UNIVERSITÉ DE STRASBOURG

N° d'ordre : 1407

École Doctorale Mathématiques, Sciences de l'Information et de l'Ingénieur

UdS – INSA – ENGEES

THÈSE

présentée pour obtenir le grade de

Docteur de l'Université de Strasbourg
Discipline : Sciences
Spécialité : Signal, image et vision

par

Akram Belghith

Indexation de spectres HSQC et d'images IRMf appliquée à la détection de bio-marqueurs.

Soutenue publiquement le 30 mars 2012

Membres du jury

Directeur de thèse :	M. Christophe Collet, Professeur à l'Université de Strasbourg
Co-Directeur de thèse :	M. Jean-Paul Armspach, Ingénieur de recherche, HDR, à l'Université de Strasbourg,
Rapporteur externe :	Mme. Danielle Nuzillard, Professeur à l'Université de Reims
Rapporteur externe :	M. Stéphane Canu, Professeur à l'Université de Rouen
Invité :	M. Izzi-Jacques Namer, Professeur, Praticien Hospitalier, à l'Université de Strasbourg
Invité :	M. Karim Elbayed, Maitre de conférences, HDR à l'Université de Strasbourg
Invité :	M. Jack Foucher, Maitre de conférences, Praticien Hospitalier, à l'Université de Strasbourg

Laboratoire des Sciences de L'Image, de l'Informatique et de la Télédétection **UMR CNRS 7005**

Laboratoire d'Imagerie et de Neurosciences Cognitives **UMR CNRS 7237**

Acknowledgments

First and foremost, many thanks to my PhD supervisors: Christophe Collet and Jean-Paul Armspach for their valuable insights, patience and guidance. I am very pleased to have had the opportunity to work with them.
I would also like to thank Stéphane Canu and Danielle Nuzillard for having accepted to review this thesis and giving me valuable comments. Thank you also to Karim El Bayed and Jack Foucher and Izzi-Jacques Namer for having made me the honour to take part in the jury.
I would like to thank the Alsace Region and ARSEP for support of this thesis.
I would like to express my gratitude to Fatma.
I also thank all MIV and LINC members, past and present, for the enjoyable work atmosphere they created.
Last but not least, I would like to thank my family for their unconditional support.

Résumé

Introduction

Des équipements médicaux de plus en plus sophistiqués apportent aujourd'hui des informations indispensables à la prise de décision du médecin, en particulier dans les domaines de la cancérologie et des pathologies neurologiques. Les progrès permis par les nouvelles technologies d'acquisition d'information médicale, que ce soit par l'amélioration du rapport signal à bruit ou le couplage inédit entre anatomie et physiologie, ont progressivement donné à l'imagerie médicale un statut indispensable dans l'élaboration du diagnostic, du pronostic et de la prise en charge thérapeutique. Par ailleurs, l'analyse de combinaisons de signaux biologiques apparait aujourd'hui prometteuse sans toutefois faire systématiquement partie de la routine clinique. Les techniques d'acquisition des signaux médicaux sont en outre en constante évolution (échographie endoscopique, Tomographie par émission de Positons TEP, spectroscopie par Résonance Magnétique Nucléaire, Imagerie par Résonance Magnétique IRM, imagerie IRM fonctionnelle IRMf) et fournissent une quantité croissante de données hétérogènes qui doivent être analysées par le médecin. En effet, d'une observation statique à une multi-observation dynamique, d'une information sur la structure des organes à l'information sur leurs fonctions, les techniques d'acquisition de signaux médicaux portent potentiellement la signature de la maladie (bio-marqueurs) bien au-delà de l'examen clinique ponctuel. Dans ce contexte, des méthodes automatiques de traitement des signaux médicaux sont régulièrement proposées pour aider l'expert dans l'analyse qualitative et quantitative en facilitant leur interprétation. Ces méthodes doivent tenir compte de la physique de l'acquisition, de l'a priori que nous avons sur ces signaux et de la quantité de données à analyser pour une interprétation plus précise et plus fiable. Parmi les nouvelles techniques d'acquisition de signaux, l'analyse des tissus biologique par spectroscopie RMN ou la recherche des activités fonctionnelles cérébrales et leurs connectivités par IRMf sont explorées pour la recherche de nouveaux bio-marqueurs (objet), que ce soit pour l'aide au diagnostic de pathologies ou pour le suivi d'effets thérapeutiques. Pour ce faire, il est également nécessaire d'améliorer les outils d'analyse associés à ces nouvelles techniques. Dans cette optique, nous proposons un nouveau schéma d'indexation et de recherche par le contenu d'objets pour la détection des bio-marqueurs.

Le scénario classique d'exploitation d'un système de recherche d'information est le suivant : un utilisateur soumet une requête et le système identifie les informations pertinentes à la requête soumise, puis les retourne à l'utilisateur. Ainsi, le but d'un système de recherche d'information est de retrouver les documents pertinents par rapport à une requête donnée. Cependant, l'évaluation de la pertinence d'un document n'est toujours pas aisée puisque la notion de la pertinence est très dépendante des préférences de l'utilisateur. La recherche traditionnelle des documents (par exemple, les images médicales, les sons respiratoires, etc) par mots-clés est l'approche la plus ancienne et la plus utilisée. Cependant, elle reste limitée par le faible pouvoir expressif des mots, par les contraintes linguistiques

(le passage d'une langue à une autre, l'ambigüité sémantique) et par le caractère objectif des annotations (deux médecins peuvent annoter différemment une image médicale). En outre, elle nécessite l'intervention humaine et est donc contraignante pour les bases de données de tailles importantes si les mots clés sont générés manuellement. De plus, notons que l'annotation ne pourra jamais décrire le contenu d'un document de façon exhaustive.

Afin de contourner ces inconvénients, l'approche d'indexation et de recherche par le contenu a été proposée (appelée désormais approche d'indexation) [Eakins96]. Elle consiste à rechercher des documents en n'utilisant que le document lui-même, c'est-à-dire son contenu sans aucune autre information. Par exemple, dans le cas des images, l'idée est de caractériser le contenu visuel des images par des descripteurs visuels et d'effectuer des recherches par similarité visuelle à partir de ces descripteurs. Par conséquent, l'approche d'indexation basée sur le contenu nous permet non seulement d'indexer automatiquement les documents et d'interroger une base de données directement à partir de leur contenu informatif, sans intervention humaine, mais aussi d'analyser objectivement son contenu. Par exemple, si on considère une tumeur cérébrale comme une requête, nous pouvons facilement identifier avec une fonction de mesure de similarité objective les tumeurs similaires appartenant à la base de données sollicitée.

Généralement, le système d'indexation nécessite:

1. Une étape d'alignement de documents,

2. Une étape de codage de document et de mesure de similarité. En effet, le codage du document consiste à calculer pour chaque document un ensemble d'attributs descriptifs compacts qui définit sa signature. Une mesure de similarité utilisant ces descripteurs permet de comparer deux documents et d'identifier ainsi les documents similaires.

Afin d'accélérer la sollicitation de grande base de données, le schéma d'indexation peut être divisé en deux phases:

1. Une phase hors ligne dans laquelle on réalise l'alignement et le codage du contenu de la base de données. Durant cette phase, l'utilisateur n'est pas encore connecté au système. Cette phase peut alors prendre le temps nécessaire à l'extraction des descripteurs. Le codage hors ligne consiste à extraire les signatures associées aux contenus de la base de données. Ces dernières sont ensuite enregistrées dans une base de données organisée comme un dictionnaire inverse (nom du document et signature) permettant ainsi de retrouver rapidement le document associé à une signature donnée.

2. Une phase en ligne dans laquelle l'utilisateur interroge la base de données à l'aide d'un document exemple. Durant cette seconde phase, le temps de réponse du système est crucial et il faut l'optimiser. Notons que les étapes de l'alignement et du codage ne concernent que le document requête. Une mesure de similarité entre la

signature de la requête et celles établies dans le dictionnaire inverse est alors calculée. Enfin, les documents appartenant à la base de données sont classés par ordre de similarité.

Bien que ce système classique d'indexation a été appliqué avec succès sur des bases de données du Web [1], il n'est malheureusement pas adapté à la tâche d'identification des bio-marqueurs. En effet, cette dernière nécessite la classification des profils de signaux médicaux (groupes) pour la détection de changements. Par exemple, si on considère deux classes de profils: la classe saine et la classe pathologique, la tâche d'identification de bio-marqueurs revient à classifier un groupe de signaux médicaux dans la classe des signaux sains ou pathologique (par exemple, le cancer ou les maladies psychologiques) et de détecter alors les différences (variations) entre eux. En fait, la classe des signaux sains peut être considérée comme la classe du "non changement" et la classe pathologique comme la classe du "changement". Par conséquent, l'ajout d'une étape de classification/détection de changement au schéma classique d'indexation nous permettrait de détecter les bio-marqueurs à partir des données médicales considérées. Nous nous focalisons dans ce travail de thèse sur :

1. Les spectres à deux dimensions HSQC (Heteronuclear Single Quantum Coherence) obtenus en Résonance Magnétique Nucléaire (RMN) et plus particulièrement en spectroscopie RMN hétéro-nucléaire HR-MAS (High-Resolution Magic Angle Spinning : RMN haute résolution par rotation de l'échantillon à l'angle magique) qui permet l'analyse directe des tissus biologiques (biopsie)
2. Les images IRMf pour les régions fonctionnelles cérébrales.

De ce fait, et contrairement au schéma d'indexation classique, le nouveau schéma d'indexation contient deux étapes supplémentaires: une étape de détection d'objets (détection de pics de spectres HSQC et des zones actives d'images IRMf) et une étape de classification d'objets (détection de changements). Chaque information médicale traitée (spectres 2D RMN ou images IRMf) est alors caractérisée par un ensemble d'objets (bio-marqueurs) que nous cherchons à extraire, aligner et coder. Le regroupement de ces objets par la mesure de leur similarité permet alors leur classification. C'est ce schéma globale d'indexation et de recherche par le contenu d'objets que nous avons adopté. Dans notre cas, ces objets sont :

- Les raies d'émission pour les spectres RMN HR-MAS 2D (i.e., un ensemble de pics est la réponse correspondant à la présence de métabolites, chaque métabolite générant différents pics d'émission traduisant la présence de petite molécule à travers des interactions Proton-Carbone 13 dans le cadre des spectres HSQC).
- Les zones actives pour les images IRMf (i.e., une zone active est la réponse d'une activité cérébrale à un stimulus).

Le schéma d'indexation proposé est alors divisé en deux phases:

1. Une phase hors ligne dans laquelle on réalise sur chaque signal traité (spectre HSQC et image IRMf) : la de détection d'objets, l'alignement d'objets, lr codage d'objets, la mesure de similarité et enfin la classification d'objets. Finalement, des profils de groupes ou de populations données (par exemple groupe de signaux normaux ou pathologiques) sont établis.

2. Une phase en ligne dans laquelle l'utilisateur interroge la base de données en utilisant une requête (nouvel individu/groupe de spectres). Les mêmes étapes que dans la phase hors ligne sont appliquées sur la requête (spectre HSQC ou image IRMf). Enfin, cette requête est assignée à un profil préalablement défini à l'étape hors ligne. Notons que contrairement au schéma d'indexation classique, l'étape de mesure de similarité d'objets vise ici à regrouper les objets similaires appartenant à un groupe de signaux médicaux donnés permettant ainsi l'attribution de ce groupe au profil approprié. En d'autres termes, la tâche d'attribution d'un nouveau groupe/individu est abordée ici au niveau de l'étape de classification et non pas au niveau de l'étape de mesure de similarité comme c'est le cas pour le schéma d'indexation classique. Figure 2.15 montre le schéma du traitement.

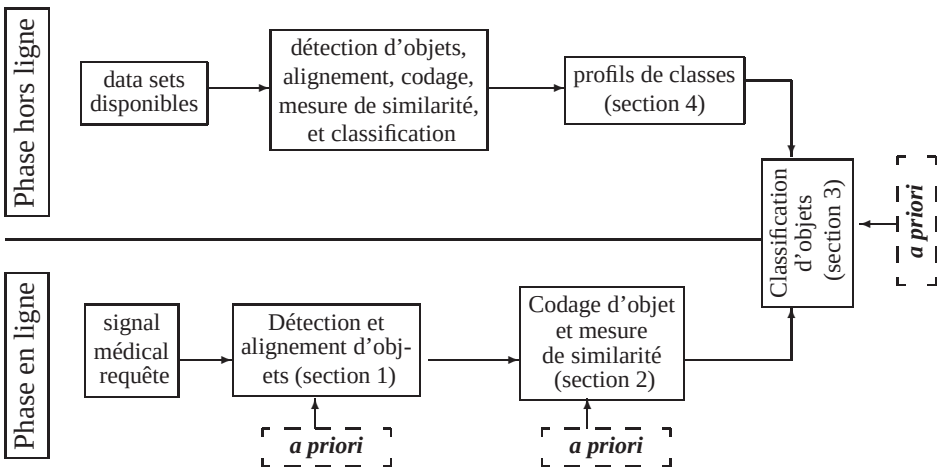


Figure 1: Schéma synoptique de la méthode d'indexation proposée.

Dans la suite, nous développons les étapes de détection-alignement d'objets, de codage

et de mesure de la similarité ainsi que l'étape de classification d'objets.

1. Détection et alignement d'objets

Le problème de la détection d'objets consiste à découper le signal en un ensemble de groupes significatifs (objets) en se basant sur les informations spatiales et/ou les intensités des pixels. La tâche d'alignement d'objet est le processus de superposition de deux ou plusieurs objets pris à des moments différents, et/ou différents points de vue, et/ou par des modalités différentes. Plus précisément, l'alignement des objets consiste à aligner géométriquement un objet par rapport à un motif de référence. Notons que la tâche de détection d'objets et celle de l'alignement sont deux étapes cruciales dans le système d'indexation car toutes les autres étapes en dépendent. Par conséquent, afin d'aboutir à un résultat de détection et d'alignement d'objets optimal, toutes les connaissances *a priori* que nous avons sur les données doivent être correctement intégrées dans les méthodes proposées de détection et d'alignement d'objets. C'est la démarche que nous suivons dans cette thèse.

Pour l'étape d'extraction et d'alignement des objets, nous proposons une nouvelle méthode basée sur l'utilisation de la théorie de l'évidence qui combine la détection et l'alignement des raies d'émission. En effet, la théorie de l'évidence permet la manipulation de l'incertitude des modèles et l'imprécision qui caractérisent les spectres HSQC. Par conséquent, nous proposons le couplage entre la théorie des ensembles flous et la théorie bayésienne pour modéliser et quantifier le degré d'imprécision des spectres qui sera ainsi exploité pour définir les fonctions de masse (i.e., une fonction qui modélise le degré de croyance sur une hypothèse donnée). En ce qui concerne les images IRMf, nous procédons, dans une première étape à l'extraction des zones actives en utilisant un algorithme de segmentation par chaînes de Markov. Ensuite, nous proposons un nouvel algorithme d'alignement des zones actives basé sur l'utilisation de la méthode d'Analyse en Composante Principale (ACP) non-linéaire pour l'estimation des symétries de réflexion. Ces symétries de réflexion sont ensuite utilisées pour l'alignement des zones actives.

1.1 Détection et alignement de pics

Nous proposons dans cette thèse une nouvelle méthode de détection et d'alignement des pics de spectres RMN HR-MAS 2D en utilisant la théorie de l'évidence. Cette théorie permet d'affecter des degrés de confiance, aussi connus sous le nom de fonctions de masses, non seulement à des hypothèses simples, mais aussi à des réunions d'hypothèses(si la connaissance disponible ne porte que sur un ensemble d'hypothèses sans plus de précision). Par exemple, un pixel de spectre peut appartenir à la classe $\{H_1, H_2\}$ où H_1 représente la classe des raies d'émission et H_2 celle du bruit. Il n'existe pas une méthode générique reconnue pour construire ces fonctions de masses et leur définition est très dépendante de l'application étudiée. Pour ce faire, nous nous intéressons à la modélisation du conflit pour la quantification de l'imprécision dans les spectres en définissant trois hypothèses triviales:

hyp_1 (l'imprécision (contradiction) portant sur les amplitudes), hyp_2 (l'imprécision portant sur les formes des pics) et hyp_3 (l'imprécision portant sur les positions des pics). La contradiction sera maximale lorsque un pixel correspond à une raie d'émission dans un spectre et correspond en même temps à un bruit dans l'autre. Les différents paramètres des spectres (localisation des pics, charachtéristiques de chaque pic: amplitude et forme) sont estimés par une procedure Monte Carlo Markov Chain MCMC [Griffin04]. Figure 6.3 représente la chaîne de traitement de la méthode de détection et d'alignement des pics.

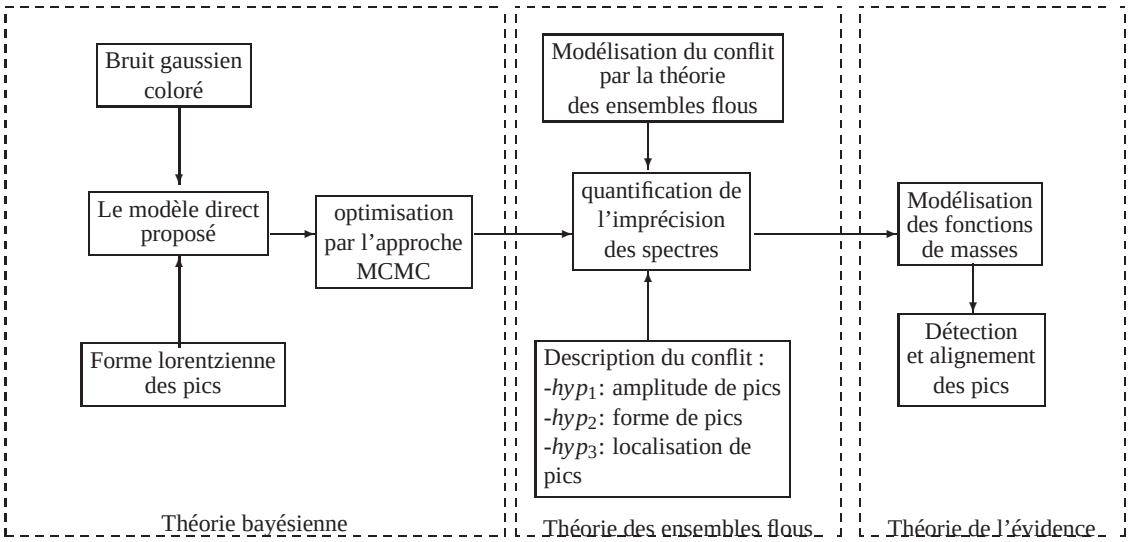


Figure 2: La chaîne de détection et d'alignement des pics.

1.2 Détection et alignement de zones actives

Nous rappelons que une image IRMf peut simplement être représentée par un ensemble de zones actives que nous cherchons à détecter et à aligner (comme pour les pics du spectre HSQC) en 3D. Chaque zone active peut être caractérisée par sa positon, sa forme et les intensités de ses voxels. Contrairement au schéma de détection et d'alignement des pics, les problèmes de détection et d'alignement des zone actives sont traités séparément. Plus explicitement, étant donnés deux objets détectés, notre objectif est de les aligner en fonction de leur pose canoniques. Pour ce faire, la détection des zones actives est d'abord effectuée en utilisant une méthode classique de segmentation par les chaînes de Markov cachées

[Bricq08] permettant d'intégrer l'information spatiale dans la procédure de segmentation.

Afin d'aboutir à un résultat satisfaisant d'alignement, nous nous appuyons sur la perception humaine dans le schéma d'alignement qui consiste à aligner un objet en fonction de ses axes de symétrie. Cette approche nous permet de trouver la pose la plus naturelle de l'objet et ensuite aligner les objets visuellement similaires de la même manière. La plupart des méthodes basées sur la perception humaine ont opté soit pour le choix de l'ACP ou l'ACP continue [Vranić01a, Vranic01b] pour estimer les plans de réflexion. Ces plans sont ensuite utilisés pour estimer le système de coordonnées cartésiennes approprié associé à l'objet. Bien que ces méthodes aient été appliquées avec succès pour l'alignement des objets 3D du Web [Vranic01b], elles sont malheureusement pas adoptées pour les objets 3D IRMf. En effet, en raison de la forme du cortex, la symétrie de réflexion sur les zones actives est plus sphérique que planaire. Nous proposons alors d'utiliser la ACP non-linéaire qui est plus adaptée à la forme de nos objets pour modéliser la symétrie de réflexion des zones actives [Bishop95]. Pour cela, nous avons développé une nouvelle méthode d'estimation des symétries de réflexion sphérique en se basant sur les réseaux de neurones qui ont montré leurs intérêts dans la modélisation de l'aspect non-linéaire des données[Hsieh98, Stamkopoulos98, Scholz05]. Figure 6.5 représente la chaîne de traitement de la méthode de d'alignement des zones actives.

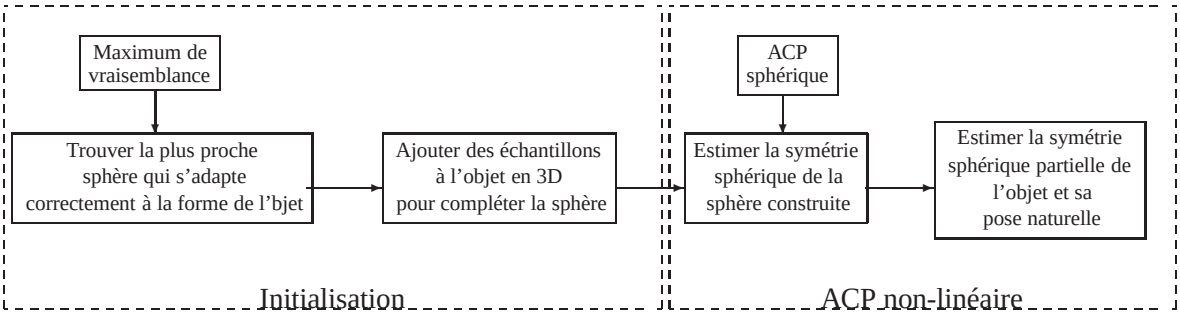


Figure 3: La chaîne d'alignement des zones actives.

2 Codage d'objets et mesure de similarité

Nous rappelons que chaque spectre est composé de plusieurs pics qui sont dispersés dans le spectre. Ces pics sont les réponses de la présence de métabolites. Par conséquent, les pics appartenant à un métabolite donné ont des propriétés communes. Afin d'aboutir à un meilleur résultat de codage et de mesure de similarité des pics, ces propriétés devraient être modélisées et injectées dans le schéma proposé. Ainsi, il est préférable de manipuler les métabolites plutôt que leurs pics séparément. Par ailleurs, dans le cas des spectres HSQC, l'étape de codage d'objets s'avère inutile puisque les pics peuvent être uniquement décrits par trois paramètres (localisation, amplitude et forme) et ils sont donc déjà représentés

d'une façon parcimonieuse. De ce fait, l'étape de codage et de mesure de similarité des pics revient à identifier les métabolites. Pour cela, nous proposons une nouvelle méthode basée sur la combinaison de la théorie bayésienne et de la théorie d'ensembles flous permettant de gérer l'incertitude et le caractère flou des observations et d'injecter notre connaissance *a priori* dans le modèle d'inférence. Concernant les images IRMf, nous proposons une nouvelle méthode de codage basée sur la transformation gaussienne généralisée permettant de décrire d'une manière fiable la topologie de surface des zones actives.

2.1 Identification de métabolites

On distingue généralement deux approches pour l'étude des métabolites. Dans la première approche, appelée l'**approche schématique**, les composants ne sont pas initialement identifiés. Seulement leurs modèles spectraux et intensités sont connus et comparés statistiquement pour identifier leurs caractéristiques spectrales appropriées qui distinguent des classes. Une fois ces caractéristiques établies, une variété d'approches peut alors être utilisée pour identifier les métabolites [Brindle02]. Dans l'autre approche, appelée l'**approche du profil ciblé**, les composants sont d'abord identifiés et évalués quantitativement en comparant le spectre NMR de la biopsie à une bibliothèque de référence spectrale obtenue de composants purs [Weljie06]. Bien que la première approche présente l'avantage de pouvoir détecter des métabolites non connus à l'avance, elle n'exploite pas de contraintes supplémentaires comme la connaissance de la composition de la biopsie, nombre de pics d'un métabolite, etc. Pour cela, nous allons définir trois critères triviaux pour modéliser ces *a priori* qui sont la localisation des pics, les paramètres de la densité de probabilité des amplitudes des pics et finalement le rapport entre les différent pics. Pour ce dernier, nous allons supposer que les rapports d'intensité des raies d'émission d'un métabolite donné sont les mêmes. Bien que cette contrainte soit théoriquement valide, elle n'est que rarement vérifiée en pratique du fait des changements des conditions d'acquisition, de la perte de la matière, etc. Pour contourner ce problème, nous allons utiliser la théorie des ensembles flous pour modéliser les erreurs introduites par ces perturbations et proposer un schéma d'annotation automatique. Le deuxième critère est les hyperparamètres de la Figure 4 représente la chaîne d'identification des métabolites.

2.2 Codage et mesure de similarité des zones actives

Parmi les méthodes de codage d'objets, le descripteur gaussien 3D (3DGD) proposé par Chaouch [Chaouch09] a montré son efficacité comparé à d'autres méthodes et a été classé premier sur la base de données de Princeton Shape Benchmark. Il fait partie de la famille des descripteurs basés sur une partition de l'espace. Le principe de ce descripteur est de caractériser et d'amplifier localement le voisinage de la surface 3D. Pour cela, les auteurs proposent d'utiliser des fonctions gaussiennes qui mesurent l'influence des points de la surface sur des points régulièrement répartis dans l'espace englobant l'objet 3D. Ce descripteur offre une caractérisation compacte, robuste et attachée à la forme 3D. Bien que cette méthode ait été appliquée avec succès sur la recherche sur Internet d'objets 3D, elle présente une lacune. En effet, elle ne fournit pas une information sur la topologie de sur-

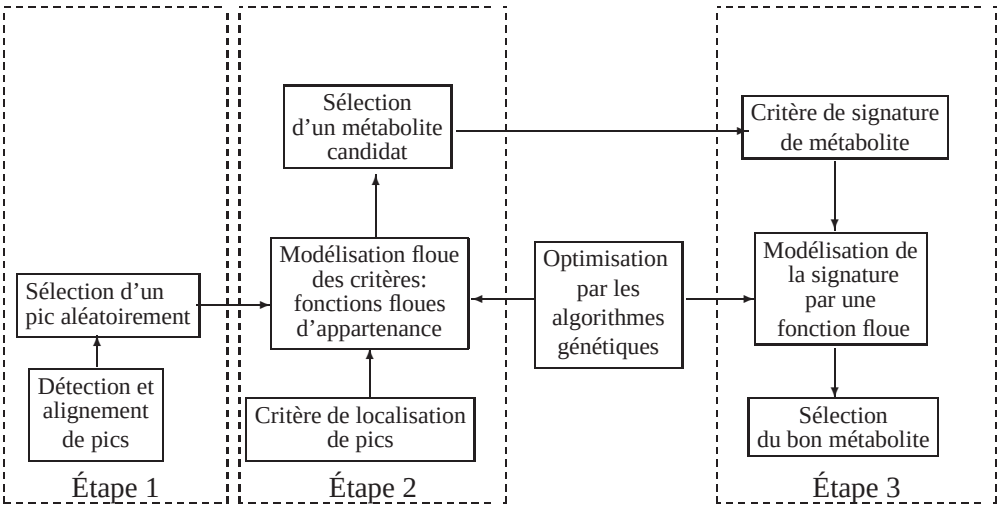


Figure 4: Chaîne d'identification des métabolites.

face d'objets. Pour cela, nous proposons un nouveau descripteur: le descripteur gaussien généralisé (3DGGD) inspiré de la méthode 3DGD. Cette méthode se base sur l'utilisation de la loi gaussienne généralisée à la place de la loi gaussienne permettant ainsi de s'adapter à la topologie de la surface de l'objet (surface plane, aiguë,..) grâce à son paramètre de forme α. La mesure de similarité peut être calculée en utilisant une distance euclidienne dans l'espace des coefficients gaussiens généralisés. Figure 5 représente la chaîne du codage des zones actives.

3 Classification d'objets

Plusieurs algorithmes de classification pour détecter les changements ont été développés dans les dernières décennies. Certains restent supervisées en raison de la difficulté de la tâche. D'autres ne le sont pas ce qui cause parfois une perte de robustesse et un temps de calcul relativement élevé. La première approche s'appuie sur des méthodes de classification supervisées afin de détecter les changements entre plusieurs acquisitions [Derrode03]. Cette tâche revient à discriminer les données entre deux classes: *changement* et *non changement*. La première nécessite une réalité de terrain afin d'en tirer une formation appropriée pour définir le processus d'apprentissage des classificateurs. Cependant, la vérité terrain est souvent difficile et coûteuse à trouver. Par conséquent, l'utilisation de méthodes de détection de changement non-supervisées est cruciale dans de nombreuses applications où la vérité terrain est hors portée [Fumera00].

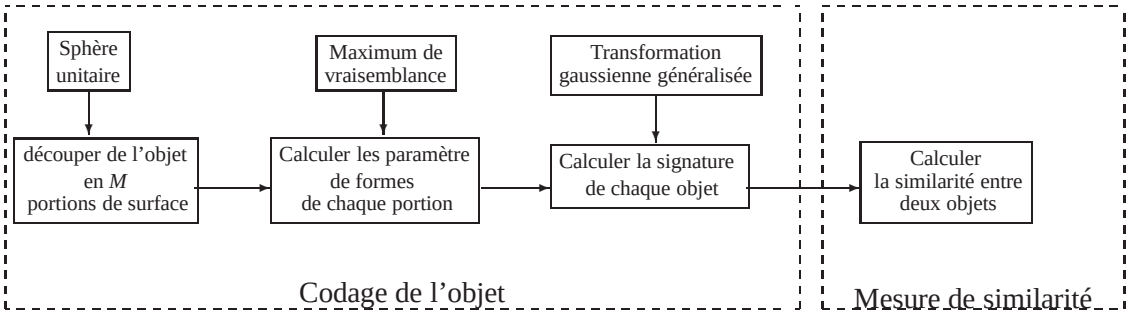


Figure 5: La chaîne du codage et de mesure de similarité des zones actives.

Dans la littérature, les méthodes à noyau sont largement utilisées pour la détection de changement. En effet, elles offrent plusieurs avantages par rapport à d'autres approches : elles réduisent la malédiction de la dimensionnalité élevée dans les données, augmentent la fiabilité et la robustesse de la méthode à la présence d'un niveau élevé de bruit et permettent une cartographie flexible entre les objets qui sont représentés par un vecteur de caractéristiques (entrées) et de l'étiquette de classe (sorties) [Shawe-Taylor04]. Cependant, l'inconvénient majeur des méthodes à noyau est le choix de la fonction noyau qui dépend fortement de l'application [Scholkopf00] .

Parmi les différentes méthodes à noyaux présentées dans la littérature (par exemple [Furey00] et [Bruzzone06]), les Descripteurs de données à vecteurs de support (SVDD) [Tax04] est adoptée ici.L'objectif de la méthode de classification SVDD consiste à cartographier les données dans un espace de grande dimension. Dans ce nouvel espace, une hypersphère entourant la plupart de l'ensemble de données appartenant à la classe d'intérêt (*cible* correspondant à la classe des *données inchangées*) et en rejetant les autres observations (qui seront considérées comme *les valeurs aberrantes*) est définie. Dans cet article, le problème de détection de changement est abordé d'une manière non supervisée. Notre objectif est de discriminer les données en deux classes : classe des données changées et classe de données inchangées.

Bien que les fonctions noyau de base sont plus ou moins appliquées avec succès pour la détection de changement, elles n'exploitent pas des contraintes supplémentaires souvent disponibles, tels que la dépendance et la distribution des données. Afin de tenir compte de ces caractéristiques dans notre schéma de détection de changement, nous proposons une nouvelle fonction noyau qui combine les fonctions noyau de base avec de nouvelles informations sur la distribution de caractéristiques et de la dépendance des données. Le défi est alors de trouver le moyen approprié pour traiter cette dépendance. Pour cela, nous avons opté pour la théorie des copules qui a prouvé son efficacité pour traiter la dépendance. La méthode proposée est notée SV3DH (SV3DH est l'acronyme de Support Vector Data Description including Dependency Hypothesis). Figure 6 représente la chaîne de classifi-

cation d'objet et de détection du changement.

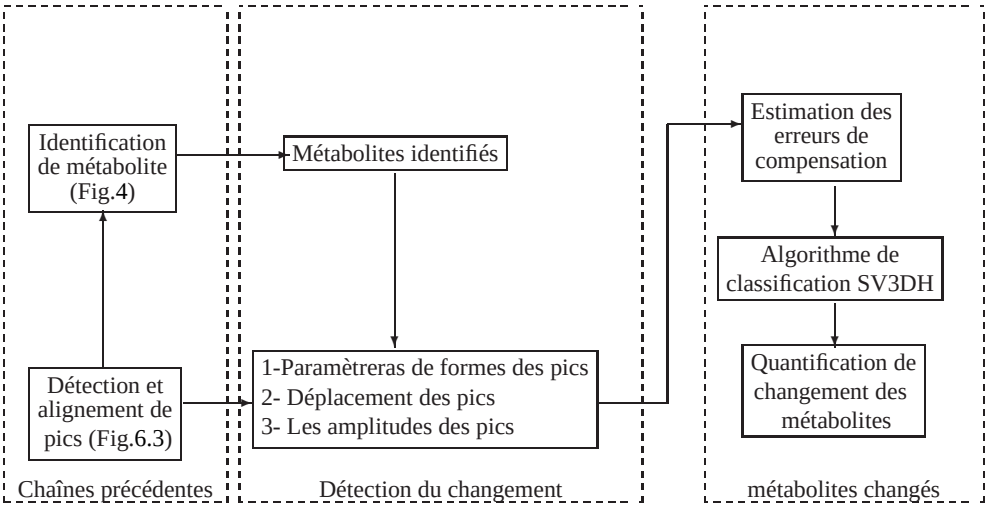


Figure 6: Chaîne de calcification et de détection du changement.

4 Résultats

Dans cette section, nous présentons les résultats expérimentaux obtenus avec les méthodes proposées sur des données simulées et réelles. Pour mettre en évidence l'intérêt de ces méthodes, nous avons comparé chaque méthode proposée avec les méthodes existantes. Concernant la détection et l'alignement de pics des spectres HSQC, et afin de montrer l'intérêt de l'utilisation de la théorie de l'évidence, nous avons comparé notre méthode avec une méthode purement bayésienne [Toews05] sur des spectres simulés avec différent Peak to signal ratio PSNR. Les résultats obtenus sont présentés dans Table 1. Nous pouvons facilement constater que la méthode proposée est meilleure que la méthode bayésienne.

En ce qui concerne l'alignement des zones actives, nous avons comparé notre méthode avec la méthode de l'ACP continue [Vranić01a] sur cinq bases de données simulées. Chacune contient cent objets 3D. Les résultats obtenus sont présentés dans Table 3.4. Nous pouvons constater que la méthode proposée est plus adaptée aux zones actives que l'ACP continue.

