.0051**
16	b+e	0.1229	0.0795	0.0083	0.2336	**0.0023**
17	b+f	0.0318	0.0486	0.0071	0.0898	**0.0024**
18	b+g	0.0338	0.0435	0.0077	0.0971	**0.0025**
19	c+d	0.0268	0.0159	0.0039	0.0889	**0.0011**
20	c+e	0.0636	0.0679	0.0052	0.1208	**0.0016**
21	c+f	0.0699	0.0502	0.0467	0.2880	**0.0187**
22	c+g	0.0874	0.0580	0.0653	0.2969	**0.0253**
23	d+e	0.0562	0.0668	0.0043	0.1242	**0.0009**
24	d+f	0.0667	0.0472	0.0468	0.2787	**0.0173**
25	d+g	0.0724	0.0689	0.0607	0.2876	**0.0186**
26	e+f	0.0489	0.0731	0.0221	0.2872	**0.0069**
27	e+g	0.0554	0.0721	0.0293	0.2941	**0.0111**
28	f+g	0.0195	**0.0137**	0.0370	0.1355	0.0205

Tableau 5.1: Résultats des simulations dans le cas de deux sources et 1000 données (l'erreur d'Amari) ; la matrice d'observation est de dimension 2×1000. Les *fdp* des sources sont indiquées dans la deuxième colonne du tableau et correspondent aux *fdp* présentées dans la figure 5.6. Les résultats représentent la moyenne de 100 essais. Les distributions sont bornées $[0\ \infty)$. Les meilleurs taux de séparation sont mis en évidence dans le tableau.

Trois sources

Dans ce cas, les sources ont été tirées aléatoirement à partir de trois distributions. Les performances de l'algorithme proposé restent comparables avec celles des algorithmes FastICA et JADE et de plus, elles sont meilleures que celles des méthodes de séparation par factorisation en matrices non-négatives, tableau 5.2. Les meilleures performances sont obtenues dans le cas des sources exponentielles.

	Fdp des sources	FastICA	JADE	NMF ALS	NMF(Lee-Seung)	PExSAS
1	c+c+c	0.1159	**0.0357**	0.1666	0.4303	0.0675
2	c+d+e	0.1538	0.0960	0.1256	0.2319	**0.0454**
3	d+e+f	0.1327	**0.0855**	0.2128	0.4402	0.1247
4	e+e+e	0.1612	0.0824	0.0853	0.2154	**0.0298**

Tableau 5.2: Résultats des simulations dans le cas de trois sources et 1000 données (l'erreur d'Amari) dans le cas d'un $RSB = 50dB$; la matrice d'observation est de dimension 3×1000. Les *fdp* des sources sont indiquées dans la deuxième colonne du tableau. Les résultats représentent la moyenne de 100 essais. Les distributions sont bornées $[0\ \infty)$. Les meilleurs taux de séparation sont mis en évidence dans le tableau.

Quatre sources

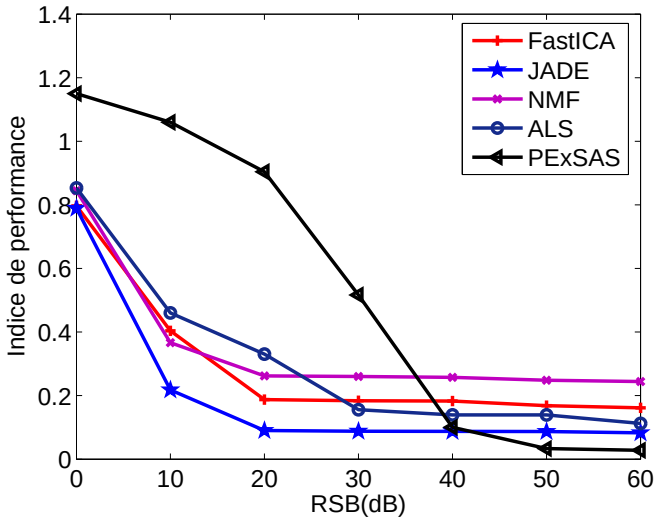
En augmentant le nombre de sources, les performances de l'algorithme proposé se dégradent, tableau 5.3 car la probabilité d'avoir des instances où une seule source est active est de moins en moins importante. Pourtant, quand les sources sont des distributions exponentielles, ses performances restent dans des limites acceptables par rapport aux algorithmes utilisés pour la comparaison.

	Fdp des sources	FastICA	JADE	NMF ALS	NMF(Lee-Seung)	PExSAS
1	c+c+c+c	0.0257	**0.0168**	0.5216	0.6976	0.6538
2	e+e+e+e	0.2443	**0.0400**	0.3607	0.6493	***0.0779***
3	c+d+e+f	**0.0518**	0.0523	0.4855	0.6660	0.6266

Tableau 5.3: Résultats de la séparation dans le cas de quatre sources et 10000 données (l'erreur d'Amari) dans le cas d'un $RSB = 50dB$; la matrice d'observation est de dimension 4×1000. Les *fdp* des sources sont indiquées dans la deuxième colonne du tableau. Les résultats représentent la moyenne de 100 essais. Les distributions sont bornées $[0\ \infty)$. Les meilleurs taux de séparation sont mis en évidence dans le tableau.

5.3.6 Influence du niveau du bruit

L'étude de l'influence du niveau de bruit sur les performances de l'algorithme proposé a été réalisée en faisant varier progressivement le niveau de bruit sur la matrice d'observations X. Les résultats sont présentés dans la figure 5.10. Dans le cas d'un RSB très grand, les performances de séparation de l'algorithme PExSAS sont supérieures à celles des autres algorithmes pour des sources exponentielles. Par contre, l'algorithme proposé est très sensible au bruit, ce qui fait qu'une méthode de réduction de bruit doit précéder la séparation dans le cas des observations noyées dans du bruit.



(a)

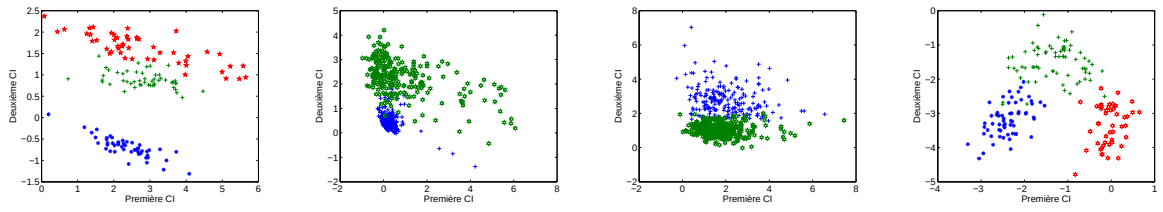
Figure 5.10: Influence du niveau de bruit sur les performances de la séparation ; comparaison avec les méthodes FastICA, JADE, NMF et NMF-ALS. Le nombre de sources est $p = 3$, le nombre des observations $m = 10$ et le nombre d'instances d'observations $n = 1000$. La distribution des sources est exponentielle.

5.4 Evaluation des méthodes de SAS dans le contexte de réduction de la dimension pour la classification non supervisée

Dans cette section nous évaluons les performances des méthodes de SAS dans le contexte de la réduction de dimension pour la classification non supervisée de données multivariées. Les bases de données utilisées sont celles présentées dans le quatrième chapitre. Parmi celles-ci nous avons choisi celles où les données sont décrites par des attributs non-négatifs : Iris, WBC, WDBC et Wine.

5.4.1 Réduction de dimension par l'ACI

Une première approche consiste à réaliser la réduction de la dimension par des méthodes de séparation utilisant le critère d'indépendance, FastICA et JADE. La nouvelle représentation 2D des données nous permet de faire une inspection visuelle sur les données ainsi que sur la forme et la distribution des classes. La figure 5.11 montre le résultat de la réduction de dimension par FastICA. Les résultats de la réduction de dimension par l'algorithme JADE sont présentés dans les annexes. Dans ce nouvel espace de représentation, l'algorithme *C-moyenne* a été utilisé pour classifier les données et les résultats sont présentés dans le tableau 5.4.

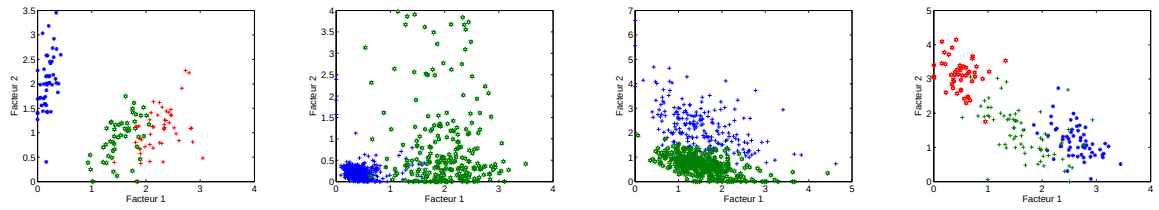


(a) (b) (c) (d)

Figure 5.11: Représentation des données dans l'espace des deux premières composantes indépendantes obtenues par FastICA pour les bases de données Iris (a), WBC (b), WDBC (c), Wine (d).

5.4.2 Réduction de dimension par la prise en compte de la non-négativité

Les méthodes de séparation par la prise en compte de la non-négativité ont été aussi testées dans le contexte de la réduction de dimension des données multivariées. Celles-ci n'utilisent pas les statistiques des données pour réaliser la séparation, mais seulement les contraintes de non-négativité de la matrice des sources et des vecteurs de mélange. Ceci fait que ces méthodes sont moins puissantes en ce qui concerne la réconstitution des sources que les méthodes d'ACI. Pourtant, dans certaines applications qui demandent des résultats non-négatifs ces méthodes fournissent des résultats plus réalistes que les méthodes d'ACI. Les résultats de la réduction de la dimension par la méthode NMF, sur les bases de données présentées au début de cette section sont presentés dans la figure A.2 et les résultats de la classification dans l'espace de facteurs non-négatifs sont presentés dans le tableau 5.4. Les résultats de la réduction de dimension par l'algorithme NMF ALS sont présentés dans les annexes.

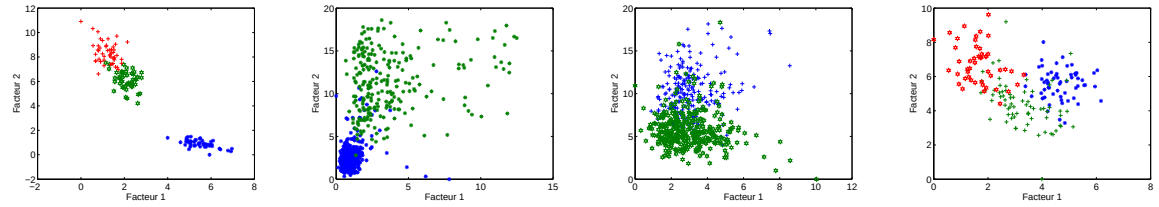


(a) (b) (c) (d)

Figure 5.12: Représentation des données dans l'espace des deux premiers facteurs obtenus par factorisation en matrices non-négatives, l'algorithme NMF pour les bases de données Iris (a), WBC (b), WDBC (c), Wine (d).

5.4.3 Réduction de dimension par l'algorithme PExSAS

Dans la troisième simulation de cette section nous testons l'algorithme PExSAS pour la réduction de dimension de données multivariées. Les résultats de classification sont comparables à ceux obtenus par les méthodes de factorisation en matrice non-négatives, voir tableau 5.4. Les avantages de cet algorithme sont l'unicité de la solution et sa simplicité.

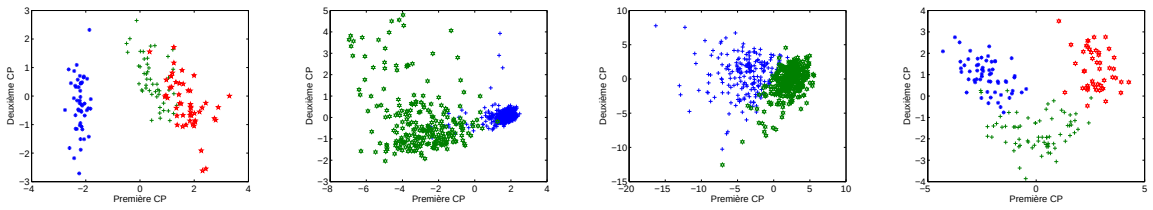


(a) (b) (c) (d)

Figure 5.13: Représentation des données dans l'espace des deux premiers facteurs obtenus par l'algorithme PExSAS pour les données Iris (a), WBC (b), WDBC (c), Wine (d).

5.4.4 Réduction de dimension par l'ACP

Nous présentons aussi pour comparaison, les résultats de la réduction de dimension obtenus par l'ACP. Les résultats de la classification dans l'espace des deux premières composantes principales sont présentés dans le tableau 5.4.



(a) (b) (c) (d)

Figure 5.14: Représentation des données dans l'espace des deux premières composantes principales pour les données Iris (a), WBC (b), WDBC (c), Wine (d).

5.4.5 Recapitulatif de résultats

Le tableau 5.4 fournit une évaluation numérique des performances de l'algorithme *C-moyenne* lorsque plusieurs méthodes de réduction de dimension sont utilisées pour chaque base de données. Les résultats représentent le taux moyen de classification après 10 essais.

	Iris	WBC	WDBC	Wine
FastICA + C-Moyenne	81.62%	94.14%	86.47%	97.17%
JADE + C-Moyenne	81.43%	96.34%	**90.51%**	**97.19%**
NMF + C-Moyenne	83.33%	96.34%	86.52%	79.28%
ALS + C-Moyenne	88.67%	**96.63%**	86.99%	79.78%
PExSAS + C-Moyenne	**97.52%**	95.17%	85.59%	84.27%
ACP + C-Moyenne	83.67%	92.46%	89.84%	94.68%

Tableau 5.4: Résultats de la classification par *C-moyenne*. Plusieurs méthodes de réduction de dimension ont été utilisées et sont indiquées dans la première colonne.

5.5 Conclusion

Dans ce chapitre nous avons proposé une nouvelle approche de séparation de sources non-négatives. L'originalité de cette méthode se trouve dans l'interprétation géométrique du modèle de mélange linéaire. Les simulations numériques ont permis de tester et de comparer la nouvelle méthode avec les méthodes de séparation de sources classiques. Les résultats de simulations montrent les avantages et les inconvénients de cette méthode. La méthode proposée est supérieure aux autres méthodes de séparation si pour chaque source il existe un instant i pour lequel elle est active et toutes les autres sont inactives. Si cette condition n'est pas satisfaite, la méthode PExSAS donne de bons résultats pour résoudre le problème de séparation de 2 et 3 sources positives et de densité bornée. Si le nombre de sources augmente, les performances de la séparation se dégradent car la probabilité d'avoir un instant où une seule source est active et les autres

inactives est de moins en moins importante. L'influence du niveau de bruit a été aussi mesurée. Les résultats montrent que la méthode proposée est plus sensible au bruit que les méthodes classiques de séparation. Ceci constitue un inconvénient et des techniques de réduction du bruit doivent être utilisées avant de réaliser la séparation dans le cas d'un bruit important.