A l'égard de l'identification des métabolites, nous avons comparé notre méthode avec la méthode Support Vector Machine SVM [Camps-Valls05] et une autre méthode de seuillage [Xia08]. Afin de valider et de souligner les avantages de l'approche proposée, nous

	Méthode évidentielle		Méthode bayésienne	
$PSNR$	ε_c	ε_H	ε_c	ε_H
$30dB$	**5.1** 10^{-3}	**9.1** 10^{-5}	0.097	6.1 10^{-3}
$28dB$	**1.21** 10^{-2}	**5.1** 10^{-4}	0.139	8.6 10^{-3}
$25dB$	**0.1098**	**2.5** 10^{-3}	0.2584	1.91 10^{-2}
$23dB$	**0.1874**	**9.35** 10^{-3}	0.3278	2.03 10^{-2}

Table 1: Les erreurs de déplacement chimiques moyennes du carbone ε_c et de l'hydrogène ε_H exprimées ppm.

	Méthode proposée	ACP continue
data set1	**0.02**$\pm$ **1.2** 10^{-4}	0.097 $\pm$ 7.85 10^{-4}
data set2	**0.038** $\pm$ **5.8** 10^{-4}	0.124 $\pm$ 2.14 10^{-3}
data set3	**0.0474** $\pm$ **9.7** 10^{-4}	0.301 $\pm$ 4.23 10^{-3}
data set4	**0.062** $\pm$ **1.03** 10^{-3}	0.832 $\pm$ 5.82 10^{-2}
data set5	**0.078** $\pm$ **4.12 1**0^{-3}	1.177 $\pm$ 7.98 10^{-2}

Table 2: Les erreurs de déplacement moyenne et l'écart type obtenus par la méthode proposée et la ACP continue.

utilisons deux mesures de performences: *rappel* et *précision* définis par:

$$rappel = \frac{TP}{TP+FN}; \quad precision = \frac{TP}{TP+FP}$$

où TP représente le nombre de vraies identifications, FN le nombre de fausses identifications négatives et FP le nombre de fausses identifications positives. Les résultats sont présentés dans Table 4.1. Nous remarquons que la méthode proposée donne de meilleurs résultats que ceux obtenus avec la méthode SVM qui ne prend pas en compte notre connaissances *a priori* sur les spectres.

	Méthode proposée		SVM		Méthode de seuillage	
$PSNR$	$rappel(\%)$	$prcision(\%)$	$rappel(\%)$	$prcision(\%)$	$rappel(\%)$	$prcision(\%)$
$30dB$	**93.87**	**95.11**	90.38	91.72	81.16	78.01
$28dB$	**92.42**	**94.82**	88.50	89.61	78.98	76.12
$25dB$	**92.84**	**94.64**	82.11	86.90	75.77	74.25
$23dB$	**89.02**	**90.18**	83.02	84.66	74.02	71.88

Table 3: Les mesures $rappel(\%)$ et $prcision(\%)$ obtenues avec: notre méthode, la méthode SVM, et la méthode de seuillage sur des données simulées.

Pour le codage et la mesure de similarité des zones actives, nous avons comparé notre méthode à la méthode (3DGD) [Chaouch09] et à la méthode de l'histogramme [Ankerst99].

	Méthode proposée		3GD		Méthode de l'histogramme	
dataset	*recall*(%)	*precision*(%)	*recall*(%)	*precision*(%)	*recall*(%)	*precision*(%)
dataset1	**91.35**	**88.02**	84.21	83.25	80.38	75.72
dataset2	**90.44**	**88.12**	85.50	82.01	81.67	77.45
dataset3	**91.74**	**87.23**	81.09	78.96	77.10	75.98
dataset4	**88.17**	**86.88**	79.69	78.64	70.18	67.41
dataset5	**90.95**	**89.03**	86.78	84.35	74.11	72.08

Table 4: Les mesures $rappel(\%)$ et $prcision(\%)$ obtenues avec: notre méthode, la méthode 3GD, et la méthode de l'histogramme sur des données simulées.

Les résultats obtenus sont présenté dans Table 4.2. Comme nous pouvons le voir, notre méthode achève un meilleur résultat de codage et de mesure de similarité.

Nous présentons maintenant les résultats expérimentaux obtenus avec la méthode SV3DH pour la détection de changements en imagerie satellitaires. Ces images ont été particulièrement sélecionnées car on dispose d'une vérité terrain ce qui n'est pas souvent le cas pour d'autres applications comme l'imagerie médicale.

Pour cela, nous avons considéré une série d'images à haute résolution (1305 x 1520 pixels) recueillies sur une zone géographique de l'Alaska. Ces images sont disponibles en ligne [lsiml]. Elles ont été acquises par le satellite Landsat-5 Thematic Mapper (TM) en 22 juillet 1985 et 13 juillet 2005, respectivement. Une zone avec 1024 x 1024 pixels est sélectionnée pour les expériences. Le satellite Landsat-5 TM fournit des imageries optiques sur sept bandes spectrales (Bandes 1-7). la vérité terrain des cartes de détection des changements est disponible dans [lsiml].

Afin de valider et de souligner les avantages de l'approche proposée, nous utilisons trois mesures de performances: le nombre de fausses détections *PFA*, le nombre de détections manquées *PMD* et l'erreur globale *PTE* :

$$PFA = \frac{FA}{N_F} \times 100\%; \quad PMD = \frac{MD}{N_M} \times 100\%; \quad PTE = \frac{MD+FA}{N_M+N_F} \times 100\%$$

Où *FA* représente le nombre de pixels inchangés et qui ont été incorrectement déterminé comme changés, N_F le nombre total de pixels inchangés, *MD* le nombre de pixels changés et qui ont été incorrectement déterminés comme inchangés, N_M le nombre total des pixels changés.

Table. 5.2 presente les résultats obtenus avec la méthode SV3DH avec deux autres méthodes: l'SVDD classique [Tax04] et le Système à Vastes Marges SVM standard [Bruzzone06]. Notons que le noyau RBF gaussien est utilisé pour les deux méthodes. Nous pouvons constater que la méthode que nous proposons fournit des résultats meilleurs par rapport aux deux autres méthodes. Cela signifie que la fonction noyau proposée améliore la discrimination des caractéristiques.

	Fausses détections	détections manquées	erreur globale
SV3DH	0.71 %	5.01 %	1.09 %
SVDD	1.87 %	6.81 %	2.01 %
SVM	1.04 %	6.31 %	1.75 %

Table 5: Le nombre de fausses détections, le nombre de détections manquées et l'erreur globale obtenus avec les méthodes SV3DH, SVDD et SVM.

Conclusion

Dans cette thèse, nous nous sommes intéressés à la recherche par le contenu d'objets 3D, et plus particulièrement à l'identification des bio-marqueurs. L'objectif de notre travail a été de proposer des méthodes rapides et efficaces permettant de classifier les signaux médicaux et de détecter les changement entre un individu/groupe requête et les différents groupes de profils appartenant à une base de signaux médicaux. L'idée clé était de proprement intégrer notre connaissances *a priori* dans les méthodes proposées. Les différentes problématiques de été étudiées à savoir l'alignement, le codage, la mesure de similarité et la classification d'objets 3D. Les différentes méthodes proposées ont été validées, dans une première étape, sur des données simulées (ou réelle quand on dispose d'une vérité terrain) afin de prouver leur apport comparées aux méthodes existantes. Dans une deuxième partie, l'ensemble des résultats obtenus sur les données réelles ont été examinés par des experts de chaque domaine (spectroscopie RMN HR-MAS et IRMf). Cette validation montre le bon comportement de nos algorithmes ainsi que leur applicabilité à grande échelle que ce soit pour les spectres HSQC (RMN HR-MAS) et les images IRMf.

Contents

CHAPTER 1

Introduction

Contents

1.1 Motivation

Technology is transforming healthcare. The better patient care and diagnosis efficiency is the goal. In this context, medical equipment and technology are increasingly sophisticated providing a new information essential to physician diagnosis, particularly in the areas of cancer and neurological diseases. Moreover, the progress in medical information acquisition technologies by improving the signal to noise ratio or by the coupling between anatomy and physiology, has gradually given to medical imaging its indispensable status for diagnosis, prognosis and therapeutic management. Furthermore, the analysis of biological signal combinations appears promising even if they are not systematically part of clinical routine today.

The medical signal acquisition techniques are constantly evolving in recent years (ultrasound endoscopy, Positron Emission Tomography PET, Nuclear Magnetic Resonance NMR spectroscopy, Magnetic Resonance Imaging-MRI, functional MRI-fMRI) and providing an increasing amount of data which should be then analyzed. Indeed, from a static observation to a dynamic multi-observation, from an information about the organs structure to an information about their functions, the signal acquisition techniques are potentially carrying the disease signature (biomarkers) well beyond the punctual clinical examination. In this context, automatic signal processing methods are regularly proposed to assist the expert in the qualitative and quantitative analysis of these images in order to facilitate their interpretation. These methods should take into account the physics of signal acquisition, the *a priori* we have on the signal formation and the amount of data to analyze for a more accurate and reliable interpretation.

Among the new signal acquisition techniques, the NMR spectroscopy for biological tissues analysis and the fMRI for functional brain activities and connectivity analysis are

explored to identify new biomarkers (objects). These biomarkers could be used to help the diseases diagnosis or to monitor therapeutic effects. To this end, it is important to improve the analysis tools associated with these techniques. In this context, we propose a new content-based object indexing and retrieval scheme for biomarkers detection.

This proposed indexing scheme consists of both an off-line and an on-line phases. In the off-line one, the medical profiles of different medical signal group or population (*e.g*, group of normal or pathological signals) are established. In the on-line phase, the assignment of a new individual/group to a given profile defined in the off-line phase is performed. Both phases are divided into three steps:

1. Object detection and alignment,
2. Object coding and similarity measurement,
3. Object classification.

1.2 Methodology

We focus in this thesis on:

- The two-dimensional 2D Heteronuclear Single Quantum Coherence HSQC spectra obtained by High-Resolution Magic Angle Spinning HR-MAS NMR for biological tissue (biopsy) analysis [Schmidt-Rohr94].
- The fMRI images for functional brain activities analysis [Engel97].

Each processed medical information (2D NMR spectra or fMRI) will be characterized by a set of objects (biomarkers) that we seek to extract, align, and code. The clustering of these objects by measuring their similarity will allow then their classification. It is this global content-based object indexing and retrieval scheme (henceforth called indexing scheme) that we adopt. In our case, these objects are:

- The emission peaks for 2D HR-MAS NMR spectra (*i.e*, a set of peaks is the response corresponding to the metabolite presences, each metabolite generating different emission peaks reflecting the presence of small molecule through the interaction between the Proton and Carbon-13 in the case of HSQC spectra).
- The active zones for fMRI (*i.e*, an active zone is the response of brain activity to a stimulus).

However, this indexing task is not trivial and requires the development of new processing tools. Therefore, we are interested in this thesis to properly model and integrate the *a priori* knowledge we have on these biological signal allowing us to propose thereafter appropriate methods to each indexing step and each type of signal. The methods we propose are:

- For the object detection and alignment step in the case of 2D HR-MAS NMR spectra, we propose a new peak detection and alignment method based on the use of evidence theory. Indeed, the evidence theory allows us to handel the uncertainty models and the imprecision that well adopted HSQC spectra. Therefore, we propose the coupling between the fuzzy set theory and Bayesian theory to model and quantify the degree of spectrum imprecision which will be used to define the mass functions (*i.e*, a function that models the belief degree on a given hypothesis). We particularly show that the use of the evidence theory for peak detection and alignment consistently achieve a higher performance compared to a pure Bayesian approach. Regarding the fMRI images, we proceed in a first step to the extraction of fMRI active zones using a Hidden Markov chain segmentation algorithm. Then, we propose a new active zone alignment algorithm relying on the use of non-linear principal component analysis (PCA) algorithm, well suited to fit the cortex shape, to estimate the planes of symmetry. These planes of symmetry will be used then to align active zones.

- The object coding step consist in calculating for each object a set of compact descriptive attributes defining its signature. A similarity measurement using the descriptors aims at comparing two objects and at grouping similar objects as well. For HSQC spectra, the object coding step is not useful since the peaks already have a parsimonious representation with three parameters (location, amplitude and shape). For the similarity measurement, we propose a new method based on the combination of Bayesian theory and the theory of fuzzy sets to handle the uncertainty and fuzziness of the observations and to integrate *a priori* knowledge in the inference process. Concerning the fMRI images, we propose a new coding method based on the generalized Gaussian transformation allowing us to reliably describe the surface topology of the active zones. In particular, we show that the proposed coding method not only provides a compact representation of the object, but also a signature faithful to its shape. We also propose a similarity measurement robust to small displacements and little variations of the objects.

- For the objects classification step, we propose a new Support Vector Data Description (SVDD) kernel function combining the features of basic kernel functions with new information about features distribution and then dependency between samples. The dependency between samples will be based on copulas theory that is used for the first time to our knowledge in the SVDD framework. We show that the use of the new kernel function increases the classification performance with respect to the basic kernel functions either on simulated or real data.

1.3 Results

The different proposed methods were validated in a first part on simulated data to demonstrate their behavior compared to existing methods. In a second part, all the results obtained on real data have been examined by experts in each domain (HR-MAS NMR spectroscopy and fMRI). This validation shows the good performance of our algorithms leading to sim-

ilar results to those obtained by physicians in a short time both for the HSQC spectra (HR-MAS NMR) and fMRI images.

1.4 Organization of the manuscript

This manuscript is divided into six chapters. The second chapter is devoted to the description of the medical images formation/acquisation on which we work in this thesis. We focus on the description of image content and on the different characteristics associated with the acquisition process. This chapter is divided into three parts: the first one describes the physical principles of Nuclear Magnetic Resonance (NMR) as well as the 1D and the 2D NMR experiments. Functional Magnetic Resonance Imaging (fMRI) is then introduced in part two. Finally, the Statement of indexing problem is studied.

In the third chapter, we describe in the first part the widely used object detection and alignment methods proposed for the indexing schemes. Then we detail the proposed peak detection and alignment method as well as the active zone alignment method in part two and three respectively.

Chapter four is devoted to the second step of the indexing scheme: the object coding and similarity measurements. To this end, we present in the first part an overview of object coding and similarity measurement methods. Then, we describe the proposed method for peak similarity measurement. In the third part, the proposed active zone coding algorithm inspired from the Gaussian transformation is presented.

The fith chapter presents the kernel-based methods for object classification task. We propose a new Support Vector Data Description (SVDD) kernel function which combines the characteristics of basic kernel functions with new information about features distribution. We pay a particular attention to check that the proposed kernel function is robust with higher performance compared to classic Support Vector Machine (SVM) and SVDD methods on both synthetic and real data sets.

In the sixth chapter, we first describe the entire work-flows of both treatment chains (HSQC spectra and fMRI data). Then, we provide an assessment of the developed methods. This is done in two stages. Initially, we proceed to a detailed study of some representative cases. Then, the results on a consistent databases is compared to a ground truth provided by experts.

This work ends with a general conclusion that provides a summary contributions of this thesis. We present likewise some perspectives.

Publications

The main publications are available at the following URL:

http://lsiit-miv.u-strasbg.fr/miv/84-BELGHITH-Akram.html

Journal articles

- A. Belghith, C. Collet, L. Rumbach, J-P. Armspach, A unified framework for peak detection and alignment: application to HR-MAS 2D NMR spectroscopy, Journal of Signal, Image and Video Processing (to appear).
- A. Belghith, C. Collet, J-P. Armspach, Change detection based on Support Vector Data Description handling dependency , Pattern recognition letter (Article in review in February 2011).
- A. Belghith, C. Collet, I-J Namer, K. Elbayed, J-P. Armspach, A statistical framework for biomarker identification of biopsies using HR-MAS 2D NMR spectroscopy, IEEE Transaction on biomedical engineering (Article submitted in February 2011).

International conferences

- A. Belghith, C. Collet, J-P. Armspach, Detection and identification of biomarker in biopsies using HR-MAS HSQC spectroscopy, 4th International Conference on Biomedical Engineering, VietNam, January 2012.
- A. Belghith, C. Collet, J-P. Armspach, A unified framework for peak detection and alignment: application to HR-MAS 2D NMR spectroscopy, 6th International Conference on Mass Data Analysis of Images and Signals in Medicine, Biotechnology and Chemistry, MDA 2011, New York, États-Unis, septembre 2011.
- A. Belghith, C. Collet, J-P. Armspach, Change detection based on Support Vector Data Description handling dependency, 18 th IEEE International Conference on Image Processing, ICIP Bruxelle, Belgique, septembre 2011.
- A. Belghith, C. Collet, J-P. Armspach, A statistical framework for biomarker identification of biopsies using HR-MAS 2D NMR spectroscopy, 8th IEEE International Symposium on Biomedical Imaging, ISBI 2011, Chicago, États-Unis, mars 2011.

International workshops

- A. Belghith, C. Collet, J-P. Armspach, L. Rumbach, I-J. Namer, K. Elbayed, A statistical framework for biomarker identification using HR-MAS 2D NMR spectroscopy, International Society for Magnetic Resonance in Medicine (ISMRM) Meeting, Montréal, Canada, mai 2011.

- A. Belghith, C. Collet, J-P. Armspach, A Unified framework for metabolite processing, Computational Surgery Conference, Houston, États-Unis, janvier 2011.

Software

This PhD work has led to the design of a free software called ***Medihsqc*** dedicated to the 2D HSQC spectrum processing. It is available online at the following URL:

http://alsace.u-strasbg.fr/ipb/tim/doku.php?contenu=Medipy/index.html

CHAPTER 2

Magnetic Resonance Signal background

Contents

The medicine has considerably progressed over the past twenty years. On the one hand, the development of medical technology has enabled a significant increase in life expectancy. On the other hand, the costs of health services are increasing particularly in France which devotes 8.7% of its GDP (Gross Domestic Product) to medical services, the highest proportion of all countries [ICID10]. Therefore, increasing the diagnosis accuracy seems to be crucial in order to reduce the time of hospitalization and improve patient survival and his life quality. In this context, the medical signal processing has found its success and has emerged as an ideal technique for biomarkers identification and analysis and hence for helping the differential diagnosis of diseases.

Medical signal acquisition has certainly contributed to the improvement of medicine from 20 to 30 years. In particular, the development of new medical signal acquisition techniques such as the two dimensional 2D Heteronuclear Single Quantum Coherence (HSQC) NMR spectroscopy and the functional Magnetic Resonance imaging fMRI which are examples of technological advancements in medicine researches. These techniques allow physicians to directly observe phenomena that previously had to be blind-evaluated or predicted. Indeed, on the one hand, the NMR spectroscopy enables the identification of metabolites in non-invasive manner. On the other hand, fMRI technique is used to measure the hemodynamic response related to neural activity in the brain.

This chapter is divided into three parts: in the first one we describe the physical principles of NMR spectroscopy, then we present the 1D as well as the 2D NMR *ex vivo* experiments. We particularly show the contribution of 2D spectra compared to 1D spectra. fMRI experiment is then briefly introduced in Section 2.2. Finally, the proposed content-based

information indexing and retrieval scheme for biomarkers identification, the main topic of this thesis, is detailed in section 2.3.

2.1 NMR spectroscopy

The Metabolomics is an exponentially growing field of 'omics' research concerned with the comparison, identification and quantification of large numbers of metabolites in biological system [Fiehn02]. This emergent science of metabolomics enables the identification of biomarker diseases that integrates biochemical changes in disease and predict human reaction to treatments. In this context, the NMR spectroscopy has emerged as an ideal platform for metabolite studying [Holzgrabe99, Beckonert07]. Indeed, in 1977, Ekstrand et al. have established the possibility of studying the metabolism by application of NMR [Ekstrand77]. In this work, a suspension of red blood cells was analyzed by liquid NMR to study the proton relaxation times of some metabolites such as lactate, pyruvate, alanine or creatine. The large amount of information becoming accessible to a better understanding of metabolism using the NMR technique was immediately apprehended by the scientific community. This has naturally led the proton NMR performed on biological fluids to take a prominent place in the field of pharmacology, toxicology and the study of pathological changes metabolism. Fluids that have been analysed by NMR are: the urine [Bales84, Foxall95, Griffin00, Melendez01], the bile [Keun02, Paczkowska03], the blood plasma [Wevers94, Nicholson95, Alum08], the cerebrospinal fluid [Dunne05, Jukarainen08, Sinclair10], the milk [Martin-Pastor00, Holmes00, Bertram07], the saliva [Silwood99, Grootveld05, Grootveld06], the gastric fluid [Lof97], the seminal fluid [Lynch94, Tomlins98] and the amniotic fluid [Joe08, Graca09, Cohn09]. Nevertheless, the first studies involving the human organ analysis were addressed with the in vivo and ex vivo NMR. In fact, since the first application of NMR spectroscopy in vivo [Ross83] and ex vivo [Mountford82], the technique of NMR has been increasingly used as a powerful tool to explore, in situ, a significant number of organs such as heart [Barba07], kidney [Tate00, Garrod01, Righi07], prostate [Swanson08], cervical [Mahon04, Sitter04] as well as the stomach [Tugnoli04, Tugnoli06, Calabrese08]. Indeed, the technique of NMR allows us to obtain metabolic information needed for clinical diagnosis of the patient with a suspected injury, to establish the prognostic or to study the diseases evolution [Howe93, Ross94, Howe03, Kwock06]. The fields of application in the medicine include as well the study of brain tumors [Opstad08, Opstad08, Wright10], the breast tumor [Beckonert03], the ovarian tumors [Odunsi05], the neurological disorders such as epilepsy [Hammen07], the acquired immune deficiency syndrom AIDS [Corr06], the Alzheimer's disease [Thompson07] multiple sclerosis disease [Brenner93, Davies95] or other neurodegenerative diseases [Apostolova07] as Parkinson's disease [Camicioli07]. In the following we describe the principles of the NMR technique.

2.1.1 NMR principles

The structure determination of almost any biological or organic molecules as well as many inorganic molecules begins with the NMR spectroscopy. Indeed, the technique of NMR

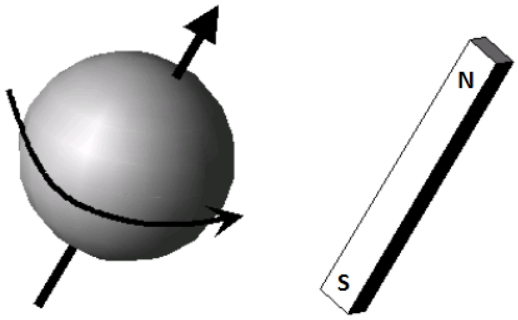


Figure 2.1: The nucleus has a magnetic moment which can be assimilated to a small magnet.

has become one of the praised methods for identifying the structure of both pure and mixed compounds as well as solid or liquid compounds. This technique often involves performing NMR experiments to deduce the molecular structure from the magnetic properties of the atomic nuclei and the surrounding electrons. The NMR technique relies on the atom nucleus behavior while spinning. Indeed, the atom nucleus can be considered as a positively charged sphere spinning on itself (Fig.2.1). As result of this spin, each nucleus processes an angular momentum p and a magnetic moment μ [Friebolin91].

Thanks to the quantum mechanics, we can express the behavior of the magnetic moment μ with the number of spin denoted I [Liboff98]. Indeed, I determines the number of possible directions that a nucleus can adopt in the presence of an external magnetic field. In fact, in the presence of a external magnetic field $\overrightarrow{B_0}$, the component of μ along $\overrightarrow{B_0}$ (oriented along z-axis) is expressed as [Friebolin91]:

$$\mu_z = \gamma.\hbar.m \tag{2.1}$$

where γ is the gyromagnetic ratio related to the treated atom and $\hbar$ is the reduced Planck constant. The magnetic quantum number m can take values $m \in \{I, I-1, I-2, ... -I\}$ where I is the spin number. For example, if we consider a proton ^1H ($I = \frac{1}{2}$) placed in a magnetic field, there are two possible states:
$m = \pm\frac{1}{2}$ or $m = -\frac{1}{2}$. Thus, we have:

$$\mu_z = \pm\frac{\gamma.\hbar}{2} \tag{2.2}$$

Fig.2.2 shows the two possible orientations for a spin with $I = \frac{1}{2}$.

The proton nucleus, abundantly present in the human body in the water molecules form, are assumed to be randomly oriented. When a sample is exposed to an intense

external magnetic field $\overrightarrow{B_0}$, the nuclear magnetic moments are orientated in the direction of this field. However, the thermal agitation opposes to this orientation [Van de Ven95]. As we saw in the previous paragraph, the component magnetic moments μ_z of spin $\frac{1}{2}$ can only adopt two possible orientations: one is parallel to $\overrightarrow{B_0}$ (Fig.2.3.a) and the other one is anti-parallel to $\overrightarrow{B_0}$ (Fig.2.3.b).

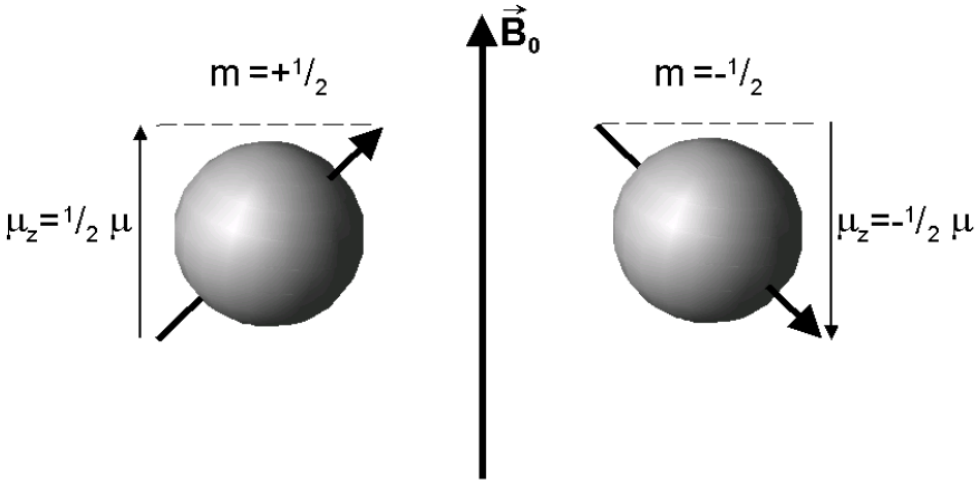


Figure 2.2: Orientations of the magnetic dipole in the presence a magnetic field $\overrightarrow{B_0}$ for a spin $\frac{1}{2}$.

Since, the number of parallel nuclei is slightly higher than that of antiparallel nuclei, the vector sum of all nuclear magnetic moments is non-zero and is then aligned among the direction of the field $\overrightarrow{B_0}$ [Friebolin91]. This amount is called the nuclear magnetization and denoted by $\overrightarrow{M} = \sum_i \overrightarrow{u_i}$ (Fig.2.4) where $\overrightarrow{\mu_i}$ the magnetic moment of the i^{th} nuclei. Since the dephasing between the different precession movements (*i.e*; changes in the orientation of the nuclei rotation axis) of the elementary magnetization $\overrightarrow{\mu_i}$ is uniformly distributed, the transverse component of the resulting magnetization is equal to zero.

Moreover, when a small magnetic field $\overrightarrow{\mu}$ is plunged into an intense magnetic field $\overrightarrow{B_0}$, we can show that $\overrightarrow{\mu}$ is animated by a precession movement around $\overrightarrow{B_0}$ which is analogous to the movement of a spintop axis about the vertical (Fig.2.5) [Fukushima81]. The speed at which this precesses occurs is given by the Larmor frequency relationship:

$$\omega_0 = \gamma.||\overrightarrow{B_0}||$$

Since this precession is around $\overrightarrow{B_0}$, it does not alter the direction or the modulus of the magnetization $\overrightarrow{M}$ [Fukushima81]. The origin of this precession movement lies in the fact that, when plunged into a magnetic field $\overrightarrow{B_0}$, a magnetic moment $\overrightarrow{\mu}$ undergoes the force $\overrightarrow{F}$: $\overrightarrow{F} = \overrightarrow{\mu} \wedge \overrightarrow{B_0}$

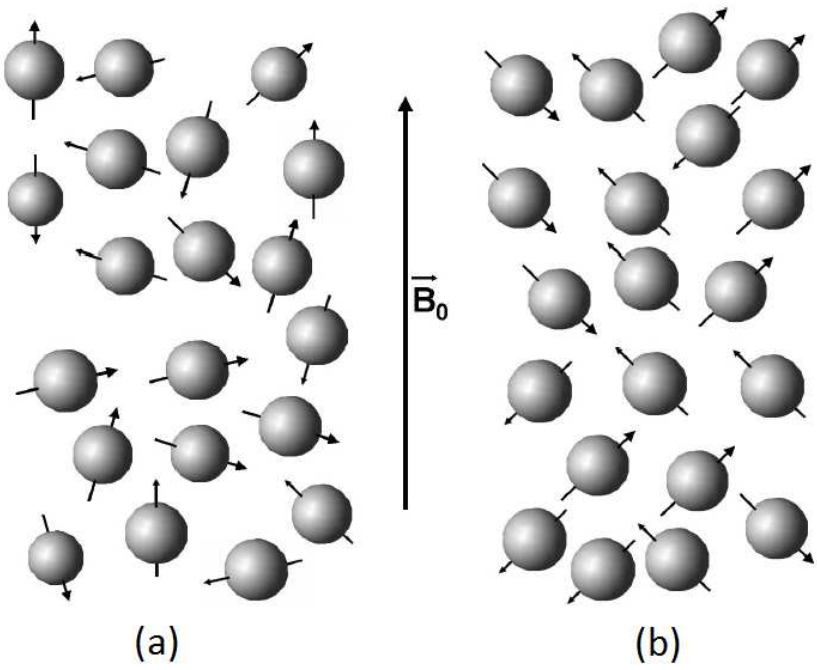


Figure 2.3: Orientation of nuclear magnetic moments: a) in the absence b) in the presence of an external magnetic field.

This force is hence the result of the vector product $\wedge$ of the magnetic moment $\overrightarrow{\mu}$ and the magnetic field $\overrightarrow{B_0}$. This vector relation can be written as three differential equations:

$$\frac{d\mu_x}{dt} = \gamma B_0 \mu_y, \qquad \frac{d\mu_y}{dt} = \gamma B_0 \mu_x, \qquad \frac{d\mu_z}{dt} = 0, \tag{2.3}$$

The solution of these equations is:

$$\mu_x = \mu_x^{t=0} sin(\gamma B_0 t), \qquad \mu_y = \mu_y^{t=0} cos(\gamma B_0 t), \qquad \mu_z = \mu_z^{t=0} \tag{2.4}$$

μ_z remains unchanged and μ_{xy} turn around $\overrightarrow{B_0}$. The vector $\overrightarrow{\mu}$ is therefore animated by a precess around $\overrightarrow{B_0}$.

We recall that the magnetization is proportional to the number of spins, so it is that we attempt to measure by NMR. However, $\overrightarrow{M}$ remains unobservable when it is parallel to $\overrightarrow{B_0}$, so it should be rotated by 90° from z-axe (Fig.2.6.a). To rotate $\overrightarrow{M}$, it is sufficient to apply another transient magnetic field $\overrightarrow{B_1}$ directed to 90° from $\overrightarrow{B_0}$. The magnetic moments are animated to precess around $\overrightarrow{B_1}$. Once a rotation of 90° was obtained, the $\overrightarrow{B_1}$ field is turned off (Fig.2.6.b). Since, the magnetization $\overrightarrow{M}$ is the sum of all nuclear magnetic moments, then it is oriented at 90° from $\overrightarrow{B_0}$ and hence can be measured (Fig.2.6.c). The resonance frequency depends on the molecular environment as well as the gyromagnetic ratio γ and $\overrightarrow{B_0}$. We call "chemical shift" the variation of resonance frequency with the

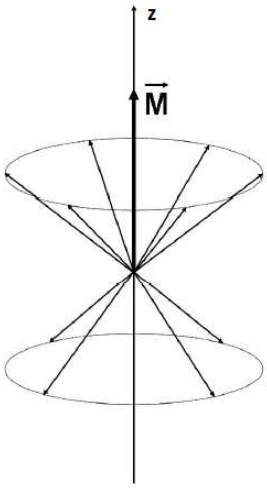


Figure 2.4: The sum of magnetic moment vector $\overrightarrow{M}$. In the presence of a magnetic field, $\overrightarrow{M}$ is nonzero and directed in the direction of this field.

shielding hyperparameter σ (the process of reducing) of $\overrightarrow{B_0}$. The expression of the chemical shift Ch_s is given by:

$$\delta = \frac{\gamma||\overrightarrow{B_0}||(1-\sigma)}{2\pi} \tag{2.5}$$

Note that chemical shift Ch_s is usually expressed in parts per million (ppm) by frequency, and it is calculated as follows:

$$\delta = \frac{\text{difference between the resonance frequency and that of a reference substance}}{\text{operation frequency of the spectrometer}} \tag{2.6}$$

Since the resonance frequency strongly depends of the structural environment of the nucleus, the NMR technique becomes the structural tool of choice for chemists [Bovey69] to study molecular structures and their associations and interactions.

As shown in the previous paragraph, the advantage of NMR precisely lies in the observation of the spin return to the equilibrium state after being irradiated with $\overrightarrow{B_1}$. This return to equilibrium is characterized by two relaxation processes. In order to describe the magnetization position during an NMR experiment, the coordinate system called "reference coordinate system" can be used. It consists of three orthogonal axes (x, y, z) relatively fixed to a reference. By convention, the z-axis is parallel to $\overrightarrow{B_0}$. The z-axis is called the longitudinal axis and the plane (xy) is called transverse plane. At any instant of an NMR experiment, the magnetization has a component that is parallel to $\overrightarrow{B_0}$ so-called longitudinal magnetization (denoted M_z) and a component that is perpendicular to $\overrightarrow{B_0}$ called transverse magnetization (denoted M_{xy}).

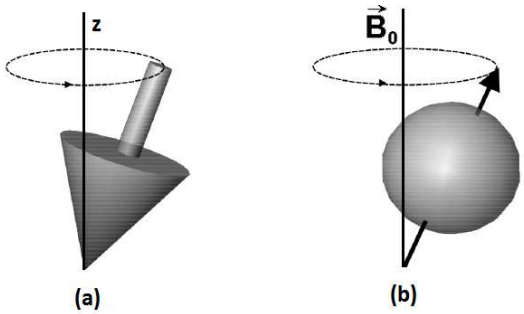


Figure 2.5: Precession movement: a) of a spintop around the vertical b) of a magnetic moment around the field $\overrightarrow{B_0}$.

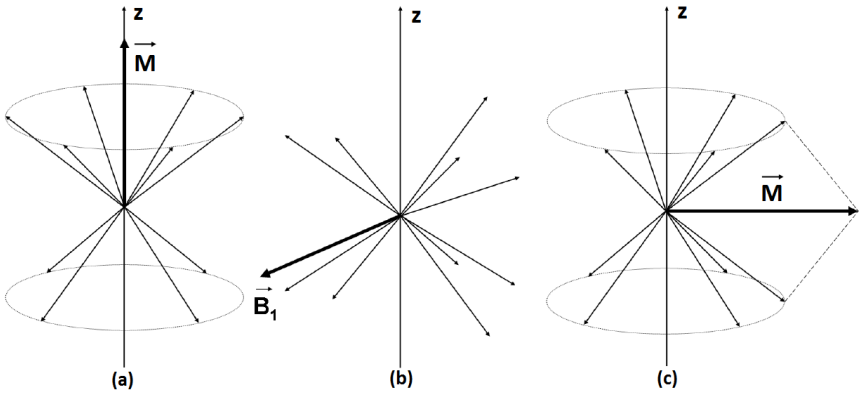


Figure 2.6: The sum of magnetic moment vector $\overrightarrow{M}$. In the presence of $\overrightarrow{B_0}$, $\overrightarrow{M}$ is non-zero and directed along z (a). The application of a field $\overrightarrow{B_1}$ perpendicular to $\overrightarrow{B_0}$ causes the 90° rotation of all the magnetic moments (b). When $\overrightarrow{B_1}$ is no longer applied, the magnetic moments return to a precession rotation around $\overrightarrow{B_0}$, and only $\overrightarrow{M}$ remains in the transverse plane (c).

After being tipped in the transverse plane, the magnetization will return to its equilibrium position, parallel to $\overrightarrow{B_0}$. The longitudinal and transversal components are differently affected by the relaxation phenomenon [Friebolin91]:

1. **the longitudinal relaxation**: The return of the longitudinal magnetization (Mz) to its equilibrium state generally happens according to a mono-exponential process. We call the characteristic time of the longitudinal magnetization decay the "longitudinal relaxation time". It is denoted by T_1. The longitudinal relaxation mechanisms are associated with fluctuations of local magnetic fields. What matters in the context of longitudinal relaxation are the fluctuations of the local fields at the Larmor frequency.

2. **the transverse relaxation**: It consist in canceling M_{xy}. In most cases the transverse magnetization mono-exponentially decreases. We call the characteristic time of the monoexponential decay the "transverse relaxation time". It is denoted by T_2. Transverse relaxation mechanisms include, besides those mentioned above, interactions between spins. The interactions between magnetic dipoles of neighbor spins generate at each spin a local magnetic field. The local field fluctuates as a result of movement of neighbor spins, or as a result of changes in their quantum states. This loss of phase coherence is an additional mechanism for the transverse relaxation.

2.1.2 NMR experiments

The 1D NMR experiment

The basic 1D NMR experiment can be decomposed as follows:

1. An excitation phase: it consists in irradiating a sample with a radio-frequency (RF) pulse whose frequency is close to the resonance frequency of the considered nucleus. The amplitude and duration of the RF pulse are calibrated to tip the magnetization to the transverse plane (90° rotation of the magnetization).

2. A detection phase: it consists in measuring the tipped magnetization.

Since the $\overrightarrow{B_0}$ is always present, the magnetization $\overrightarrow{M}$ rotates around it at the speed ω_0. Due to the rotation of $\overrightarrow{M}$, the magnetic flux through the coil (proportional to the component of M along the axis of the coil) periodically varies and then induces, in the coil, an electrical signal directly proportional to $\overrightarrow{M}$. This signal is called the FID for "Free Induction Decay [Van de Ven95]. As $\overrightarrow{M}$ returns gradually to its equilibrium position parallel to $\overrightarrow{B_0}$, the FID signal decreases over time (Fig.2.7.a).

In order to determine the nuclei frequencies of each molecules examined by the NMR, a Fourier Transform (FT) is then performed on the recorded FID signal (Fig.2.7.b). As a matter of fact, each metabolite is presented by a set of peaks with specific characteristics (peak frequency and amplitude) (Fig.2.8).

The Magic angle spinning MAS NMR technique

In traditional NMR techniques, the spectrum resolution is poor (*i.e*; peaks with large

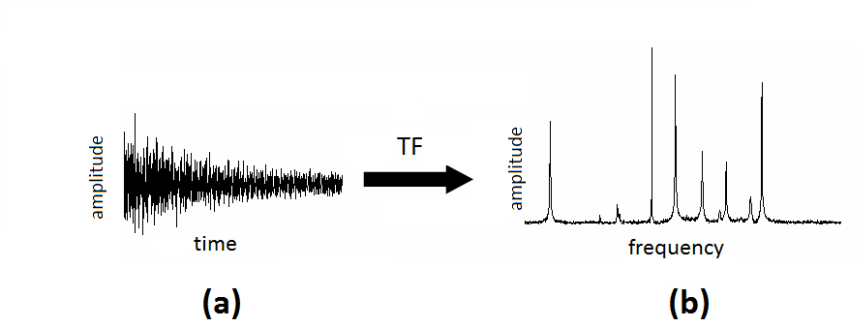


Figure 2.7: (a) The NMR signal detected and (b) the spectrum obtained after Fourier transform.

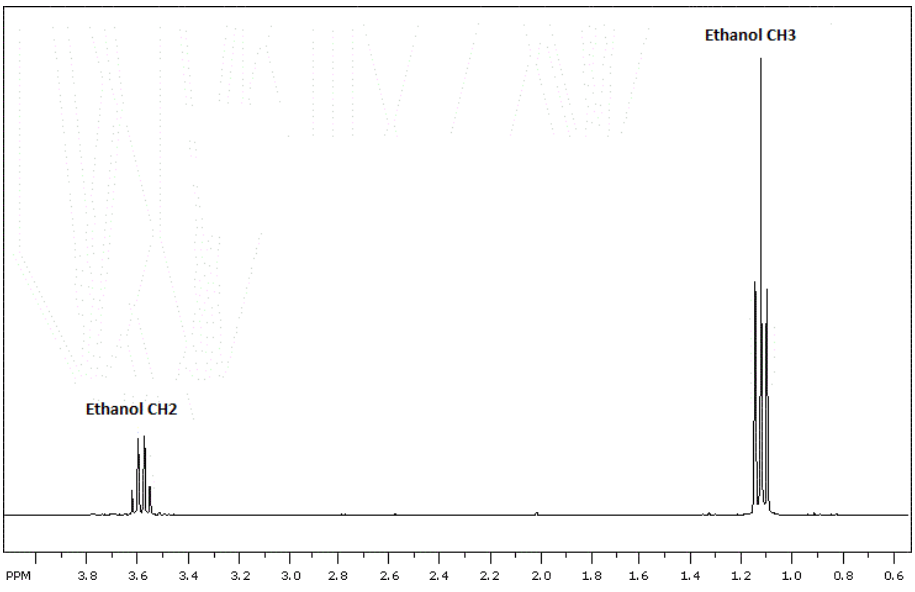


Figure 2.8: The ethanol peaks.

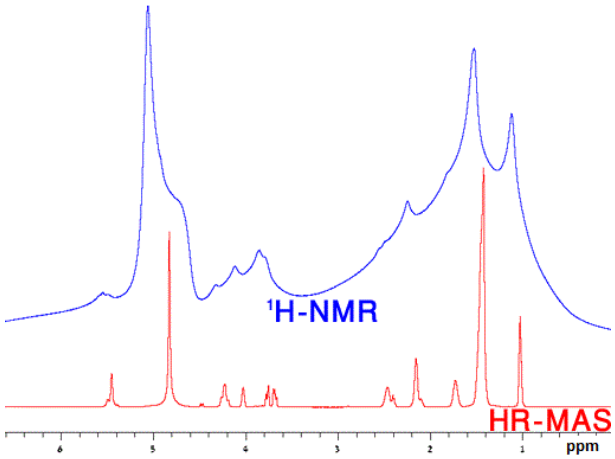


Figure 2.9: An example of a 1D spectrum with the NMR technique in blue and the HR-MAS technique in red.

widths) for semisolid or solid samples due to residual dipole interactions, chemical shift anisotropy and susceptibility within the tissue sample. These factors can be minimized by low-speed Magic Angle Spinning MAS [Griffin04]. In MAS, samples rotate rapidly at 54.7° to a magnetic field (angle between the rotation axis and the NMR magnetic field). Indeed, only at the magic angle, the nuclear dipole-dipole interaction between nuclei moments averages to zero. The use of MAS improves the quality of NMR spectra by eliminating broad peaks and obtaining enough information for easier molecule identification (Fig.2.9)

The 2D NMR experiment

Although the 1D HR-MAS NMR spectroscopy is more or less successfully applied to identify the structure of solid or liquid compounds, this technique suffers from several shortcomings. Indeed, 1D NMR spectra of complex biological samples typically have high spectral overlap, which significantly limits the number of metabolites that can be uniquely identified and quantified. To overcome this drawback, the two dimensional 2D NMR HR-MAS should be recommended. This technique offers more detailed and unequivocal assignments of biologically metabolites in intact tissue samples and enables accurate identification of a large number of metabolites that are not resolvable in a 1D NMR spectroscopy. More precisely, 2D NMR offers two distinct advantages:

1. It reduces the overcrowding of resonance lines. Indeed, as the spectral information is spread out in two frequencies (better than a single frequency) the 2D NMR spectrum technique can reduce spectral overlap and allows the identification of some molecules that remains unresolvable in 1D NMR spectra. In other words, a 2D spectrum peak is no longer characterized by one frequency but by two frequencies allowing an easier discrimination.

2. It offers the ability to correlate pairs of resonance such as the proton ^1H and the carbon ^{13}C. Indeed, since almost metabolites contain carbon and proton, it is interesting to determinate which protons are connected to which carbons (*i.e*; which proton ^1H is correlated with which carbon ^{13}C).

For a typical 2D spectrum, we can distinguish four time intervals or periods, as shown in Fig.2.10 [Schmidt-Rohr94]: τ_p the preparation period, t_1 the evolution period, τ_M the mixing period and t_2 the detection period. In the first period, the magnitude environment is prepared. During the t_1 period, the magnetization evolves freely, so that M_{xy} precesses at its Larmor frequency, and each nuclear magnetization is featured according to its Larmor frequency. During the period τ_M, another RF pulse signal or pulse signal sequence is injected into the system to enable the mixing of the nuclear magnetizations used to produce *xy*. Finally, t_2 is the usual data acquisition period in which an FID is acquired as in 1D experiment.

This procedure is then repeated many times, with different durations of the evolution period t_1 and keeping all other settings constant. For each value of t_1 period, the signal that is acquired during t_2 is stored. Once the experiments are achieved, we obtain a 2D time domain signal $s(t_1,t_2)$ (FID 2D). In order to obtain the 2D spectrum, a 2D Fourier Transform is performed on the obtained 2D FID signal. Among various spectrum types, the most frequently applied is the Heteronuclear Single Quantum Coherence better known by its acronym, HSQC [Berger04]. In a HSQC experiment, the chemical shift range of the proton ^1H spectrum is plotted on one axis, while the chemical shift range of the ^{13}C spectrum for the same sample is plotted on the second axis. Indeed, since almost metabolites contain carbon and proton, the addition of a second dimension (^{13}C or ^1H) improves the frequency discrimination and enables the identification of a large number of metabolites that are not resolvable in a standard 1D ^1H or 1D ^{13}C NMR spectrum. Therefore, the 2D HSQC spectrum has become arguably one of the widely used technique to elucidate the relationships between clinically relevant cell processes and specific metabolites in order to identify diseases such as the multiple sclerosis disease [Tiberio06] and tumor identification [Piotto09]. Along these lines, the relevant information characterizing the spectra is the **metabolite peaks** which could be considered for biomarkers identification and analyzing.

When the data are subjected to a Fourier transform, the resulting spectrum plot shows the chemical shift of ^1H plotted along *x*-axis and the chemical shift of ^{13}C plotted along the *y*-axis. Fig.2.11 shows an example of a HSQC spectra of colon biopsy.

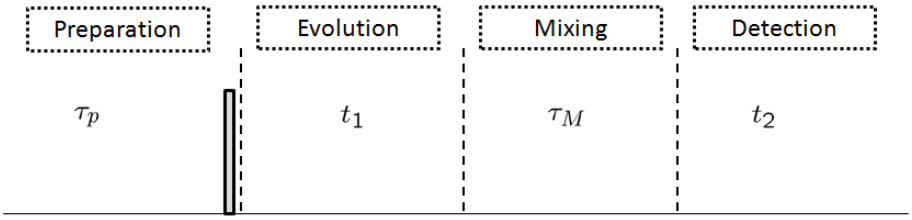


Figure 2.10: The schematic representation of a basic 2D NMR experiment in term of four periods: τ_p evolution, t_1 mixing, τ_M mixing and t_2 detection. For a given experiment, τ_p and τ_M are usually fixed periods while t_1 and t_2 are variable time periods.

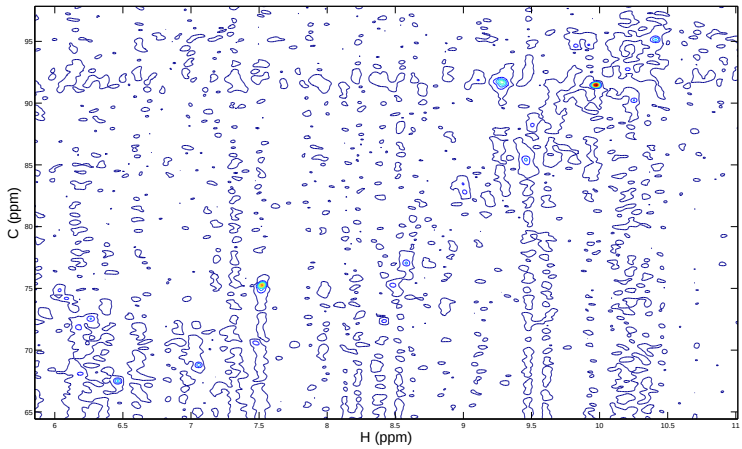
2.2 Functional Magnetic Resonance Imaging

Since twenty years, functional brain imaging techniques allow an in vivo analysis of neural and hemodynamic phenomena associated with brain activity (activation imaging). The motor, sensory, or cognitive functions can be assigned to one or several anatomical areas of the cortical that are activated as networks [Aster05].

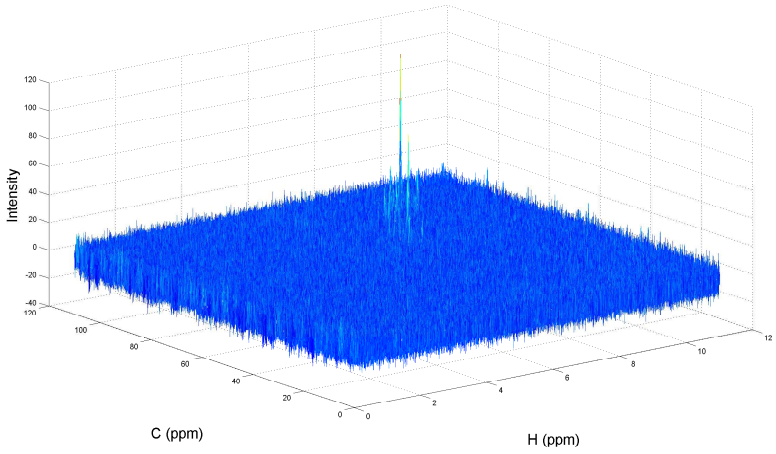
The general basis of the human brain functional organization have been further refined through the Positron Emission Tomography (PET) technique. This technique measures changes in brain perfusion during a cognitive task thanks to the use of water labeled with the oxygen 15 as a marker [Kaplan99]. However, the significant technical limitations as well as the high cost of this technique explain the fact that the activation by PET imaging is currently restricted to research centers. Moreover, the danger of radiation limits its use on children.

The development of MRI in the early 1980s has revolutionized the study of neuroanatomical human brain thanks to its relatively easy use and particularly the lack of constraints related to irradiation. The MRI imaging has been very fertile in fundamental works on brain perfusion study at rest, using exogenous contrast agents [Villringer88], intrinsic contrast agents such as blood velocity (MRI angiography) [Rosen90], or diffusion phenomena [Le Bihan91]. However, the application of these techniques has been rapidly supplanted by the discovery of the BOLD (Blood Oxygenation Level Dependent) contrast associated with deoxyhemoglobin whose ease of implementation and high accuracy have promoted the success of the fMRI technique [Engel97, Vazquez98, Rombouts09].

Indeed, fMRI is an indirect imaging of neuronal activity through the detection of local perfusion changes. In early 1890, Roy and Sherrington had suggested the existence of a spatial relationship between neuronal activity and brain perfusion [Roy90]. Neuronal activity causes a small increase in local cerebral metabolism and hence an oxygen consumption as well. Very rapidly (within hundreds of milliseconds), this phenomenon is followed by an increase in local brain perfusion and hence an increase in the intake of oxyhemoglobin (ar-



(*a*)



(*b*)

Figure 2.11: An example of a HSQC spectrum displayed as (a) a contours plot (b) as 3D plot.

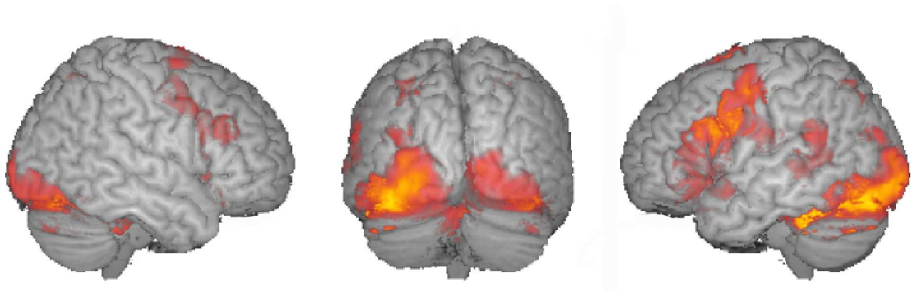


Figure 2.12: An example of fMRI acquisitions. The active zones are presented with the red color.

terial blood) which exceeds oxygen consumption [Fox86]. It is possible to measure blood perfusion with MRI image, nevertheless, it has been shown that the use of BOLD contrast leads to a more accurate measurement [Lee95].

The difference in magnetic susceptibility between vessels filled red blood cells responsible for interstitial deoxyhemoglobin and diamagnetic induces a local gradient magnetic field that extends beyond the vascular wall. In this perivascular gradient, whose size depends on the diameter of the vessel and the concentration of deoxyhemoglobin, the spins of the protons undergo interstitial diffusion leading to a reduction in the value of T2 weighted image (T2 weighted image is a type of MRI image [Mintorovitch91]). Ergo, blood vessels have a fairly dark color that is easily detected in the T2 weighted image. As a result, the relevant information that typifies the fMRI images is the **the active zones** which could be used to identify and analyze biomarkers. Fig. 2.12 shows an example of a fMRI acquisitions. The active zones are presented in red color. Fig. 2.13 display the superposition of the activation map (the different activation zones) on the anatomic brain image.

Furthermore, the development of the BOLD technique has facilitated the use the fMRI in medical researches. For example, many studies have attempted to show that fMRI has the ability to identify in primary visual [Ganis04], somato-motor [Nakata08] and even auditory brain areas [Downar01] by the use of simple stimuli (flashing lights, simple movements of the hand). These initial results, the arrival of the echo-planar sequences and the relative ease of access to technology have contributed to the broad development of fMRI. This technique is now suitable for studying specific neurophysiological issues of patients and some clinical applications have been developed. Indeed, clinical applications of fMRI are booming. The two most advanced applications are: the presurgical sensorimotor functional mapping [Zhang09] and the study of hemispheric dominance of language [Binder97]. Other researchs explored more diseases such as epilepsy disease, the pathophysiology of various neurological disorders (MS [Rocca02, Reddy02], pathology of the basal ganglia [Ferrandez03], ...), psychiatric (schizophrenia) [Surguladze10], or cortical plasticity after injury or after surgery.

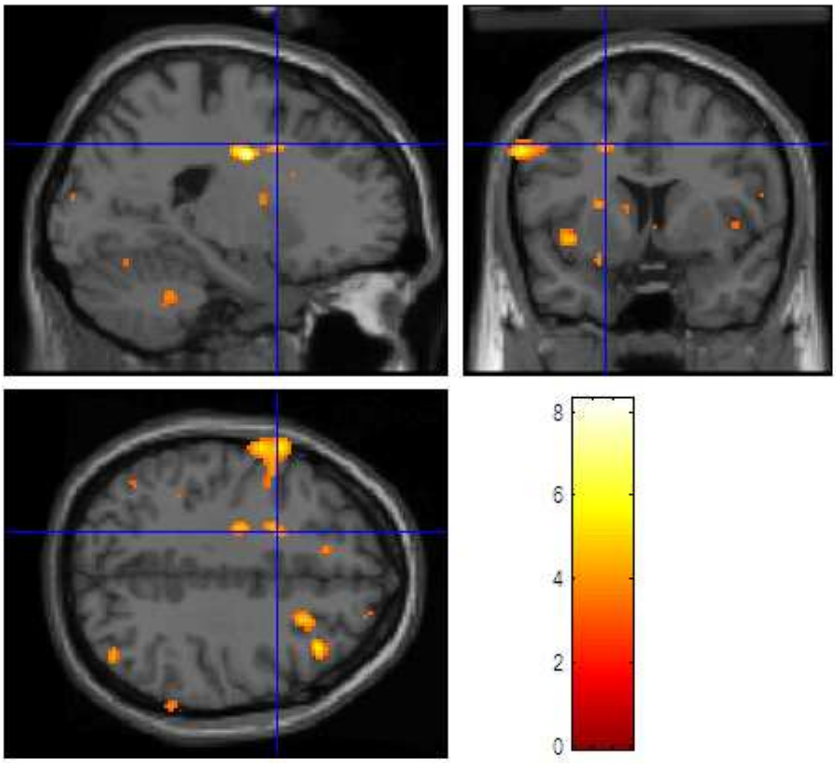


Figure 2.13: Superposition of the activation map on the anatomic brain image. The active zones are presented with the red color.

2.2.1 fMRI experiment

The investigators hypothesize that individual responsiveness to specific antidepressant medications can be predicted based on patterns of brain network activity as visualized by functional connectivity analysis of fMRI data. Indeed, brain networks consist of groups of nerve cells from different brain regions that communicate with one another. Using fMRI, researchers can monitor which groups of nerve cells are active at any given moment. Functional connectivity studies take this analysis one step further. By comparing levels of activity among different groups of brain cells, researchers can determine which areas are communicating with one another. During an fMRI session, patients undergoes an activation paradigm. The activation paradigm is the sequence of stimuli that are used to highlight the desired brain activity given the context. As a matter of fact, two major contexts can be schematically distinguished:

1. The search context: in this context we want to highlight, in a homogeneous group of subjects, the common neurophysiological or neuropsychological behavior in response to a certain type of stimulations to understand better the operation of any normal or pathological cortical network. Therefore, elaborated paradigms must be perfectly reproducible from one patient to another. Nonresponder subject may be excluded or analyzed separately. The data analysis could be done subject by subject, by applying the same criteria of analysis, allowing the study of inter-individual variation.

2. The clinical context: in this context the problem is quite different. Indeed, it consists in analyzing for a particular patient behavior in response to a particular stimuli, depending on the pathology, diagnostic purposes and/or pretreatment. These studies can only be performed individually and a false negative results may have adverse consequences for the patient. The paradigm, which has previously been tested and calibrated in healthy subjects and possibly similar patients, should be relatively simple, robust and easy to perform in hospitals. Given the variability of aptitude among patients, the compliance and/or performance of the subject must be taken into account in the data analysis step.

However, recent functional connectivity analysis have shown that certain brain networks, called resting state networks, are especially active when the brain is at rest (*i.e*; no activation paradigm is performed) [De Luca06, Scholvinck10]. These networks are particularly used to analyze alzheimer [Rombouts05] and schizophrenia diseases [Liang06]. Note that in this thesis, we particularly focus on Nuto study network healthy patients.

2.3 Statement of the indexing problem

In the previous sections, two medical information acquisition techniques were presented: the 2D HSQC HR-MAS spectroscopy and the fMRI imaging. Moreover, we described the relevant information that typifies each signals: the *peaks* for the 2D HSQC spectra and the *active zones* for the fMRI images. Thus, each medical information (HSQC spectrum

or fMRI image) consists of a set of relevant information or **objects** that could be used to identify new biomarkers particularly for cancer and neurological diseases. To this end, it is important to improve the tools of analysis associated with these techniques. In the same vein, we propose a new content-based object indexing and retrieval scheme for biomarkers detection.

The classical content-based information indexing and retrieval scheme

A classical indexing and information retrieval scheme operates as follows: a user submits a query and the system identifies the relevant information to the submitted query and then returned it to the user. The most ancient and widely used indexing and document (*e.g.*, medical images, respiratory sounds, etc) retrieval scheme is the indexing by keywords where the document is described by a set of keywords (*e.g*; a word, a phrase, or an alphanumerical term). However, this technique remains limited by the low expressive power of words, by the language constraints (the transition from one language to another, semantic ambiguity) and the subjective nature of the annotations (two physicians can differently annotate a medical image). Moreover, the indexing technique by keywords requires human intervention which is a binding task particularly on large databases if the keywords are generated manually. In addition, the keyword annotation can never exhaustively describe the contents of a document (*e.g*, a tumor in a MRI image : location, type, ... etc).
In order to overcome these drawbacks, the content-based information indexing and retrieval approach (henceforth called indexing approach) was proposed [Eakins96]. Indeed, content-based means that the search will analyze the actual contents of the information rather than the meta-data such as keywords, tags, and/or descriptions associated with the information. For example, the term 'content' might refer to colors, textures, or any other information that can be derived from an image itself. Therefore, the content-based information indexing and retrieval approach allows us not only to automatically index the documents and query a database directly from their information content without human intervention but also to objectively analyze the database content. For example, if we consider a cerebral tumor as a query, we would be able to easily identify with an objective similarity measurement function the similar tumors belonged to the requested database.
Basically, the indexing scheme requires:

1. Document alignment step,

2. Document codifying and similarity measurement steps. The first one aims at codifying different documents into a compact description whereas the second one consists in establishing the object similarity measurement procedure.

In order to accelerate the large database queries, the indexing scheme can be divided into two phases (Fig. 2.14):

1. An off-line phase in which the content database alignment and coding is performed. Indeed, the off-line coding consists in extracting the signatures associated with the contents of the database. The latter are then stored in a reverse dictionary (file name

and signature), which allows to quickly find the document associated with a given signature.

2. An online phase in which the user queries the database using a document request. The online alignment and coding steps only concern the document request. A similarity measurement between the request document signature and those calculated in the reverse dictionary is then performed. Finally, the documents belonging to the database are classified by order of similarity.

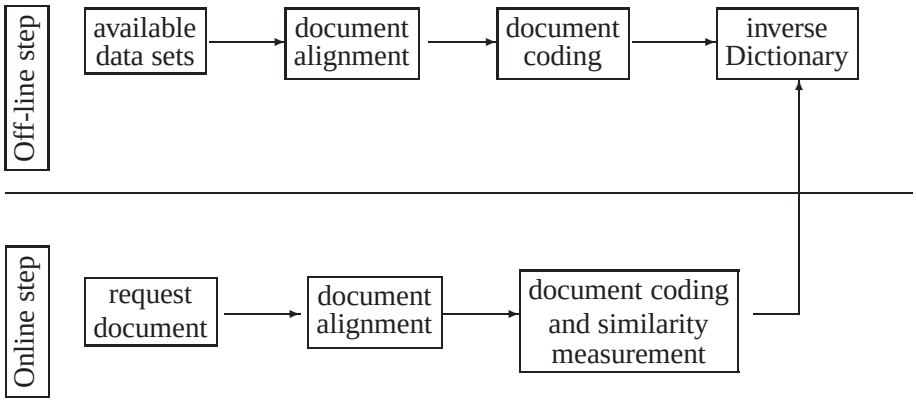


Figure 2.14: Overview diagram of the classical indexing scheme.

Although this indexing scheme was successfully applied on web databases document retrieval [Chaouch09], it is not suited to the biomarker identification task. Indeed, the later requires the classification of medical signal profiles (groups) for change detection. For example, if we consider two profile classes: the healthy class and the pathologic class, the biomarker identification consists in classifying a group of medical signals into the healthy and the pathologic classes (*e.g* cancer or psychological diseases) and to detect then the differences (changes) between them. As a matter of fact, healthy class can be considered as the "unchanged" class whereas the pathologic class as "changed" class. Consequently, adding a classification step to the classical indexing scheme would allow us to detect biomarkers from both HSQC spectra and fMRI images.

The proposed content-based object indexing and retrieval scheme

Unlike the classical indexing scheme (Fig. 2.14), the proposed indexing scheme contains two additional steps: an object detection step (detection of HSQC peaks and active zones fMRI images) and an object classification (change detection) step. The new indexing scheme is likewise divided into two phases (Fig. 2.15):

1. An off-line phase in which we perform on each medical signal: an object detection and alignment step, an object coding and similarity measurement step and finally an object classification step (*e.g*, healthy or pathological profile).

2. An online phase in which the user queries the database using a request (new individual/ group of medical signals). The same steps as in the off-line phase are applied on the medical signal request. Finally, the later is assigned to a previously defined profile (in the off line step). Note that unlike the classical indexing scheme, the object similarity step aims here at clustering the similar objects belonging to a given medical signal group allowing then the assignment of this group to the appropriate profile. In other word, the assignment of a new group/individual task is addressed using the classification step and not the similarity measurement step as in the classical indexing scheme.

Fig. 2.15 shows the overview diagram of the proposed indexing scheme.

Conclusion

This chapter presented a brief recall of the NMR principles and how 1D/2D spectra are constructed from acquired signals. It also detailed the main basis of fMRI imaging techniques and its medical applications today and in the future. We particularly described the relevant information that typifies each signals: the *peaks* for the 2D HSQC spectra and the *active zones* for the fMRI images. Finally a first contribution of this work which consists in adding of a classification step to the classical indexing scheme for biomarkers identification was presented. In the next chapters, different steps of the indexing scheme (object detection and alignment Chap.3, object coding and similarity measurement Chap.4 and object classification Chap.5) are described.

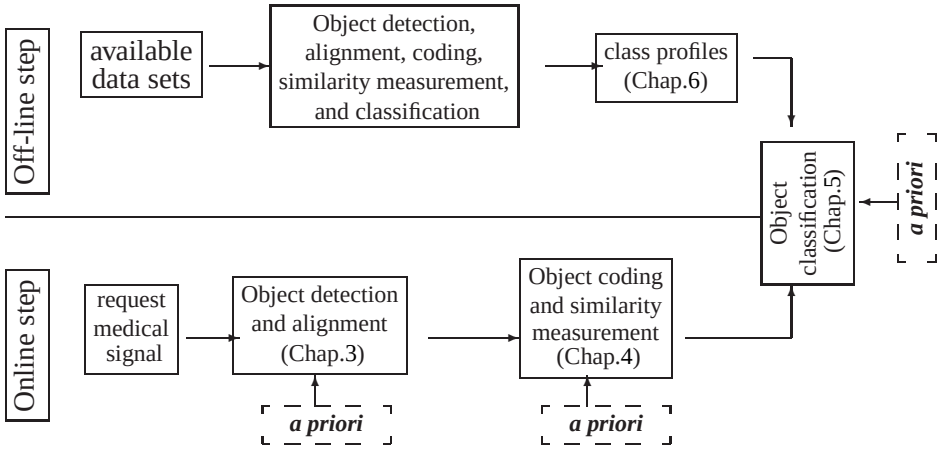


Figure 2.15: Overview diagram of the proposed classification framework

CHAPTER 3

Object detection and alignment

Contents

Symbols:

$T_g()$	Geometric transformation
$T_p()$	Photometric transformation
X	Theoretical spectrum
X_{ref}	Reference theoretical spectrum
Y	Observed spectrum
Y_{ref}	Reference observed spectrum
B	Noise
Γ_b	Noise covariance matrix
$h(i,j)$	Peak shape filter
$\gamma^Y = (\gamma_1^Y, \gamma_2^Y)$	Shape hyperparameters of peaks within the spectrum Y
$hyp_{1,2,3}$	Hypotheses modeling the spectrum imprecision
$\hat{X}$	An estimation of the theoretical spectrum X
f_{hyp_1}	S-membership function associated to hyp_1
f_{hyp_2}	S-membership function associated to hyp_2
f_{hyp_3}	S-membership function associated to hyp_3
H_1	Peak absence hypothesis
H_2	Peak presence hypothesis
$\mu_{i,j}^{[hyp_1]}$	Imprecision degree of the pixel located at (i,j) associated to hyp_1
$\mu_{i,j}^{[hyp_2]}$	Imprecision degree of the pixel located at (i,j) associated to hyp_2
$\mu_{i,j}^{[hyp_3]}$	Imprecision degree of the pixel located at (i,j) associated to hyp_3
$\mu_{i,j}$	Combined imprecision degree of the pixel located at (i,j)
$m_{i,j}(Y)$	Mass function associated to the pixel located at (i,j) in Y
$\mathfrak{m}_{i,j}$	Combined mass function associated to the pixel located at (i,j)
$Bel_{i,j}$	Belief function associated to the pixel located at (i,j)
$d_{i,j}$	Local displacement vector associated to the pixel located at (i,j)
$\hat{d}_{i,j}$	Estimation of $d_{i,j}$
d	Global peak displacement vector
$\hat{d}$	Estimation of d
V	Peak neighborhood
Ω	frame of discernment
$(\hat{X}_c, \hat{X}_h)$	Estimate coordinates of X in ppm
$(\hat{X}_{rc}, \hat{X}_{rh})$	Estimate coordinates of X_{ref} in ppm
ε_c	Carbon mean error chemical shift in ppm
ε_h	Hydrogen mean error chemical shift in ppm
$\mathrm{E}[X]$	Expected value of X

Symbols:

P	Observed object coordinates
P'	Added samples coordinates
$\hat{P}$	Estimation of P
Z	Projection of P with Neural Network PCA
$z_i^{[l]}$	Weight of the i^{th} node belonging to the l^{th} layer
W_l	Network weight matrix connecting l^{th} layer with $(l+1)^{th}$ layer
w_{ij}	Connection weight from node i to node j
$\Phi_{extr}(.)$	Neural Network extraction function
$\Phi_{gen}(.)$	reconstruction function
(r, θ, ϕ)	Spherical coordinate system

Acronyms:

HR-MAS	High Resolution Magic Angle Spinning
EEG	Electro-Encephalo-Graphy
HSQC	Heteronuclear Single Quantum Coherence spectrum
fMRI	functional Magnetic Resonance Imaging
PCA	Principal Component Analysis
PC	Principal Components
MI	Mutuel Information
DS	Dempster Shafer
SSD	Sum of Squared Difference
PRST	Planar-Reflective Symmetry Transform
PSF	Point Spread Function
PSNR	Peak Signal to Noise Ration
MCMC	Monte Carlo Markov Chain
MLP	Multi-Layer Perceptron
NNPCA	Neural Network Principal Component Analysis
ANNs	Artificial Neural Networks
GA	Genetic Algorithm

Introduction

The problem of object detection consists in dividing signal into meaningful groups (objects) based on the spatial arrangement and/or pixel intensity. The object alignment task is the process of overlaying two or more objects taken at different times, and/or from different viewpoints, and/or by different modalities and intrinsic variabilities. More precisely, object alignment consists in geometrically aligning an object with reference pattern. The object detection and alignment task is a crucial step in the indexing scheme since all other steps depend on it. Consequently, in order to lead to an optimal object detection and alignment results, all *a priori* knowledge that we have on the data need to be properly integrated in the proposed object detection and alignment methods. That is the way we take in this thesis where we pay a particular attention to include *a priori* knowledge for two types of data (2D HSQC spectra and fMRI images).

In this chapter we present section 3.1) a brief overview of the object detection and alignment methods used in the indexing schemes. In section 3.2, we present the scheme for peak detection and alignment based on the use of evidence theory. We particularly show that the proposed evidential scheme for peak detection and alignment consistently achieves a higher performance compared to the existing schemes on both synthetic and real spectra.

Regarding the fMRI images, we present (section 3.3) the proposed active zone detection and alignment algorithm. The detection step is addressed using a Markov chain segmentation algorithm whereas the alignment step relies on the use of non-linear Principal Component Analysis (PCA) algorithm, which would be well suited to fit the cortex shape, to estimate the non-linear planes of symmetry. We notably show that the use of the non-linear PCA allows us to get more accurate object-alignment results compared to the classical PCA alignment algorithm.

3.1 Overview of object detection and alignment methods

3.1.1 Object detection

In the literature, we can distinguish two approaches to address the problem of object detection: the signal segmentation based approach and feature based approach.