Les méthodes de SAS ont été aussi testées dans le contexte de la réduction de dimension. Les résultats de simulations montrent que la réduction de dimension faite par des méthodes SAS peut améliorer les résultats de classification par rapport au cas où la réduction de dimension a été faite par l'ACP. Les meilleurs taux de classification dépendent de la méthode de réduction de dimension utilisée ; de bonnes performances sont obtenues si les méthodes de séparation par ACI sont employées. Pourtant, les méthodes de séparation par la prise en compte de la non-négativité et la méthode PExSAS donnent des résultats comparables aux méthodes de séparation par l'ACI. Le fait que les facteurs obtenus sont positifs constitue leur principal avantage dans des applications où la contrainte de non-négativité est imposée par les lois de la physiques (spectroscopie, imagerie multispectrale etc). Dans le chapitre suivant, nous allons utiliser les méthodes de SAS dans deux applications : la segmentation d'une image multispectrale et l'analyse des séries temporelles d'images médicales, et l'utilité des méthodes de séparation des sources par factorisation en matrices non-négatives sera mise en évidence.

Chapitre 6

Application à la segmentation des images multivariées

Les résultats théoriques obtenus dans les chapitres 4 et 5 ont été appliqués à la segmentation d'images multivariées par classification des pixels ainsi qu'à l'analyse de séries temporelles d'images médicales. Pour la première application (la segmentation d'une image multivariée de microscopie), deux directions sont suivies : la première où les métriques non euclidiennes sont employées afin d'améliorer les résultats de la classification obtenus en utilisant la métrique euclidienne comme mesure de similarité ; aucune méthode de réduction de dimension n'est utilisée et la méthode *C-moyenne* a été retenue pour la classification. Les résultats sont présentés dans la sous section 6.1.2. La deuxième direction consiste à utiliser les techniques de SAS afin de réduire la dimension de données. L'objectif est de réduire la complexité des calculs des algorithmes de classification, mais aussi de mettre en évidence des structures cohérentes, ayant un sens physique réel, présentes dans l'ensemble de données. Pour ceci, des méthodes de séparation par ACI (JADE), des méthodes par la prise en compte de la non-négativité ainsi que la méthode proposée PExSAS ont été employées. Les résultats sont présentés dans la sous-section 6.1.7. Pour la classification des pixels dans le nouvel espace de dimension réduite, les méthodes *C-moyennes*, *Mean-Shift* et *Parzen-Watershed* ont été retenues. Dans la deuxième application (l'analyse de série temporelles d'images médicales) les méthodes de séparation par la prise en compte de la non négativité et la méthode PExSAS sont mises en oeuvre.

6.1 Segmentation des images de microscopie

6.1.1 Description des données et problématique

Pour cette application, l'image multispectrale fournie est une image de grain d'orge de la variété Clarine acquise en fluorescence par microscopie confocale à balayage laser. Une coupe transversale au milieu du grain a été observée avec un objectif X10. Le tableau suivant décrit précisément les conditions d'acquisition :

Séquence de 19 images	Longueur d'onde d'excitation (en nm)	Longueur d'onde d'émission (ou réfléchie) (en nm)
01	633(laser rouge)	> 665
02	543(laser vert)	> 570
03	543	575 et 640
04	543	> 665
05	488(laser bleu)	> 515
06	488	entre 515 et 565
07	488	> 570
08	488	entre 575 et 640
09	488	> 665
10	364(laser UV)	> 397
11	364	entre 450 et 490
12	364	> 515
13	364	entre 515 et 565
14	364	> 570
15	364	entre 575 et 640
16	364	> 665
17	488 (laser bleu)	entre 510 et 525
18	364	entre 400 et 435
19	364	entre 510 et 525

Tableau 6.1: Conditions d'acquisition des composantes multispectrales

La figure 6.1 présente les 19 composantes de l'image multispectrale.

Les grains de céréales sont constitués de plusieurs tissus qui se superposent. La partie centrale contient de l'amidon et les couches externes servent de protection au grain. Les tissus externes ont la propriété d'être naturellement fluorescents. Les composés chimiques responsables de la fluorescence des tissus sont la cutine, l'acide férulique et la lignine. Les pseudo-spectres de référence de l'acide férulique et de la lignine sont présentés dans la figure 6.2, le deuxième respectivement le troisième spectre. L'acide férulique est présent dans les grains de céréales sous deux formes hybrides. Dans la figure 6.2, seulement une de ces formes est présente. Nous ne

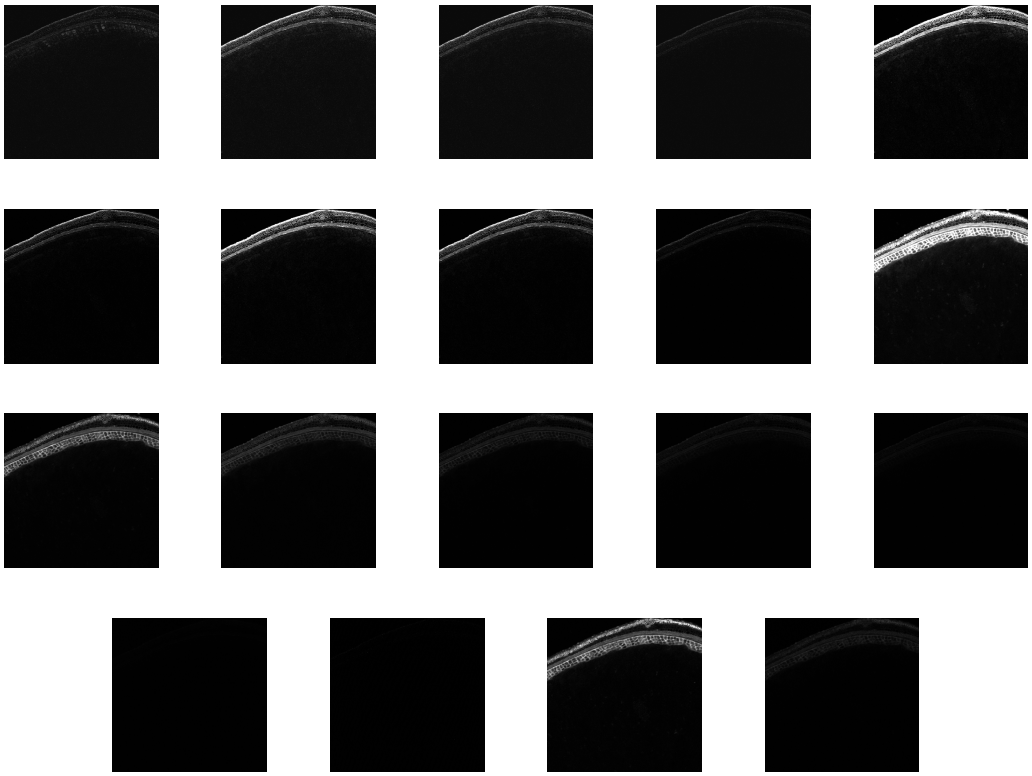


Figure 6.1: Image multispectrale représentant une coupe transversale dans un grain d'orge.

disposons pas du spectre de la cutine. La distribution des tissus dans la couche externe du grain d'orge est présenté dans la figure 6.3. L'acide férulique est localisé dans les parois des cellules à aleurones ainsi que dans les glumelles ; la lignine se retrouve dans les glumelles tandis que la cutine est retrouvée seulement dans les grains d'orge dans le tissu externe qui sert de protection contre l'humidité. L'objectif est ici d'identifier les tissus à partir des constituants fluorescents présents dans la couche externe du grain de céréale. La microscopie par fluorescence permet de visualiser l'autofluorescence des tissus. Pour cela, plusieurs excitations laser associées à un jeu de filtres placés en réception permettent d'obtenir une pseudo image multi-spectrale à 19 composantes. Lorsque la composition chimique d'une partie de l'image est homogène, les pixels correspondants sont regroupés dans une même classe et localisés par leur carte de concentration.

6.1.2 Classification par des métriques non-euclidiennes

Les résultats de la segmentation par la méthode *C-moyenne* sont présentés dans cette sous-section ; le paramètre de la métrique est considéré comme étant un paramètre de la méthode de classification. La métrique euclidienne L_2, la métrique de Manhattan L_1 ainsi qu'une métrique

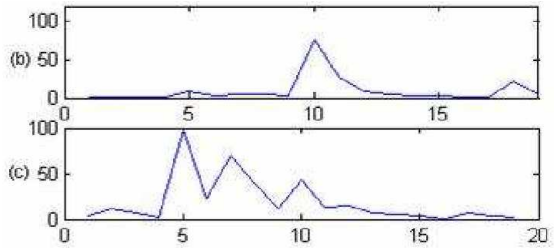


Figure 6.2: Signatures spectrales des composés chimiques responsables de la fluorescence des tissus : premier pseudo-spectre - acide férulique, deuxième pseudo-spectre - lignine ; on ne dispose pas du pseudo-spectre de la cutine.

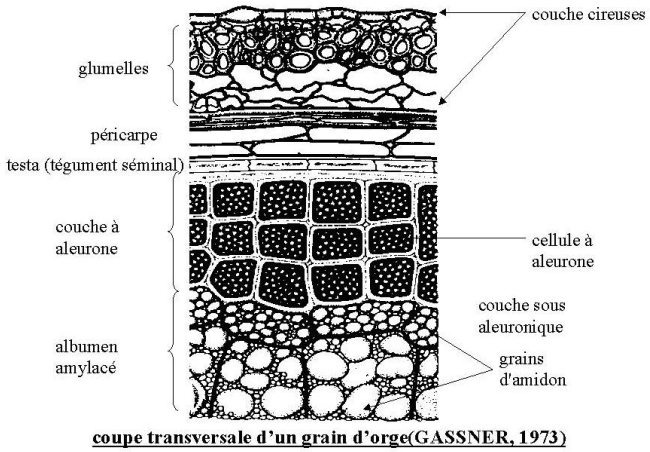


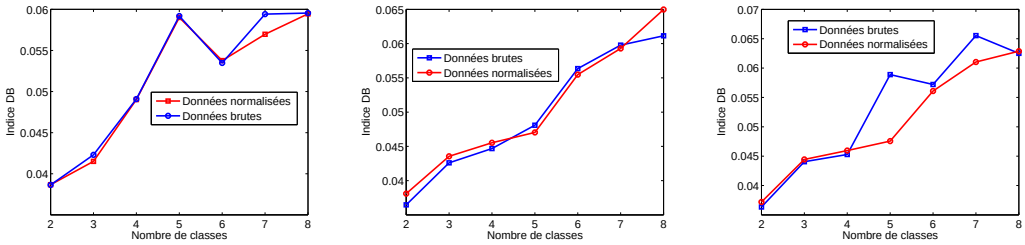
Figure 6.3: Coupe transversale dans un grain d'orge.

fractionnaire $L_{0.7}$ ont été prises en compte et les résultats sont évalués et comparés à l'aide de 2 indices : *l'indice Davies-Bouldin (DB)* et *l'indice compacité-séparabilité (CS)*. Dans une première démarche, nous testons l'effet de la normalisation des données sur les résultats de la classification ; ensuite nous choisissons la métrique optimale (parmi les trois métriques proposées) comme étant la métrique pour laquelle les indices de validité montrent les plus faibles valeurs. Le choix du nombre de classes est discuté et quelques résultats sont présentés.

Comparaison des résultats : données brutes vs. données normalisées

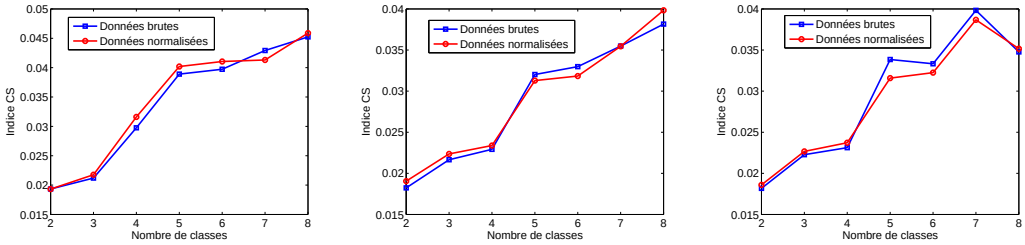
Les résultats obtenus sur les données brutes ont été comparés avec ceux obtenus sur les données normalisées dans l'intervalle [0 1]. Comme nous l'avons dit précédemment, la classification a été réalisée en considérant 3 métriques différentes : la métrique euclidienne L_2, la métrique de Manhattan L_1 et la métrique fractionnaire $L_{0.7}$. Les résultats sont comparés à l'aide des indices

DB figure 6.4 et CS figure 6.5. Les résultats montrent que l'influence de la normalisation des données sur les résultats de la classification est très faible pour cette application.



(a) (b) (c)

Figure 6.4: Indice DB : comparaison entre les résultats obtenus sur les données brutes et normalisées pour les métriques L_2 (a), L_1 (b) et $L_{0.7}$ (c). Les figures ne montrent pas de différence importante entre les résultats : on observe une faible amélioration des résultats dans le cas où les données normalisées sont utilisées.



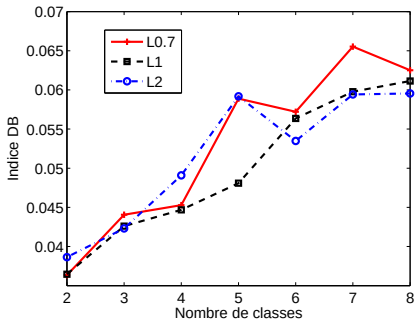
(a) (b) (c)

Figure 6.5: Indice CS : comparaison entre les résultats obtenus sur les données brutes et normalisées pour les métriques L_2 (a), L_1 (b) et $L_{0.7}$ (c).

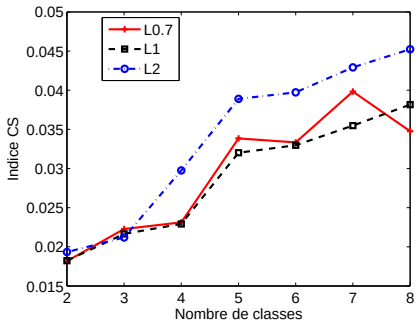
Choix de la métrique

La deuxième démarche consiste à choisir la métrique optimale. Les indices DB et CS ont été utilisés pour comparer les résultats et pour évaluer la meilleure métrique. Ceux-ci sont inversement proportionnels à la distance interclasse ; pour un nombre de classes fixé et pour différentes valeurs du paramètre r, la plus faible valeur de ces indices indique la métrique la plus discriminante. L'indice DB figures 6.6 (a) et 6.7 (a) montre que la norme L_1 donne les meilleurs résultats pour une classification en 4 et 5 classes ; sinon, les résultats sont comparables avec ceux obtenus avec la norme L_2 et $L_{0.7}$. Par contre, l'indice CS figures 6.6 (b) et 6.7 (b) montre la supériorité des normes L_1 et $L_{0.7}$ par rapport à la norme euclidienne dans le cas de données brutes ainsi que dans le cas de données normalisées. Pour cette application, nous proposons

d'utiliser la métrique de Manhattan comme mesure de similarité.