- **The segmentation approach**: it aims at partitioning a signal into distinct meaningful entities sharing together the same properties, by defining boundaries between different objects. Several clustering techniques have been proposed in the literature [Pappas92, Tao07, Zhang08]. Basically, clustering algorithms don't use the training data. So, to compensate this deficiency, these algorithms seek to alternate between signal segmentation and cluster property characterization. In other words, clustering methods try to get trained by the utilization of the available data. Three commonly used clustering algorithms are the K-means [MacQueen67], the Markov model [Fjortoft03] and the active contour [Kass88]. The latter approximates the ob-

ject shape as a flexible 2D curve or a 3D surface that can be deformed to better fit the object shape. The notion of active contour was firstly proposed in Kass [Kass88]. The underlying idea is to create a link between a low-level process that only uses the available signal of intensities and a high-level process that integrates mechanisms for semantic interpretations.

- **The features based approach**: it aims at providing a sparse local representation of data. Its goal is to describe regions by keeping distinctive information and, at the same time, to provide robustness to small transformations and noise. The region descriptors offer an elegant solution to deal with occlusion and cluttered background since they only store salient information of the region and therefore they are not distracted by other parts of regions. The most used feature detection algorithms are the Harris point [Harris88], the Hessian detector [Beaudet78] and the scale-invariant point of interest (*e.g*, the Laplacian-of-Gaussian (*LoG*) [Lindeberg98]).

In this work, we adopt the feature based approach for peak detection (section 3.2) and the segmentation approach for the active zone detection (section 3.3).

3.1.2 Object alignment

Cost function minimization based approach

In this section, we focus on the alignment methods based on the optimization of an error function. This error function is constructed from a reference object and another object we need to align. The most used transformation models are:

1. The geometric distortion: it is parameterized by a geometric transformation $T_g()$ which models the viewpoint changes of the camera, the displacement of the object with respect to the reference object. This transformation acts either on the spatial coordinates of the objects [Berg05, Schneiderman98] or on the object shape using the deformable model [McInerney96, Leymarie93],

2. The photometric transformation: it is parameterized by a transformation $T_p()$ which models changes in brightness or noise measure modifying the pixel intensities of the object [Seo05].

Cost function minimization based approach aims at estimating the parameters of the transformations $T_g()$ and $T_p()$ by the definition of an optimization criterion (cost function). As for transformation models, the latter is chosen according to the context of the application. For example, we can opt to be robust to photometric transformations without estimating them (*e.g.*; shape based alignment methods), or to establish a noise model in order to take it into account in the alignment scheme.
To estimate the hyperparameters of $T_g()$ and $T_p()$, it is necessary to develop a strategy for solving the established cost function. Unfortunately, this task is generally complex and often cannot solve it exactly.

Cost function definition

Cost function establishment generally consists in defining a criterion for direct comparison of objects. This criterion can be based on light intensities [Belongie02], mutual information [Viola95], etc, and its estimation leads to an optimization task. The existing methods often consider iterative approaches because of the nonlinear nature of the cost function.
Among the existing cost functions, the most commonly used is the squared error SSD (Sum of Squared Difference) proposed in [Lucas81], which is optimal in the sense of maximum likelihood under the assumption that the measurements are corrupted by a centered white Gaussian noise. Several variations of this SSD function have been proposed in the literature in order to increase the alignment accuracy results [Keller04] [Zimmermann09] or to reduce computing time [Hager98, Baker04].
Furthermore, in the case of significant geometric distortions, some studies propose the kernel-based cost function allowing to be more robust in situations where the geometric distortion of the observed scene is not fully addressed by the basic alignment algorithm [Comaniciu03, Hager04].
Finally, in order to address the alignment task with a Bayesian framework, methods based on the maximization of mutual information between images [Dowson08, Dame09] have been proposed. Indeed, mutual information measures the statistical dependence between the intensities of compared object. This dependency is assumed to be maximized when both objects are aligned. Nevertheless, it may turn out that mutual information is not adapted to object with thin structures [Roche01].

Cost function optimization

Once the cost function determined, the optimization step consists in minimizing the cost by minimizing model hyperparameters in a supervised or an unsupervised way. The former mainly aims at approximating the cost function in order to lead to a linear system since these cost functions are generally non-linear. Generic iterative optimization methods such as gradient descent (used in [Amberg09]), or the Newton algorithm [Shum02, Xiao08] are the most commonly used ones. These methods consist in estimating the extrema of the cost function (*e.g.*; points where the gradients of the cost function are null).
The second approach relies on a learning step. Indeed, the supervised methods require the modeling of most expected transformations between two patterns. Once these transformations are modeled, one uses a reference pattern to generate a series of reference errors which can be obtained by calculating differences between the reference pattern and the transformed one [La Cascia00]. These errors allow us to establish the link between a given observed error and the corresponding transformation [Jurie02, Bayro-Corrochano07]. Although these methods are not time-consuming, they require a high prior knowledge that is not always available.

Canonical system based approach

The canonical system based approach consists in estimating the canonical pose of a given object (*e.g.*;. estimate the axes of the cartesian coordinate system associated to the pattern). Thus, it is possible, if necessary, to define the sign or the direction of these axes in order to solve the problems of reflections. Methods based on this approach do not operate directly on object intensities but on object shapes (2D curves or 3D surfaces). In this section we present a brief overview of methods based on the canonical system and particularly those applied on the 3D objects. Generally, a 3D object can be presented in different manners. We can distinguish:

- Surface representation (3D mesh representation): the object is represented by its border. In the case of a polyhedral object, the border is composed of a set of planar polygonal facets. In the literature, the triangular mesh is the most popular form of polyhedral surfaces. Therefore, the object surface is composed of a set of interconnected triangles where each triangle consists of three vertices and a gravity center. The 3D triangular mesh presentation enables a compact encoding of object and a suitable object display according to the desired resolution [Kos01].

- Volumetric representation (voxel representation): the object is represented by a union of disjoint elementary unit volumes called voxels. Unlike the surface representation, it is particularly useful for representing data density (point cloud). The fitness of this representation depends on the number of voxels [Gibson97].

- Algebraic representation: the object is described by an equation (*e.g.*; $f(x,y,z) = 0$). The algebraic presentation enables a sparse object representation. However, such equation is not always available particularly for object with complex shape [Kang01].

The two widely used methods to estimate the object canonical pose (the axes of the cartesian coordinate system associated to the pattern) from these possible definitions of the object, are:

Principal component analysis PCA

The Principal Component Analysis, also known as "PCA" is commonly used in data analysis to find the principal axes of an object [Jolliffe86]. For 3D objects, it is used to calculate the three coordinate axes necessary for the 3D model. These three axes constitute then the new cartesian coordinate system associated to the pattern. Several variants of basic PCA method have emerged to address the problem of 3D object alignment. Indeed, although the basic PCA is not time-consuming and robust in the case of the object volumetric representation, it is not adapted to the object mesh representation [Vranić01a] (the object is presented with a set of connected triangular forming its continuous surface). Thus, improvements have been proposed to overcome these problems. On the one hand, Paquet et al. [Paquet00] propose to weight the triangle facet gravity centers by their surface. On the other hand, Vranic and Saupe [Vranić01a, Vranic01b] extend the work of Paquet et al and propose the PCA in the continuous case, noted "continuous PCA", and thus allow a

better robustness in the case of a 3D mesh.

To estimate the alignment of a 3D model it is necessary to calculate a PCA on all centers of the triangle facets of an object. To achieve this, the objects must be composed of polygons whose vertices have always coplanar three points. The PCA aims at finding a basis where the projection of an object is symmetrically invariant.

In order to determine the orientation of an object, with respect to its faces, the continuous PCA estimates the three main axes of the 3D object. To this end, it calculates the covariance matrix on the facets vertices of the 3D mesh. The idea is to find the axis which maximizes the variance of the point cloud. The maximum variance is then obtained for the eigenvector associated with the largest eigenvalue. Similarly, the vector that contains the largest remaining inertia is the second eigenvector while the third expresses the residual inertia. To conclude, to define the 3D object orientation, it suffices to diagonalize the covariance matrix of normalized faces. The eigenvectors matrix stands for then the rotation matrix defining the 3D model alignment.

Orientation by the axes of symmetry

To orient a 3D object, Podolak et al. [Podolak06] propose to calculate the symmetry planes of the model. For this, they define the notion of symmetry for the plane intersecting the object through the calculation of the "Planar-Reflective Symmetry Transform" (PRST). The PRST aims at associating to each plane a scalar value measuring its symmetry. The more this scalar value is great, the more the associated plane is symmetry. Then, they choose as the first axis, the normal of the plane that has the maximal symmetry. The second is the plane of the maximal symmetry orthogonal to the plane previously selected. Finally, the last axis is similarly obtained by finding the maximal symmetry plane that is perpendicular to the normal of the two selected planes.

3.1.3 Retained approaches

In the previous sections, we described two strategies to address the object detection task: the segmentation approach and the features based approach. We presented likewise two approaches to address the object alignment task: the cost function minimization based approach and the canonical system based approach.

A spectrum is composed of several peaks which are the responses of metabolite presence. Each peak can be characterized by its locations (chemical shifts), its amplitude and its shape. These peaks are scattered within the spectrum and hence they can not be presented by a curve. Therefore, the canonical system based approach is unappropriate in the peak case and for this reason we have opted for the cost function minimization based approach to address the peak alignment task. Note that the peak detection step is simultaneously performed with the alignment step.

Concerning the active zone of the fMRI images, it can be characterized by its location

and its shape. Unlike peaks, an active zone is a dense form (consisting of neighboring voxels) and hence may be presented by a 3D surface. Therefore, the segmentation approach is adopted for active zone detection. In order to be robust to photometric transformation (a delicate task for the cost function minimization based approach), we have chosen the canonical system based approach for the fMRI active zone alignment. Note that the active zone detection step is separately addressed from the alignment step.

3.2 Peak detection and alignment algorithm (HSQC)

In this section, we propose a new method able to simultaneously detect and align different peaks. The peak characteristics theoretically invariable for the same metabolite between samples are in practice corrupted by a noise: a location imprecision is added to the spectra in practical cases. We will model this imprecision and the *uncertainty* always present on the observed HR-MAS 2D data so as to obtain an optimal peak alignment results.

The notions of uncertainty and imprecision are distinct and they must be now clearly defined [Shafer76]. On the one hand, the uncertainty presents the belief or the doubt we have on the existence or on the validity of the data [Gilks96] (presence or absence of a peak in the case of HSQC spectra). In the other hand, when we have not enough knowledge on the data, we describe it with vague terms but its realization is certain: in this case we speak about *imprecision* (a modification of the peak shape and location in the case of HSQC spectra). In order to take into account both imprecision and uncertainty of the spectra, we propose the use of the evidence theory [Shafer76] which can be well suited to deal with raw data through the definition of a mass function. This mass function allows us to quantify the reliability of a given hypothesis. An overview of Evidence theory and its application on real cases is presented in (Appendix C).

Moreover, the evidential peak alignment scheme proposed in this thesis is based on the fuzzy set theory [Bezdek99] to model and quantify the imprecision degree presented in the spectra. In particular, we show that this modeling, used in the mass function definition, increases the performance of the alignment scheme with comparison to the Bayesian scheme.

3.2.1 Evidential peak detection and alignment method

Spectra modeling

In this work, the spectrum is considered as a random field. To model 2D HSQC spectrum formation, we consider a 2D spectrum realization Y such that $Y = y(i,j)_{i=1...M,\ j=1...N}$ where $(M \times N)$ is the spectrum sizes. It corresponds to the observation of a theoretical 2D spectrum image realization X such that $X = x(i,j)_{i=1...M,\ j=1...N}$, considered as a random field as well, through a nuclear magnetic resonance system. In our case, X consists of the various peaks corresponding to the metabolites present in the biopsy. If the nuclear magnetic resonance system was linear and shift-invariant, the relation between $y(i,j)$ and $x(i,j)$

on the same location should be expressed as a convolution product [Dobrosotskaya08]:

$$y(i,j) = \sum_{k_1,k_2} x(k_1,k_2)\, h(i-k_1, j-k_2) + b(i,j). \tag{3.1}$$

where h is the Point Spread Function (PSF) of the nuclear magnetic resonance system, and $B = b(i,j)_{i=1...M,\ j=1...N}$ is a realization of a random field corresponding to an additive noise modeling both acquisition noise and degradation of the biopsy tissues.

In the case of the 2D spectra, h is assumed to be a Lorentzian filter [Lowry08] whose continuous expression is parameterized by $\gamma^Y = (\gamma_1^Y, \gamma_2^Y)$:

$$h(k_1,k_2;\gamma_1^Y,\gamma_2^Y) = \frac{1/\gamma_1^Y}{((1/\gamma_1^Y)^2 + k_1^2)} \frac{1/\gamma_2^Y}{((1/\gamma_2^Y)^2 + k_2^2)} \tag{3.2}$$

Note that in some studies (*e.g*; [Schanda05, Feliz06]), the peak shape is assumed to be gaussian.

Imprecision quantification

Before step estimation, we have to define three assumptions hyp_1, hyp_2 and hyp_3 in order to model and manage conflicts.

Assumption hyp_1

Let H_1 be the hypothesis corresponding to the absence of a peak located at (i,j), and H_2 the hypothesis of presence of a peak (detection) at the same position. We are interested with the *a posteriori* probabilities of the hypotheses $H_k, k \in \{1,2\}$ of the observation Y to quantify the contradiction degree. The estimation of these probabilities $p_{i,j}(H_k/Y)$ at every position (i,j) will be presented in Appendix A Eq. A.6.

For a given hypothesis H_k estimated at the location (i,j) in both images Y_{ref} (reference image) and Y (image to align with respect to Y_{ref}), we will assume that the more the *a posteriori* probability are close the more the imprecision on the data is small.

Let us take the extreme case where $p_{i,j}(H_2/Y_{ref}) = 1$ and $p_{i,j}(H_2/Y) = 1$. The contradiction in this case is absent because the peak appears at the same position in both images. This is based on the assumption that the higher is the conflict, the higher is the imprecision.

Assumption hyp_2

Let us denote by $\gamma^{Y_{ref}}(i,j) = (\gamma_1^{Y_{ref}}(i,j), \gamma_2^{Y_{ref}}(i,j))$ the shape parameters of a peak at position (i,j) belonging to Y_{ref}, and $\gamma^Y(i_2,j_2) = (\gamma_1^Y(i_2,j_2), \gamma_2^Y(i_2,j_2))$ the shape parameter of a peak at position (i_2,j_2) belonging to Y. The more the parameters of both peaks are close, the more the imprecision on the data is small.

Assumption hyp_3

We will model that the more the peaks are far, the more the contradiction is large. Indeed, the peak position variations are limited by a fuzzy neighborhood around the expected position. Outside of this neighborhood, two peaks can not be assigned as corresponding.

Membership function

These hypothesis are defined to quantify the imprecision in the data which may be modeled using the fuzzy set theory. This is based on the assumption that the concept of the imprecision is an ambiguous one, *i.e*, each data item is considered as imprecise with a certain degree of membership in this fuzzy set denoted $E_{imprecise}$ (e.g., the imprecise data set). In our case, the degree of membership $\mu_{i,j}$ denotes how much the pixel with specific *a posteriori* probability is imprecise, given different hypothesis.

The link between hard domain and fuzzy domain can be given with an *S*-membership function f whose expression is given in Eq. 3.3. Note that the range $[a,c]$ defines the fuzzy region.

$$f(x;a,b,c) = \begin{cases} 0 & x < a \\ \frac{(x-a)^2}{(b-a)(c-a)} & a \leq x < b \\ 1 - \frac{(x-c)^2}{(c-b)(c-a)} & b \leq x < c \\ 1 & \text{otherwise} \end{cases} \tag{3.3}$$

where $a < b < c$.

To calculate $\mu_{i,j}$ for each observed coefficient $y(i,j)$, we will define in the next subsection, a *S*-membership function associated to each hypothesis.

Imprecision modeling through member function

Let us describe in this part the three *S*-membership functions associated with the three hypotheses:

Modeling of hyp_1

hyp_1 expresses the contradiction between two *a posteriori* probabilities for the same hypotheses of peak presence/absence. The modeling of hyp_1 requires the definition of a *S*-membership degree, $\mu_{i,j}^{[hyp_1]} \in E_{imprecise}$ for every pixel of Y_{ref} and Y.

To measure the distance between two variables we generally use their ratio, however this approach leads sometimes to obtain undefined fraction (dividing by zero). To avoid such difficulty, it is better to manipulate the exponential of these two variables before computing their ratio.

Therefore, membership degree computing in $E_{imprecise}$ means here calculating the ratio of the exponential of the *a posteriori* probabilities, then finding its projection by the *S*-membership function (defined in Eq. 3.3).

This function allows us to quantify, from this ratio of exponential, the membership

to the fuzzy set $E_{imprecise}$. The proposed $\mu_{i,j}^{[hyp_1]}$ [1] defined by the exponential ratio of the smallest and the biggest probability of the couple $(p_{i,j}(H_2/Y_{ref}), p_{i,j}(H_2/Y))$ in order to keep a ratio smaller than one (< 1), is given by:

$$\mu_{i,j}^{[hyp_1]} = f_{hyp_1}\left(\frac{exp^{min\left(p_{i,j}(H_2/Y_{ref}),p_{i,j}(H_2/Y)\right)}}{exp^{max\left(p_{i,j}(H_2/Y_{ref}),p_{i,j}(H_2/Y)\right)}}; a_1, b_1, c_1\right) \tag{3.4}$$

Note that the probabilities $(p_{i,j}(H_2/Y_{ref}), p_{i,j}(H_2/Y))$ are estimated using a Monte Carlo Markov Chain procedure (Appendix A, Eq.A.6).

Modeling of hyp_2

hyp_2 models the contradiction between the shape parameters of two peaks belonging to two spectrum images. The modeling of hyp_2 requires the definition of a membership degree $\mu_{i,j}^{[hyp_2]} \in E_{imprecise}$ using a *S*-membership function f_{hyp_2} expressed as:

$$\mu_{i,j}^{[hyp_2]} = f_{hyp_2}\left(\frac{min\left(\gamma_1^{Y_{ref}}(i,j).\gamma_2^{Y_{ref}}(i,j), \gamma_1^{Y}(i_2,j_2).\gamma_2^{Y}(i_2,j_2)\right)}{max\left(\gamma_1^{Y_{ref}}(i,j).\gamma_2^{Y_{ref}}(i,j), \gamma_1^{Y}(i_2,j_2).\gamma_2^{Y}(i_2,j_2)\right)}; a_2, b_2, c_2\right) \tag{3.5}$$

Note that since the peak shape parameters γ_1^Y and γ_2^Y are strictly positive, the exponential function is unuseful in this case.

Modeling of hyp_3

hyp_3 models a neighborhood where the possibility to assign two peaks is highly encouraged. The modeling of hyp_3 requires the definition of a membership degree $\mu_{i,j}^{[hyp_3]} \in E_{imprecise}$ using an *S*-membership function f_{hyp_3}:

$$\mu_{i,j}^{[hyp_3]} = f_{hyp_3}((i-i_2)^2 + (j-j_2)^2; a_3, b_3, c_3) \tag{3.6}$$

where (i, j) stands for the position of the peak belonging to Y_{ref} and (i_2, j_2) stands for the position of the peak belonging to Y.
In practice, the values of the coefficients (a_1, b_1, c_1, a_2, b_2, c_2, a_3, b_3, c_3) are automatically estimated by the Genetic Algorithm (Appendix B) using a training datasets (spectra with known peak alignment results). Once these coefficients are estimated, they will used to measure the imprecision of new spectra and no coefficients reestimation is needed.

Overall membership degree function

Our aim is now to propose the overall membership degree $\mu_{i,j} \in E_{imprecise}$. To this end, we simply opted for the average operator as fusion operator:

$$\mu_{i,j} = \frac{1}{3}.(\mu_{i,j}^{[hyp_1]} + \mu_{i,j}^{[hyp_2]} + \mu_{i,j}^{[hyp_3]}) \tag{3.7}$$

[1] $\mu_{i,j}^{[hyp_1]}$ is the membership function degree to the fuzzy set $E_{imprecise}$ related to the first hypothesis hyp_1.

In conclusion, we proposed in this section, an imprecision modeling scheme of spectrum images through three intuitive assumptions, mathematically translated and fused in order to obtain the overall membership degree $\mu_{i,j}$. We need *a posteriori* probability and peak shape parameters to estimate $\mu_{i,j}$: a proposed estimation scheme is presented in Appendix A with a Bayesian approach (Monte Carlo Markov Chain MCMC optimization). The quantification of $\mu_{i,j}$ allows us now to define the mass functions, crucial step in the evidence theory (Appendix C). This mass function will be then used to detect and align peaks.

The peak detection and alignment scheme

Proposed mass function

Determination of the proposed mass function requires the *a posteriori* probability and the imprecision degree $\mu_{i,j}$ already defined. Indeed, two extreme situations appear: 1) The first one is characterized by the total imprecision absence ($\mu_{i,j} = 0$), in this case only the mass functions of the simple hypotheses are non-zero. 2) The second situation is characterized by the total ignorance ($\mu_{i,j} = 1$): all the mass functions of the simple hypotheses are null. The expression of the proposed non-normalized mass function $m_{i,j}(Y)$ for a given observation Y is:

$$m_{i,j}(\{H_1\};Y) = (1-\mu_{i,j}).p_{i,j}(H_1/Y) \tag{3.8}$$

$$m_{i,j}(\{H_2\};Y) = (1-\mu_{i,j}).p_{i,j}(H_2/Y) \tag{3.9}$$

$$m_{i,j}(\{H_1,H_2\};Y) = \mu_{i,j}.max(p_{i,j}(H_1/Y), p_{i,j}(H_2/Y)) \tag{3.10}$$

The normalization step consists in having:

$$m_{i,j}(\{H_1\};Y) + m_{i,j}(\{H_2\};Y) + m_{i,j}(\{H_1,H_2\};Y) = 1 \tag{3.11}$$

When we have two or several sources on the same frame of discernment Θ built by various hypotheses ($\Omega = \{H_1, H_2, \{H_1,H_2\}\}$ in our case), we can associate for every image Y a mass function $m_{i,j}(Y)$ which quantifies knowledge brought by the observation. The combination rule of Dempster-Shafer (DS) consists in supplying a single mass function from all the mass functions $m_{i,j}(Y)$ associated to each observation Y (Appendix C). The combined mass function $\mathfrak{m}_{i,j}$ is then calculated using the DS combination (Appendix C) as follows:

$$\begin{aligned}\mathfrak{m}_{i,j}(\{H_2\}) =& m_{i,j}(\{H_2\};Y)\ .\ m_{i,j}(\{H_2\};Y_{ref}) \\ &+\ m_{i,j}(\{H_1,H_2\};Y)\ .\ m_{i,j}(\{H_1,H_2\};Y_{ref})\end{aligned} \tag{3.12}$$

$$\begin{aligned}\mathfrak{m}_{i,j}(\{H_1\}) =& m_{i,j}(\{H_1\};Y)\ .\ m_{i,j}(\{H_1\};Y_{ref}) \\ &+\ m_{i,j}(\{H_1,H_2\};Y)\ .\ m_{i,j}(\{H_1,H_2\};Y_{ref})\end{aligned} \tag{3.13}$$

$$\mathfrak{m}_{i,j}(\{H_1,H_2\}) = m_{i,j}(\{H_1,H_2\};Y)\ .\ m_{i,j}(\{H_1,H_2\};Y_{ref})$$

This mass function will be used in the following paragraph to estimate the chemical shift of a detected peak. We hence be able to propose a method realizing simultaneous detection and alinement of peaks.

Proposed cost function

In order to model the peak chemical shifts, we are interested in recovering a displacement vector $d = \{d_{i,j}\}_{i=1...M,\ j=1...N}$ where $d_{i,j} = \begin{bmatrix} d_i \\ d_j \end{bmatrix}$ is a local displacement vector associated to the peak at the location (i,j). Adopting an evidence strategy previously defined, we formulate $\hat{d}$ as:

$$\hat{d} = \underset{d/d_{i,j}\in V}{\operatorname{argmax}} [Bel_{i,j}(\{H_2\}/Y_{ref},(Y+d)] \tag{3.14}$$

where V is the neighborhood selected according to hyp_3 and $Bel_{i,j}$ is the belief function which is derived from the mass function $\mathfrak{m}_{i,j}$ and expressed as (Appendix C):

$$Bel_{i,j}(\{H_2\}) = \mathfrak{m}_{i,j}(\{H_2\}) + \mathfrak{m}_{i,j}(\{H_1,H_2\}) \tag{3.15}$$

To maximize the cost function Eq. 3.14, we need the *a posteriori* probabilities as well as the parameters of the shape filters. An analytical solution of this problem is unfortunately impossible, and we decide to use a MCMC procedure to realize such optimization (see Appendix A). An overview diagram of the peak detection and alignment chain in Fig.3.1.

3.2.2 Peak detection and alignment validation

This part describes some peaks detection and alignments results which are obtained with the proposed evidential alignment scheme. This method was applied on synthetic spectra designed to fit the characteristics of the HSQC HR-MAS spectra as well as on some real spectra. More results on real spectra will be presented in chapter 6. In order to validate and emphasize the benefit of the proposed approach, we have retained the following criteria for the estimaion validation:

1. **Accuracy** : it defines the accuracy level of estimated parameters. It represents the difference between computed and theoretical value known from a ground truth. In our case, we use the mean chemical shift error function. For each 2D spectrum image, we calculate the bias for the carbon chemical shift ε_c and the hydrogen chemical shift ε_h where

$$\varepsilon_c = \frac{1}{N_p}\sum_{p=1}^{N_p}[\hat{X}_c(p) - X_{r_c}(p)] \quad \varepsilon_h = \frac{1}{N_p}\sum_{p=1}^{N_p}[\hat{X}_h(p) - X_{r_h}(p)] \tag{3.16}$$

   where $(\hat{X}_c(p),\hat{X}_h(p))$ stands for the estimated coordinates of the a peak X_p whereas $(X_{r_c}(p),X_{r_h}(p))$ stands for the theoretical location of the peak X and N_p is the number of peaks.

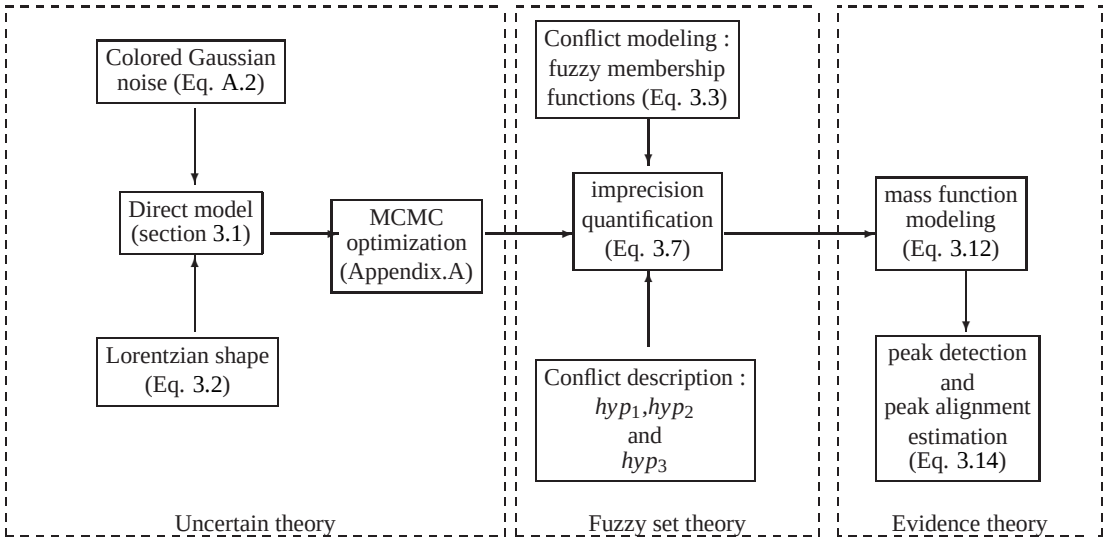


Figure 3.1: Overview diagram of the peak detection and alignment chain.

2. **Validation of some registration properties** : in this study, we just retain the transitivity property : for any three spectra, the generated transformation from the second to the third spectrum composed with the generated registration transformation (*i.e.*, alignment) from the first to the second one should be equal to the alignment from the first to the third one. This transitivity property can be formalized, for every detected peak at location (i, j), as :

$$\hat{d}_{1\to 2}(\hat{d}_{2\to 3})(i,j) = \hat{d}_{1\to 3}(i,j) \tag{3.17}$$

3. **Robustness** : evaluates the performance of the method in special cases such as the presence of pathology or different noise level in the data. The principle is that even if we have different initial conditions, the system converges toward a unique solution. In our case, this means we obtain same alignment results (same number of detected peaks and value of mean error chemical shift).

The main advantage of using simulated data is that we perfectly know the characteristics of the data such as the number of peaks presented in every spectrum and the peak chemical shifts values. For this, we firstly generate a synthetic theoretical 2D spectrum image X_{ref} with ($M = 500$ pixels by $N = 500$ pixels) which contains $N_p = 200$ peaks.

This synthetic spectrum will be used as reference to register other synthetic spectra. Three other synthetic theoretical 2D spectrum images X_1, X_2 and X_3 are generated from X_{ref} by modifying the location and the shape hyperparameters of each peak of X_{ref}. The values of chemical shift vectors are assumed randomly distributed following a Gaussian distribution with zero mean and variance matrix $\begin{bmatrix} 0.02 & 0 \\ 0 & 0.25 \end{bmatrix}$. Note that this variance matrix was chosen to fit the real chemical shift vectors. The hyperparameters of the shape filter h for each peak are randomly generated from a Gaussian distribution with mean equals to 1 and variance 0.1. In order to simulate the peak shape modification, we randomly add a white gaussian error of variance 0.005 to each peak shape hyperparameters. A zero mean correlated noise B with covariance matrix Γ_b was added to each spectrum to obtain the synthetic spectra used in the simulation Y_{ref}, Y_1,Y_2 and Y_3.

The Peak Signal to Noise Ratio PSNR of Y_{ref} is set to 30dB where the PSNR is expressed as:

$$PSNR = 10\, log_{10}(max(X_{ref})^2 / \mathrm{E}[(B)^2]) \tag{3.18}$$

This value of PSNR was chosen to fit at best the real spectra. Indeed, the PSNR of the real spectrum is $\sim$ 30dB that corresponds ta a strongly noised observation.

In order to emphasize the robustness of the proposed approach to the high level of noise, we have processed to the detection and the alignment of peaks contained in Y_1, by taking Y_{ref} as reference, with different values of PSNR (30, 28, 25 and 23 dB). The different peak detection and alignment results are presented in Table 4.1. Since the pixel resolution is 0.16ppm (*resp* 6.8 10^{-3} ppm) for the *y*-axis, *i.e.* the ^{13}C chemical shift axis, (*resp* *x*-axis, *i.e.* the ^{1}H chemical shift axis), as one can remark, even with a PSNR=25dB, we obtained

PSNR	ε_c	ε_h	nb of missed peaks	nb of false peak assignments
30dB	0.0064	$9\ 10^{-5}$	9/200	4/200
28dB	0.0107	$7\ 10^{-4}$	9/200	5/200
25dB	0.149	0.0052	11/200	8/200
23dB	0.291	0.019	13/200	10/200

Table 3.1: Detection and alignment error on the synthetic spectrum Y_1. The mean chemical shift errors are expressed in ppm.

a sub-pixel precision for the mean chemical shift errors. In fact, with a PSNR=25 dB, the $\varepsilon_c < 0.16$ and $\varepsilon_h < 6.8\ 10^{-3}$ (Eq. 3.16). Figure. 3.2(a) shows an example of missed peak. As one would suspect, the missed detections correspond to weak events which are strongly noised. Indeed, the average amplitude of missed peaks is $\sim \frac{1}{10}$ maximum simulated noise amplitude. Fig. 3.3 shows an example of x-axis and y-axis projection of a missed and a detected peak. As one can remark, the missed peak is completely burred in the noise. The false peak assignments correspond to events which are strongly imprecise : the shapes, locations and *a posteriori* probabilities of the right peaks and the estimated assignment peak are too close. Figure. 3.2(b) shows an example of false peak alignment errors, the right location of the peak is presented with a continuous arrow and the estimated peak location is presented with a dotted arrow. As one can see, the characteristics of the estimated peak (location, shape and amplitude) are too close to the assignment peak characteristics. In this case, the imprecision is so great that a distinction between these two peaks turns out to be difficult and sometimes impossible.

In order to emphasize the benefit of the proposed approach, two different alignment methods were applied to the synthetic spectrum Y_3 with different values of PSNR: a Bayesian method [Toews05] and our alignment method. The peak alignment results are presented in Table 3.2. We can easily observe that the proposed method performed better than the Bayesian method. Indeed, even with a PSNR=25dB, we obtained a sub-pixel precision for the mean chemical shift errors which is two times smaller compared to that obtained by the Bayesian method. This can be explained by the fact that we took into account in our alignment scheme both uncertainty (the *a posteriori* probability) and imprecision in the spectra (conflict information). It is important to note that the Bayesian scheme provides only tools to handle the uncertainty and, for this reason, the use of evidence theory was proposed. Note that the Bayesian method only addresses the alignment step. Therefore, we have used the detection results obtained by the proposed method to perform the alignment step with the Bayesian Method.

Now we consider the problem of transitivity property validation. Since the missed detections correspond to weak events which are strongly noised, the missed peaks are the same for each spectrum. The alignment error using transitivity property are presented in Table 3.3. As one can see, we obtained the same alignment results which validate the

	Proposed scheme		Bayesian scheme	
PSNR	ε_c	ε_h	ε_c	ε_h
$30dB$	$5.1\ 10^{-3}$	$9.1\ 10^{-5}$	0.097	$6.1\ 10^{-3}$
$28dB$	$1.21\ 10^{-2}$	$5.1\ 10^{-4}$	0.139	$8.6\ 10^{-3}$
$25dB$	0.1098	$2.5\ 10^{-3}$	0.2584	$1.91\ 10^{-2}$
$23dB$	0.1874	$9.35\ 10^{-3}$	0.3278	$2.03\ 10^{-2}$

Table 3.2: The mean chemical shift errors ε_c ε_h on the synthetic spectrum Y_3 expressed in ppm obtained by the proposed and the Bayesian methods.

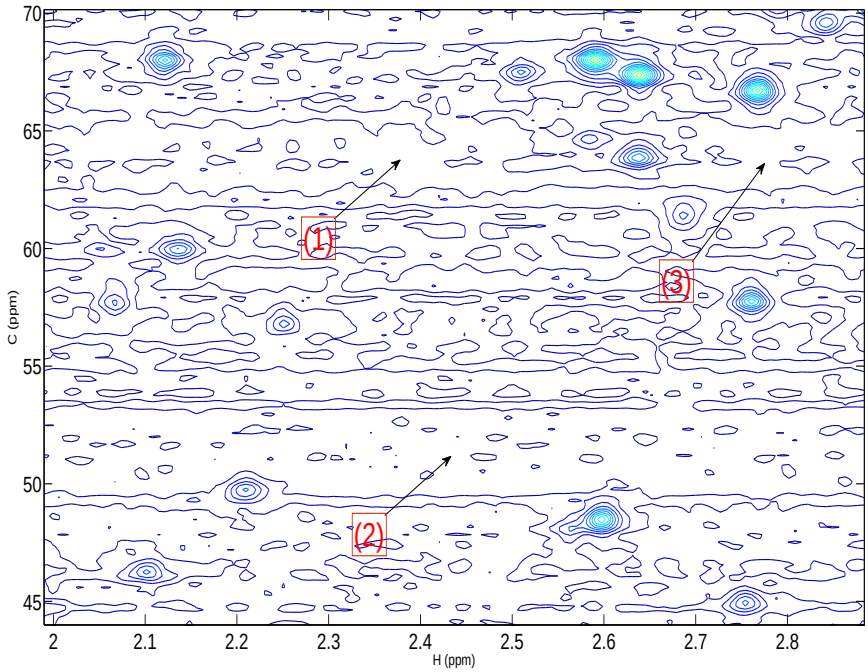
	$Y_{ref\to1\to2}$		$Y_{ref\to2}$	
PSNR	ε_c	ε_h	ε_c	ε_h
$30dB$	0.0059	$8.7\ 10^{-5}$	0.006	$8.5\ 10^{-5}$
$28dB$	0.011	$5.3\ 10^{-4}$	0.010	$4.2\ 10^{-4}$
$25dB$	0.149	0.0012	0.0984	0.0011
$23dB$	0.2245	0.0121	0.2278	0.0131

Table 3.3: Alignment error using transitivity property

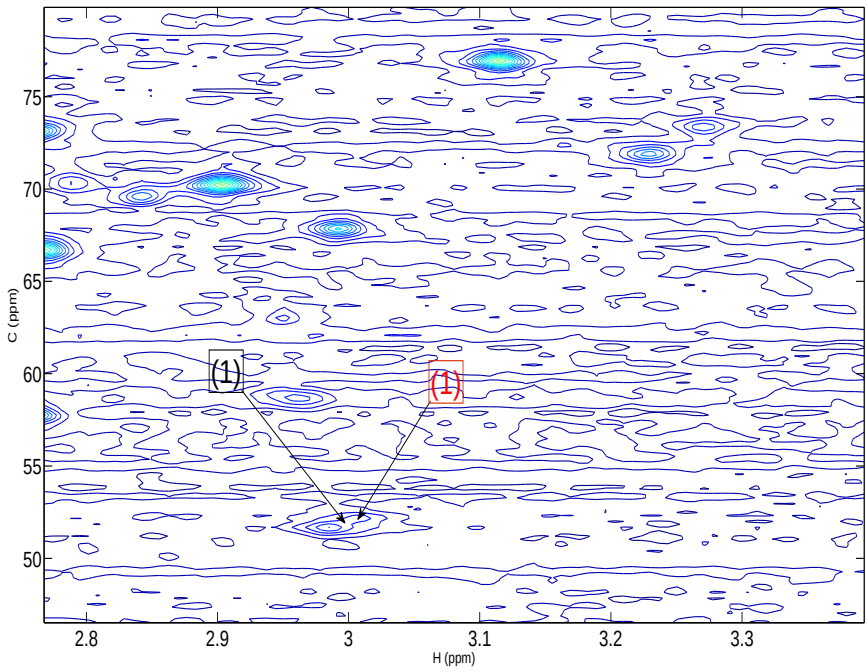
transitivity property.

Figure 3.4 presents some results of peak alignment on the same region of simulated spectra. We can see that all the peaks were correctly detected and aligned. In addition, we can easily remark from Fig. 3.4 that a manual extraction of peaks (7,8; 13,14; 24 and 29) seems difficult even impossible.

Figure 3.5 displays the peaks detection and alignment results on the same zone of a healthy spectrum and cancerous spectrum. Some peaks like peak number 10 are visually very difficult to detect due to the high noise level, yet our method is actually able to detect and align it. Note that the spectra are presented as contour plots with the same number of level which explain the absence of a presentation the peak number 10 in Fig.3.5 (b).



(*a*)



(*b*)

Figure 3.2: Example of (a) a missed peak, (b) a false peak assignment : real location in dotted arrow and estimated position in continuous arrow.

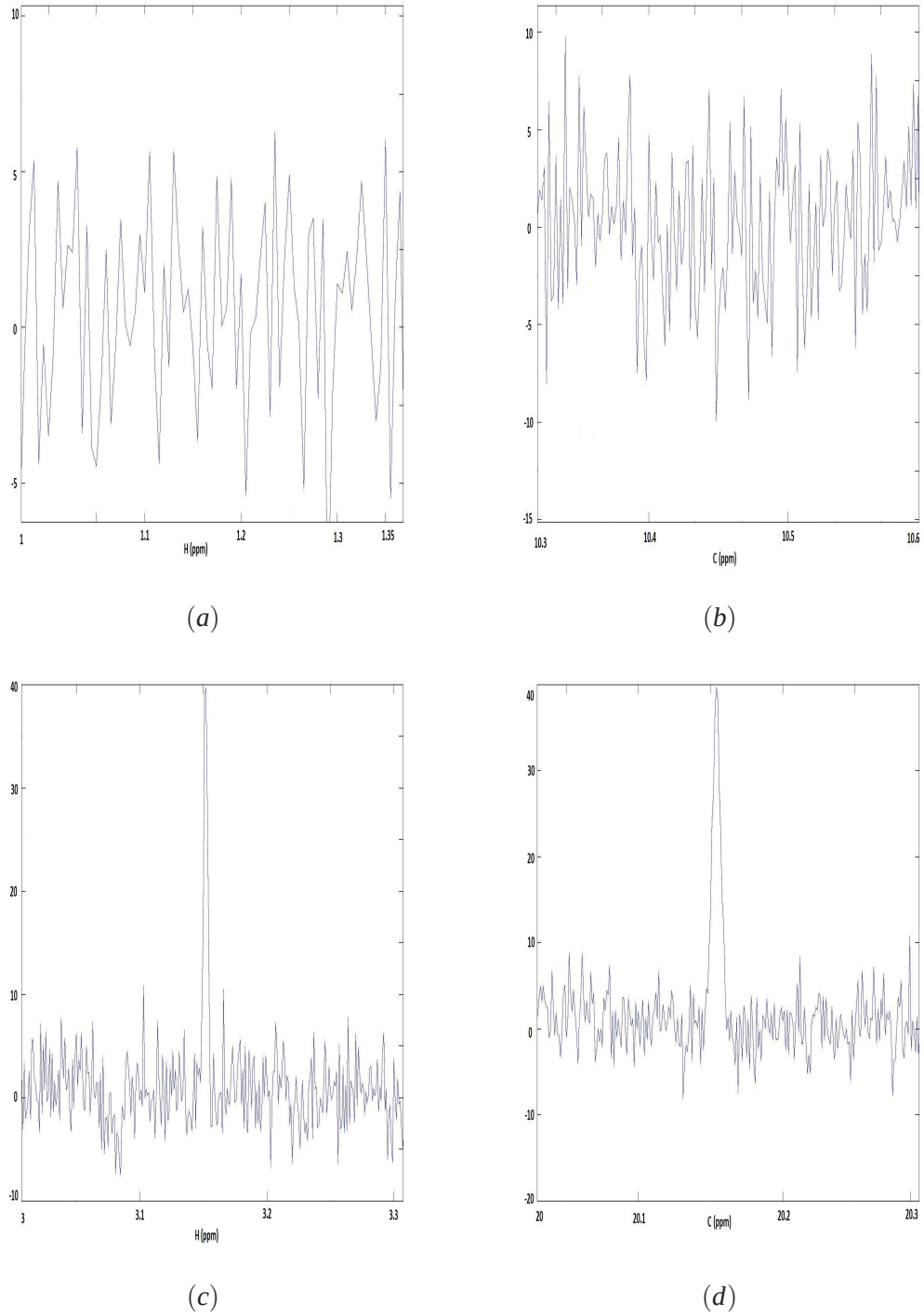
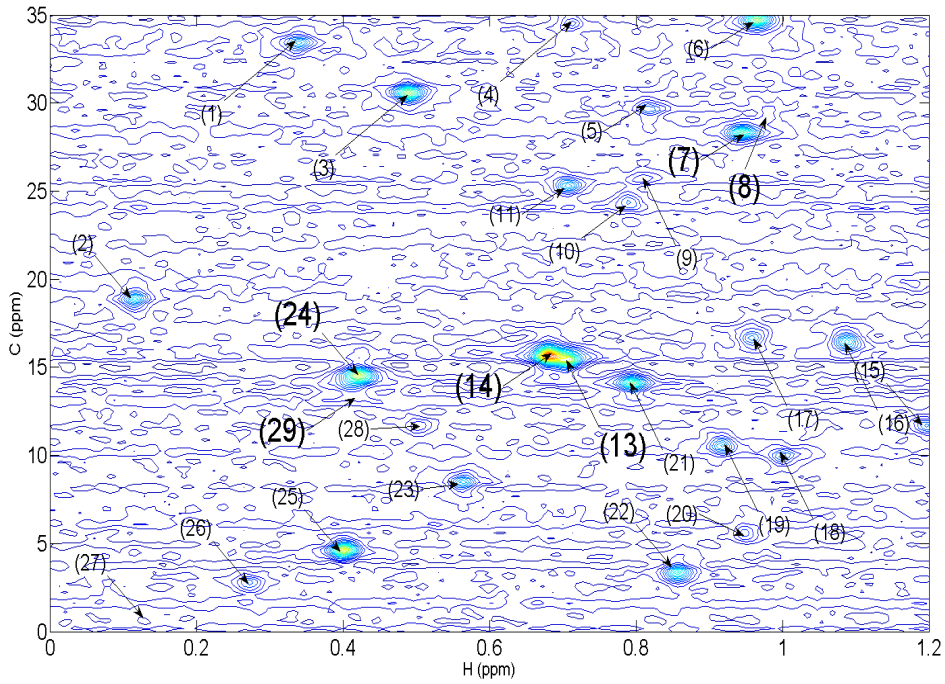
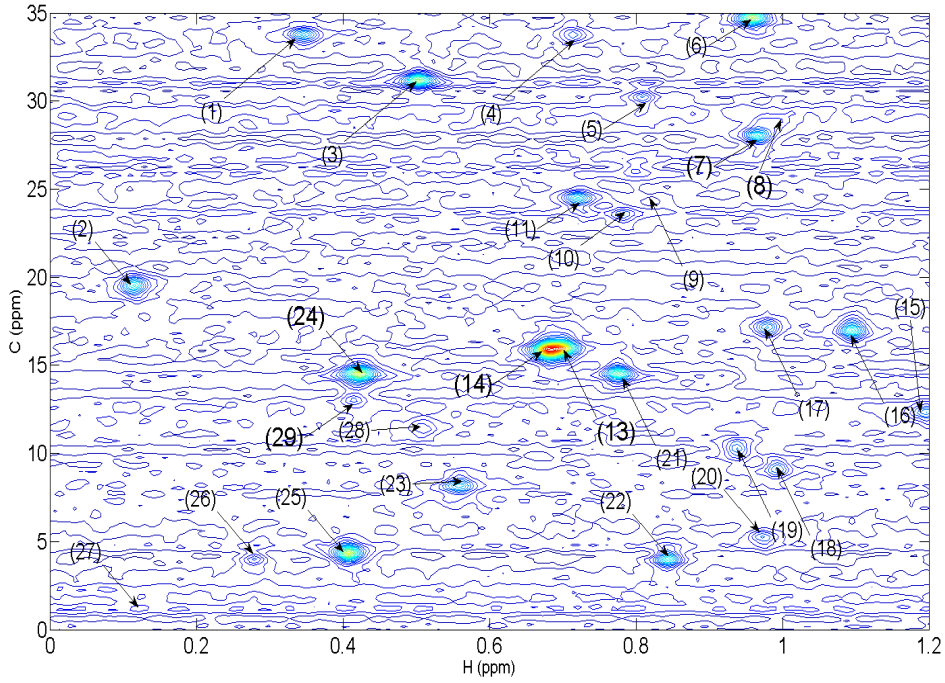


Figure 3.3: An example of the x and y projections of a missed and a detected peak (a) x-axis projection of a missed peak, (b) y-axis projection of a missed peak, (c) x-axis projection of a detected peak, (d) y-axis projection of a detected peak

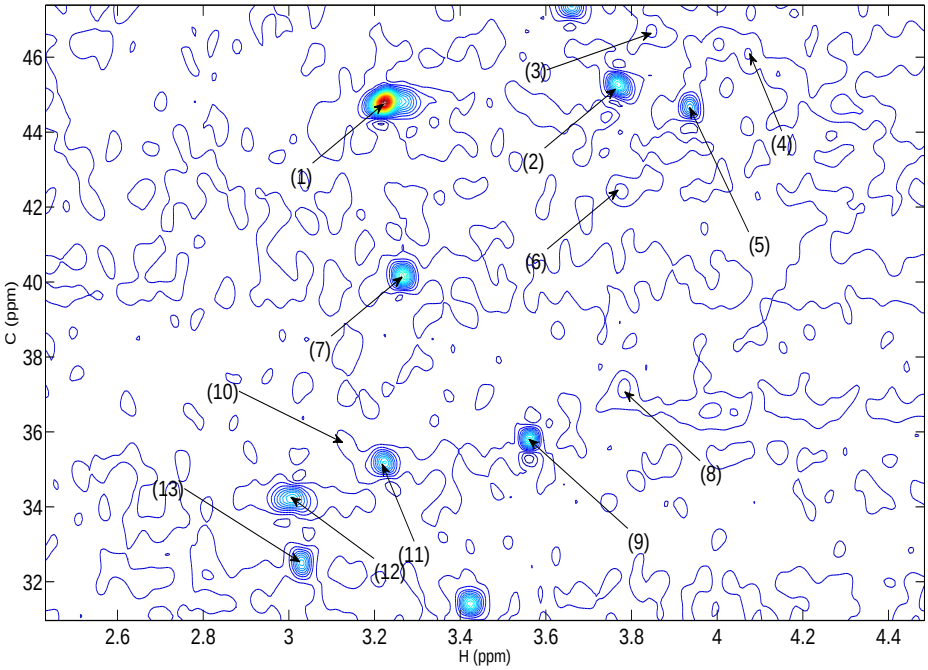


(*a*)

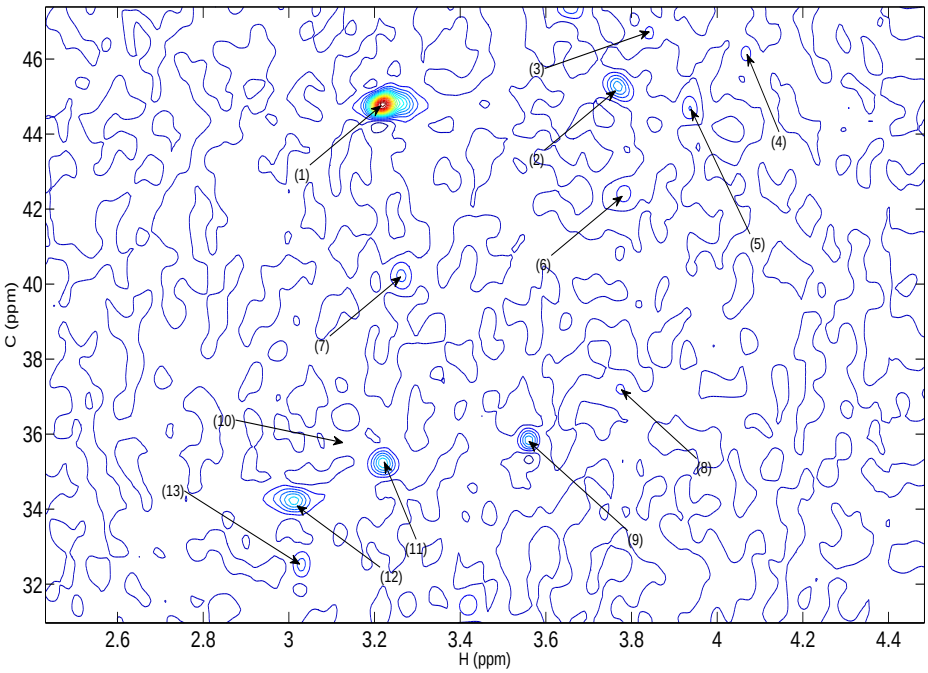


(*b*)

Figure 3.4: The detection and alignment results on (a) the reference synthetic observed spectrum Y_{ref} (PSNR = 30dB) and (b) the synthetic observed spectrum Y_1 (PSNR = 30dB)



(*a*)



(*b*)

Figure 3.5: The detection and alignment results on (a) a real healthy spectra (b) a real cancerous spectra.

3.3 Active zone detection and alignment (fMRI)

We recall that an fMRI image can merely be presented by a set of active zones that we have to detect and align (as for HSQC spectrum peaks) but in 3D. Each active zone may be characterized with its location, its shape and its voxel intensities. In contrast to the peak alignment scheme, the detection and the alignment of the active zone is separately addressed. More explicitly, given two detected objects, our aim is to align them according to their canonical poses. To this end, the active zone detection is firstly performed using a Hidden Markov chain segmentation HMC approach [Bricq08] which allows us to integrate spatial information into the segmentation method. Concerning the alignment method, it is important to note that the alignment can be effectively exploited in the indexing procedure only if all the following criteria are verified:

- 3D rotation invariance: objects with similar shape should be identically aligned whatever their initial orientations;
- 3D anisotropic transformation invariance: an aligned object that has been undergoes a narrowing or lengthening of a reasonable size following the alignment directions should maintain the alignment result;
- Weak time-consuming.

Note that since the active zone detection is addressed using a classical HMC segmentation method, we present in this thesis only the active zone alignment method. In order to lead to a satisfactory alignment result, we integrate human perception in the scheme of alignment. Generally, one seeks to align an object according to its symmetry axis. This approach allows us both to find the most object natural pose and align visually similar objects in the same manner. Most methods opted for this approach use either the PCA/continuous PCA [Vranić01a, Vranic01b] or the reflection measurement [Simari06] to find the mirror planes (planes of reflection). These planes are then used to estimate the appropriate cartesian coordinate system associated with the object. Although these methods were successfully applied on 3D internet object searching, they are unfortunately not currently adopted to the 3D fMRI objects. Indeed, due the cortex shape, the reflection symmetry of the active zones is more spherical than planar. We propose to use non-linear PCA that is well suited to model the reflection symmetry of the cortical active zones.

3.3.1 Partial spherical object alignment method

The characterization of a 3D object by reflection symmetry has aroused a lot of works [Bustos04, Podolak06, Simari06, Mitra06]. Mainly relying on the studies of human perception, these works have motivated our choice and led us to consider this reflection symmetry. In the literature, the mirror symmetry is the most used symmetry. However, the main drawback of this approach lies in its high time-consuming which makes its utilization very constraining [Vranić01a]. To overcome this, some authors [Chaouch08] propose the use of the PCA/continuous PCA algorithms to calculate all reflection symmetries that may

be characterized by the object main direction. Indeed, the PCA/continuous PCA seeks to describe the data variance with a set of orthogonal features better known as the Principal Components (PC) which can be estimated by the correlation matrix. These PC will be used as a new coordinate system on which the original data will be projected. In the linear (basic) PCA, the PC is either lines (2D object) or planes (3D object). The use of the planar PC as reflection symmetries has found lot of success in many applications particularly the 3D internet object searching [Chaouch08].
The main drawback of this approach lies in its high time-consuming which makes its utilization very constrained. To overcome this, the authors [Vranić01a, Chaouch08] propose the use of the PCA/continuous PCA algorithms to calculate all reflection symmetries that may be characterized by the object principle direction. Indeed, the PCA/continuous PCA seeks to describe the data variance with a set of orthogonal features better known as the Principal Components (PC) which can be estimated by the correlation matrix. These PC will be used as a new coordinate system on which the original data will be projected. In the linear (basic) PCA, the PC is either lines (2D object) or planes (3D object). The use of the planar PC as reflection symmetries has found lot of success in many applications particularly the 3D internet object searching [Chaouch08].

Nevertheless, the basic PCA/continuous PCA is very well not adapted for fMRI active zones alignment task due to the cortex shape. Indeed the reflection symmetry in our case is likely more spherical than planar. Our aim is then to properly integrate this *a priori* knowledge in the proposed scheme. An elegant way to address this task is the use of nonlinear PCA. This nonlinear behavior was firstly presented in [Lingoes67, Kruskal74]. Then, many varieties of the nonlinear PCA were proposed such as the probabilistic nonlinear PCA [Lawrence05], the kernel nonlinear PCA [Ge09] and the neural network PCA [Kramer91, Scholz07]. Among all these works, we pay a particular attention to this latter: the Neural Network PCA (NNPCA) [Kramer91] has proved its high accuracy to estimate the non-linear PC in many field such as the meteorology [Hsieh98], the EEG Electro-Encephalo-Graphy signal [Stamkopoulos98], metabolism [Scholz05]. Indeed, the NNPCA relies on Multi-Layer Perceptron (MLP) [Bishop95] with an auto associative topology allowing an identity mapping (*e.g*; the neural network input P should be equal to the neural network output $\hat{P}$ where P is the observed object voxel coordinates). To this end, the square reconstruction error $||P-\hat{P}||^2$ should be minimized. The neural network is generally composed of five layers [2-1-2] interconnected with four network weight W_l, $l = 1...4$ (Fig. 3.6). The layer in the middle is called the bottleneck layer which consists of a number of nodes lower than that of the input/output layers. This bottleneck layer, leading to a data compression and a decompression steps, makes the minimization of the square error not trivial. The two first layers constitute the non-linear extraction function $\Phi_{extr}(.)$ allowing us to perform a nonlinear projection of the observed object coordinates $P = p_1...p_N$ (*i.e*; p_i is the cartesian coordinates of the i^{th} voxel of the object (the size of p_i is three)), N is the number of object voxels, into the second layer ($\Phi_{extr}(P) = Z$) in order to obtain the matrix of score (the matrix of the nonlinear PC). In the other hand, the two last layers constitute the reconstruction function $\Phi_{gen}(.)$ allowing us to perform a nonlinear reconstruction of the observation ($\Phi_{gen}(Z) = P$) thanks to the estimated score matrix (Fig. 3.6).

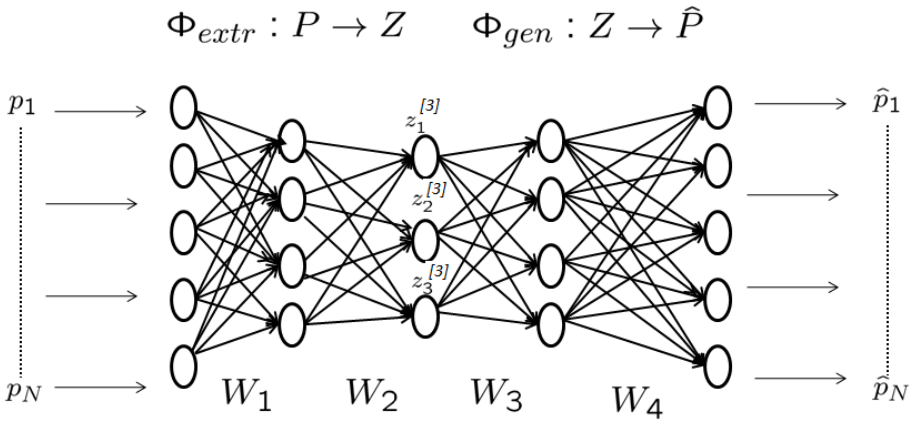
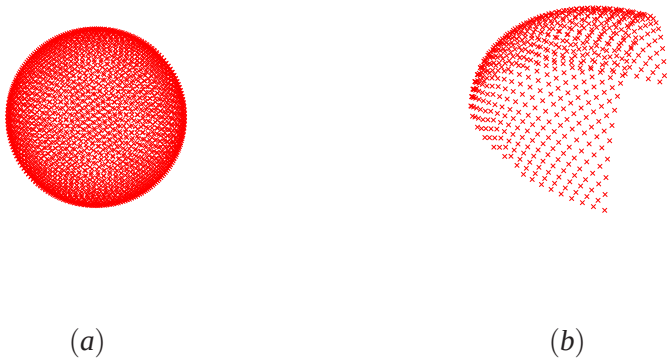


Figure 3.6: A standard MLP network with an auto associative topology. This network is composed of 3 parts [2-1-2]. The first two layers aim at compressing the original signals P to only three components $z_1^{[3]}$, $z_2^{[3]}$ and $z_3^{[3]}$ (the bottleneck layer) thanks to Φ_{extr} function. The last two layers aim at reconstructing the data to $\hat{P}$ thanks to Φ_{gen} function.

Our aim now is to properly integrate the *a priori* knowledge that we have on the data into this network. In other words, we should adopt the network to the spherical shape of the active zone. To this end, we can distinguish two cases. The first one consists of an entire spherical shape (Fig.3.7.a) whereas the second one consists of a partial spherical shape (Fig.3.7.b) which is the case of the fMRI active zones. In the next part, we describe the proposed methods to adopt the network for both shape cases: entire spherical shape and partial spherical shape.

Entire spherical shape case



(a) (b)

Figure 3.7: Example of (a) an entire sphere, (b) a partial sphere.

The developed method was inspired from the work proposed in [Kirby96] where authors adapt the network to the case of circular data (2D data). We extend this work to the 3D data for entire spherical shape. To this end, the triplet $(z_1^{[3]}, z_2^{[3]}, z_3^{[3]})$ [2] (Fig. 3.6) are constrained to lie on a unit sphere:

$$(z_1^{[3]})^2 + (z_2^{[3]})^2 + (z_3^{[3]})^2 = 1$$

Generally, a sphere can be characterized with the triplet (r, θ, ϕ) where $r \geq 0$ is the distance from the origin, $\theta \in [0, 2\pi[$ the azimuth angle and $\phi \in [-\frac{\pi}{2}, -\frac{\pi}{2}[$ the elevation angle. The three nodes of the bottleneck layer can be described by two hyperparameters (θ and ϕ):

$$z_1^{[3]} = cos(\theta)cos(\varphi), \qquad z_2^{[3]} = sin(\theta)cos(\varphi) \quad \text{and} \quad z_3^{[3]} = sin(\varphi) \tag{3.19}$$

To calculate the radius r of the sphere, a forward propagation is performed. We recall that the link between the layer number 2 and the bottleneck layer is assured by the matrix W_2. Let $z_m^{[2]}$ $m \in [1...M]$, M the number of nodes in the second layer, the value of the m^{th} node of the layer number two. Each node m of this layer is connected to each node of the bottleneck layer with a weighted connection w_{im} where $i \in \{1, 2, 3\}$. The radius r is expressed as:

$$r = \sqrt{(v_1)^2 + (v_2)^2 + (v_3)^2} \tag{3.20}$$

where

$$v_1 = \sum_{m=1}^{M} w_{1m} z_m^{[2]}, \quad v_2 = \sum_{m=1}^{M} w_{2m} z_m^{[2]} \quad \text{and} \quad v_3 = \sum_{m=1}^{M} w_{3m} z_m^{[2]}$$

In order to obtain a spherical constraint, we should have:

$$z_1^{[3]} = \frac{v_1}{r}, \qquad z_2^{[3]} = \frac{v_2}{r} \quad \text{and} \quad z_3^{[3]} = \frac{v_3}{r} \tag{3.21}$$

To estimate the network weight W_l, $l = 1...4$, a backward propagation is performed to minimize the error function E with respect to all network weight W_l, $l = 1...4$:

$$E = \sum_{n=1}^{N} \sum_{j=1}^{3} \left(\hat{p}_n^j - p_n^j\right)^2 \tag{3.22}$$

where (p_n^1, p_n^2, p_n^3) is the cartesian coordinate of the object voxel p_n and $(\hat{p}_n^1, \hat{p}_n^2, \hat{p}_n^3)$ the reconstructed coordinates of p_n. This error function can be minimized using the gradient optimization algorithm [Nocedal99] and the derivation of the network weights W_l, $l = 1...4$ are obtained by standard back propagation [Fahlman88].

Partial spherical shape case

In this paragraph, we adapt the method previously presented (the entire spherical shape) to the case of the fMRI active zones (the partial sphere shape). Unfortunately this task is not trivial since the fMRI active zones do not retain a unique shape and consequently the partial sphere modeling the reflection symmetry changes from an active zone to another one. In order to overcome the problem of shape variability, we proceed as follows:

[2] $z_i^{[3]}$ is the value of the ith node of layer number three.

1. The first step consists of finding the closest sphere that properly fits the 3D object shape. To this end, we used the Inverse Non-Linear PCA method proposed in [Scholz05] to initialize the reflection symmetry of the 3D object. Indeed, this method allows the estimation of the surface that models the reflection symmetry with no constrains on the expected surface shape (in our case a spherical shape). Then, we use the gradient descent algorithm to estimate the hyperparameters of the closest partial sphere (the triplet $(\hat{r}, \hat{\theta}, \hat{\phi})$) that properly fits the estimated surface.

2. The second step consists of adding samples to the 3D object to complete the sphere based on the partial sphere hyperparameters estimated in the pervious step. Therefore we obtain two 3D objects. The first one (P) is composed of the original samples coordinates and the second one, denoted by $P' = p'_1 ... p'_{N'}$, is composed of the added samples coordinates.

3. The last step consists in applying the proposed method to the constructed sphere. However, only P contains the useful data and therefore samples of P' should be penalized in the network training step. To this end, it is sufficient to modify the error function (Eq. 3.22) by introducing a new hyperparameter $0 < \xi < 1$ according to:

$$E = \sum_{n=1}^{N} \sum_{j=1}^{3} \left(\hat{p}_n^j - p_{n=1}^j \right)^2 + \xi \sum_{n}^{N'} \sum_{j=1}^{3} \left(\hat{p'}_n^j - {p'}_n^j \right)^2 \tag{3.23}$$

   By this way, the error introduced by P' is penalized and it partially contributes to the gradient.

An overview diagram of the proposed active zones alignment chain is presented in Fig. 3.8. Finally, two objects are aligned according to their estimated partial spherical symmetries. Indeed, it is sufficient to first estimate two partial spherical symmetries for each object (Fig. 3.9) and than to superpose them to find the common object pose (Fig. 3.10).

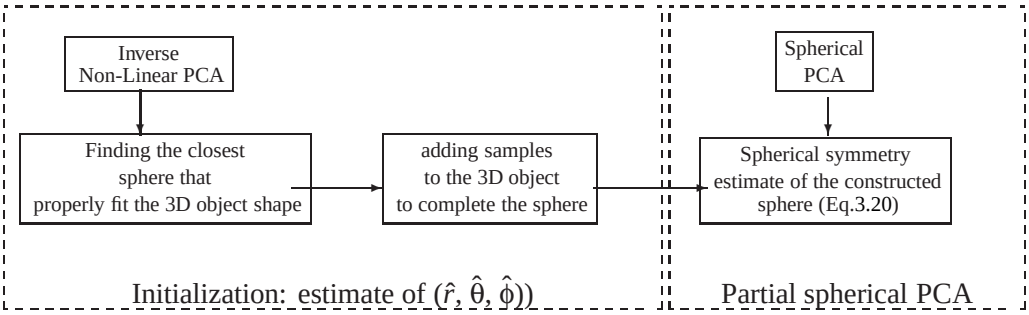


Figure 3.8: Overview diagram of the proposed active zones alignment chain.

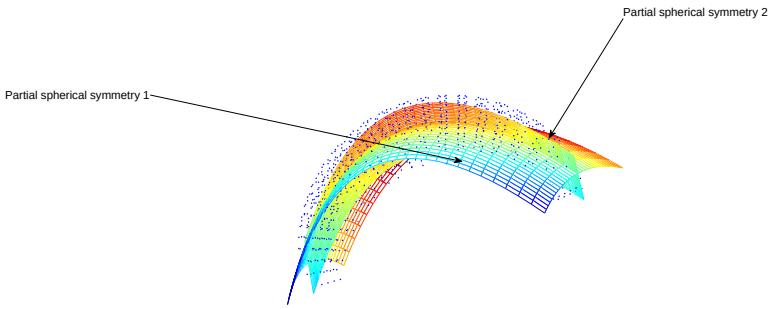
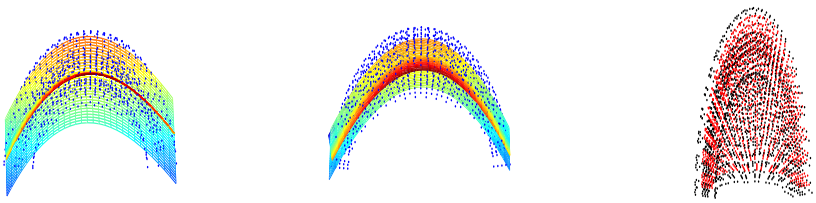
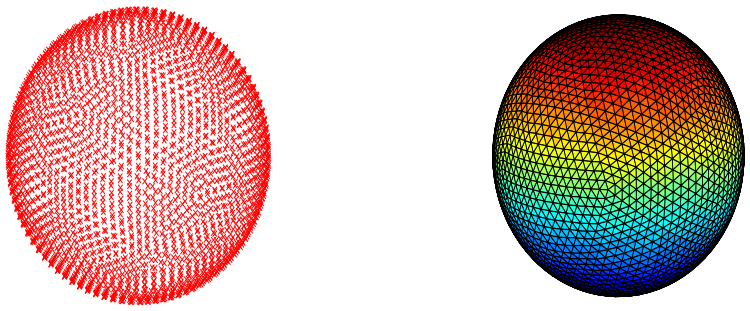


Figure 3.9: Two reflection symmetries estimation of a 3D object.



(a) (b) (c)

Figure 3.10: The two estimated reflection symmetries of (a) object 1 and (b) object 2, (c) the superposition result of the estimated reflection symmetries.



(a) (b)

Figure 3.11: (a) A cloud point representation of the reference sphere, (b) A triangular mesh representation of the reference sphere

3.3.2 Active zone alignment validation

In this part, we provide some 3D object alignment results obtained with the proposed method. This method was applied on both spherical 3D objects and partial 3D objects designed to fit the characteristics of the active zones as well as on some real fMRI active zone. More results on real spectra will be presented in chapter 6. In order to emphasize the benefit of the proposed method and particularly the use of the non-linear PCA, we have compared our algorithm with the continuous PCA method [Vranić01a].

The goal of our experiments is to assess the performance of the proposed alignment scheme. Let us start with spherical toys problem to demonstrate the effects of the proposed and the continuous PCA strategies. To construct our toy data set, we firstly generated a reference sphere denoted by S_r (Fig 3.11.a). Note that we have opted for the 3D triangular mesh representation for all artificial toys (Fig 3.11.b). This sphere will be used as reference to generate twenty nine other spheres by adding a Gaussian noise with zero mean and variance σ_o to its cartesian coordinates (x_r, y_r, z_r) such that the PSNR $\in [5dB, 20dB]$ (Fig. 3.12).

Fig. 3.13 shows some results of spherical symmetry reflection estimation on synthetic toys with different PSNR values. In order to facilitate the visual interpretation of the results we have presented the cross sections of different spheres. As one can see, even with a small PSNR, we can correctly estimate the reflection symmetries.