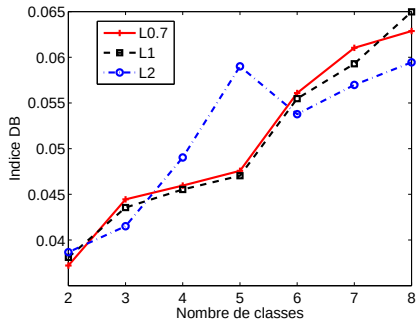


(a)

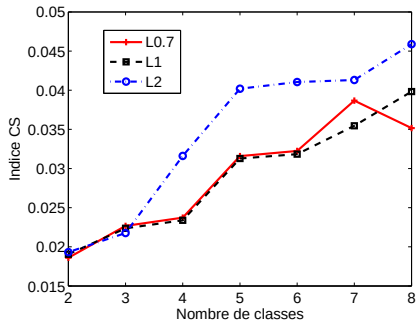


(b)

Figure 6.6: Critères de validation du nombre de classes - données brutes : l'indice DB et l'indice CS. Comparaison entre différentes normes.



(a)



(b)

Figure 6.7: Critères de validation du nombre de classes - données normalisées : l'indice DB et l'indice CS. Comparaison entre différentes normes.

Choix du nombre de classes

Le choix du nombre de classes dans un problème de classification non supervisée doit être réalisé tout en respectant les informations fournies par un ou plusieurs indices de validité. Les indices DB et CS sont des indicateurs bien adaptés à la méthode *C-moyennes* et ils sont présentés dans le premier chapitre. Ces deux indices indiquent le nombre optimal de classes comme étant la valeur pour laquelle ces fonctions présentent leur minimum global. Pour cette application, la valeur minimale de ces fonctions indique une partition en deux classes comme étant optimale. Or cette solution n'est pas correcte car les trois classes correspondant aux composés fluorescents ne sont pas retrouvées. Ceci représente une des limites de ces indices ainsi que de la méthode *C-moyenne* qui favorise des classes plus peuplées au détriment de classes moins peuplées.

Nous faisons donc appel aux connaissances *a priori* sur l'ensemble de données : la couche externe d'un grain d'orge contient 3 composés fluorescents : la cutine, la lignine et l'acide

férulique. Ces informations imposent une partition en minimum 4 classes, la quatrième classe représentant le fond de l'image. Une partition en 5 classes est aussi envisageable en prenant en compte les différentes concentrations de l'acide férulique. Les figures 6.8 et 6.9 présentent les résultats de la classification en 4 et 5 classes pour les différentes normes utilisées par l'algorithme *C-moyenne*.

Classification en 4 classes

Les 4 classes recherchées doivent correspondre aux trois composés fluorescents présents dans la couche externe du grain d'orge et au fond de l'image. Les résultats obtenus en utilisant les normes L_1 et $L_{0.7}$ indiquent les 4 classes recherchées : la cutine dans le tissu externe, l'acide férulique dans la couche à aleurone et la lignine dans les glumelles (le tissu situé au milieu de la couche externe). Par contre, les parois des cellules aleurones (là où se trouve la plupart de l'acide férulique) ne sont pas bien mises en évidence. Si la métrique euclidienne est utilisée, la classe correspondant à la cutine n'est pas mise en évidence.

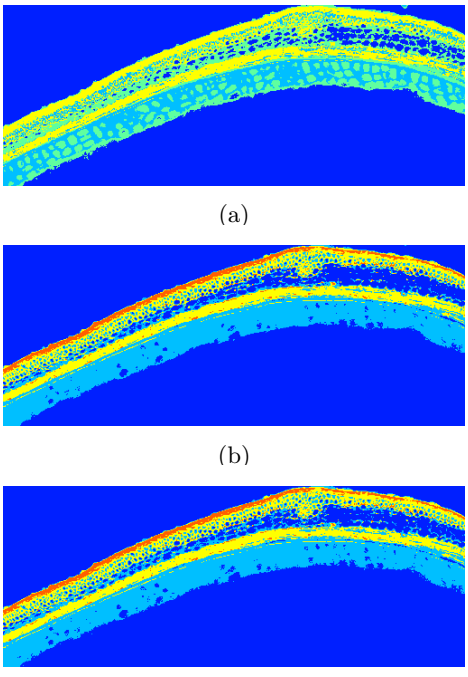


Figure 6.8: Résultats de la segmentation pour 4 classes : a) C-moyenne + L_2, b) C-moyenne + L_1 et c) C-moyenne + $L_{0.7}$.

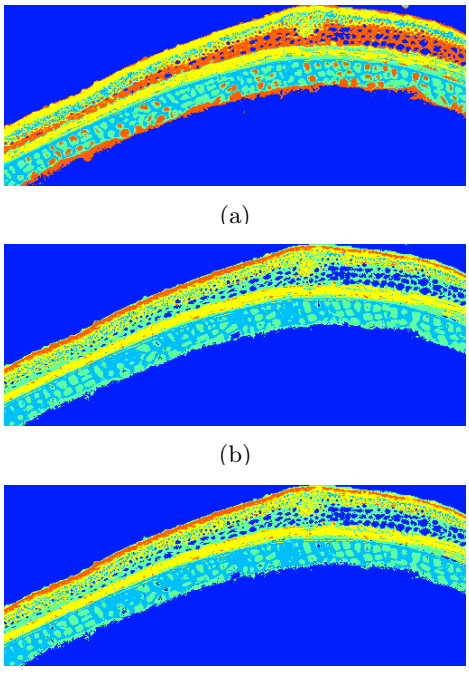


Figure 6.9: Résultats de la segmentation pour 5 classes : a) C-moyenne + L_2, b) C-moyenne + L_1 et c) C-moyenne + $L_{0.7}$.

Classification 5 classes

Les 5 classes recherchées correspondent aux composés fluorescents : cutine, lignine, acide férulique (avec 2 classes pour l'acide férulique en fonction de sa concentration) et la cinquième classe correspond au fond de l'image. Les normes L_1 et $L_{0.7}$ donnent des résultats similaires en indiquant les classes recherchées, tandis que la norme L_2 ne met pas en évidence la classe correspondant à la cutine.

6.1.3 Segmentation par réduction de dimension et classification

Cette deuxième approche est résumée en deux étapes : la première - la réduction de dimension par une méthode de SAS, et la deuxième - la classification non supervisée des pixels dans l'espace de sources. Pour reduire la dimension des données nous avons utilisé une méthode de séparation par ACI - JADE, une méthode de séparation par la prise en compte de la non-négativité - NMF et la méthode géométrique proposée - PExSAS. Pour la classification de pixels nous avons utilisé les algorithmes *C-moyennes*, *Parzen-Watershed* et *Mean-Shift*. L'estimation de la dimension intrinsèque des données a été réalisée en prenant en compte les connaissances *a priori* sur les données.

Estimation de la dimension intrinsèque des données

Une méthode classique pour l'estimation de la dimension intrinsèque d'un ensemble de données multivariées consiste à choisir le nombre de valeurs propres supérieures à un certain seuil [JD88] ; ceci revient à choisir les vecteurs propres concentrant la plupart de l'énergie de l'ensemble des données. En pratique, un pourcentage supérieur à 80% de l'énergie totale des données est suffisant pour pouvoir justifier la réduction de dimension. Pourtant, dans certains cas cette valeur peut s'avérer insuffisante car certaines composantes de faible énergie peuvent contenir de l'information utile pour l'analyse. Certains scientifiques [PBJD79] ont affirmé leurs réserves concernant l'estimation de la dimension intrinsèque des données multivariées par cette méthode. La distribution énergétique des composantes principales obtenue après l'ACP est présentée dans le tableau 6.2. Ces résultats nous indiquent la dimension intrinsèque des données égale à 2 car plus de 90% de l'énergie totale est concentrée dans les deux premières composantes principales.

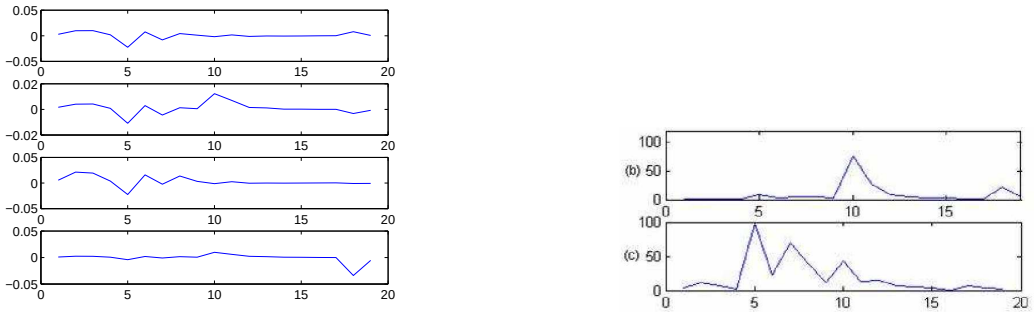
Composante principale	Distribution d'énergie
CP1	73.0692%
CP2	20.1714%
CP3	3.0561%
CP4	1.5576%
CP5	0.4478%
...	...

Tableau 6.2: Distribution d'énergie des composantes principales.

Dans cette application nous ne suivons pas ce principe pour estimer la dimension intrinsèque, mais nous allons la considérer comme étant égale au nombre de sources. Pourtant, l'estimation du nombre de sources est généralement réalisée par le même principe (*e.g.* en choisissant le nombre de vecteurs propres de la matrice de covariance pour lesquels les valeurs propres correspondantes sont supérieures à un certain seuil). Nous décidons donc d'utiliser les connaissances *a priori* et de considérer le nombre de sources et implicitement la dimension intrinsèque des données égale au nombre de composés chimiques purs présents dans la couche externe de grain d'orge. Nous poursuivons donc une séparation en 4 sources.

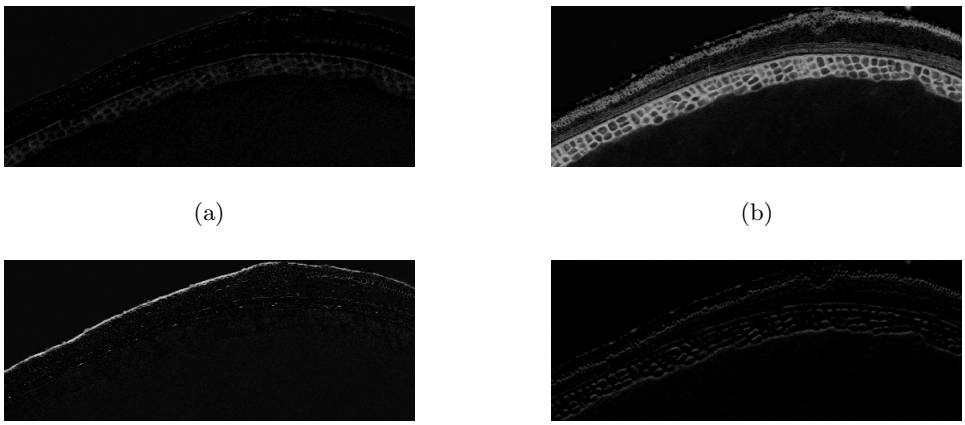
6.1.4 Classification dans l'espace des composantes indépendantes

Les résultats de la séparation par l'algorithme JADE sont présentés dans les figures 6.10 (a) (les vecteurs de mélange) et 6.11 (les sources). Les résultats obtenus par JADE ne nous permettent pas d'obtenir l'identification des spectres référence, figure 6.10 (b).



(a) (b)

Figure 6.10: Réduction de la dimension par ACI (algorithme JADE) : a) vecteurs de mélange, b) pseudo-spectres de référence.

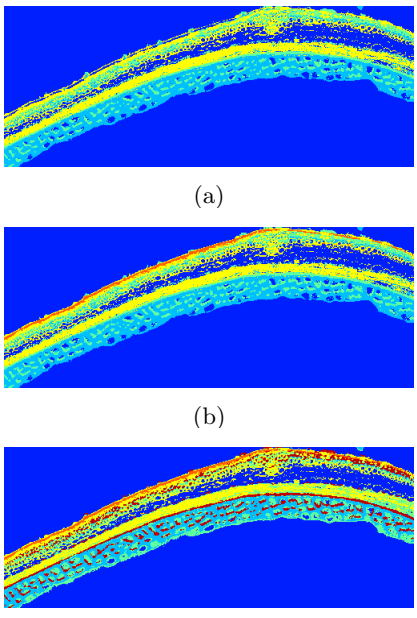


(a) (b)

(c) (d)

Figure 6.11: Réduction de dimension par ACI (algorithme JADE) : les sources.

JADE + C-moyenne L'indice *DB* présenté dans les annexes propose une partition en 2 classes. Cette solution n'est pas satisfaisante. Les résultats de la classification en 4, 5 et 6 classes sont présentés dans la figure 6.12.

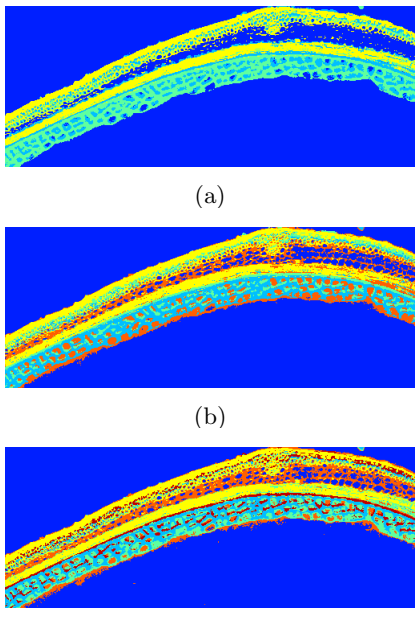


(a)

(b)

(c)

Figure 6.12: Résultats classification JADE + C-moyennes : a) 4 classes, b) 5 classes et c) 6 classes.



(a)

(b)

(c)

Figure 6.13: Résultats classification NMF + C-moyennes : a) 4 classes, b) 5 classes et c) 6 classes.

Les quatre classes recherchées correspondant à la cutine, la lignine, l'acide férulique et au fond de l'image sont mises en évidence par une partition en 5 classes. Les parois des cellules aleurones où la plupart de l'acide férulique est concentrée ne sont pas bien identifiés. La cinquième classe correspond à l'intérieur des cellules aleurones qui contiennent des faibles quantités d'acide férulique.

JADE + Mean-Shift Dans l'espace des sources obtenues par la méthode JADE, le critère de stabilité propose une classification en 2 classes. La solution n'est pas satisfaisante. Nous décidons de classifier les données par la méthode *Parzen-Watershed* dans l'espace des deux premières composantes indépendantes pour réduire l'information redondante.

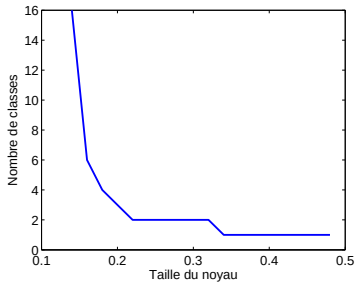


Figure 6.14: Indice de stabilité du nombre de classes : JADE + Mean-Shift.