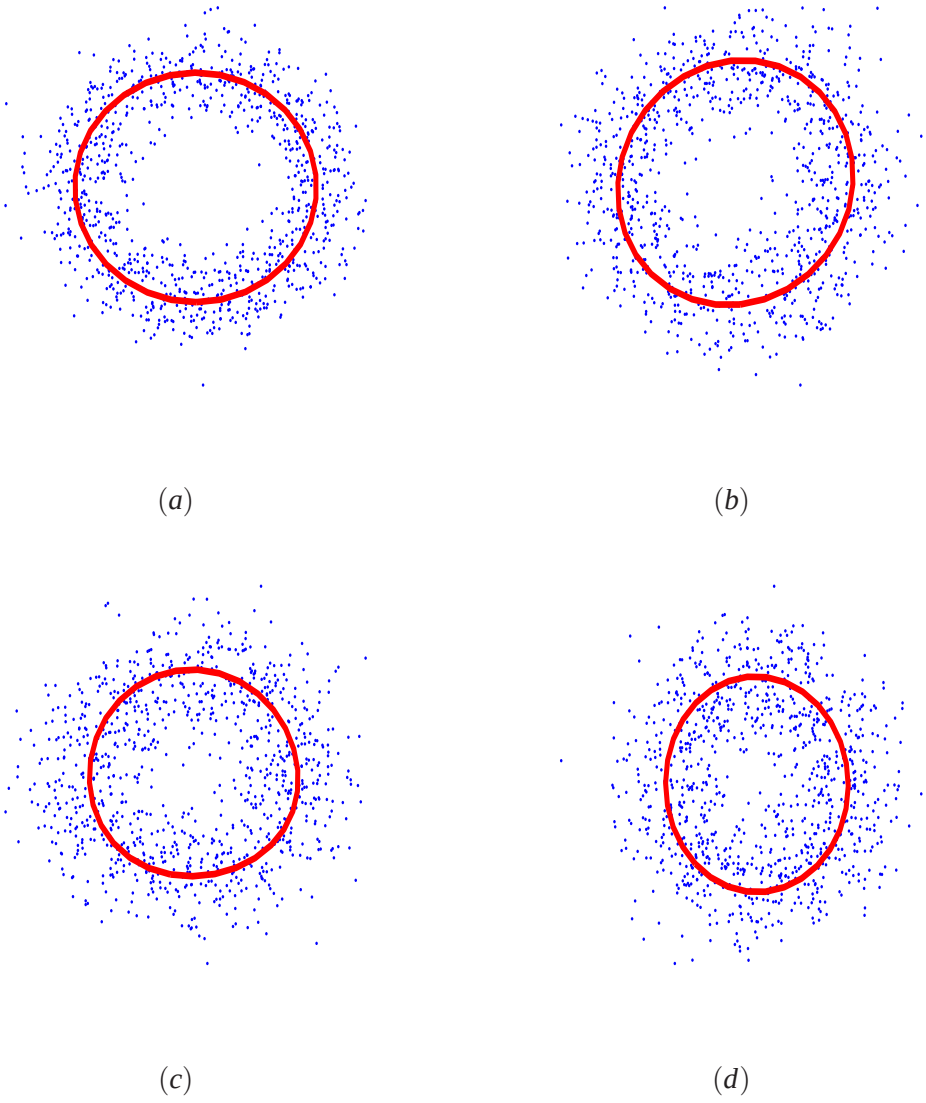
We now address the problem of 3D partial spherical object alignment. To this end, we generated five partial spheres by removing 55%, 60%, 65%, 70% and 75% of S_o and by adding a Gaussian noise with zero mean and variance σ_{oi} as that the $PSNR \sim 15db$. Then,

Figure 3.12: Two synthetic spherical toys generated from the reference sphere S_o

	Proposed method	Continuous PCA
data set1 (55% of S_o)	$0.02 \pm 1.2\ 10^{-4}$	$0.097 \pm 7.85\ 10^{-4}$
data set2 (60% of S_o)	$0.038 \pm 5.8\ 10^{-4}$	$0.124 \pm 2.14\ 10^{-3}$
data set3 (65% of S_o)	$0.0474 \pm 9.7\ 10^{-4}$	$0.301 \pm 4.23\ 10^{-3}$
data set4 (70% of S_o)	$0.062 \pm 1.03\ 10^{-3}$	$0.832 \pm 5.82\ 10^{-2}$
data set5 (75% of S_o)	$0.078 \pm 4.12\ 10^{-3}$	$1.177 \pm 7.98\ 10^{-2}$

Table 3.4: The mean shift errors and the standard deviation obtained by the proposed and the continuous PCA methods.

for each partial sphere S_{oi}, $i = 1...10$, we generated forty nine other spheres by rotating S_{oi}. The rotation angle is between 20 and 130 degrees. At the end of this procedure we obtained five data sets each one composed of 50 toys. Fig.3.14.(a) shows an example of a generated toy whereas Fig.3.14.(b) shows an example of a fMRI active zone. The toy alignment results for each data set are presented in Table 3.4. We can easily observe that the proposed method performed better than the continuous PCA method. Fig. 3.15 shows an example of real active zones alignment. As one can see, our method performs the best alignment result.



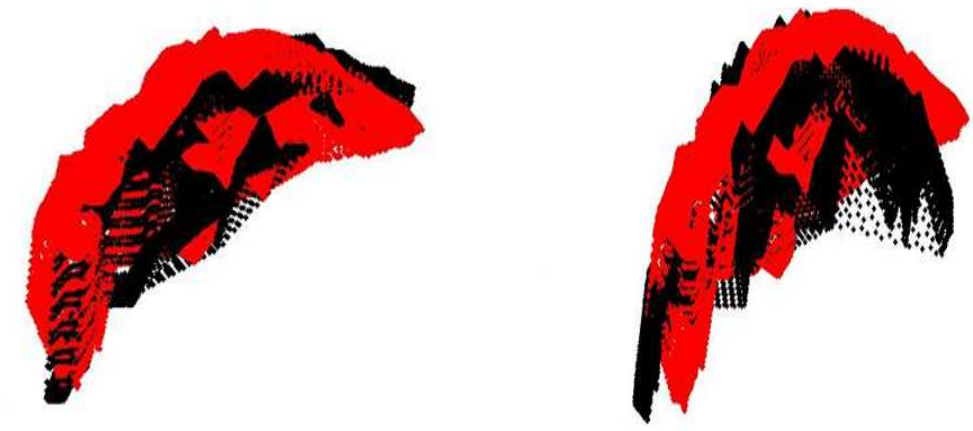
(a) (b)

(c) (d)

Figure 3.13: Spherical symmetry reflection estimation on synthetic toys with (a) PSNR = 15dB, (b) PSNR = 12dB, (c) PSNR = 10dB and (d) PSNR = 8dB

(*a*) (*b*)

Figure 3.14: An example of (a) a synthetic toy, (b) a real active zone



(*a*) (*b*)

Figure 3.15: Alignment result of real active zones obtained with (a) the proposed method, (b) the continuous PCA method.

Conclusion

In this chapter, we addressed the object detection and alignment task.

Concerning the 2D HSQC spectrum, we proposed a new peak detection and alignment methods which combined the modeling of the knowledge by means of the evidence theory and integrate the fuzzy theory to quantify the imprecision degree presented in the spectra. The handling of both imprecision and uncertainty by the evidence theory increased the robustness of the proposed alignment scheme with comparison to the Bayesian method. In addition, we have used the deconvolution model to achieve a better fit of the HSQC spectrum. The synthetic validation of the proposed approach has shown its efficiency and particularly its robustness to the high level of noise, one of the delicate issues in HSQC spectra and its ability to align peaks even if they are manually difficult to separate. The proposed method offers not only a powerful automated tool for peaks detection and alignment but also a parametric representation of the NMR 2D spectrum which will be used for spectrum indexation in the next chapter.

On the other hand, the second method, dedicated to the fMRI active zone alignment, relies on the canonical system approach. This approach allows us to find the most object natural pose and to visually align similar objects in the same manner. In order to integrate our *a priori* knowledge, we proposed a new method for spherical symmetry estimation based on the non-linear PCA that is well suited to model the reflection symmetry of the cortical active zones. The synthetic validation of the proposed active zone alignment scheme has shown that the modeling of the partial spherical symmetry has increased the robustness of the proposed registration scheme with comparison to the continuous PCA method which does not take into consideration this *a priori* knowledge.

CHAPTER 4

Object coding and similarity measurement

Contents

Symbols:

X	Candidate peak amplitudes
R	Reference peak amplitudes
Y	Observed spectrum
Y_{ref}	Reference observed spectrum
r_o	Observed peaks ratio
r_e	Expected peaks ratio
$\mathcal{G}$	Gamma distribution
(α, β)	Hyperparameters of the Gamma distribution
Γ_X	Inter-metabolite correlation matrix
I	Identity matrix
Φ	Standard Gaussian cumulative distribution
B	Noise
Γ_b	Noise covariance matrix
c_g	Gaussian copula
$h(i, j)$	Peak shape filter
$\gamma^Y = (\gamma_1^Y, \gamma_2^Y)$	Shape hyperparameters of peaks within the spectrum Y
$hyp_{1,2,3}$	Hypotheses modeling the spectrum imprecision
$\hat{X}$	An estimation of the theoretical spectrum X
f	Π -membership function used for candidate metabolites selection
f_1	S-membership function associated to criterion 1
f_2	S-membership function associated to criterion 2
f_M	Π -membership function associated to criterion 3
g	Trapezoidal membership function associated to global criterion
μ_M	Metabolite reliability degree associated to f
μ_{cr1}	Metabolite reliability degree associated to criterion 1
μ_{cr2}	Metabolite reliability degree associated to criterion 2
μ_{cr3}	Metabolite reliability degree associated to criterion 3
M_e	Metabolite belonging to exception list
$recall$	Recall measurement
$Precision$	Precision measurement
$\mathtt{E}[X]$	Expected value of X
$\mathtt{X}$	Theoretical spectrum
$\mathtt{X}_{ref}$	Reference theoretical spectrum
Y	Observed spectrum
Y_{ref}	Reference observed spectrum
M	A metabolite
$\hat{M}$	An estimation of M
C_M	Candidate Metabolite set

Symbols:

O	A 3D Object
T	Object triangle mesh
$M_{o,q,r}$	The geometric moments of order $(o+q+r)$
A_p	Area of the triangle associated to $p \in T$
$g_{i,j,k}$	Gravity center of object portion number (i, j, k)
$c_{i,j,k}$	Object Gaussian descriptor associated to $g_{i,j,k}$
$G(.)$	Gaussian transformation
$d(.,.)$	The euclidian distance
c^i	Descriptors set of object O^i
$\Delta(O^1, O^2)$	Object similarity measurement between O^1 and O^2
t_m	The m^{th} surface object portion
g_m	The m^{th} point belonging the unit shpere
$GG_t(.)$	Generalized Gaussian transformation

Acronyms:

HR-MAS	High Resolution Magic Angle Spinning
HSQC	Heteronuclear Single Quantum Coherence spectrum
fMRI	functional Magnetic Resonance imaging
PCA	Principal Component Analysis
3DGD	3D Gaussian Descriptor
3DGGD	3D Generalized Gaussian Descriptor
PSNR	Peak Signal to Noise Ration
MCMC	Monte Carlo Markov Chain
GA	Genetic Algorithm
ML	Maximum Likelihood

Introduction

The Object coding and similarity measurement is the second step in the proposed object indexing scheme (Fig. 2.15). On one hand, the object coding task consists in coding different objects into a compact description. This compact description allows us to accelerate large database queries. On the other hand, the similarity measurement task consists in establishing the object similarity measurement procedure. In other word, this task returns to find and to group the most similar objects within a given medical signal group/population query.

In this chapter we present in section 4.1 an overview of the most used object coding and similarity measurement methods. In section 4.2, we develop a novel scheme for peak similarity measurement. Indeed, the step of object encoding is unnecessary since the peaks can only be described by three parameters (location, amplitude and shape) and are therefore already parsimoniously presented. For the similarity measurement, we propose a new method based on the combination of Bayesian theory and the fuzzy sets theory allowing us to handle the uncertainty and fuzzyness that characterize the observations and to inject our *a priori* knowledge into the inference model.

In section 4.3, we propose a new coding method based on generalized Gaussian transformation to reliably describe the topology of the active zones. In particular, we show that the proposed method provides not only a compact representation of the object in its space but also a signature faithful to its shape. We also propose a similarity measurement robust to small displacements and variations of objects. We show that the use of the proposed algorithms allow us to get more accurate object coding and similarity measurement results compared to the existing schemes.

4.1 Overview of object coding and similarity measurement algorithms

4.1.1 Object coding

Shape distributions

Osada et al. [Osada02] propose five distribution forms to code a 3D object. The object is assumed to be triangular meshing (Fig.4.1). Therefore, the object surface is composed of a set of inter-connected triangles where each triangle consists of three vertices and a gravity center. The considered measurements are:

- The angles between three points on the surface (Fig.4.2.(b)),
- The distance between the mass center and a point of the object (Fig.4.2.(c)),
- The distance between two points (Fig.4.2.(d)),
- The area square root of a triangle formed by three points (Fig.4.2.(e)),
- The volume cube root of the tetrahedron formed by four points (Fig.4.2.(f)).

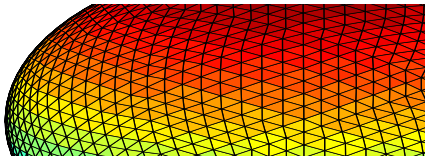


Figure 4.1: An object triangular mesh.

All considered points are randomly selected on the object triangular facets. The main advantages of this approach are the descriptor compactness and the calculation rapidity. Moreover, since the considered measurement does not depend on the object coordinate system, the shape distributions method is invariant to geometric transformations (rotation and translation). However, its use is more suited to search objects of similar overall shape since it is not able to discriminate small variations in the object meshing [Ohbuchi03].

Geometric moments

The geometric moment approach consists in projecting a characteristic function h, that models a 3D object, on the family of functions $x^o y^q z^r$, $(o,q,r) \in \mathbb{N}^3$. The geometric moments of order $(o+q+r)$ are denoted by M_{oqr} and calculated as follows:

$$M_{oqr} = \int_{p \in T} h(p)\, x^o y^q z^r \, dp \tag{4.1}$$

where p is a point belonging to the triangular mesh T of the object. In particular, the geometric moments of order one and two are used to calculate the normalization hyperparameters of the 3D object (the object gravity center and the three principal axes). In the context of coding, the description of a 3D shape by the geometric moments was proposed in [Paquet99]. In this work, the geometric moments are obtained by the following equation:

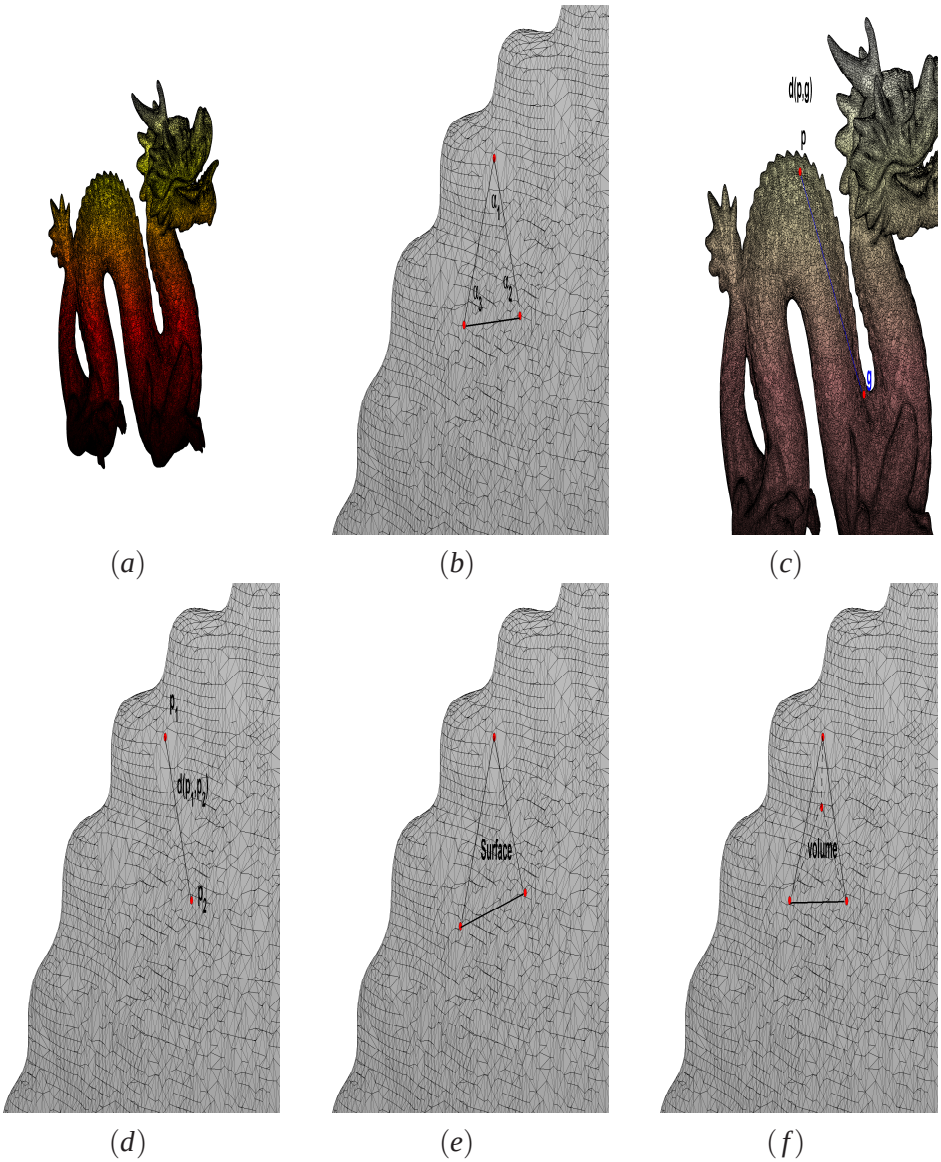
$$M_{oqr} = \sum_{p \in T} A_p \, (x_p - x_g)^o (y_p - y_g)^q (z_p - z_g)^r \tag{4.2}$$

where A_p (*resp* $g_p = (x_p, y_p, z_p)$) is the area (*resp* the gravity center) of the triangle associated to $p \in T$ and $g = (x_g, y_g, z_g)$ is the object gravity center.
Note that the use of geometric moments is not the best object coding method. Indeed, a comparative study in terms of performance made by Varnic and Saupe [Vranic00] on different bases of 3D objects (*e.g*; the Princeton Shape Benchmark) which contains a database of 3D object models collected from the World Wide Web (Fig. 4.3) shows the limits of this method particularly for complex 3D shape coding [Shilane04].

Shape histogram

This method, proposed by Ankerst et al. [Ankerst99], consists in uniformly partitioning the object space into three representations:

1. SHELLS: This partitioning allows to overcome any object rotation through concentric shells around the center of the object (Fig.4.4.(b)). Moreover, to cover the entire



(a) (b) (c)

(d) (e) (f)

Figure 4.2: The five considered measurements of the shape distribution method applied on the (a) Dragon object: (b) 1-the angles between three points on the surface, (c) 2-the distance between the mass center and a point of the object, (d) 3-the distance between two points, (e) 4-the area square root of a triangle formed by three points and (f) 5-the volume cube root of the tetrahedron formed by four points.

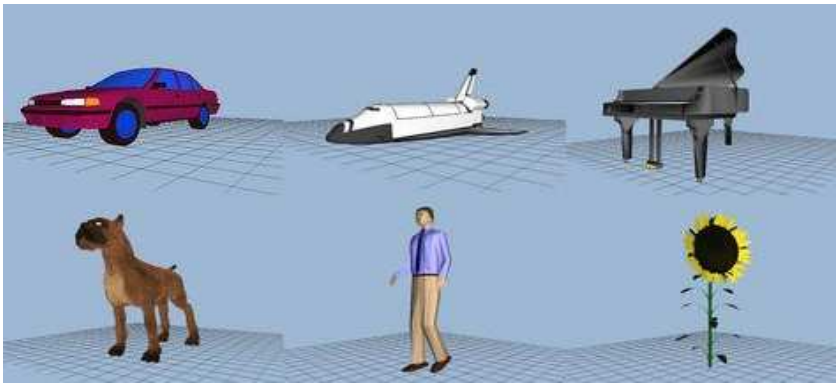


Figure 4.3: Examples of Princeton Shape Benchmark 3D objects.

model, the last shell has no top boundary.

2. SECTORS: This partitioning allows to project the facets of a regular polyhedron on the unit sphere through an angular decomposition with homogeneous size (Fig.4.4.(c)) .

3. SECSHEL: This partitioning is the combination of SHELLS and SECTORS partitions (Fig.4.4.(d)).

Finally, for each bin an histogram of mesh triangle center within this bin is calculated and the union of all histograms constitutes the signature of the object.

The 3D Gaussian Descriptor

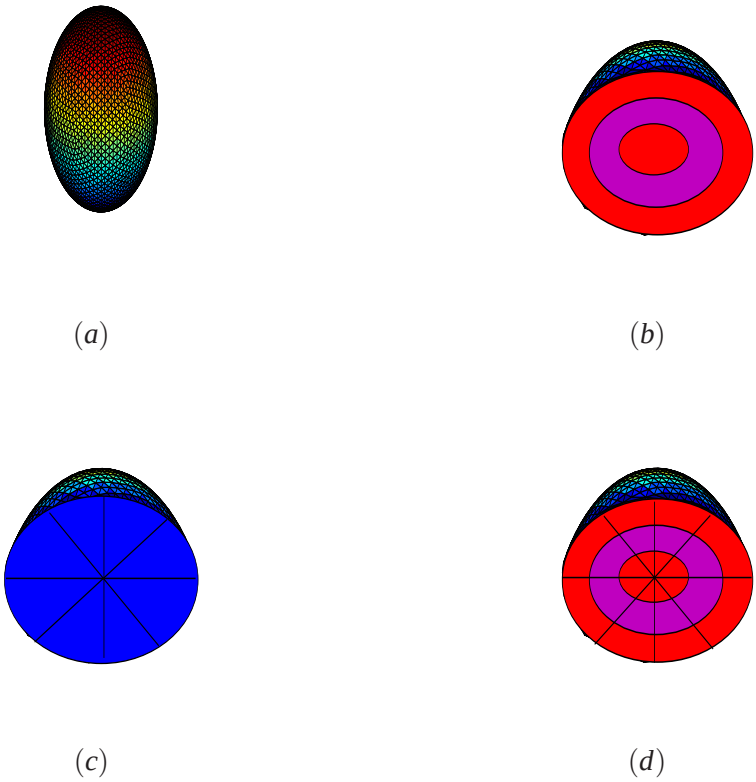
The 3D Gaussian Descriptor (3DGD) introduced by Chaouch [Chaouch09] relies on the object Gaussian transformation which is derived from the Gaussian law. Indeed, the Gaussian transform is a real application defined on a set of defined point in space and obtained by a summation over the surface of the 3D object. The Gaussian transformation denoted G on a point q of the space is given by the following expression:

$$G(q,T,\sigma) = \int_{p\in T} \exp\left(\frac{-(p-q)^2}{2\sigma^2}\right) dp \tag{4.3}$$

where d is the euclidian distance, $p \in T$ (T the object triangular mesh) and $\sigma > 0$ is a positive real.

In order to code a 3D object, author proposes the partitioning of the object into N^3 cells and then the calculation for each one's center $g_{i,j,k}$, $i = 1...N,\ j = 1...N,\ k = 1...N$ (Fig. 4.5). Then, the author assigns to each cell a *characteristic value* $c_{i,j,k}$ as following:

$$c_{i,j,k} = \sum_{p\in T} A_p \exp\left(\frac{-(p-g_{ijk})^2}{2\sigma^2}\right) \tag{4.4}$$



(a) (b)

(c) (d)

Figure 4.4: The shape histogram descriptors: (a) Spherical space, (b) 3 shell bins, (c) 6 sector bins, (d) combination of 3 shell bins and 6 sector bins.

where A_p is the area of the triangular facet associated to p (Fig. 4.6).

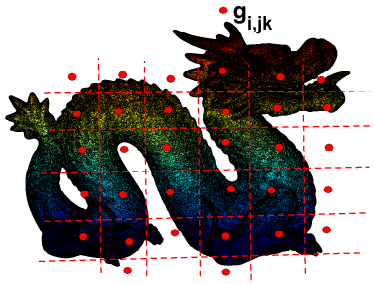


Figure 4.5: The 3D object partitioning.

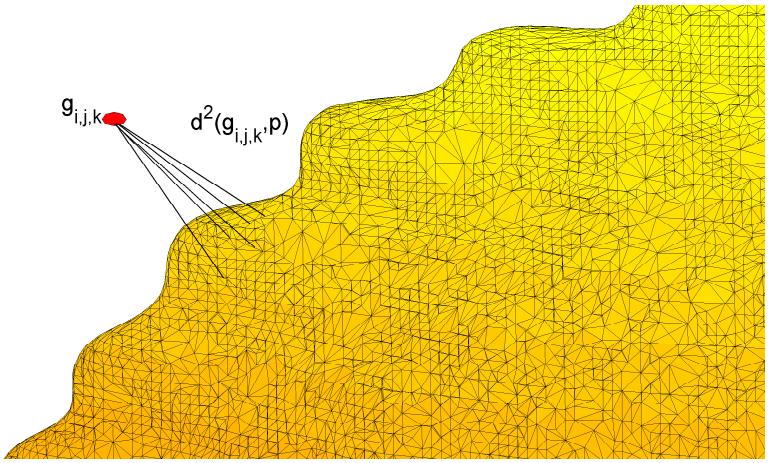


Figure 4.6: Contribution of object surface point in the local description of $g_{i,j,k}$.

Thus, the 3D object is codified with $c = \left[c_{i,j,k}\right]_{i,j,k \in N}$. Note that the 3D Gaussian Descriptor (3DGD) has showed its effectiveness compared to other methods such as the shape histogram method and was ranked first on the Princeton Shape Benchmark database [Chaouch09].

4.1.2 Similarity measurement

The similarity measurement consists in establishing the similarity between two codified objects, it means the most similar objects within the database toward the object we are

looking for (query object).

Threshold approach

The threshold schemes are often used to accommodate differences between object descriptors. To this end, it is sufficient to experimentally set a threshold *thres* and to evaluate the differences between two objects O^1 and O^2 as following:

$$\Delta(O^1,O^2) = \begin{cases} 1 & \text{if } d(c^1,c^2) \leq thres \\ 0 & \text{otherwise} \end{cases} \tag{4.5}$$

where c^1 and c^2 are the object descriptors of O^1 and O^2 respectively and $d(.,.)$ is the euclidian distance. Thus, O^1 and O^2 are assigned together if the function $\Delta(O^1,O^2)$ is equal to 1. The threshold based approach is a rudimentary technique but it found some success in applications such as in metabolite identification [Zheng07, Xia08] thanks to its simplicity and speed. Nevertheless, thresholds set may strongly affect the robustness of the similarity measurements.

Distance minimization approach

The similarity between two objects O^1 and O^2 can be calculated using the euclidian distance as follows:

$$\Delta(O^1,O^2) = d(c^1,c^2) \tag{4.6}$$

Thus, the most similar object O^s to query object O within the database is the given by:

$$O^s = \underset{O_i}{\operatorname{argmin}} \left(d(O,O^i)\right) \tag{4.7}$$

In order to be less sensitive to small displacements or minor geometric variations, [Chaouch09] introduced a new similarity measurement that minimizes the distance between adjacent pairs of components:

$$\Delta(O^1,O^2) = \frac{1}{N^3} \sum_{i,j,k=1}^{N} min \left(\begin{cases} (c^1_{i,j,k} - c^2_{i,j,k})^2 \\ (c^1_{i,j,k} - c^2_{i',j',k'})^2, \quad i',j',k' \in V(i,j,k) \end{cases} \right) \tag{4.8}$$

where $V(i,j,k)$ is the $3 \times 3 \times 3$ neighborhood of the descriptor $c^2_{i,j,k}$.

Supervised learning approach

Supervised learning approach relies on the link function establishment between object descriptors and expected results (similar or no similar objects) via a training step. The widely common used methods are the Support Vector approach SVM [Bruzzone06] and the Artificial Neural Networks [Bate98]. Thus, the similarity measurement task is equivalent to a classification task. In other words, for each object O^i belonging to the database, a classification function $\mathfrak{f}(.)$ is established using a training dataset which contains objects

with known similarity results. Therefore, a test object O^t is assigned to the object O^i (*i.e.*; O^t and O^i are considered as similar) if $\mathfrak{f}(O^t) = 1$. As one can remark, the supervised learning approach is a generalization of the threshold approach since it allows a non-linear behavior of the assignment function $\mathfrak{f}(.)$.

4.1.3 Retained approaches

We recall that each spectrum is composed of several peaks that are scattered within the spectrum. These peaks are the responses of metabolite presence. Therefore, peaks belonging to a given metabolite have common properties. In order to improve the peak coding and similarity measurement task, these properties should be modeled and injected into the proposed scheme. Thus, it is better to manipulate metabolites than single peaks. Moreover, since a metabolite is described by a set of few peaks with specific characteristics (locations, shapes and amplitudes) the metabolite coding step is no longer necessary (the metabolite is already compactly presented) and finally the peak coding and similarity measurement task is simply equivalent to a *metabolite identification* task (*e.g*, identify the most similar metabolite to a given metabolite).

Concerning the fMRI image, unlike peaks, an active zone is a dense form (consisting of neighboring voxels) and therefore a compact representation of this active zone should be carried out. Among different object coding methods, the 3D Gaussian Descriptor (3DGD) introduced by [Chaouch09] has showed its effectiveness compared to other methods and was ranked first on the Princeton Shape Benchmark database. For this reason, the Gaussian transformation is retained in this work for fMRI active zones

4.2 Metabolite similarity measurement (HSQC)

In the literature, we distinguish two methods [Zheng07, Xia08] that deal with the metabolite similarity measurement. In these works, two metabolites are considered as similar if their peaks appear at the same locations. To this end, these methods use threshold schemes to accommodate the chemical shift differences between different metabolites. Nevertheless, the choice of the thresholds may strongly affect the robustness of the similarity measurement method. To overcome this problem, we propose the use of the fuzzy set theory which is well appropriate to handle fuzzy situations [Waltz90] and hence avoid a binary reasoning (similar or not similar). Moreover, in order to lead to a more robust, more accurate and efficient similarity measurement method, additional constraints, other than peaks location, such as the knowledge of the biopsy compositions or peak amplitudes should be integrated into the assignment scheme. These *a priori* knowledge are available in a metabolites library denoted the "corpus". Indeed, the corpus contains different metabolites expected to be present within the spectra as well as the characteristics of each metabolites (expected peak locations and peaks amplitudes). The corpus can be obtained by two ways. In the first one, a reference spectrum is manually annotated (*i.e*; the spectrum annotation task is equivalent to metabolites identification task) and then the annotation results are used to identify metabolite presented in other spectra obtained from the same type of biopsy. In

the second one, the corpus does not depend on the biopsy. For example, it can be the annotation results of spectra obtained from pure compounds. Indeed, since the metabolites are independent of the treated tissue type, the corpus origin has no influence on the metabolite identification scheme. Moreover, in order to increase the metabolite similarity measurement accuracy, it is better to consider a multi-spectra metabolite identification.

4.2.1 Metabolite identification

The proposed fuzzy metabolite similarity measurement scheme is divided into three steps:

1. Randomly choose a peak denoted the reference peak,
2. Find the candidate metabolites that may contain the selected reference peak,
3. Find the right metabolite candidate according to different criteria exposed in the following.

Candidate metabolites selection

The first selection criterion, in the metabolite similarity measurement scheme, is the peak location. Indeed, each metabolite is composed of one or several peaks at very specific frequency coordinates (carbon-^{13}C chemical shift in x axis and proton-^{1}H chemical shift in y axis of the spectrum image). However, we recall that peaks can slightly be shifted from their expected positions. To overcome this, we assume that the peak membership to a metabolite is a fuzzy concept. In fact, a given peak may belong to several metabolites with a membership degree denoted μ_M. The value of this membership degree depends on both the expected and the measured peak location.

For the transformation from the hard to fuzzy domain, a Π membership function denoted f is used. The expression of f is given below:

$$f(x;a_1,b_1) = \frac{1}{1+\left(\frac{x-a_1}{b_1}\right)^2} \tag{4.9}$$

Fig 4.7.(a) shows an example of Π membership function for a given pair (a_1,b_1). For a given metabolite M from the corpus, the membership degree μ_M, using a Π membership function, is expressed as:

$$\mu_M = f((i-i_M)^2+(j-j_M)^2;a_1,b_1) \tag{4.10}$$

where (i,j) is the peak measured position and (i_M,j_M) is the expected peak position (available from the corpus). The hyperparameters (a_1,b_1) are automatically estimated with the Genetic Algorithm procedure (see Appendix B). Indeed, the GA aims at estimating the model hyperparameters using a training dataset that contains metabolites with known similarity results. Once the candidate metabolites (set of metabolites such that $(a_1-b_1)<\mu_M<(a_1+b_1)$ denoted henceforth C_M) have been selected, we address in the next part the problem of the right candidate identification.

Right metabolite identification

Our challenge in this study is to properly model the metabolite profile using the *a priori* knowledge in order to lead to a optimal selection of the right metabolite. To this end, we define three criteria to be respected in the metabolite scheme identification:

Criterion 1

Theoretically, and with respect to the reproducibility principle, the ratio of two peaks belonging to the same metabolite must be the same for any observation. But this rule is not perfectly verified in practice (due to the degradation of the tissues used in the biopsy during acquisition and/or the acquisition conditions that are not necessarily the same for all observations). Therefore, the modeling of the ambiguity introduced by these disturbances is essential to avoid false negative identification (assigning a peak to a wrong candidate). Let us denote r_o as the observed ratio between the reference peak and a new candidate peak and r_t as the expected ratio given by the corpus. The more these ratios are close, the more the metabolite identification is reliable. To model this reliability, we define an *S* type function f_1 as the membership function. The expression of f_1 is :

$$f_1(x;a_2,b_2,c_2) = \begin{cases} 0 & x < a_2 \\ \frac{(x-a_2)^2}{(b_2-a_2)(c_2-a_2)} & a_2 \leq x < b_2 \\ 1 - \frac{(x-c_2)^2}{(c_2-b_2)(c_2-a_2)} & b_2 \leq x < c_2 \\ 1 & \text{otherwise} \end{cases} \tag{4.11}$$

Where $a_2 < b_2 < c_2$ are the hyperparameters of the f_1 function. Fig 4.7.(b) shows an example of *S* membership function for a given triplet (a_2,b_2,c_2).
The proposed reliability degree denoted μ_{cr1} is then given by Eq. 4.12. The hyperparameters of this function are estimated using the Genetic algorithm (see Appendix B).

$$\mu_{cr1} = f_1(min(r_t,r_o)/max(r_t,r_o);a_2,b_2,c_2) \tag{4.12}$$

Criterion 2

In this method, we consider simultaneously *N* biopsies (multivariate analysis). We assume that the peak amplitudes follow a Gamma distribution. Indeed, the major advantage of this distribution is that the shape parameters allow the fitting of spectral data that possibly present some sparsity and/or a background [Dobigeon09]. The criterion 2 models the likelihood between the observed peaks and a given metabolite belonging to the corpus. The Gamma distribution [Dobigeon09] $\mathcal{G}$ is expressed as:

$$\mathcal{G}(y_i;\alpha,\beta) = y_i^{(\alpha-1)} \frac{\beta^\alpha}{\Gamma(\alpha)} exp^{(-\beta y_i)} \quad y > 0 \tag{4.13}$$

where y_i stands for the amplitude, and $\alpha,\beta > 0$ represent the shape and the inverse scale parameters respectively. The likelihood term has to be expressed in a multidimensional

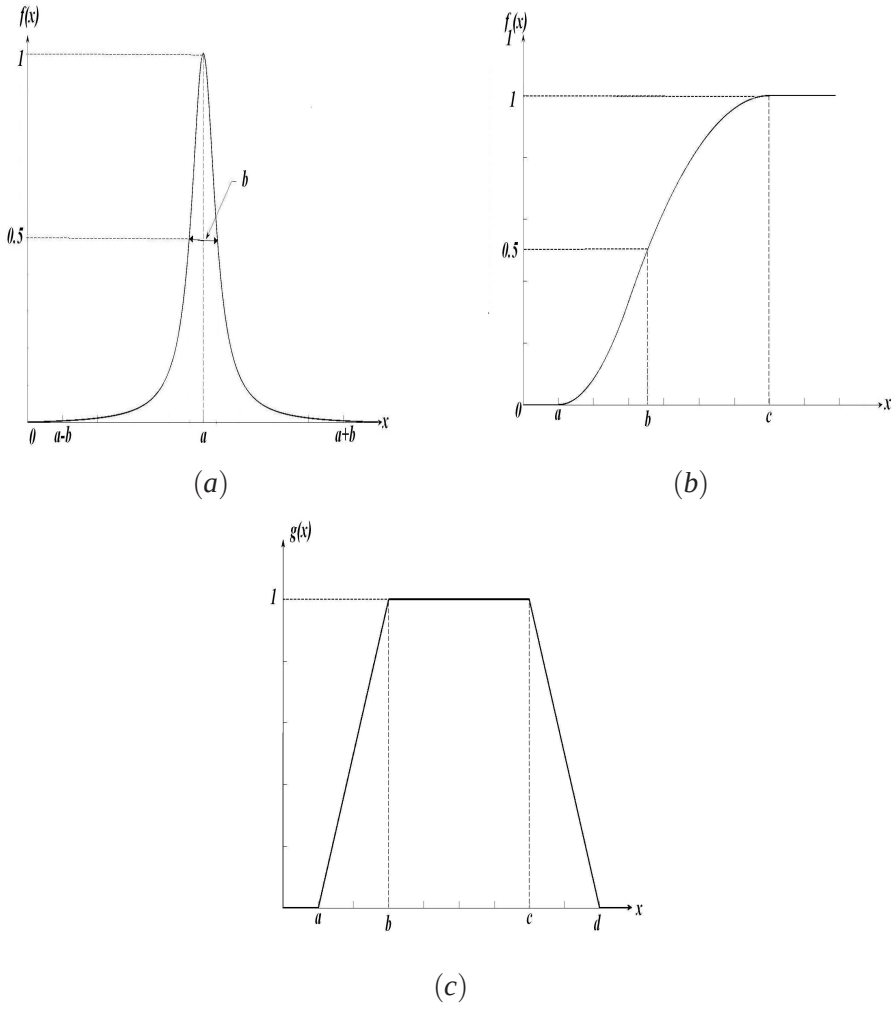


Figure 4.7: Different fuzzy membership functions used in the metabolite identification scheme: (a) Π membership function, (b) *S* membership function and (c) trapezoidal membership function. The hyperparameters of these fuzzy membership functions are estimated with the Genetic algorithm using a set of training spectrum databases. Each database contains several spectra with known metabolite identification results. The simulations show that these fuzzy membership function hyperparameters are almost the same for all databases.

way. Nevertheless, an analytical expression of the multidimensional Gamma distribution is not available. To overcome this problem, we propose the use of the copula theory which offers an elegant way to model the dependency between the different observations (the metabolite realizations over several spectra) and hence to access to the multidimensional Gamma distribution [Joe97].