JADE + Parzen-Watershed Les résultats de la classification dans l'espace des deux premières sources obtenus par l'algorithme JADE sont présentés dans la suite. Le critère de stabilité propose comme solution 4, 3 et 2 classes, figure 6.15. Les résultats de la classification en 4 classes sont présentés dans la figure A.4. La classe correspondant à la cutine n'est pas mise en évidence.

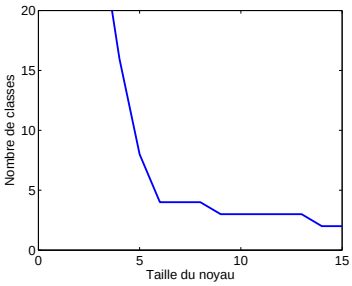


Figure 6.15: Indice de stabilité du nombre de classes : JADE + Parzen-Watershed.

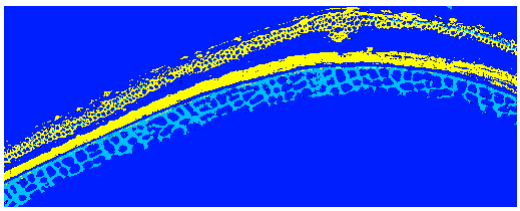
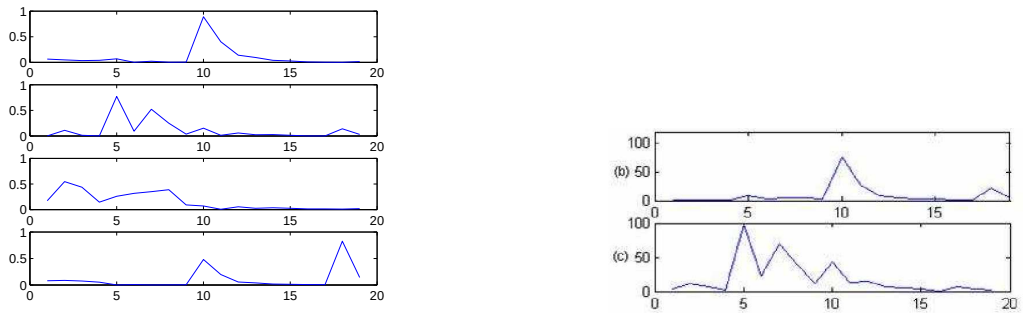


Figure 6.16: Résultats de la classification : JADE + Parzen-Watershed.

6.1.5 Classification dans l'espace des composantes non négatives

Les données et les résultats recherchés sont et doivent être positifs et donc, il est préférable d'appliquer des méthodes de SAS qui tiennent compte des contraintes physiques du problème traité, comme la positivité des données [END+, GEV+, BNDB04]. L'algorithme NMF a été utilisé pour la réduction de la dimension des images multi-spectrales dans [NL07, NLB+07].

Les résultats de la séparation par l'algorithme NMF sont présentés dans les figures 6.17 (les vecteurs de mélange) (a) et 6.18 (les sources). Dans la figure 6.17 (a), le premier et le deuxième spectre correspondent respectivement à l'acide férulique à la lignine. Le troisième spectre correspond à la cutine et il est identifiable par sa localisation spatiale (le tissu externe du grain d'orge, figure 6.18 (c)) tandis que le quatrième spectre correspond à la deuxième forme hybride de l'acide férulique. La distribution spatiale de l'acide férulique est identique pour les deux formes hybrides (couche aleurone et glumelles), figures 6.18 (a) et (d), sauf que les concentrations sont différentes.

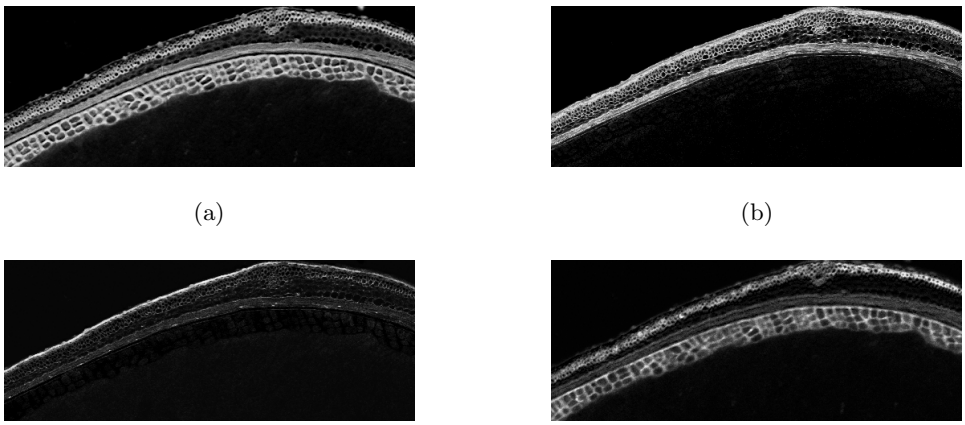


(a) (b)

Figure 6.17: Réduction de dimension par NMF : a) vecteurs de mélange, b) pseudo-spectres de référence : premier spectre - acide férulique, deuxième spectre - lignine.

NMF + C-moyenne L'indice DB, voir annexes, propose une partition en 2 classes. Cette solution n'est pas satisfaisante. Les résultats de classification en 4, 5 et 6 classes sont présentés dans la figure 6.13.

NMF + Mean-Shift Dans ce cas, l'indice de validité, figure 6.19 propose 3 solutions: 5, 4 et 2 classes. Parmi celles-ci, une classification en 5 et 4 classes correspond à nos besoins. Les résultats de la classification en 5 classes sont présentés dans la figure 6.20. Les classes sont bien identifiables, mais la cutine est divisée en 2 sous-classes. Une partition en 5 classes est préférable



(a) (b)

(c) (d)

Figure 6.18: Réduction de dimension par NMF : les sources représentant la répartition spatiale des composés chimiques. a) première forme hybride de l'acide férulique, b) lignine, c) cutine, d) deuxième forme de l'acide férulique.

à une partition en 4 classes car la cutine est mieux mise en évidence.

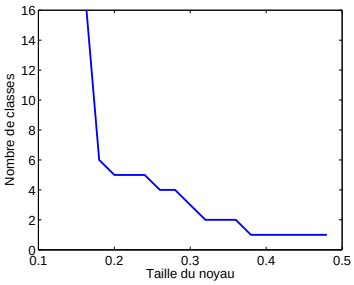


Figure 6.19: Indice de stabilité du nombre de classes : NMF + Mean-Shift.

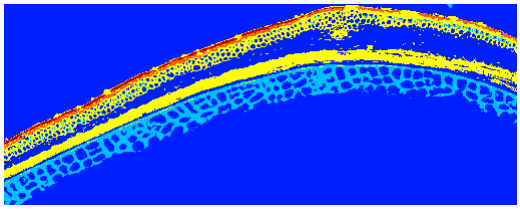


Figure 6.20: Résultat de la classification : NMF + Mean-Shift.

NMF + Parzen-Watershed Dans l'espace des deux premières sources obtenues par l'algorithme NMF, le critère de stabilité propose une classification en 5 et 3 classes, figure 6.21. Les résultats de la classification en 5 classes sont présentés dans la figure A.5. La classe correspondant à la cutine est divisée en deux sous-classes. La lignine et l'acide férulique sont bien mis en évidence.

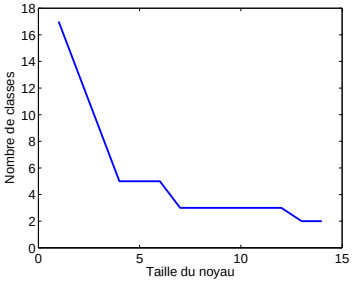


Figure 6.21: Indice de stabilité du nombre de classes : NMF + Parzen-Watershed.

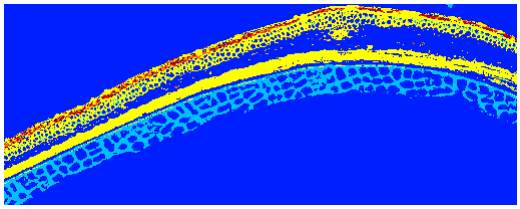
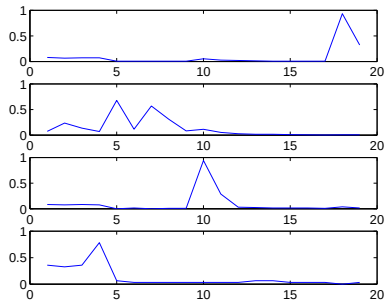


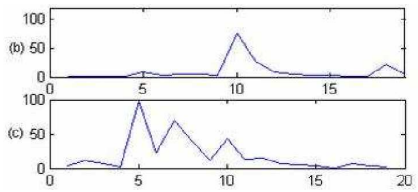
Figure 6.22: Résultats de la classification : NMF + Parzen-Watershed.

6.1.6 Classification dans l'espace des composantes géométriques

Les résultats de la séparation par l'algorithme PExSAS sont présentés dans les figures 6.23 (les vecteurs de mélange) (a) et 6.24 (les sources). Les résultats sont similaires avec ceux obtenus par la méthode NMF. Les spectres des deux formes hybrides de l'acide férulique (figure 6.23 (a), troisième et quatrième spectre) ainsi que la lignine (figure 6.23 (b), troisième et quatrième spectre) sont bien identifiés et ils sont quasiment identiques avec ceux estimés par la NMF. Par contre, les spectres correspondant à la cutine, estimés par les deux méthodes (NMF et PExSAS) sont différents.



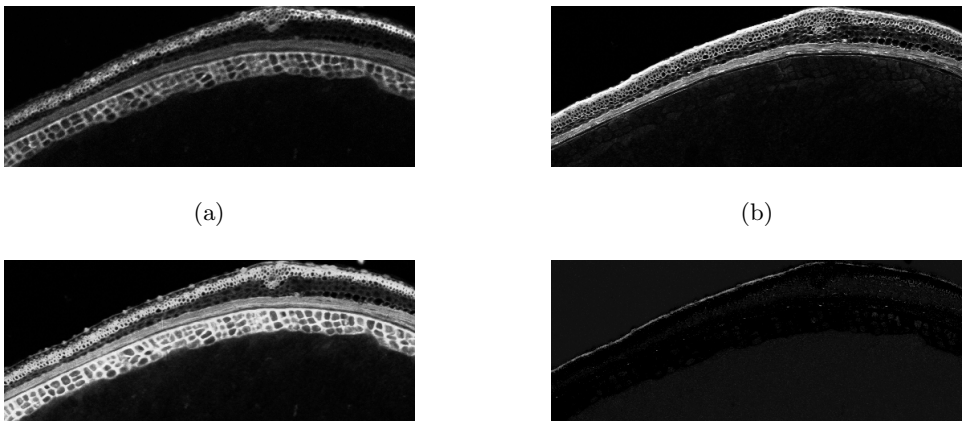
(a)



(b)

Figure 6.23: Réduction de dimension par PExSAS : a) vecteurs de mélange, b) pseudo-spectres référence : premier spectre - acide ferulique, deuxième spectre - lignine.

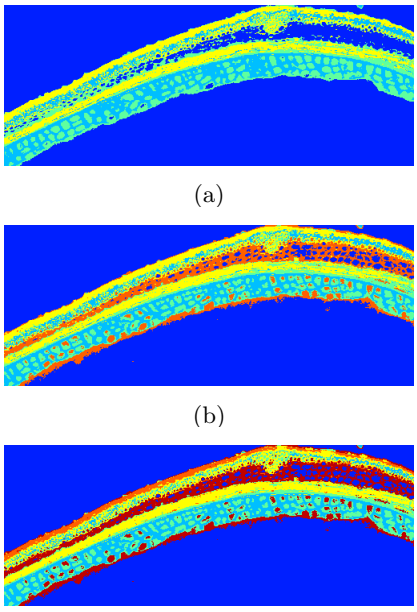
PExSAS + C-moyenne L'indice DB, voir annexes, propose toujours une partition en 2 classes. Les résultats de classification en 4, 5 et 6 classes sont présentés dans la figure 6.25.



(a) (b)

(c) (d)

Figure 6.24: Réduction de dimension par PExSAS : sources représentant la répartition spatiale des composés chimiques. (a) deuxième forme de l'acide férulique, (b) lignine, (c) première forme hybride de l'acide férulique, (d) cutine.

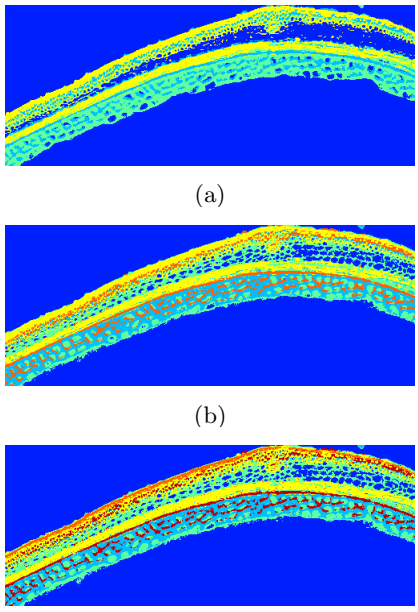


(a)

(b)

(c)

Figure 6.25: Résultats de la classification PExSAS + C-moyennes : a) 4 classes, b) 5 classes et c) 6 classes.



(a)

(b)

(c)

Figure 6.26: Résultats de la classification ACP + C-moyennes : a) 4 classes, b) 5 classes et c) 6 classes.

Une partition en 6 classes dans l'espace de sources obtenues par la méthode PExSAS met très bien en évidence les trois classes correspondant aux composés fluorescents présents dans les

grains de céréales. Une sixième classe est mise en évidence regroupant des tissus qui présentent une faible réponse à l'excitation.

PExSAS + Mean-Shift L'indice de validité, figure 6.27 propose 3 solutions: 5, 3 et 2 classes. Parmi celles-ci, une classification en 5 classes correspond à nos besoins. Les résultats de la classification en 5 classes sont présentés dans la figure 6.28. Les classes sont bien identifiables, mais la cutine est divisée en 2 sous-classes.

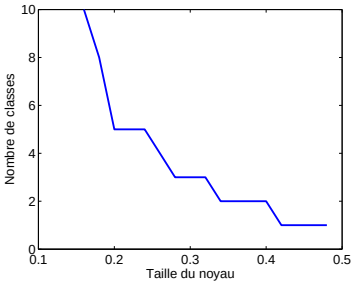


Figure 6.27: Indice de stabilité du nombre de classes : PExSAS + Mean-Shift.

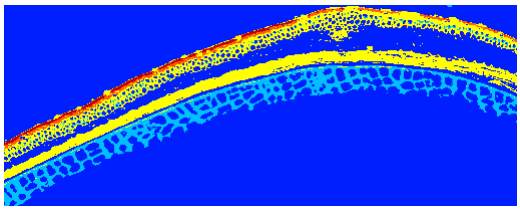


Figure 6.28: Résultat de la classification : PExSAS + Mean-Shift.

PExSAS + Parzen-Watershed Les résultats de la classification dans l'espace des deux premières sources obtenus par l'algorithme PExSAS sont présentés dans la suite. Le critère de stabilité propose comme solution 4 et 2 classes, figure 6.29. Les résultats de la classification en 4 classes sont présentés dans la figure A.6. Les trois classes correspondant aux composés fluorescents présents dans les tissus de grain d'orge sont bien mises en évidence.

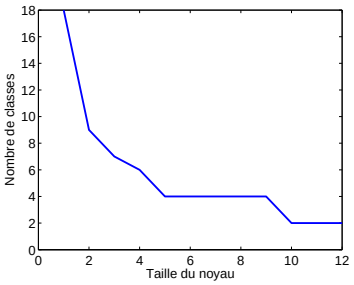


Figure 6.29: Indice de stabilité du nombre de classes : PExSAS + Parzen-Watershed.

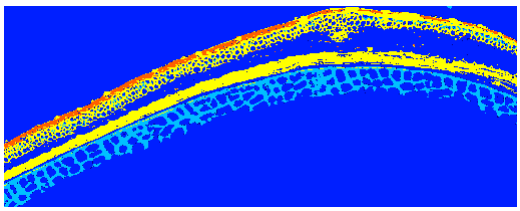
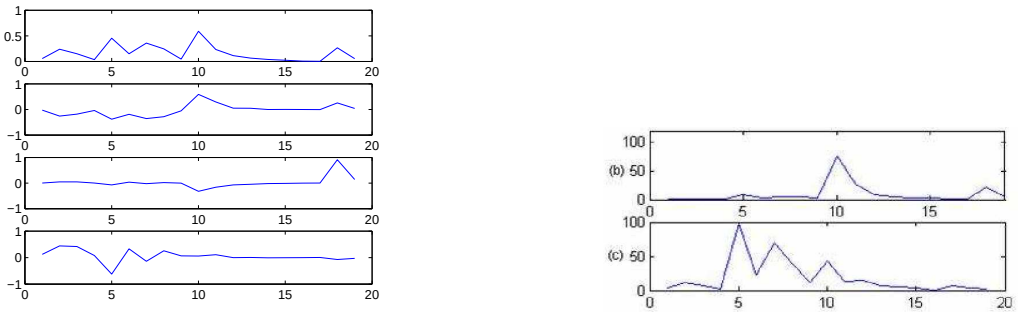


Figure 6.30: Résultats de la classification : PExSAS + Parzen-Watershed.

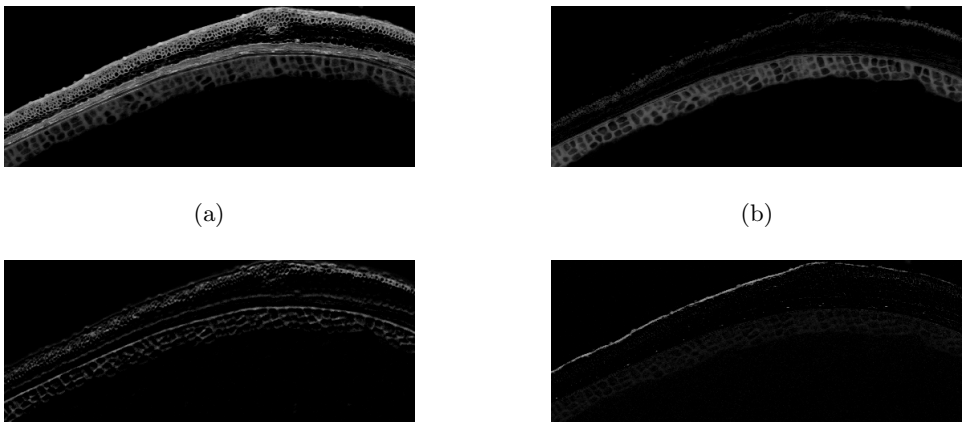
6.1.7 Classification dans l'espace des composantes principales

Pour comparaison, nous allons présenter les résultats de la séparation faite sous la contrainte d'orthogonalité des sources, réalisée par l'ACP: figure 6.31 (a) - les quatre premiers vecteurs propres, figure 6.32 les quatre premières composantes principales. L'identification des spectres de référence n'est pas possible.



(a) (b)

Figure 6.31: Réduction de dimension par ACP : les quatre premiers vecteurs propres.



(a) (b)

(c) (d)

Figure 6.32: Réduction de dimension par ACP : les quatre premières composantes principales.

ACP + C-moyenne L'indice DB est présenté dans les annexes. La solution proposée par cet indice (2 classes) n'est pas satisfaisante. Les résultats de classification pour 4, 5 et 6 classes sont présentés dans la figure 6.26.

Les trois classes recherchées sont mises en évidence par une partition en 6 classes. Par contre, une sixième classe indésirable est mise en évidence dans la couche aleurone.

ACP + Mean-Shift Dans l'espace des quatre premières composantes principales obtenues par l'ACP, le critère de stabilité propose 2 solutions: 3 et 2. Aucune de ces deux solutions n'est conforme avec les connaissances dont on dispose sur les données. Nous décidons de classifier les données par la méthode *Parzen-Watershed* dans l'espace des deux premières composantes indépendantes pour réduire l'information redondante.

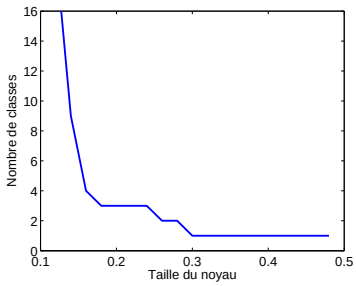


Figure 6.33: Indice de stabilité du nombre de classes : ACP + Mean-Shift.

ACP + Parzen-Watershed Les résultats de la classification dans l'espace des deux premières composantes principales obtenues par l'ACP sont présentés dans la suite. Le critère de stabilité propose comme solution 5, 4 et 2 classes, figure 6.34. Les résultats de la classification en 4 classes sont présentés dans la figure A.7. La classe correspondant à la cutine n'est pas très bien mise en évidence.

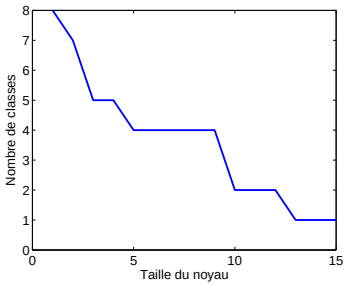


Figure 6.34: Indice de stabilité du nombre de classes : ACP + Parzen-Watershed.

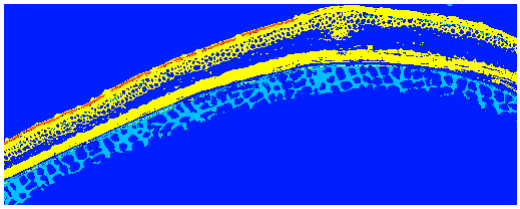


Figure 6.35: Résultats de la classification : ACP + Parzen-Watershed.

6.2 Analyse des séries temporelles d'images médicales

6.2.1 Description des données et problématique

La pyélographie intraveineuse est une procédure utilisée en médecine pour visualiser le comportement du système urinaire (les reins, les urètres et la vessie) et identifier un éventuel comportement anormal de ses organes. Afin d'établir un éventuel diagnostic, le médecin exige de savoir si le système urinaire du patient présente des obstructions. La procédure de la pyélographie intraveineuse est décrite dans la suite : une injection intraveineuse avec un marqueur à rayons X est administrée au patient. Le marqueur est éliminé du système sanguin par les reins ; il devient visible quelques instants après l'administration. Les rayons X sont captés à des intervalles de temps réguliers au fur et à mesure que le marqueur passe dans le système urinaire. Ceci offre une visualisation compréhensible de l'anatomie du patient ainsi que des informations sur le fonctionnement du système urinaire.

Une fois le marqueur administré, le système urinaire du patient est observé pendant environ 20 minutes et des images sont acquises chaque 7,5 secondes. On dispose ainsi d'une image multivariée correspondant à 160 instants temporels qui résument la circulation du marqueur dans le système urinaire. La taille de l'image multivariée est 128x128x160 pixels et l'acquisition a été réalisée avec une caméra *e.cam Siemens*. Le diagnostic est établi en analysant les fonctions relatives correspondant à chaque compartiment physiologique du système urinaire.

Pour cette application, deux séries temporelles d'images médicales ont été acquises : la première qui présente un comportement normal et la deuxième qui présente quelques anomalies de comportement de la circulation du marqueur dans le système urinaire. Le but est de visualiser les cinétiques de chaque compartiment ainsi que les éventuelles obstructions et de mettre en évidence les régions où celles-ci apparaissent.

Les obstructions sont identifiables à partir des fonctions relatives ou cinétiques correspondant à chaque compartiment ; celles-ci sont mises en évidence par des méthodes d'extraction d'attributs. Les méthodes NMF-ALS et PExSAS sont mises en oeuvre car elles respectent les contraintes de non-négativité imposées par l'application : tout d'abord les sources signalent la présence ou l'absence du marquer dans les compartiments du système urinaire, d'où la non-négativité des sources (il s'agit d'une localisation spatiale de la présence du marquer qui met en évidence les différentes compartiments du système urinaire). Les coefficients de mélange signalent la localisation temporelle du marquer dans le système urinaire. Le marquer est soit

présent à un instant donné et dans ce cas sa contribution est positive, soit absent et dans ce cas, sa contribution est nulle. Les compartiments du système urinaire sont identifiés en utilisant une méthode de classification non supervisée des pixels. Les méthodes d'extraction d'attributs nous permettent d'obtenir les signatures temporelles de chaque organe mais aussi de réduire la dimension des données. La classification a été réalisée dans l'espace de dimension réduite.

6.2.2 Analyse de la première série temporelle d'images medicales

Extraction des signatures

Les cinétiques recherchées correspondent au passage du marqueur dans les organes du système urinaire (reins, urètres et vessie) ainsi qu'au passage du marqueur du système circulatoire sanguin dans le système urinaire. Le nombre de sources recherché est égal à 4. La cinétique correspondant au passage du marqueur du système circulatoire sanguin dans le système urinaire est active pendant quelques secondes au début de l'enregistrement, figures 6.36 d) et 6.37 a) ; celle des reins est active au début du passage du marqueur dans le système urinaire, figures 6.36 b) et 6.37 d) ; au fur et à mesure que les reins se vident, le marqueur passe dans les urètres, figures 6.36 a) et 6.37 c) et ensuite dans la vessie, figures 6.36 c) et 6.37 b). Les résultats de l'extraction des cinétiques pour les deux méthodes utilisées sont présentés respectivement dans les figures 6.36 (pour la méthode NMF-ALS) et 6.37 (pour la méthode PExSAS). Ces résultats correspondent aux vecteurs de mélange dans un problème de séparation de sources. Les sources visualisent les compartiments correspondant à chaque cinétique, elles sont présentées dans les figures 6.38 (pour la méthode NMF-ALS) et 6.39 (pour la méthode PExSAS).

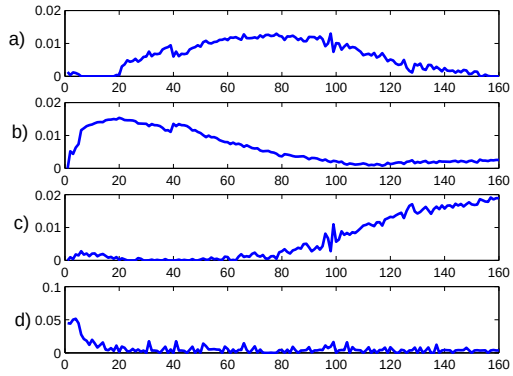


Figure 6.36: Cinétiques obtenues par la méthode NMF-ALS : a) urètres, b) reins, c) vessie, d) circulation sanguine.

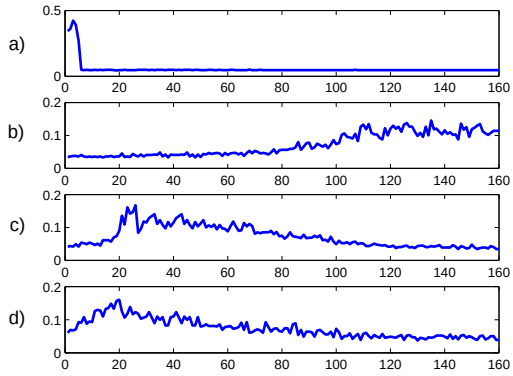
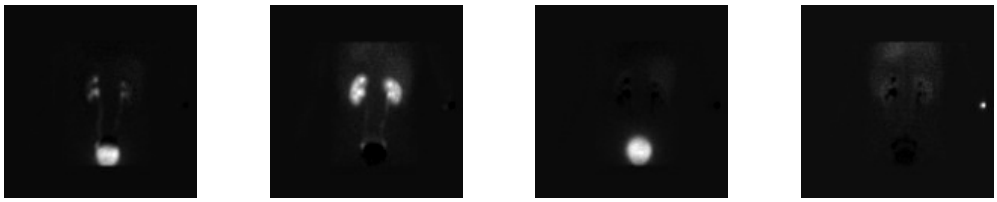


Figure 6.37: Cinétiques obtenues par la méthode PExSAS : a) circulation sanguine, b) vessie, c) urètres, d) reins.



(a) (b) (c) (d)

Figure 6.38: Compartiments correspondant à chaque cinétique pour la méthode NMF-ALS : a) urètres, b) reins, c) vessie, d) circulation sanguine.

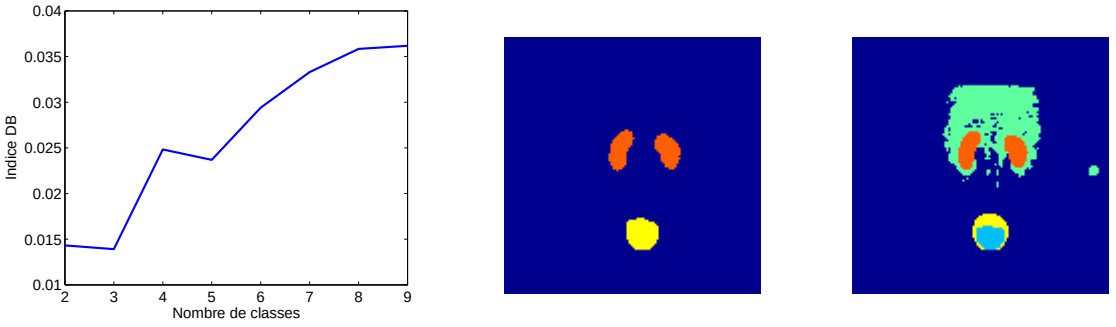
(a) (b) (c) (d)

Figure 6.39: Compartiments correspondant à chaque cinétique pour la méthode PExSAS : a) circulation sanguine, b) vessie, c) urètres, d) reins.