Several studies show the effectiveness of the Gaussian copula c_g to handle dependency [Joe97]. In order to properly take into account the metabolite dependency into the similarity measurement scheme, we adopt this copula: $\forall\, \mathbf{y} = (y_1, \cdots, y_N) \in \mathbb{R}^N$,

$$c_g(\mathbf{y}, \Gamma) = |\Gamma|^{-\frac{1}{2}} \exp\left[-\frac{\tilde{\mathbf{y}}^T(\Gamma^{-1} - I)\tilde{\mathbf{y}}}{2}\right] \tag{4.14}$$

where $\tilde{\mathbf{y}} = (\Phi^{-1}(y_1), \cdots, \Phi^{-1}(y_N))^T$ with $\Phi(.)$ the standard Gaussian cumulative distribution, Γ is the inter-spectra correlation matrix and I the $N \times N$ identity matrix. Let us now denote $R = (R_1, ..., R_i, ..., R_N)$ the amplitude of reference peak over the N considered spectra and $X = (X_1, ..., X_i, ..., X_N)$ the amplitude of a candidate peak over the N spectra and let a_i be the theoretical ratio between R_i and X_i: $X_i = a_i \times R_i$, $i = 1...N$. In our case, R_i follows a gamma distribution with hyperparameters α_i and β_i. Under these assumptions, (X_i/M) follows a gamma distribution with hyperparameters α_i and $\frac{\beta_i}{a_i}$. Using the Gaussian copula c_g, the likelihood is then given by :

$$P(X/M) = f_{ga}(X_1; \alpha_1, \frac{\beta_1}{a_1}) \times ... \times f_{ga}(X_N; \alpha_N, \frac{\beta_N}{a_N}) \times c_g(X, \Gamma_X)$$

where Γ_X is the inter-metabolite correlation matrix. The hyperparameters of $P(X/M)$ are estimated using an MCMC procedure [Smith93]. We use this last expression to build the reliability degree: considering an S membership f_2 as, the proposed reliability degree denoted μ_{cr2} is given by:

$$\mu_{cr2} = f_2(P(X/M); a_3, b_3, c_3) \tag{4.15}$$

Thus, the higher the likelihood value $P(X/M)$ is great, the more μ_{cr2} is close to 1 and the more the probability of the metabolite M to be the right metabolite is large.

Criterion 3

This criterion deals with the variations of the observed peak chemical shifts from their theoretical positions defined in the corpus. We use the same membership function as in Eq.4.9. The proposed reliability degree denoted by μ_{cr3} is hence expressed as:

$$\mu_{cr3} = f_M(((i - i_M)^2 + (j - j_M)^2); a_1, b_1) \tag{4.16}$$

where (i, j) stands for the observed peak position and (i_M, j_M) for the theoretical peak position.

Global criterion

We combine the 3 reliability degrees previously defined, to expect the best candidate in the following way:

$$\hat{M} = \underset{M \in C_M}{argmax}(\prod_{k=1}^{K} \mu_{cr1}(X_k).\mu_{cr2}(X_k).\mu_{cr3}(X_k)) \tag{4.17}$$

As one can remark, given the cost function definition (Eq. 4.17), a solution of the metabolite identification always exists which is not necessarily correct (*i.e*; a metabolite can be not present in a spectrum). In order to reduce the number of false positive identification, the reliability of the membership degrees for all different criteria of $\hat{M}$ should also be quantified. In other word, not all solutions of (Eq. 4.17) are acceptable. Therefore, we propose a last fuzzy decision function that takes as input the criterion value of $\hat{M}$ and as output their reliability. This reliability is quantified by the membership degree of $\hat{M}_B$ to the fuzzy set: the metabolites biomarkers set. We opted for the trapezoidal function denoted by g as a membership function. The expression of g is given by :

$$g(x; a_4, b_4, c_4, d_4) = \begin{cases} (\frac{x-a_4}{b_4-a_4}) & a_4 \leq x < b_4 \\ 1 & b_4 \leq x < c_4 \\ (\frac{d_4-x}{d_4-c_4}) & c_4 \leq x < d_4 \\ 0 & \text{otherwise} \end{cases} \tag{4.18}$$

where $a_4 < b_4 < c_4 < d_4$.

The allure of g for a given quadruple (a_4, b_4, c_4, d_4) is presented in Fig. 4.7.(c). Another time, the Genetic algorithm was used to estimate this quadruple. The decision to identify $\hat{M}$ as a right metabolite depends on the output of the function g. As one can remark, g function depends on the quadruple (a_4, b_4, c_4, d_4) as well as on (a_2, b_2, c_2), (a_3, b_3, c_3) and (a_1, b_1). The decision function is expressed as:

$$g(\prod_{k=1}^{K} \mu_{cr1}(X_k).\mu_{cr2}(X_k).\mu_{cr3}(X_k); a_4, b_4, c_4, d_4, a_1, b_1, a_2, b_2, c_2; a_3, b_3, c_3) \tag{4.19}$$

If $g\left(\prod_{k=1}^{K} \mu_{cr1}(X_k).\mu_{cr2}(X_k).\mu_{cr3}(X_k)\right) = 1$ then $\hat{M}$ is selected. It is important to note that under some circumstances, such as the changes in the nucleus environment features, some peaks of a given metabolite may not be present. Thus, such exceptions should be taken into account in the annotation scheme due to *a priori* knowledge. To this end, an exception list is made by physicians. This list contains the peak set that may not be present for each metabolite. Let us denote by X_k, $k = 1...K$ the peaks of a given metabolite M_e belonging to the exception list and by X_{ke} the peak which could not be present. The exception handling is defining in Alg.1.

Algorithm 1 Exception handling algorithm

Input=X_{ke} and M_e.

1-Calculate the membership function $\mu_{cr3}(X_{ke})$ (Eq. 4.16),

2-If $\mu_{cr3}(X_{ke}) \neq 0$, no change on M_e composition is made,

3-If $\mu_{cr3}(X_{ke}) = 0$, the new composition of M_e is X_k, $k = \{1...K\} - \{ke\}$,

Output=new composition of M_e.

An overview diagram of the metabolite identification chain is presented in Fig. 4.8.

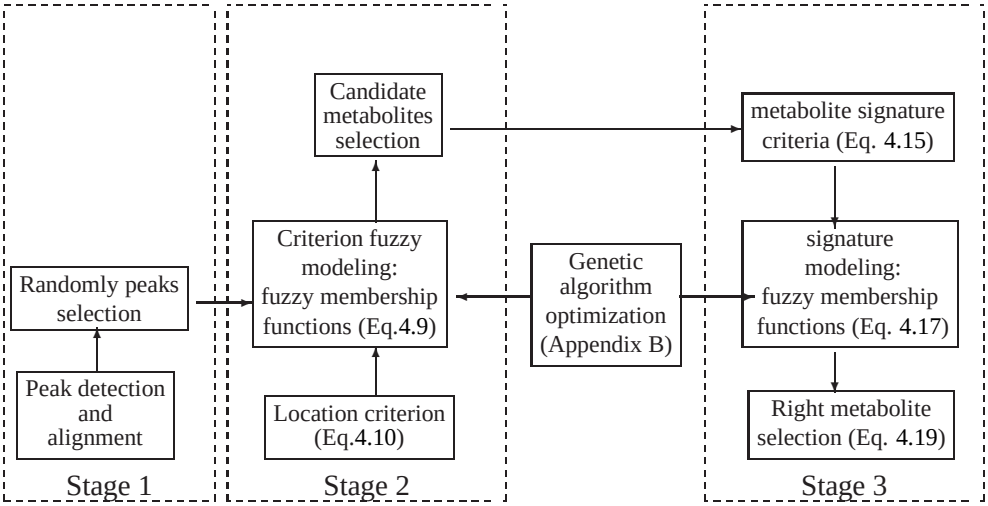


Figure 4.8: Overview diagram of the metabolite identification chain.

4.2.2 Metabolite identification validation

This part describes some biomarker identification results obtained with the proposed scheme. This method was applied on synthetic spectra designed to fit the characteristics of the ^{1}H - ^{13}C HSQC HR-MAS spectra as well as on some real spectra. More results on real spectra are presented in chapter 6.

In order to validate and emphasize the benefits of the proposed approach, we have retained the *recall* and the *precision* measurements for the synthetic data validation:

$$recall = \frac{TP}{TP+FN}; \quad precision = \frac{TP}{TP+FP}$$

where TP stands for the number of true positive identifications, FN the number of false negative identifications and FP the number of false positive identifications.

The main advantage of using simulated data is that we perfectly know the characteristics of the data such as the number of peaks presented in every spectrum and the peaks chemical shifts values. For this, we firstly generate a synthetic theoretical 2D spectrum image $\mathtt{X}_{ref}$ ($M = 500$ pixels by $N = 500$ pixels) which contains $N_p = 500$ peaks corresponding to two hundred metabolites $N_m = 200$. The positions of different peaks and the hyperparameters of the shape filter for each peak are randomly generated. This synthetic

spectrum will be used as reference to register other synthetic spectra. Nine other synthetic theoretical 2D spectrum images $\mathtt{X}_i$, $i = 1...9$ are generated from $\mathtt{X}_{ref}$ by applying a displacement vector at each peak of $\mathtt{X}_{ref}$. The values of chemical shift vectors are assumed to be random and following a Gaussian distribution with zero mean and a variance matrix $\Gamma_d = \begin{bmatrix} 0.02 & 0 \\ 0 & 0.25 \end{bmatrix}$. The new shape peak hyperparameter for each peak is calculated by adding a zero-mean Gaussian random noise of variance 0.005. A spatially zero mean correlated noise B was added to each spectrum to obtain the synthetic spectra used in the simulation Y_{ref}, Y_i, $i = 1...9$.

Figure 4.9 shows the synthetic theoretical 2D spectrum image $\mathtt{X}_{ref}$ and the reference synthetic spectrum Y_{ref} with PSNR = 30dB where:

$$PSNR = 10\, log_{10}(max(X_{ref})^2/\mathtt{E}[(B)^2]) \tag{4.20}$$

This value of PSNR was chosen to fit at best the real spectra. In fact, the PSNR of the real spectrum is $\sim$ 30dB. Concerning the used *corpus*, it summarizes the ground truth of Y_{ref}. It is important to note that all the spectra are presented as contour plots with the same number of level.

In order to emphasize the benefit of the proposed approach, three different similarity measurements methods were applied to the synthetic spectra Y_i, $i = 1...9$ with different values of PSNR: our identification method , an SVM method [Camps-Valls05] and a threshold method [Xia08] . The metabolites identification results of Y_i are presented in Table 4.1. First, as one can see, the proposed method is enough robust to a hight level of noise. In fact, even with a PSNR= 23dB, the *recall* and the *precision* measurements are still close to 90%. Secondly, we can easily observe that the proposed method performed better than the SVM method which does not take into account the *a priori* knowledge.

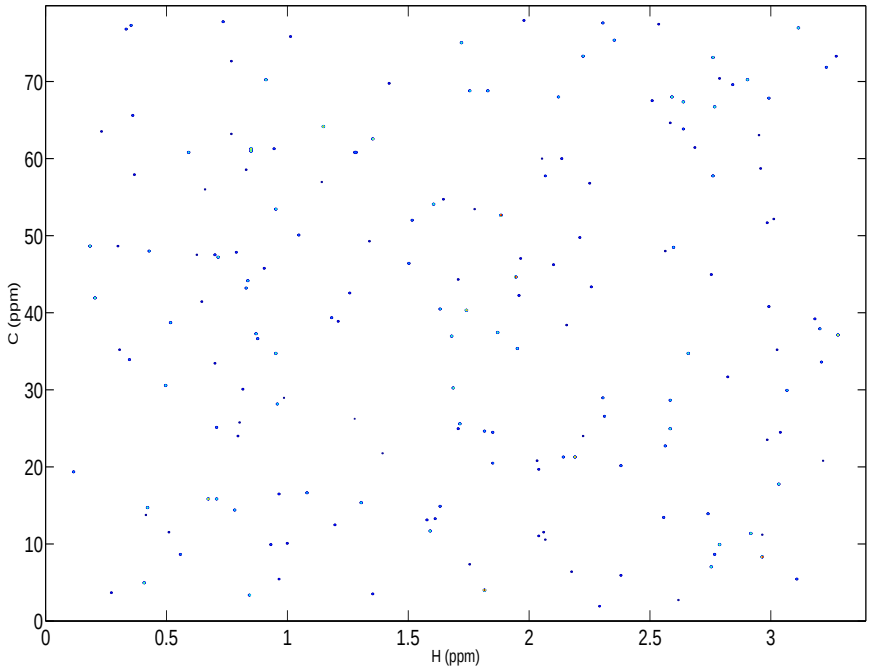
	Proposed method		SVM		Threshold method	
PSNR	*recall*(%)	*precision*(%)	*recall*(%)	*precision*(%)	*recall*(%)	*precision*(%)
30*dB*	**93.87**	**95.11**	90.38	91.72	81.16	78.01
28*dB*	**92.42**	**94.82**	88.50	89.61	78.98	76.12
25*dB*	**92.84**	**94.64**	82.11	86.90	75.77	74.25
23*dB*	**89.02**	**90.18**	83.02	84.66	74.02	71.88

Table 4.1: The average *recall*(%) and *precision*(%) obtained with: our identification method, the SVM method, a threshold method on synthetic spectra.

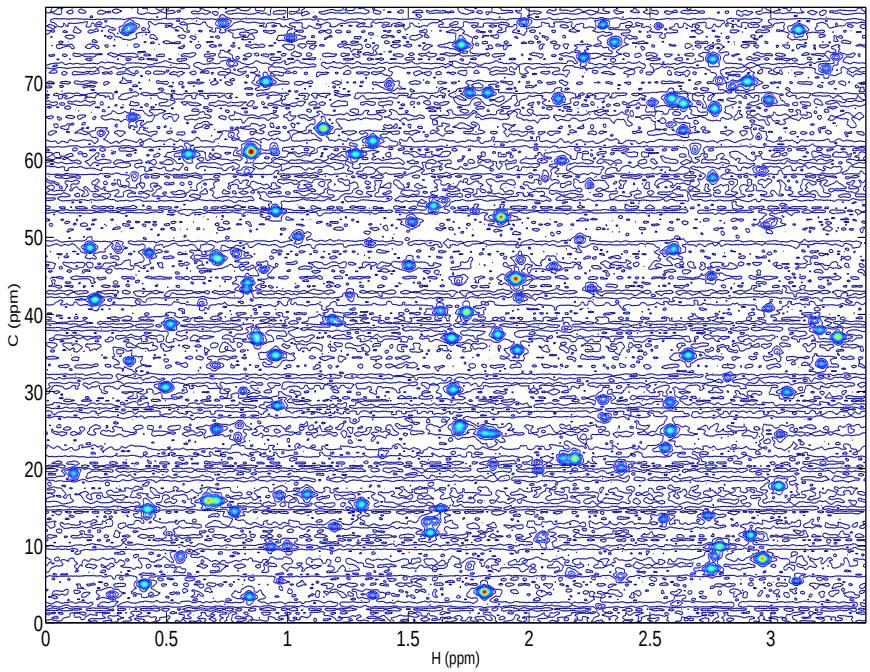
Figure 4.10 presents the metabolites identifications results on the same region of Y_1 and Y_2. Each peak p belonging to a given metabolite M is labeled with (p,M). As we can see, some peaks (like peak number (6,1), (3,2) and (9,1)) are visually very difficult to identify, yet our method is able to properly assign them.

The performance of the proposed method was evaluated in the case of missing peaks. The missing peaks were simulated by removing peaks of each metabolite randomly with 0%, 20% and 50% probabilities. We can distinguish two cases. If the modified metabolite belongs to the exception list, the exception handling algorithm (Algorithm 1) is then performed. In the other case, we assume that the peak absence of a given metabolite is the result of the metabolite absence. We have applied our annotation scheme on the synthetic spectra. The results show that every modified metabolite which does not belong to the exception list was not identified and indeed every metabolite belonging to the exception list was identified. In other words, we obtained the same results as those presented in Table 4.1.

Figure 4.11 displays the metabolites identification results on the same zone of a healthy and a cancerous spectra. As we can see, some metabolites (like metabolite number (5,2)) are visually very difficult to identify. Yet our method is actually able to correctly identify them even with a high noise level.

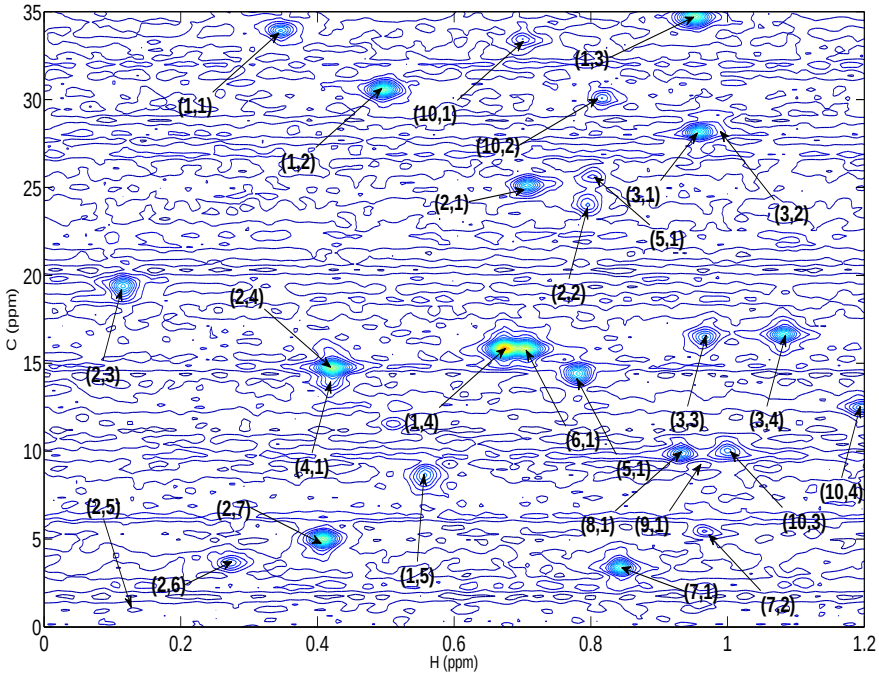


(a)

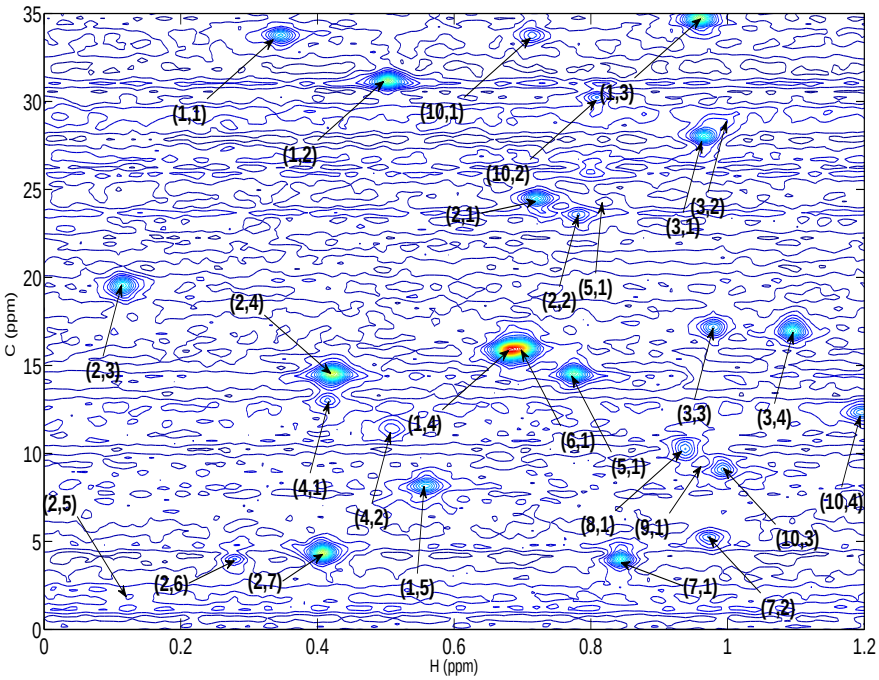


(b)

Figure 4.9: (a) The synthetic theoretical 2D spectrum X_{ref}, (b) a contours plot of the reference synthetic observed spectrum Y_{ref} (PSNR = 30dB).

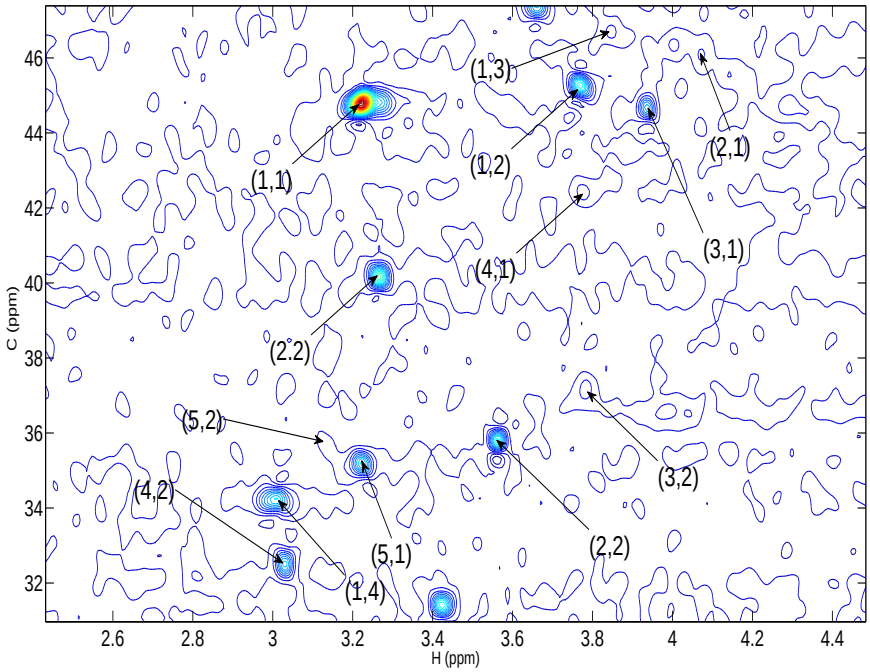


(a)

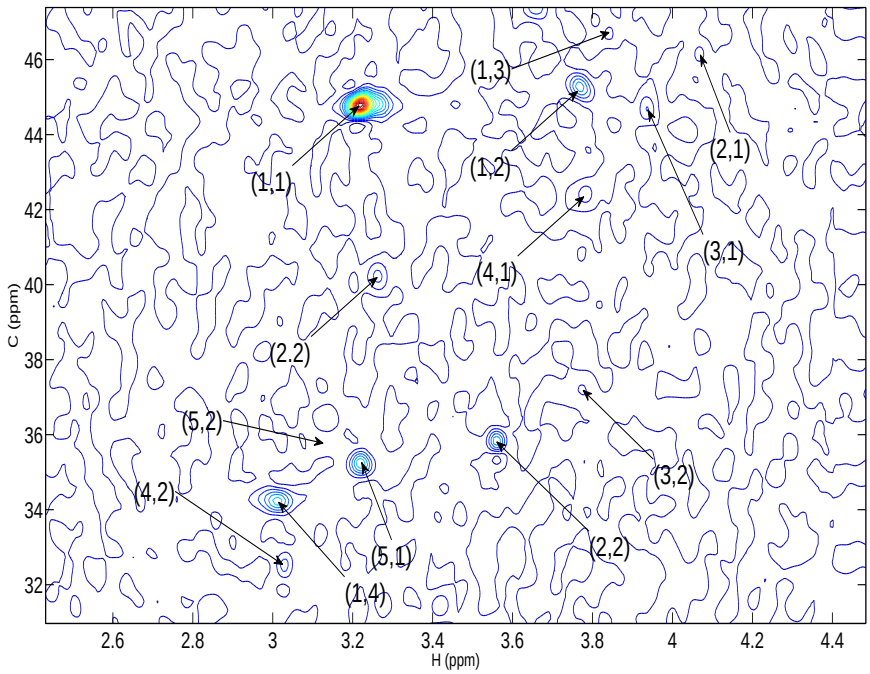


(b)

Figure 4.10: Metabolites identification results on synthetic spectra (a) Y_1 (b) Y_2. Every peak p belonging to a given metabolite M is labeled with (p, M).



(*a*)



(*b*)

Figure 4.11: Metabolites identification results on (a) a real healthy spectrum (b) a real cancerous spectrum. Every peak p belonging to a given metabolite M is labeled with (p,M).

4.3 Active zone coding and similarity measurement (fMRI)

Among object coding methods, the 3D Gaussian Descriptor (3DGD) proposed by Chaouch [Chaouch09] has showed its effectiveness compared to other methods and was ranked first on the Princeton Shape Benchmark database. The 3DGD method relies on the object Gaussian transformation which is derived from the Gaussian law. Although this method was successfully applied on 3D internet object searching, it presents a shortcoming. Indeed, it does not provide an information about the object surface topography. In order to lead to a more accurate fMRI active zone objects coding and similarity measurement result such information should be taken into account the proposed scheme. To this end, we propose a new descriptor: the 3D Generalized Gaussian Descriptor (3DGGD) inspired from the 3DGD method.

4.3.1 The 3D Generalized Gaussian Descriptor

In order to code a 3D object, we recall that author of [Chaouch09] proposed the partitioning of the object into N^3 cells and then the calculation for each one's center $g_{i,j,k}$, $i = 1...N,\ j = 1...N,\ k = 1...N$. Then, the author assigned to each cell a *characteristic value* $c_{i,j,k}$ as following (Fig. 4.6):

$$c_{i,j,k} = \sum_{p \in T} A_p \exp\left(\frac{-(p - g_{ijk})^2}{2\sigma^2}\right) \tag{4.21}$$

where A_p is the area of the triangular associated to p. Thus, the 3D object is codified with $c_{i,j,k}$.

Our goal is now to adapt the partitioning object step to the partial spherical shape of the fMRI active zones and to introduce the surface topology into the coding step. Concerning the first task (object partitioning), it is sufficient to put the active zone into a unit sphere modeled by M points g_m, $m = 1...M$ (Fig.4.12). Then, we assign to each point of the sphere a portion t_m, $m = l...M$ of the object surface. Concerning the second task, $c_{i,j,k}$ is independent of the topology of the cell surface. In other words, this method treats in the same way a flat surface (Fig.4.13.(a)) or a surface with reliefs (Fig.4.13.(b)).

An elegant way to integrate such information is the use of the Generalized Gaussian function. Fig.4.14 shows the shape of Generalized Gaussian function with different values of the shape parameter α. This parameter α allows us to adapt our function to the topology of t_m surface (Fig.4.15). In other words, the more the t_m surface is flat the more α is great. The shape parameter α for each t_m portion is estimated using the Maximum Likelihood (ML) algorithm.

The 3DGG descriptor $c_{GG}(m)$, $m = 1...M$ (Fig.4.12), is given by:

$$c_{GG}(m) = \sum_{p \in t_m} A_p \exp\left[-(\eta(\alpha_m)|p_m - g_m|)^{\alpha_m}\right] \tag{4.22}$$

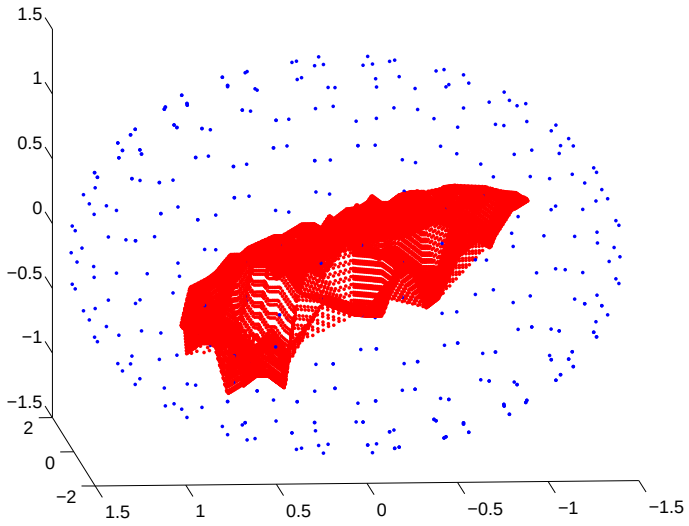
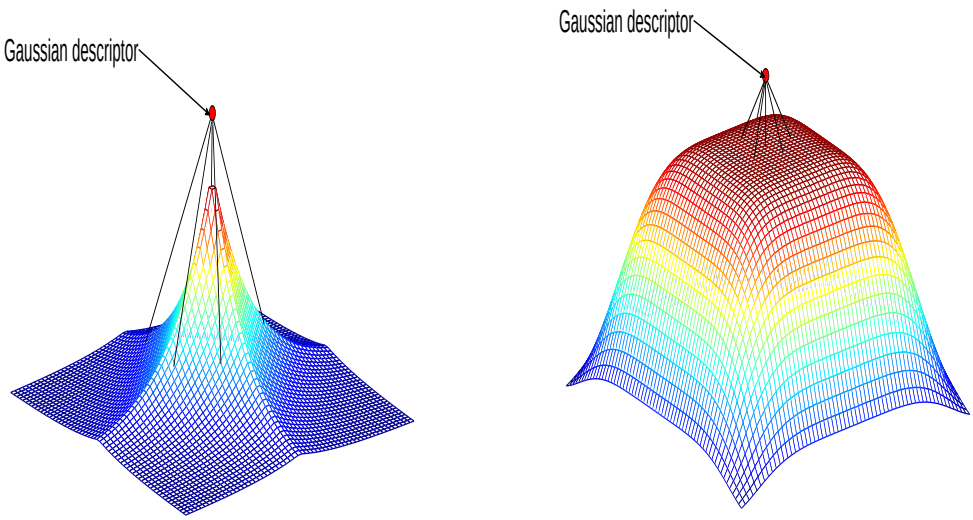


Figure 4.12: Spherical partitioning of a fMRI active zone.



(a) (b)

Figure 4.13: Contribution of an: (a) object flat surface, (b) object acute surface, points in the local description of $g_{i,j,k}$ using a Gaussian function.

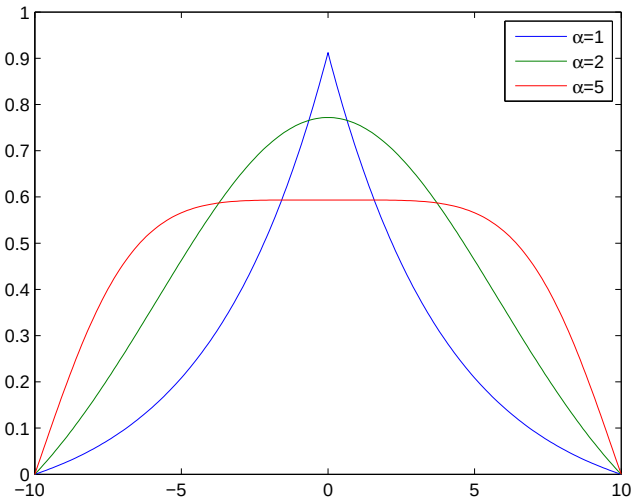
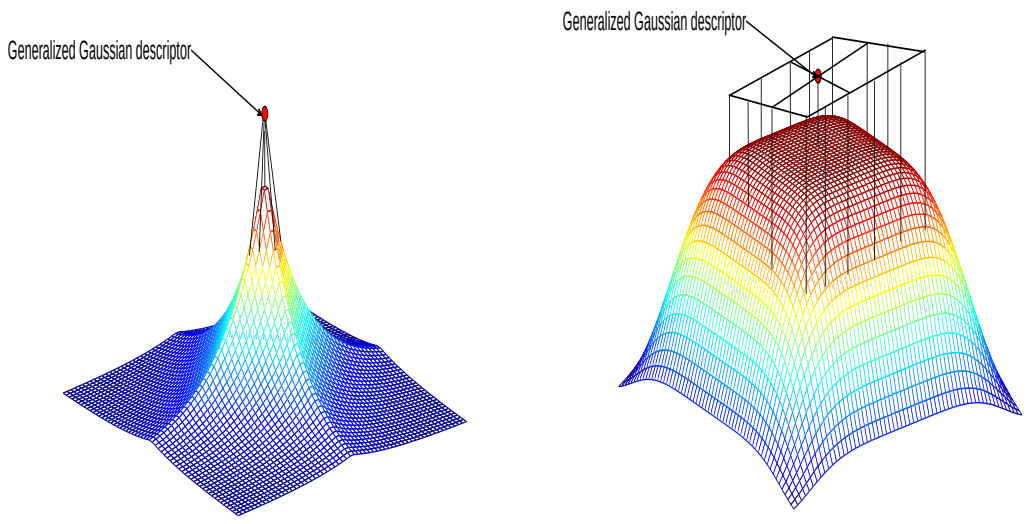


Figure 4.14: The shape of Generalized Gaussian function with different values of the shape parameter α. The Gaussian function corresponds to $\alpha = 2$.