Classification des pixels

Pour la mise en évidence des compartiments du système urinaire, nous avons testé plusieurs méthodes de classification non supervisée sur les images obtenues après la séparation. Les meilleurs résultats ont été obtenus avec la méthode *C-moyenne*. La méthode *Mean-Shift* et *Parzen Watershed* ne donnent pas de résultats satisfaisants. L'indice de validité DB présente 2 minima locaux indiquant une partition en 3 et 5 classes. Les résultats de la classification sont présentés dans la figure 6.40, b et c.

Une partition en 3 classes met bien en évidence les reins et la vessie ; les uretères ne sont pas mises en évidence. Une partition en 5 classes présente une sursegmentation de la vessie ; par contre, la présence du marquer radioactif dans le système sanguin est mise en évidence par la classe codée en vert. Les uretères ne sont toujours pas mis en évidence.



(a) (b) (c)

Figure 6.40: Résultats de la classification par la méthode *C-moyennes* : a) indice DB, b) partition en 3 classes, c) partition en 5 classes.

6.2.3 Analyse de la deuxième série temporelle d'images medicales

Extraction des signatures

Pour la deuxième série temporelle d'images médicales, des résultats satisfaisants ont été obtenus par la méthode NMF-ALS, figure 6.36. Les 4 cinétiques sont bien mises en évidence ainsi que les obstructions qui apparaissent dans le fonctionnement des compartiments correspondant à chaque cinétique. Dans les images source, figure 6.38 c), on peut observer que les reins ne sont pas vides lorsque le marqueur est déjà dans la vessie.

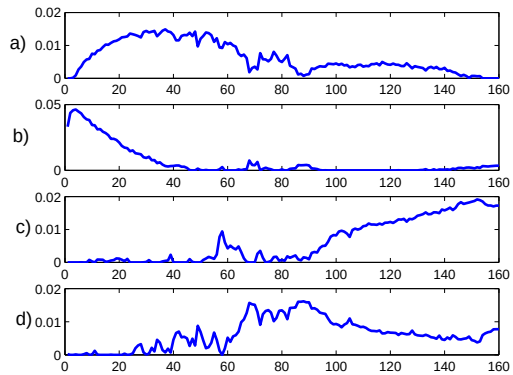


Figure 6.41: Cinétiques obtenues par la méthode NMF-ALS : a) reins, b) circulation sanguine, c) vessie, d) urètres.

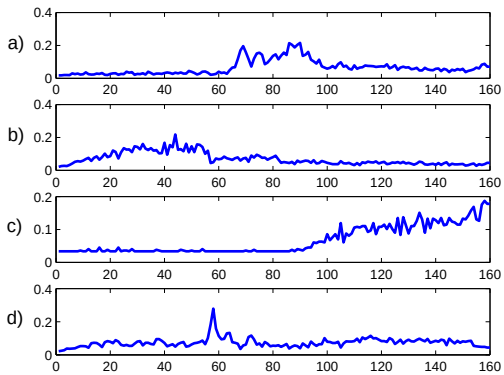


Figure 6.42: Cinétiques obtenues par la méthode PExSAS : a) urètres, b) reins, c) vessie, d) zone d'obstruction.

(a) (b) (c) (d)

Figure 6.43: Compartiments correspondant à chaque cinétique pour la méthode NMF-ALS : a) reins, b) circulation sanguine, c) vessie, d) les urètres ne sont pas visibles parce que la zone d'obstruction présente une intensité lumineuse beaucoup plus importante.

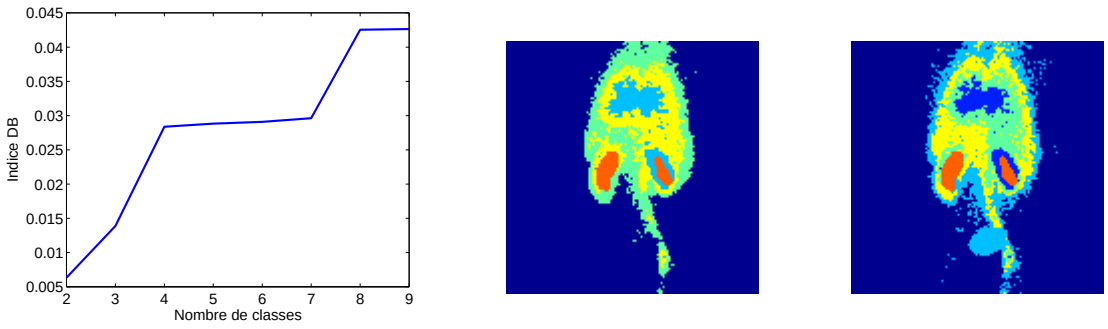
(a) (b) (c) (d)

Figure 6.44: Compartiments correspondant à chaque cinétique pour la méthode PExSAS : a) urètres, b) reins, c) vessie, d) zone d'obstruction.

Classification des pixels

Les résultats de la classification par la méthode *C-moyenne* pour la mise en évidence des compartiments du système urinaire ainsi que des régions d'obstruction sont présentés dans la figure 6.45. Les images obtenues après la séparation par la méthode NMF-ALS ont été utilisées pour la classification. La méthode *Mean-Shift* a été également utilisée pour la classification mais les résultats ne nous permettent pas la visualisation des organes. L'indice DB, figure 6.45 a) ne nous offre pas d'information sur le nombre réel de classes. Les résultats de la classification en 5 et 6 classes sont présentés dans la figure 6.45 b et c.

Les résultats de la classification ne sont pas très concluants. Les classes codées en vert et jaune indiquent la présence du marquer dans le système sanguin. La vessie, codée en bleu claire n'est pas mise en évidence par une classification en 5 classes mais elle est relevée par une partition en 6 classes. Dans ce cas, les pixels correspondant à la vessie sont regroupés avec des pixels qui se trouvent dans des régions très éloignées par rapport à la vessie. Pour éviter ce problème une solution est de prendre en compte l'information spatiale des pixels. La région



(a) (b) (c)

Figure 6.45: Résultats de la classification par la méthode *C-moyennes* : a) indice DB, b) partition en 5 classes, c) partition en 6 classes.

d'obstruction n'est pas mise en évidence par aucune des deux partitions proposées.

6.3 Conclusion

Ce chapitre présente deux applications à l'imagerie multivariée : la première a pour but l'identification des tissus de la couche externe des grains de céréale par la segmentation d'une image multivariée de microscopie ; la deuxième a comme but l'extraction des cinétiques du système urinaire et l'identification des compartiments correspondant à chaque cinétique ainsi que l'identification des éventuelles anomalies, par l'analyse des séries temporelles d'images médicales.

Pour la première application, des méthodes de classification non supervisée (à centre mobile et à densité) ont été mises en oeuvre. Pour les méthodes à centre mobile, la mesure de similarité proposée est une des métriques r ; la valeur du paramètre r a été choisie en utilisant les indices de validité D.B. et de C.S., puis elle a été confirmée par simulation. Ceci montre que, pour des problèmes de classification non supervisée de données multivariée, il est nécessaire d'accorder une attention particulière au problème du choix de la métrique utilisée comme mesure de similarité. Une comparaison entre les indices de validité obtenus pour différentes métriques peut nous indiquer la meilleure métrique parmi les métriques testées.

Des méthodes de *séparation aveugle de sources* ont été aussi employées afin de permettre la réduction de la complexité des algorithmes utilisés ainsi que la compréhension des données et l'interprétation des résultats. Les méthodes de SAS qui prennent en compte les contraintes réelles sont très utiles dans ce genre d'application car elles mettent en évidence la présence de facteurs pertinents dans l'ensemble de données ; dans ce cas, il s'agit de la présence des composés

naturellement fluorescents présents dans la couche externe des céréales. Les méthodes utilisées identifient ces composés à partir des vecteurs colonne de la matrice de mélange qui représentent leurs spectres ; la localisation de chacun est mise en évidence par les vecteurs ligne de la matrice source. La réduction de la dimension réalisée par les méthodes de SAS nous permet également d'utiliser des méthodes reposant sur le calcul des densités des attributs qui peuvent identifier des classes de forme non convexe.

Pour la deuxième application, l'extraction des cinétiques a été réalisée par des méthodes de SAS prenant en compte les contraintes des applications (la non-négativité des cinétiques). Les méthodes NMF-ALS ainsi que la méthode géométrique PExSAS ont été mises en oeuvre. La méthode PExSAS fournit de bons résultats pour la première série temporelle d'images car la condition d'avoir des instants temporels dont une seule source est active est satisfaite (les compartiments du système urinaire se remplissent et se vident l'un après l'autre). Pour la deuxième série d'images, cette condition n'est pas satisfaite car des obstructions apparaissent dans le système urinaire et donc, cette méthode ne donne pas de résultats satisfaisants. Les résultats obtenus par la méthode NMF-ALS mettent en évidence les obstructions qui apparaissent dans le système urinaire.

Afin de mettre en évidence les compartiments correspondant à chaque cinétique, des méthodes de classification non supervisée ont été mises en oeuvre sur les résultats obtenus après la séparation.

Conclusion et Perspectives

> Le doute est un état mental désagréable, mais la certitude est ridicule.
>
> Voltaire

Cette thèse apporte une contribution dans le domaine de l'analyse de données multivariées ; les travaux effectués concernent la classification des ensembles de données de grande taille décrivant des phénomènes physiques réels. Le domaine d'applicabilité est l'imagerie multivariée mais il est clair qu'il peut être aisément élargi.

Un des problèmes qui a retenu notre attention est le choix de la mesure de similarité. Il existe dans la littérature une multitude de mesures de similarité et peut être autant de doutes concernant leur utilisation dans des problèmes de classification non supervisée ; la plus utilisée par les algorithmes de classification est la distance euclidienne qui est un cas particulier des *métriques de Minkowski*, ou plus généralement, de la famille des métriques r. La pertinence de cette métrique comme mesure de similarité pour des données multidimensionnelles est mise en doute par la mise en évidence du *phénomène de concentration* ; une conséquence de ce phénomène est le fait que, pour des données de grande dimension, cette métrique perd ses capacités discriminatoires. Différents travaux de recherche affirment que des métriques moins concentrées doivent être utilisées comme alternative à la métrique euclidienne lorsque des données de grande dimension sont classifiées. Des fonctions permettant d'établir le degré de concentration de chaque métrique pour un ensemble de données quelconque ont été définies. Nous avons étudié ce phénomène et nous avons testé l'hypothèse de la supériorité des métriques moins concentrées dans des problèmes de classification de données multivariées. Nos résultats l'infirment, mais l'étude de ces résultats nous permet d'observer que, pour des classes de forme gaussienne, la distance interclasse peut être utilisée comme indice pour établir la métrique optimale dans un problème de classification. Nous décidons d'utiliser des indices de validité classiques comme

l'indice *Davies-Bouldin* ou *compacité-séparabilité* pour choisir la métrique optimale dans un problème de classification des pixels pour la segmentation des images de microscopie.

Une deuxième contribution est liée à la réduction de la dimension des données multivariées. Cette étape est essentielle lorsqu'on traite des données de grande dimension car le *phénomène de l'espace vide* empêche la découverte de structures cohérentes dans l'ensemble des données. Nous nous appuyons sur les méthodes de réduction par extraction d'attributs. Les bénéfices de ces méthodes sont multiples : la réduction du temps de calcul, la réduction de l'espace de stockage nécessaire, l'identification des facteurs pertinents ainsi que la visualisation et la compréhension des données. Pour ceci, les méthodes de *séparation aveugle de sources* ont été rappelées, étudiées et comparées dans le contexte de la classification non supervisée ; nous avons proposé également une méthode de séparation aveugle de sources basée sur une interprétation géométrique du modèle de mélange linéaire. Cette méthode est très efficace lorsque pour chaque source il existe un instant où elle-même est active et toutes les autres ne sont pas actives. La méthode est testée et comparée avec d'autres méthodes de séparation de sources avec des bons résultats. Pourtant, elle présente des limites : elle est applicable seulement pour des sources de densité de probabilité bornée et ses performances se dégradent lorsque la probabilité d'avoir des instants actifs pour seulement une source diminue.

Deux applications sont également présentées : la segmentation des images multivariées dont une première approche a été d'utiliser des métriques non euclidiennes pour la mise en évidence des tissus de la couche externe d'un grain d'orge. L'étape de réduction de la dimension n'est pas utilisée et les pixels ont été considérés comme des points dans l'espace des attributs. Dans une deuxième approche, nous avons considéré que chaque pixel est une combinaison linéaire de plusieurs spectres correspondant aux composés chimiques responsables de la fluorescence des tissus. Des méthodes de séparation de sources ont été mises en oeuvre conjointement avec des algorithmes de classification non supervisés. Les méthodes de séparation par la prise en compte des contraintes des applications se montrent les plus efficaces, et permettent aussi une bonne interprétation des résultats.

Dans une deuxième application, des séries temporelles d'images médicales décrivant le comportement du système urinaire ont été analysées. Le but était d'extraire les cinétiques et de mettre en évidence les compartiments correspondant à chaque cinétique ainsi que les éventuelles anomalies du système urinaire. La méthode NMF ainsi que la méthode PExSAS proposée dans le 5-ème chapitre ont été mises en oeuvre. La méthode *C-moyenne* a été utilisée pour la mise

en évidence des compartiments du système urinaire.

D'autres sujets liés à l'analyse des données multivariées qui peuvent rendre les résultats obtenus plus compréhensibles et améliorer les résultats obtenus ne sont pas traités dans cette thèse. Nous allons les présenter dans les perspectives de ce travail.

L'estimation de la dimension intrinsèque est une des questions ouvertes liées à l'analyse des données multivariées. Même si des méthodes existent déjà dans la littérature, il n'existe pas encore de moyen reconnu à l'unanimité par la communauté scientifique pour résoudre ce problème. Dans les applications présentées nous avons considéré la dimension intrinsèque comme étant le nombre de sources (les pixels ont été considérés comme des mélanges linéaires de plusieurs sources). Mais établir le nombre de sources dans un problème de séparation pose encore de réels problèmes. Pourtant, l'ACP est utilisée souvent pour résoudre les deux problèmes mentionnés : l'estimation de la dimension intrinsèque des données multivariées et le choix du nombre de sources dans un problème de séparation. Ceci est un lien direct entre les deux problèmes mentionnés qui nous permet d'affirmer que, dans certains cas, l'estimation de la dimension intrinsèque et le choix du nombre de sources est un et même problème. Le développement des méthodes permettant d'estimer de manière automatique (sans connaissance *a priori*) la dimension intrinsèque des données multivariées dans un problème de classification et/ou le nombre de sources dans un problème de séparation de sources est une des perspectives de ce travail. Néanmoins, la prise en compte de l'information *a priori* (là où en dispose) peut éviter l'utilisation de ces méthodes car souvent, celles-ci ne donnent pas des solutions précises.

Dans la deuxième application présentée dans le chapitre 6, nous avons vu que les méthodes de classification basées sur l'estimation de la *fdp* ne parviennent pas à identifier les compartiments du système urinaire. Un des problèmes est lié à notre avis à la densité différente des classes. Pour éviter ce problème nous avons commencé à développer des méthodes à densité avec noyau adaptatif prenant en compte le nombre d'objets situés dans une fenêtre autour d'un point de l'espace. Ceci constitue une deuxième perspective de ce travail de recherche.

Partie III

Annexes

Annexe A

A.1 Réduction de dimension des données Iris, WBC, WDBC et Wine

A.1.1 Méthode JADE

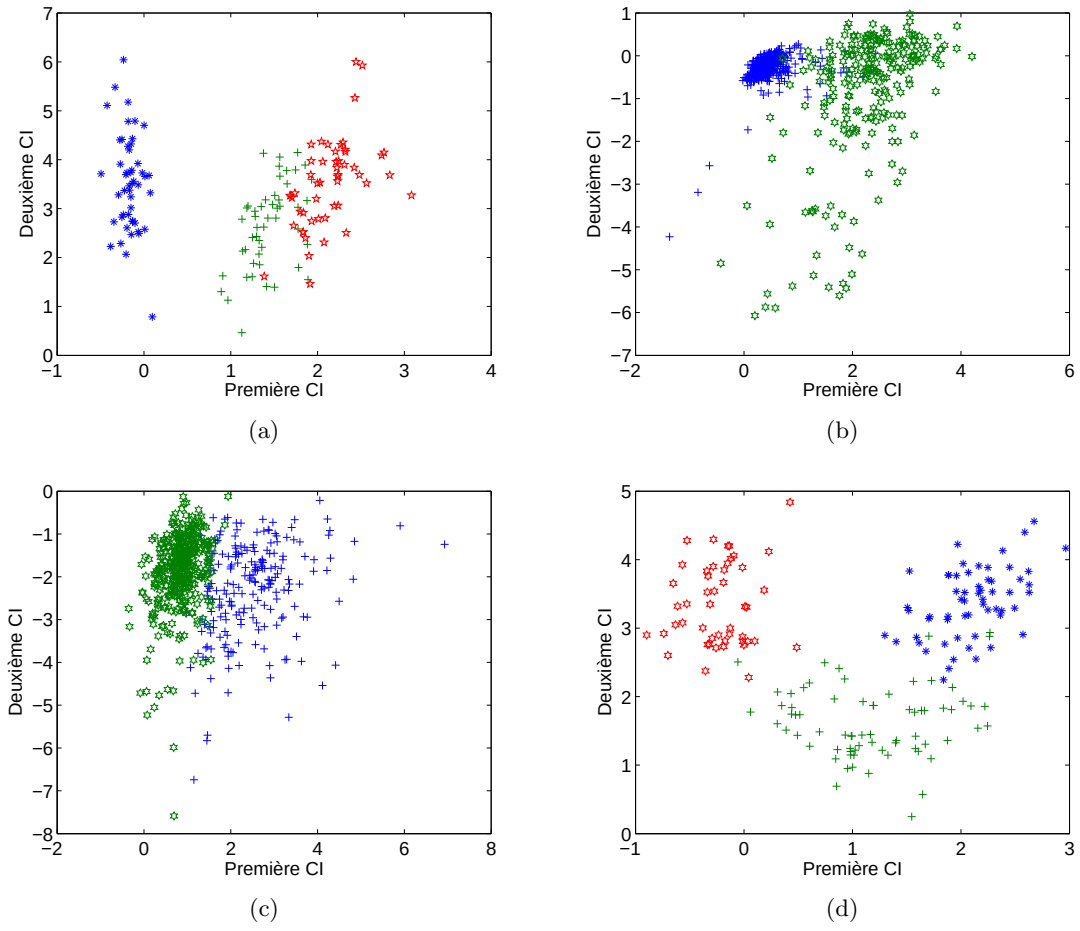


Figure A.1: Représentation des données dans l'espace des deux premières composantes indépendantes obtenues par JADE pour les bases de données Iris (a), WBC (b), WDBC (c), Wine (d).

A.1.2 Méthode NMF-ALS

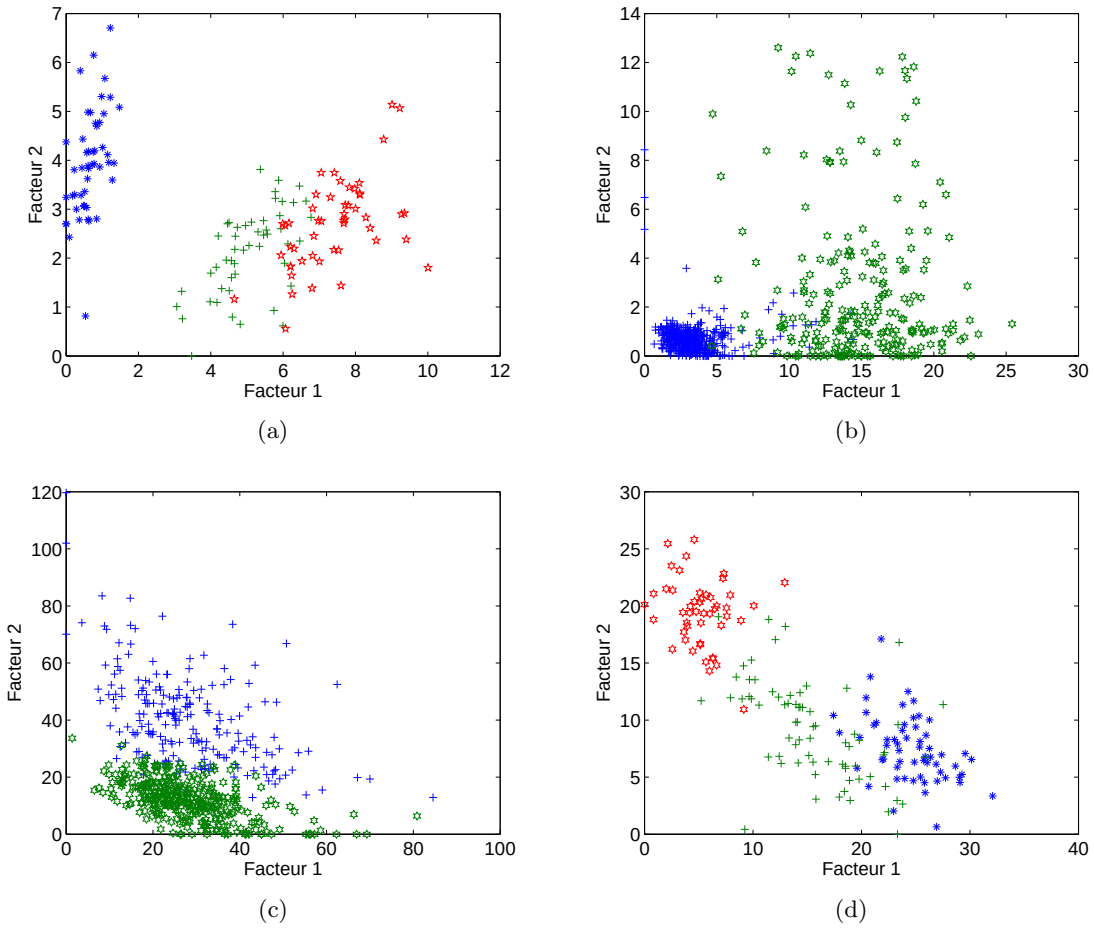


Figure A.2: Représentation des données dans l'espace des deux premiers facteurs obtenus par factorisation en matrices non-négatives, l'algorithme NMF ALS pour les bases de données Iris (a), WBC (b), WDBC (c), Wine (d).

A.2 Application à l'imagerie de microscopie

A.2.1 Indice DB

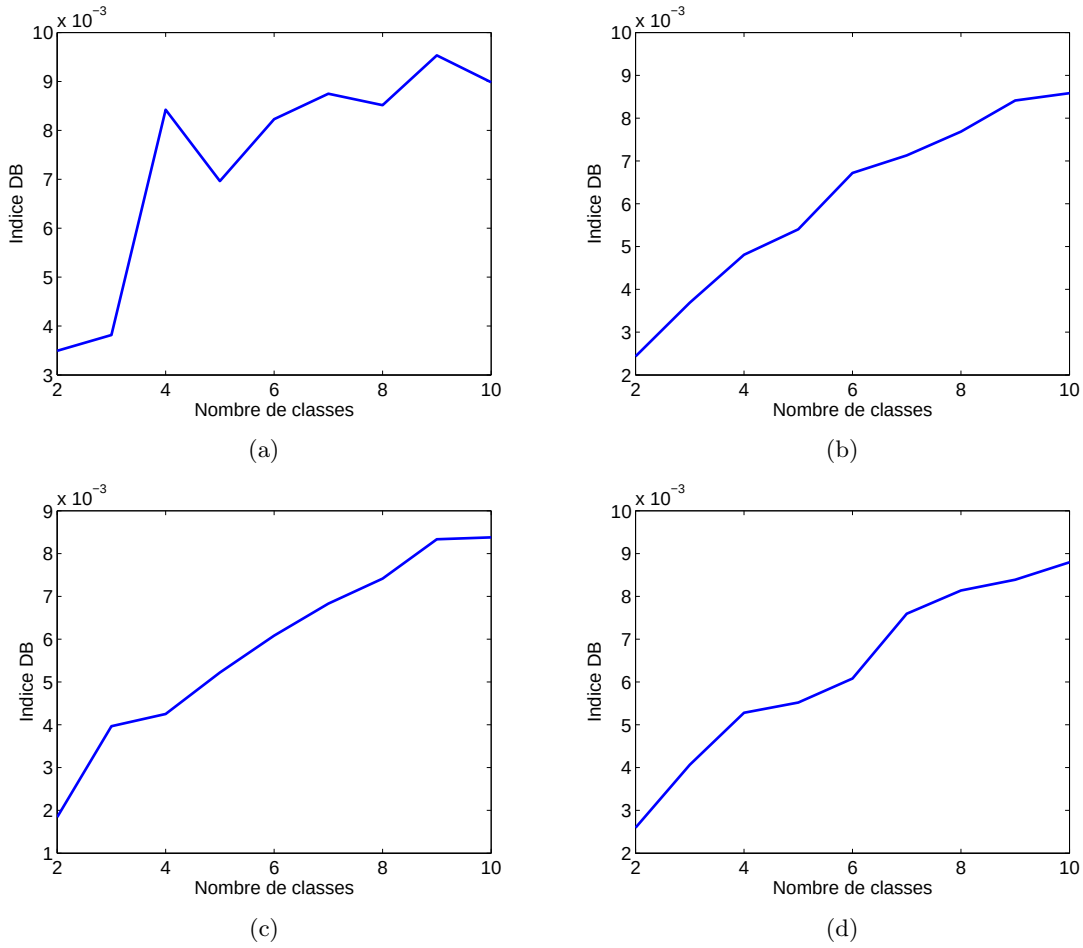
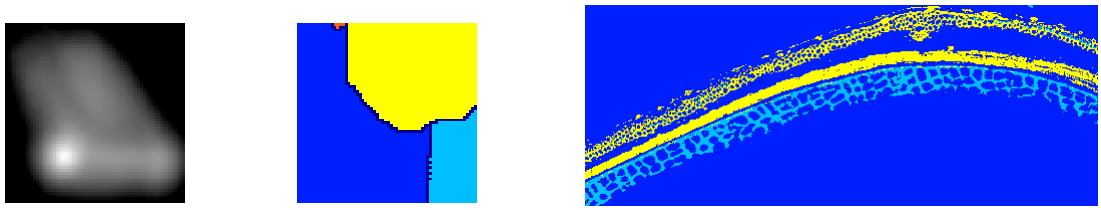


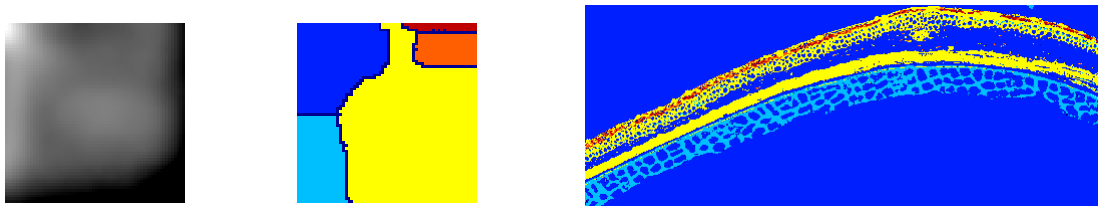
Figure A.3: Indice DB : a) JADE + C-moyennes, b) NMF + C-moyennes, c) PExSAS + C-moyennes, d) PCA + C-moyennes,

A.2.2 Illustration de la méthode Parzen-Watershed pour la segmentation de l'image de microscopie



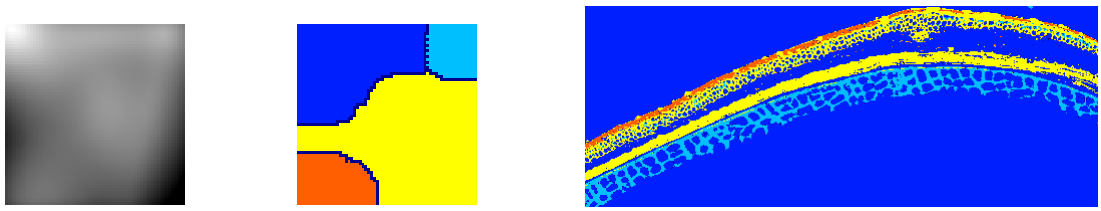
(a) (b) (c)

Figure A.4: Illustration de la méthode Parzen-Watershed dans l'espace des deux premières composantes indépendantes obtenues par la méthode JADE : a) la *fdp*, b) les zones d'influence, c) résultat de la classification.



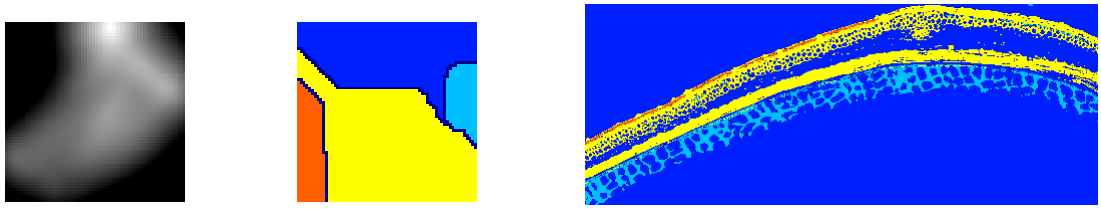
(a) (b) (c)

Figure A.5: Illustration de la méthode Parzen-Watershed dans l'espace des deux premières composantes obtenues par la méthode NMF : a) *fdp*, b) zones d'influence, c) résultat de la classification.



(a) (b) (c)

Figure A.6: Illustration de la méthode Parzen-Watershed dans l'espace des deux premières composantes obtenues par la méthode PExSAS : a) *fdp*, b) zones d'influence, c) résultat de la classification.



(a) (b) (c)

Figure A.7: Illustration de la méthode Parzen-Watershed dans l'espace des deux premières composantes principales obtenues par l'ACP : a) *fdp*, b) zones d'influence, c) résultat de la classification.

Bibliographie

[ACY96] S. Amari, A. Cichocki, and H. H. Yang. A new learning algorithm for blind signal separation. *Proc. of Neural Information Processing Systems, NIPS 96*, 8:757–763, 1996.

[AHK01] C.C. Aggarwal, A. Hinneburg, and D.A. Keim. On the surprising behavior of distance metrics in high dimensional space. *Lecture Notes in Computer Science*, 1973:420–434, 2001.

[Bel61] R. Bellmann. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1961.

[Bez81] J. C. Bezdek. *Pattern recognition with fuzzy objective function algorithms*. New York, Plenum Press., 1981.

[BGRS00] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is nearest neighbor meaningful? In *Proceedings of 26th International Conference on Very Large Data Bases, VLDB 2000*, 2000.

[BHHSV01] A. Ben-Hur, D. Horn, H.T. Siegelmann, and V. Vapnik. Support vector clustering. *Journal of Machine Learning Research*, 2:125–137, 2001.

[BNDB04] A. Bijaoui, D. Nuzillard, and T. Deb Barma. Classification and pixel demixing. In *5rd Int. ICA' 04*, pages 96–103, Granada, Spain, 2004.

[BS95] A. Bell and T. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 9:1129–1159, 1995.

[Car89] J.-F. Cardoso. Source separation using higher order moments. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 2109–2112, Glasgow, UK, 1989.

[Car92] J.-F. Cardoso. Iterative techniques for blind source separation using only fourth-order cumulants. In *Proc. EUSIPCO*, pages 739–742, Brussels, Belgium, 1992.

[Car99] J.-F. Cardoso. High order contrasts for independent component analysis. *Neural Computation*, 11(1):157–192, 1999.

[Che95] Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17:790–799, 1995.

[CJ07] P. Comon and C. Jutten. *Séparation de sources*, volume 1. Hermès - Lavoisier, 2007.

[CM81] D. Coomans and D.L. Massart. Potential methods in pattern recognition. part 2 clupot-an unsupervised pattern recognition technique. *Anal. Chim. Acta*, 133, 1981.

[CM99] D. Comaniciu and P. Meer. Mean shift analysis and applications. In *Proceedings of the International Conference on Computer Vision*, 1999.

[CM00] O. Carvalho and P. Meneses. Spectral correlation mapper (scm): An improving spectral angle mapper. In *Ninth JPL Airborne Earth Science Workshop*, pages 65–74, Jet Propulsion Laboratory, Pasadena, CA, Jan 2000.

[Com94] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36:287–314, 1994.

[CRMP07] M. Castella, S. Rhioui, E. Moreau, and J.-C. Pesquet. Quadratic higher order criteria for iterative blind separation of a mimo convolutive mixture of sources. *IEEE Transactions on Signal Processing*, 55(1):218–232, 2007.

[CS93] J.-F. Cardoso and A. Souloumiac. Blind beamforming for non gaussian signals. *IEE Proceedings-F*, 140(6):362–370, 1993.

[DAD04] K. Doherty, R. Adams, and N. Davey. Non-euclidean norms and data normalisation. In *Proc. 12th Euro. Symposiumon Artificial Neural Networks*, pages 181–186, Brugges, Begium, 2004.

[DAD07] K. A. Doherty, R. G. Adams, and N. Davey. Unsupervised learning with normalised data and non-euclidean norms. *Appl. Soft Comput.*, 7(1):203–210, Jan. 2007.

[DB79] D.L. Davies and D.W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2), 1979.

[Dem94] P. Demartines. *Analyse de données par réseaux de neurones auto-organisés*. PhD thesis, 1994.

[DH73] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley and Sons, Inc., New York, New York, 1973.

[DLM07] C. De Luigi and E. Moreau. Optimal joint diagonalization of complex symmetric third-order tensors. application to separation of non circular signals. In *ICA*, pages 25–32, 2007.

[Dun74] J. C. Dunn. Well separated clusters and optimal fuzzy partitions. *J. Cybern.*, 4:95–104, 1974.

[EKSX96] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Knowledge Discovery and Data Mining*, 1996.

[END+] A. Elhafid, D. Nuzillard, M.-F. Devaux, N. Petrochilos, and F. Belloir. Extraction des signatures de composés purs constituant la couche externe du grain d'orge à partir d'images de fluorescence. In *GRETSI'05*.

[Eve74] B.S. Everitt. *Cluster Analysis*. John Wiley & Sons. Inc. New York, 1974.

[FH75] K. Fukunaga and L.D. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Information Theory*, 21:32–40, 1975.

[Fod02] I. K. Fodor. A survey of dimension reduction techniques. Technical report, LLNL technical report, 2002.

[For65] E.W. Forgy. Cluster analysis of multivariate data: efficiency vs interpretability of classifications. *Biometrics*, 21(3):768–769, 1965.

[Fre92] C. Frelicot. *Un systeme adaptatif de diagnostic predictif par reconnaissance de formes floue*. PhD thesis, Universite Technologique de Compiègne, 1992.

[FWV05] D. Francois, V. Wertz, and M. Verleysen. Non-euclidean metrics for similarity search in noisy datasets. In *Proc. European Symp. Artificial Neural Networks (ESANN '05)*, 2005.

[FWV07] D. Francois, V. Wertz, and M. Verleysen. The concentration of fractional distances. *IEEE Transactions on Knowledge and Data Engineering*, 17(7):873–886, 2007.

[GEV+] C. Gobinet, A. Elhafid, V. Vrabie, R. Huez, and D. Nuzillard. About importance of positivity constraint for source separation in fluorescence spectroscopy. In *EUSIPCO'05*.

[GP86] J. C. Gower and Legendre P. Metric and euclidean properties of dissimilarity coefficients. *Journal of Classification*, 3:5–48, 1986.

[HAK99] A. Hinneburg, C.C. Aggarwal, and D.A. Keim. What is the nearest neighbor in high dimensional spaces? *Lecture Notes in Computer Science*, 1540:217–235, 1999.

[HB65] D. Hall and G. Ball. Isodata : a novel method of data analysis and pattern classification. Technical report, Stanford Research Institute, 1965.

[HBV96] M. Herbin, N. Bonnet, and P. Vautrot. A clustering method based on the estimation of the probability density function and on the skeleton by influence zones. application to image processing. *Pattern Recognition Letters*, 17, 1996.

[HBV01] M. Herbin, N. Bonnet, and P. Vautrot. Estimation of the number of clusters and influence zones. *Pattern Recognition Letters*, 22, 2001.

[HK98] A. Hinneburg and D.A. Keim. An efficient approach to clustering in large multimedia databases with noise. In *Proceedings of the Knowledge Discovery and Data Mining*, 1998.

[HK01] A. Hyvärinen and E. Karhunen, J. et Oja. *Independent Component Analysis*. John Wiley, New York, 2001.

[HO97] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997.

[HO00] A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4–5):411–430, 2000.

[Hoy02] P. O. Hoyer. Non-negative sparse coding. In *Neural Networks for Signal Processing XII (Proc. IEEE Workshop on Neural Networks for Signal Processing)*, pages 557–565, Martigny, Switzerland, 2002.

[Hoy04] P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004.

[HR05] P. Howarth and S.M. Rüger. Fractional distance measures for content-based image retrieval. In *Proc. 27th European Conf. Information Retrieval Research (ECIR '05)*, pages 447–456, Mar. 2005.

[HS76] L. Hubert and J. Schula. Quadratic assignment as a general data analysis strategy. *British Journal of Mathematical Psychology*, 29:190–241, 1976.

[HW79] J. A. Hartigan and M. A. Wong. A k-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.

[Hyv99a] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.

[Hyv99b] A. Hyvärinen. Survey on independent component analysis. *Neural Computing Surveys*, 2:94–128, 1999.

[JD88] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, Inc., Englewood Cliffs, New Jersey, 1988.

[Jen96] J. R. Jensen. *Introductory Digital Image Processing: A Remote Sensing Perspective*. Prentice Hall, Upper Saddle River, NJ, 1996.

[JMF99] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Comput. Surv.*, 31(3):264–323, September 1999.

[Jol02] I.T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics, 2002.

[KC03] J. Karvanen and A. Cichocki. Measuring sparseness of noisy signals. In *Proceedings of the Fourth International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Nara, Japan, 2003.

[Kin67] B. King. Step-wise clustering procedures. *Journal of the American Statistical Association*, 69:86–101, 1967.

[KLB+93] F. A. Kruse, A. B. Lefkoff, J. W. Boardman, K. B. Heiedbrecht, A.T. Shapiro, P. J. Barloon, and A. F. H. Goetz. The spectral image processing system (sips) - interactive visualization and analysis of imaging spectrometer data. *Remote Sens. Environ.*, 44:145–163, May-June 1993.

[Knu98] K. Knuth. Bayesian source separation and localization. In *A. Mohammd-Djafari, editeur : proceedings of International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MaxEnt'98)*, American Institute of Physics (AIP), 1998.

[Kul59] S. Kullback. *Information theory and statistics*. John Wiley, 1959.

[LJB06] Csaba Legány, Sándor Juhász, and Attila Babos. Cluster validity measurement techniques. In *AIKED'06: Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases*, 2006.

[LS99] D.D. Lee and H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[LS00] D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. *Proc. Neural Information Processing Systems*, 13:556–562, 2000.

[Mac67] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. Le Cam and J. Neyman, editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

[Mah36] P.C. Mahalanobis. On the generalized distance in statistics. In *Proceedings of the National Institute of Science of India 12*, pages 49–55, 1936.

[MC85] G. W. Milligan and M. C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, June 1985.

[MD99] A. Mohammad-Djafari. A bayesian approach to source separation. In *Proceedings of International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MaxEnt'99)*, pages 221–244, American Institute of Physics (AIP), 1999.

[MJ96] J. Mao and A.K. Jain. A self-organizing network for hyperellipsoidal clustering (hec). *IEEE Transactions on Neural Networks*, 7(1):16–29, 1996.

[Mou05] S. Moussaoui. *Séparation de sources non-négatives. Application au traitement des signaux de spectroscopie.* PhD thesis, Université Henri Poincaré, Nancy 1, 2005.

[MPO01] A. Mansour, C.G. Puntonet, and N. Ohnishi. A simple ica algorithm based on geometrical approach. *Sixth International Symposium on Signal Processing and its Applications*, 1:9–12, 2001.

[Mur84] F. Murtagh. A survey of recent advances in hierarchical clustering algorithms which use cluster centers. *Comput. J.*, 26:354–359, 1984.

[NL06] D. Nuzillard and C. Lazar. Comparison of two unsupervised methods of classification for segmenting multi-spectral images. In *International Conference on Acoustics, Speech, and Signal Processing 2006*, Toulouse, France, may 2006.

[NL07] D. Nuzillard and C. Lazar. Partitional clustering techniques for multi-spectral image segmentation. *Journal of Computers (JCP)*, 2(10):1–8, dec 2007.

[NLB+07] D. Nuzillard, C. Lazar, P. Billaudel, S. Curila, and F. Belloir. Pré-traitement par séparation aveugle de sources pour la segmentation d'images multi-spectrales. In *COmpression et REprésentation des Signaux Audiovisuels 2007*, pages 209–213, Montpellier, France, nov 2007.

[Paa97] P. Paatero. Least squares formulation of robust non-negative factor analysis. *Chemomerics and Intelligent Laboratory Systems*, 37:23–35, 1997.

[Paa99] P. Paatero. The multilinear engine-a table-driven least squares program for solving multilinear problems, including the n-way parallel factor analysis mode. *Journal of Computational and Graphical Statistics*, 8(4):1–35, 1999.

[Par62] E. Parzen. On the estimation of a probabitity density function and mode. *Annals Math. Stats.*, 33, 1962.

[PBJD79] K. Pettis, T. Bailey, A.K. Jain, and R. Dubes. An intrinsec dimension estimator from near-neighbor information. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI*, 1:25–37, 1979.

[Pet06] Matthieu Petremand. *Détection des galaxies à faible brillance de surface, segmentation hyperspectrale dans le cadre de l'observatoire virtuel*. PhD thesis, Université Louis Pasteur - Strasbourg I, 2006.

[Pha96] D.-T. Pham. Blind separation of instantaneous mixture sources via an independent component analysis. *IEEE Transactions on Signal Processing*, 44(11):2768–2779, 1996.

[PM01] J.-C. Pesquet and E. Moreau. Cumulant-based independence measures for linear mixtures. *IEEE Transaction on Information Theory*, 47(5):1947–1956, 2001.

[PMJ95] C. G. Puntonet, A. Mansour, and C. Jutten. Geometrical algorithm for blind separation of sources. In *Actes du XVème colloque GRETSI*, pages 273–276, Juan-Les-Pins, France, 1995.

[PP98] C.G. Puntonet and A. Prieto. Neural net approach for blind separation of sources based on geometric properties. *Neurocomputing*, 18:141–164, 1998.

[PPJ+95] C.G. Puntonet, A. Prieto, C. Jutten, M. Rodriguez-Alvarez, and J. Ortega. Separation of sources: a geometry-based procedure for reconstruction of n-valued signals. *Signal Processing*, 46(3):267–284, 1995.

[PT94] P. Paatero and U. Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5:111–126, 1994.

[Pun95] A Puntonet, C.and Prieto. An adaptive geometrical procedure for blind separation of sources. *Neural Processing Letters*, 2, 1995.

[Rob98] S.J. Roberts. Independent component analysis : Source assessment and separation, a bayesian approach. *IEE Proceedings on Vision, Image and Signal Processing*, 145(3):149–154, 1998.

[Ser82] J. Serra. *Image Analysis and Mathematical Morphologie*. Academic Press, New York, 1982.

[Smi02] Lindsay I Smith. A tutorial on principal components analysis. February 2002.

[SPST+01] B. Schölkopf, J. C. Plat, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high dimensional distribution. *Neural Computing*, 13:1443–1471, 2001.

[SS73] P. H. Sneath and R. R. Sokal. *Numerical Taxonomy.* Freeman, London, UK, London, UK, 1973.

[ST83] D. W. Scott and J. R. Thompson. Probability density estimation in higher dimensions. In *J.E. Gentle (ed.), Computer Science and Statistics: Proceedings of the Fifteenth Symposium on the Interface*, pages 173–179, Amsterdam, New York, Oxford, North Holland-Elsevier Science Publishers, 1983.

[TWB05] T. N. Tran, R. Wehrens, and L. M. Buydens. Clustering multispectral images: a tutorial. *Chemometrics and Intelligent Laboratory Systems*, 77(1–2):3–17, May 2005.

[UCI] UCI. http://archive.ics.uci.edu/ml/datasets.html.

[Ver03] M. Verleysen. Learning high-dimensional data. In *Limitations and Future Trends in Neural Computation 186*, pages 141–162, 2003.

[War63] B. W. Jr. Ward. Hierarchical grooping to optimize an objective function. *Journal of the American Statistical Association*, 58:236–244, 1963.

[YGB92] R. H. Yuhas, A. F. H. Goetz, and J. W. Boardman. Discrimination among semi-arid landscape endmembers using spectral angle mapper (sam) algorithm. In *Summaries of the 4th Annual JPL Airborne Geoscience Workshop*, pages 147–150, Jet Propulsion Laboratory, Pasadena, CA, Jan 1992.