(a) (b)

Figure 4.15: Contribution of an: (a) object flat surface, (b) object acute surface, points in the local description of $g_{i,j,k}$ using a Generalized Gaussian function.

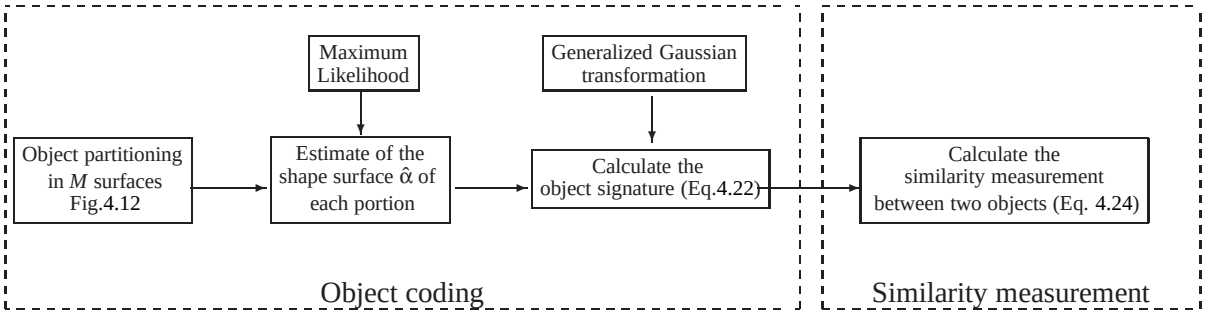


Figure 4.16: Overview diagram of the proposed active zones coding and similarity measurement chain.

where $\eta(\alpha_m) = \left[\frac{\Gamma(3/\alpha_m)}{\sigma^2\Gamma(1/\alpha_m)}\right]^{\frac{1}{2}}$, $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}\exp(-t)dt$ and g_m is the point belonging to the unit sphere and associated to t_m.

To compare two objects O^1 and O^2, we firstly calculate for each one the 3DGGD $c^1 = [c^1{}_{GG}(m)]_{m\in M}$ and $c^2 = [c^2{}_{GG}(m)]_{m\in M}$. The similarity can be calculated using the euclidian distance as follows:

$$\Delta(O^1, O^2) = d(c^1, c^2) \tag{4.23}$$

In order to be less sensitive to small displacements or minor geometric variations, we introduce a new similarity measure that minimizes the distance between adjacent pairs of components:

$$\Delta(O^1, O^2) = \frac{1}{M}\sum_{m=1}^{M} min\left(\begin{cases}(c^1_{GG}(m) - c^2_{GG}(m))^2 \\ (c^1_{GG}(m) - c^2_{GG}(m'))^2, \quad m' \in V_m\end{cases}\right)$$

where V_m is the 1-order neighborhood of the descriptor $c^2_{GG}(m)$. An overview diagram of the proposed active zones coding and similarity measurement chain is presented in Fig. 4.16.

4.3.2 Active zone coding and similarity measurement validation

In this part, we provide some 3D object coding and similarity measurement results obtained with the proposed method. As for the alignment method validation, this method was applied on partial 3D objects designed to fit the characteristics of the active zones as well as on some real fMRI images. More results on real active zones are presented in chapter 6. In order to emphasize the benefit of the proposed method and particularly the use of the Generalized Gaussian function to model the surface topology, we have compared our algorithm to (3DGD) method [Chaouch09] and the Shape histogram method [Ankerst99]. The goal of these experiments is to assess the performance of the proposed coding and similarity measurement scheme. To construct our toy data set, we firstly generated five partial

	3DGGD method		3DGD method		shape histogram method	
dataset	*recall*(%)	*precision*(%)	*recall*(%)	*precision*(%)	*recall*(%)	*precision*(%)
dataset1 (55% of S_o)	**91.35**	**88.02**	84.21	83.25	80.38	75.72
dataset2 (60% of S_o)	**90.44**	**88.12**	85.50	82.01	81.67	77.45
dataset3 (65% of S_o)	**91.74**	**87.23**	81.09	78.96	77.10	75.98
dataset4 (70% of S_o)	**88.17**	**86.88**	79.69	78.64	70.18	67.41
dataset5 (75% of S_o)	**90.95**	**89.03**	86.78	84.35	74.11	72.08

Table 4.2: The average *recall*(%) and *precision*(%) obtained with: our coding and similarity measurement method 3DGGD, the 3DGD method, the shape histogram method on different simulated datasets.

spheres, PS_i $i = 1...5$, by removing 55%, 60%, 65%, 70% and 75% of an entire sphere S_o. Each PS_i is then partitioned into $M = 20$ portions. From each PS_i, we generate nineteen other partial spheres PS_{ij}, $j = 1...10$ by only modifying three portions of PS_i surface. Finally, from each PS_{ij} we generate fifty other 3D object by adding a Gaussian noise such that the $PSNR \in [10dB, 20dB]$. At the end, we obtain five datasets. Each dataset contains one thousand 3D objects. These simulated datasets allow us to evaluate the performance of the method in difficult cases such as the capacity to discriminate objects with minor changes (only three surface portions) and its robustness to different values of Peak Signal to Noise Ratio PSNR. Note that we have opted for the 3D triangular mesh representation for all artificial toys. For the performance assessment, we retained the same *recall* and the *precision* measurements as the metabolite identification validation.

Figure.4.17 shows some simulated 3D objects. The object coding and similarity measurement results for each data set are presented in Table 4.2. First, as one can see, the proposed method performs the best coding and similarity measurement results compared to the 3DGD and the Shape histogram methods which do not take into account the surface topology information. Secondly, the *recall* and the *precision* measurements are still close to 90% for all datasets which proves that the proposed method is well adapted to this type of 3D objects (partial spherical object). Figure. 4.18 shows an example of two simulated 3D objects wrongly assigned with the 3DGD method. As one can see, it is very difficult to visually observe the differences. Note that these two objects were correctly indexed by our method. Figure. 4.19 shows an example of real active zones assignment.

Conclusion

In this chapter, we presented two new object coding and similarity measurement methods. The first method, dedicated to the 2D HSQC spectrum metabolite identification, is based on the use of the fuzzy set theory to deal with the ambiguity which is in the heart of such an identification task. The use of the metabolite likelihood measure as metabolite signature

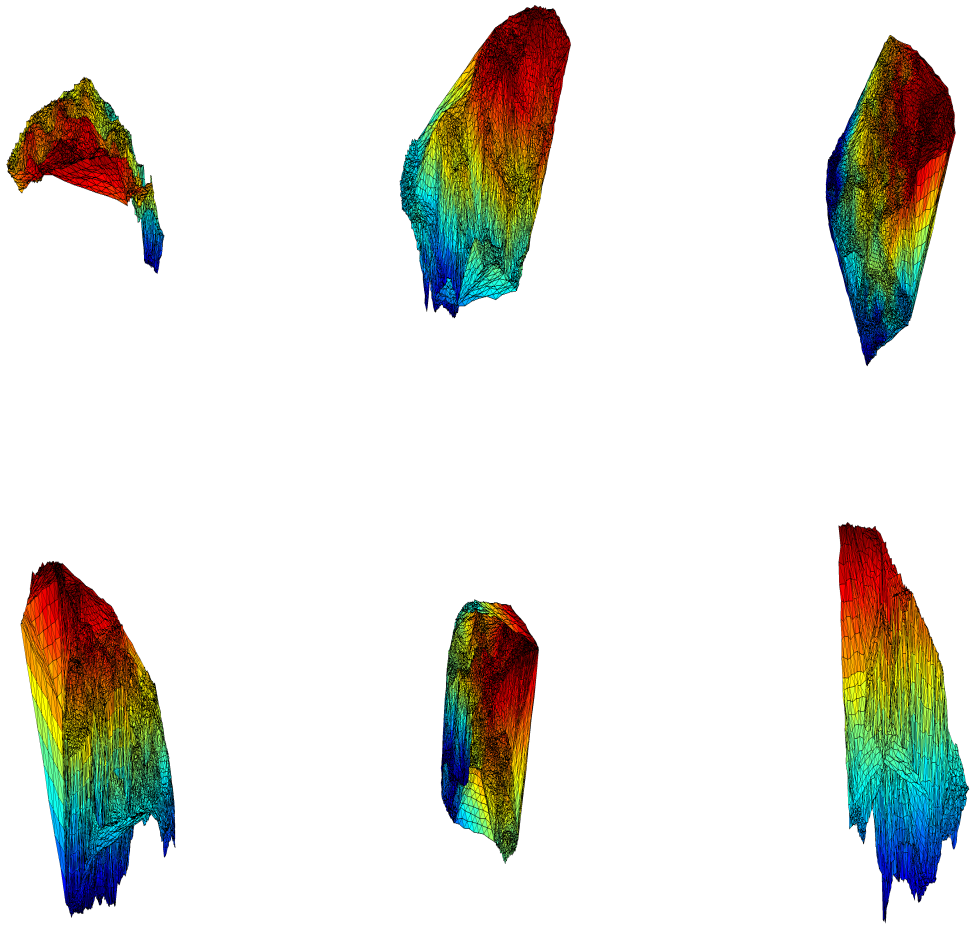


Figure 4.17: Six simulated 3D objects.

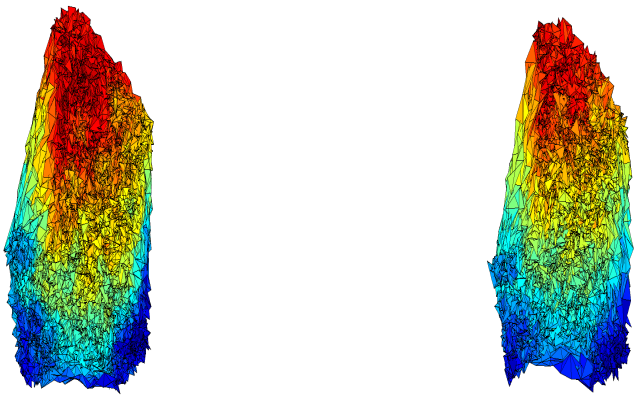


Figure 4.18: An example of two simulated 3D objects wrongly assigned with the 3DGD method.

Figure 4.19: Assignment result of real active zones obtained with the proposed method.

has increased the robustness of the proposed identification scheme with comparison to the SVM and the threshold methods which do not take into account the *a priori* knowledge.
In the other hand, the second method, dedicated to the fMRI active zone object, relies on the Partition-Space approach. This approach allows us to code different objects into an appropriate description and to calculate the similarity between two objects. In order to integrate the surface topology information into the coding and similarity measurement scheme, we proposed a new descriptor: the 3D Generalized Gaussian Descriptor (3DGGD). The synthetic validation of the proposed active zone coding and similarity measurement scheme has shown that the modeling of the surface topology has increased the robustness of the proposed coding and similarity measurement scheme with comparison to the 3DGD method which does not take into consideration this *a priori* knowledge.

CHAPTER 5

Object classification

Contents

Symbols:

x_i	Object features vector
r_i	Classification result of x_i
f	Separating function
(w, b)	Hyperplane hyperparameters
(w^*, b^*)	Optimal hyperplane hyperparameters
β	Lagrange multipliers
γ	Lagrange multipliers
α	Lagrange multipliers
α^*	Optimal Lagrange multipliers
ζ_i	Slack variable associated to x_i
C	Regularization parameter
C_1	Regularization parameter
C_2	Regularization parameter
$\mathcal{K}$	Kernel function
$\varphi(.)$	Mapping function
H	Hilbert space
(R, a)	Hypersphere hyperparameters
(R^*, a^*)	Optimal hypersphere hyperparameters
C_G	Gaussian copula
Φ	Standard Gaussian cumulative distribution
Γ	Correlation matrix
I	Identity matrix
μ	Membership degree to the target class
C_{ht}	Hard class of target population
C_{ft}	Fuzzy class of target population
C_{ho}	Hard class of outlier population
C_{fo}	Fuzzy class of outlier population
$\mathcal{G}$	Gamma distribution
$\mathcal{N}(,)$	Standard normal distribution
GG	Generalized Gaussian distribution

Acronyms:

HR-MAS	High Resolution Magic Angle Spinning
HSQC	Heteronuclear Single Quantum Coherence spectrum
fMRI	functional Magnetic Resonance imaging
SVM	Support Vector Machine
SVDD	Support Vector Data Description
SV3DH	Support Vector Data Description with Dependency Handling
ML	Maximum Likelihood
MCMC	Monte Carlo Markov Chain
LOO	Leave One Out
RBF	Radial Basis Function

Introduction

The Object classification is the third step in the proposed object indexing scheme (Fig. 2.15). This step aims at comparing a request (individual/group of medical signal) to defined groups belonging to the available database (*e.g*, healthy group or pathological group) for change detection. Indeed, detecting the changes between groups is equivalent to discriminating the data into two classes: *changed* and *unchanged* (or unimportant changed) data classes (the later will be the class of interest in the following). In this case, the corresponding classifier is know a one-class classifier (target class and reject class). The classification method may either be supervised due to the difficulty of the task or unsupervised but the cost is sometimes a loss of robustness and/or higher computing time. In the supervised case, the process requires to be able to access to a ground truth in order to derive a suitable training set for the learning process of the classifiers. However, the ground truth is usually difficult and expensive to find (which is unfortunately our case) . Consequently, the use of unsupervised change-detection methods is crucial in many applications where ground truth is out of reach.

Among all object classification/change detection methods, we pay attention in this thesis to the kernel based classification methods. Indeed, the kernel-based methods offer several advantages compared to other approaches: they reduce the curse of high dimensionality in data and increase the reliability and the robustness of the method to a high level of noise [Li06]. In this chapter, we present in section 5.1, a brief overview of the object kernel-based methods. In section 5.2, we propose a new kernel function which combines the characteristics of basic kernel functions with new information about features distribution and then dependency between samples. This kernel function is then used to map the data into a high dimensional features space where an hypersphere encloses most patterns belonging to the "unchanged" class. The dependency between samples will be based on copulas theory that will be used for the first time to our knowledge in the support vector data description (SVDD) framework. In section 5.3, we pay a particular attention to check that the proposed kernel function is robust with higher performance compared to classic Support Vector Machine (SVM) and Support Vector Data Description (SVDD) methods.

5.1 Kernel-based classifiers

In the literature, two very interesting and widely-used unsupervised change-detection methods are the Bayesian methods [Fumera00] and the kernel methods [Ben-Hur02]. Although the former approach is relatively simple, it exhibits a major drawback: it requires a large amount of knowledge about the class of interest which is not always available, particularly, in highly complex applications like the medical one [Sanchez-Hernandez07]. Moreover, when only weak changes occurred between the two considered data set, the probability density function (pdf) of the *changed data* may be confused with the *unchanged data* pdf (*e.g.*, the Hidden Markov Model method generally tries to regularize bad classification results due to this ill-posed problem and the presence of outliers in the data [Belghith09]).

Although these drawbacks, Bayesian methods offer efficient tools to include an *a priori* through a posteriori pdf.

Furthermore, the kernel methods are more flexible. Indeed, the kernel-based function offers several advantages compared to other approaches: they reduce the curse of high dimensionality in data, increase the reliability and the robustness of the method to a high level of noise and allow flexible mapping between objects (inputs) represented by a feature vector and class label (outputs)[Shawe-Taylor04]. Among all these advantages, the kernel based change-detection method is not time-consuming and then allows to develop real time applications. The mainly used kernel-based methods are the Support Vector Machine (SVM) and the Support Vector Data Description (SVDD).

5.1.1 Support vector Machine SVM

Let $\{\mathbf{x}_i\}_{i=1...K}$, $\mathbf{x}_i \in \mathbb{R}^N$ be the vector containing the N features of a given object and $\{r_i\}_{i=1...K}$, with $r_i \in \{\pm 1\}$, the corresponding output of $\{\mathbf{x_i}\}_{i=1...K}$. The SVM algorithm aims at classifying $\{\mathbf{x_i}\}_{i=1...K}$ into two classes: class of targets (*i.e.*; unchange or $r_i = +1$ class) and the outliers (*i.e.*; change or $r_i = -1$ class). In the supervised case, the purpose of SVM algorithm is to predict the label r_i from a set of observations called training set composed by objects $\{\mathbf{x_i}\}_{i=1...K}$ with known classification results $\{r_i\}_{i=1...K}$. Thus, the problem is to find a separating function f that assigns the label 1 (respectively -1) to each object $\mathbf{x_i}$ such that $f(\mathbf{x_i}) \geq 0$ (respectively $f(\mathbf{x_i}) < 0$). The separating surface (or the separating hyperplane) is then given by the equation $f(\mathbf{x_i}) = 0$.

Linear classifier

Suppose the training data

$$(\mathbf{x}_i, r_i)_{i=1...K}, \quad \mathbf{x}_i \in \mathbb{R}^N, \ r_i \in \{-1, 1\}$$

can be separated by a hyperplane :

$$\mathbf{w}^T \mathbf{x_i} + b = \sum_{j=1}^{N} w_j x_i(j) + b = 0 \tag{5.1}$$

where $\mathbf{w} \in \mathbb{R}^N$ and $b \in \mathbb{R}$ (called the *bias*) are the hyperparameters of the hyperplane.

We say that this set of vectors is separated by the optimal hyperplane if it is separated without error and the distance between closest vector to the hyperplane is maximal [Vapnik00]. The separating hyperplane can be described by the following form:

$$r_i \left(\mathbf{w}^T \mathbf{x_i} + b\right) \geq 1, \qquad i = 1...K \tag{5.2}$$

To find this hyperplane, one has to solve the following quadratic programming problem:

$$\min_{\mathbf{w}} \frac{1}{2} \mathbf{w}^T \mathbf{w} \tag{5.3}$$

subject to constraints (Eq. 5.2).

This is a classical optimization problem with inequality constraints. Such an optimization problem can be solved by the saddle point of the Lagrange function [Vapnik00]:

$$L(\mathbf{w},b,\alpha) = \frac{\mathbf{w}^T\mathbf{w}}{2} - \sum_{i=1}^{K}\alpha_i\left[r_i\left(\mathbf{w}^T\mathbf{x_i}+b\right)-1\right] \tag{5.4}$$

where $\alpha \geq 0$ is the Lagrange multipliers. To find this point, one has to minimize this function over $\mathbf{w}$ and b and to maximize it over the Lagrange multipliers $\alpha \geq 0$. At the saddle point, the solution $(\mathbf{w}^*, b^*, \alpha^*)$, should satisfy the condition:

$$\frac{\partial L(\mathbf{w}^*,b^*,\alpha^*)}{\partial \mathbf{w}} = 0 \tag{5.5}$$

$$\frac{\partial L(\mathbf{w}^*,b^*,\alpha^*)}{\partial b} = 0 \tag{5.6}$$

Rewriting theses equations, one obtains the following properties of the optimal hyperplane:

1. The coefficient α^* for the optimal hyperplane should satisfy the constraints:

$$\sum_{i=1}^{K}\alpha_i^* r_i = 0 \qquad \alpha^* \geq 0 \tag{5.7}$$

2. $\mathbf{w}^*$ is a linear combination of the vectors of the training set:

$$\mathbf{w}^* = \sum_{i=1}^{K}\alpha_i^* r_i \mathbf{x}_i \tag{5.8}$$

3. Only the so-called Support Vectors (SV) can have nonzero coefficients $\alpha_\mathbf{i}^*$ in the expansion of $\mathbf{w}^*$. The support vectors are the vectors for which in inequality (Eq. 5.2), equality is achieved.

$$\mathbf{w}^* = \sum_{i\in SV}\alpha_i^* r_i \mathbf{x}_i \tag{5.9}$$

   This fact follows from the classical Kuhn-Tucker theorem, according to which necessary and sufficient conditions for optimal hyperplane are that the separating hyperplane satisfies the conditions [Vapnik98]:

$$\alpha_i^*\left[r_i\left(\left(\mathbf{w}^*\right)^T\mathbf{x_i}+b^*\right)-1\right] = 0 \quad i = 1...K \tag{5.10}$$

Putting (Eq.5.8) into (Eq.5.4) and taking into account the Kuhn-Tucker conditions, one obtains the functional:

$$W(\alpha) = \sum_{i=1}^{K}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{K} r_i r_j \alpha_i \alpha_j \mathbf{x}_i^T\mathbf{x}_j \tag{5.11}$$

It remains to maximize this functional in the nonnegative quadrant $\alpha_i \geq 0$ and under the constraint:

$$\sum_{i=1}^{K} \alpha_i r_i = 0 \tag{5.12}$$

This can be achieved by the use of standard quadratic programming methods [Bazaraa06].

Let the vector $\alpha^* = (\alpha_1^*, ..., \alpha_K^*)$ be the solution to this quadratic optimization problem. The separating function function f is given by:

$$f(\mathbf{x}) = sign\left(\sum_{i \in SV} \alpha_i^* r_i \mathbf{x}^T \mathbf{x}_i + b^*\right) \tag{5.13}$$

where $sign(.)$ is the sign function and

$$b^* = r_i - \mathbf{w}^{*T} \mathbf{x_s} \tag{5.14}$$

where x_s is a given support vector.

Non-linear classifier

In the previous paragraph, patterns belonging to the training set are assumed to be linearly separable with a plane separating surface. However, the assumption of linear separability case is too restrictive for many particular applications, especially when data are noisy. The optimal margin algorithm is generalized [Cortes95] to nonseparable problems by the introduction of non-negative slack variables denoted by $\zeta_{i, i \in \{1,...,K\}} \geq 0$ in the statement of the optimization problem (Fig.5.1).

The changed objective functional with penalty parameter C (a regularization parameter that controls the trade-off between the marge $\mathbf{w}$ and the number of learning errors) is:

$$\frac{\mathbf{w}^T \mathbf{w}}{2} + C \sum_{i=1}^{K} \zeta_i \tag{5.15}$$

subject to the inequality constraints:

$$r_i \left(\mathbf{w}^T \mathbf{x_i} + b\right) \geq 1 - \zeta_i, \qquad i = 1...K, \xi_i \geq 0 \tag{5.16}$$

In analogy with what was done for the separable case, the use of the Lagrange multipliers leads do the following optimization problem:
Maximize

$$W(\alpha) = \sum_{i=1}^{K} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{K} r_i r_j \alpha_i \alpha_j \mathbf{x}_i^T \mathbf{x}_j \tag{5.17}$$

subject to:

$$\sum_{i=1}^{K} \alpha_i r_i = 0 \ \textit{and} \ 0 \leq \alpha_i \leq C \tag{5.18}$$

As one can remark, the only difference from the separable case is that now the α_i have an upper bound of C.

However, even with the introduction of the slack variables ζ_i, the training set may require a decision surface more complicated than a simple linear hyperplane. To take into account non-linear separator, the linear SVM can be generalized by the introduction of the kernel functions (Fig. 5.2) [Boser92]. Indeed, the kernel function allows the mapping of data set defined over the input I into a higher dimensional Hilbert space H (feature space) where the patterns are assumed to be linearly separated. The mapping function is denoted by $\varphi : \mathcal{X} \rightarrow H$. If a given algorithm can be expressed in the form of dot products in the input space, its non-linear kernel version only needs the dot products among mapped samples. Kernel methods compute the similarity between training samples using pairwise inner products between mapped samples [Halmos82]. The bottleneck for any method based on kernel function is the proper definition of a kernel function that accurately reflects the similarity among samples. In the early years of kernel machine learning research, researchers considered kernels satisfying the conditions of Mercer's theorem (e.g., [Rousseau03]).

Definition 1 *Let $\mathcal{X}$ be a closed set of $\mathbb{R}^N$. A symmetric function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ which for all $g(.) \in L_2(\mathcal{X})$ (square integrable function):*

$$\int_{\mathcal{X}} \int_{\mathcal{X}} \mathcal{K}(x,y) g(x) g(y) dx dy \geq 0 \tag{5.19}$$

is said to be a Mercer kernel [Minh06].

In [Hofmann08], authors show that the positive definite kernels are the right class of kernels to consider.

Definition 2 *Let $\mathcal{X}$ be a nonempty set $\mathbb{R}^N$. A symmetric function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ which for all $\mathbf{x_i} \in \mathcal{X}$ and real numbers $a_i \in \mathbb{R}$:*

$$\sum_{ij} a_i a_j \mathcal{K}(\mathbf{x_i}, \mathbf{x_j}) \geq 0 \tag{5.20}$$

is said to be a positive definite kernel [Minh06].

The most common used kernel are:

- the linear kernel $\mathcal{K}(\mathbf{x_i}, \mathbf{x_j}) = \mathbf{x_i}^T \mathbf{x_j}$,
- the polynomial kernel $\mathcal{K}(\mathbf{x_i}, \mathbf{x_j}) = (\mathbf{x_i}^T \mathbf{x_j} + 1)^d$, $d > 0$
- the Radial Basis Function (RBF), $\mathcal{K}(\mathbf{x_i}, \mathbf{x_j}) = exp(-(\mathbf{x_i} - \mathbf{x_j})^T (\mathbf{x_i} - \mathbf{x_j})/2\sigma^2)$, $\sigma > 0$

Once a valid kernel $\mathcal{K}$ has been chosen, to find the coefficient α_i in the separable case (analogously in the non-separable case) it is sufficient to:
maximize

$$W(\alpha) = \sum_{i=1}^{K} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{K} r_i r_j \alpha_i \alpha_j \; \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) \tag{5.21}$$

with $\alpha_i \geq 0$ and under the constraint:

$$\sum_i \alpha_i r_i = 0 \tag{5.22}$$

and the decision function is

$$f(\mathbf{x}) = sign\left(\sum_{i=1}^{K} \alpha_i^* r_i \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) + b^* \right) \tag{5.23}$$

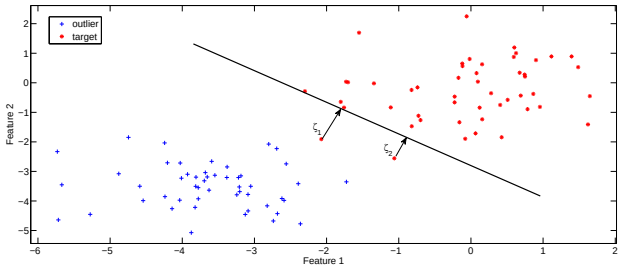


Figure 5.1: Non-linear classifier separation by a hyperplane with slack variables ζ_i.

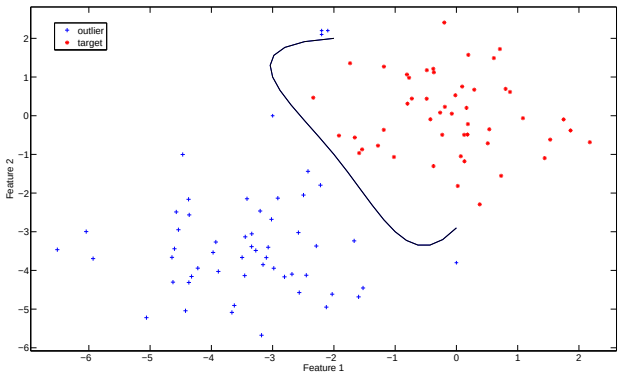


Figure 5.2: Kernel classifier separation by a complex decision surface.

5.1.2 Support Vector Data Description SVDD

The aim of SVDD classifier method consists in mapping the data into a high dimensional feature space. In this new space, an hypersphere enclosing most of the data set belonging to the class of interest (*target* class corresponding to *unchanged data*) and rejecting the other observations (that will be considered like *outliers*) is defined (Fig. 5.3) [Tax04]. This amounts to draw a minimum volume hypersphere in the kernel feature space that includes all or most of the target objects which are available in the training set. By analogy with the SVM problem (which consists in estimating the hyperparameters of the hyperplane $(\mathbf{w}, b)$, the sphere is characterized by its center a and its radius $R > 0$. Thus, the problem is to find a decision function f that assigns the label 1 (respectively -1) to each object x_i such that $f(x_i) \leq R$ (respectively $f(x_i) > R$).
In the following the target objects are enumerated by indices i and j. Thus, minimizing the volume of the sphere returns to minimizing R^2 with the constraints [Tax04]:

$$(\mathbf{x}_i - \mathbf{a})^T (\mathbf{x}_i - \mathbf{a}) \leq R^2 \qquad \forall i \tag{5.24}$$

To allow the possibility of outliers in the training set, the distance from $\mathbf{x}_i$ to the center $\mathbf{a}$ should not be strictly smaller than R^2, but larger distances should be penalized. Therefore we introduce slack variables $\zeta_i \geq 0$ and the minimization problem changes into:

$$\min_{R,\mathbf{a},\zeta_i} \left\{ R^2 + C \sum_i \zeta_i \right\} \tag{5.25}$$

with constraints that almost all objects belonging to the target class are within the sphere:

$$(\mathbf{x}_i - \mathbf{a})^T (x_i - \mathbf{a}) \leq R^2 + \zeta_i \qquad \forall i, \quad \zeta_i \geq 0 \tag{5.26}$$

As for the SVM case, the saddle point of the primal Lagrangian $L(R, \mathbf{a}, \zeta_i, \alpha_i, \gamma_i)$ [Tax04]:

$$L(R, \mathbf{a}, \zeta_i, \alpha_i, \gamma_i) = R^2 + C \sum_i \xi_i - \sum_i \alpha_i \left\{ R^2 + \xi_i - \left(\mathbf{x_i}^T \mathbf{x_i} - 2\mathbf{a}^T \mathbf{x_i} + \mathbf{a^T a} \right) \right\} - \sum_i \gamma_i \xi_i \tag{5.27}$$

where $\alpha_i \geq 0$ and $\gamma_i \geq 0$ are the Lagrangian multipliers. Again, one should find an optimal saddle point $(R^*, a^*, \zeta^*, \alpha^*, \gamma^*)$ by minimizing L with respect to (a, R, ζ) and maximizing L with respect to non-negative (α, γ). In analogy with what was done for the SVM case, a solution in dual space is found using standard conditions for an optimum of a constrained function:

$$\frac{\partial L}{\partial R} = 0, \; i.e, \qquad \sum_i \alpha_i^* = 1 \tag{5.28}$$

$$\frac{\partial L}{\partial \mathbf{a}} = 0, \; i.e, \qquad \mathbf{a}^* = \sum_i \alpha_i^* x_i \tag{5.29}$$

$$\frac{\partial L}{\partial \zeta_i} = 0, \; i.e, \qquad \alpha_i^* + \gamma_i^* = C \tag{5.30}$$

From the last equation $\alpha_i^* + \gamma_i^* = C$ and because $\alpha_i \geq 0$, $\gamma_i \geq 0$, Lagrange multipliers γ_i can be removed when we demand that $0 \leq \alpha_i \leq C$. The use of the dual variables Lagrangian leads do the following optimization problem :
Maximize

$$W(\alpha) = \sum_i \alpha_i \mathbf{x_i}^T \mathbf{x_i} - \sum_{i,j} \alpha_i \alpha_j \mathbf{x}_i^T \mathbf{x}_j \tag{5.31}$$

subject to:

$$0 \leq \alpha_i \leq C \tag{5.32}$$

Eq.5.29 shows that the center of the sphere is a linear combination of the objects. Only the support vectors $\mathbf{x_s}$ are needed in the description. R^2 is the distance from the center of the sphere a to (any of the support vectors on) the boundary. Support vectors which fall outside the description $(\alpha_i = C)$ are excluded. Therefore:

$$R^2 = \mathbf{x_s}^T \mathbf{x_s} - 2\sum_i \alpha_i \mathbf{x_s}^T \mathbf{x_i} + \sum_{i,j} \alpha_i \alpha_j \mathbf{x_i}^T \mathbf{x_j} \tag{5.33}$$

To test an object z, the distance to the center of the sphere has to be calculated. A test object z belongs to the target class when this distance is smaller or equal than the radius R:

$$f(\mathbf{z}) = \left(\mathbf{z}^T \mathbf{z} - 2\sum_i \alpha_i^* \mathbf{z}^T \mathbf{x_i} + \sum_{i,j} \alpha_i^* \alpha_j^* \mathbf{x_i}^T \mathbf{x_j} \right) \leq R^2 \tag{5.34}$$

When negative examples (objects which should be rejected) are available, they can be incorporated in the training to improve the description. In contrast with the training (target) examples which should be within the sphere, the negative examples should be outside it. This data description now differs from the normal Support Vector Classifier in the fact that the SVDD always obtains a closed boundary around one of the classes (the target class). In the following the target objects are enumerated by indices i, j and the negative examples by l, m. Again we allow for errors in both the target and the outlier set and introduce slack variables $\zeta_i \geq 0$ and $\zeta_l \geq 0$ and the minimization problem changes into [Tax04]:

$$\min_{R,\mathbf{a},\zeta_i,\zeta_l} \left\{ R^2 + C_1 \sum_i \zeta_i + C_2 \sum_l \zeta_l \right\} \tag{5.35}$$

and the constraints

$$(\mathbf{x}_i - \mathbf{a})^T (\mathbf{x}_i - \mathbf{a}) \leq R^2 + \zeta_i, \quad (\mathbf{x}_l - \mathbf{a})^T (\mathbf{x}_l - \mathbf{a}) \geq R^2 - \zeta_l, \quad \zeta_i \geq 0, \quad \zeta_l \geq 0, \quad \forall i, l \tag{5.36}$$

where C_1 and C_2 are two regularization parameters. The saddle point of the primal Lagrangian $L(R, \mathbf{a}, \zeta_i, \zeta_l, \alpha_i, \alpha_l, \gamma_i, \gamma_l)$:

$$\begin{aligned} L(R,\mathbf{a},\zeta_i,\zeta_l,\alpha_i,\alpha_l,\gamma_i,\gamma_l) =& R^2 + C_1\sum_i \xi_i + C_2\sum_l \xi_l - \sum_i \gamma_i\xi_i - \sum_l \gamma_l\xi_l \\ & - \sum_i \alpha_i\left\{R^2+\xi_i-\left(x_i^T x_i - 2a^T x_i + a^T a\right)\right\} \\ & - \sum_l \alpha_l\left\{\left(x_l^T x_l - 2a^T x_l + a^T a\right) - R^2 + \xi_l\right\} \end{aligned} \tag{5.37}$$

where $\alpha_i \geq 0$, $\alpha_l \geq 0$,$\gamma_i \geq 0$ and $\gamma_l \geq 0$ are the Lagrangian multipliers. Setting the partial derivatives of L with respect to R, $\mathbf{a}$, ξ_i and ξ_l to zero gives the constraints:

$$\sum_i \alpha_i^* - \sum_l \alpha_l^* = 1 \tag{5.38}$$

$$a^* = \sum_i \alpha_i^* x_i - \sum_l \alpha_l^* x_l \tag{5.39}$$

$$0 \leq \alpha_i \leq C_1, \quad 0 \leq \alpha_l \leq C_2, \quad \forall i,l \tag{5.40}$$

The use of the Lagrange multipliers leads do the following optimization problem [Tax04]: Maximize

$$\begin{aligned} W(\alpha_i,\alpha_l) =& \sum_i \alpha_i \mathbf{x_i}^T\mathbf{x_i} - \sum_l \alpha_l \mathbf{x_l}^T\mathbf{x_l} - \sum_{i,j}\alpha_i\alpha_j \mathbf{x}_i^T\mathbf{x}_j \\ & - \sum_{l,m}\alpha_l\alpha_m \mathbf{x}_l^T\mathbf{x}_m + 2\sum_{l,j}\alpha_j\alpha_l\mathbf{x}_j^T\mathbf{x}_l \end{aligned} \tag{5.41}$$

subject to:

$$0 \leq \alpha_i \leq C_1, \quad 0 \leq \alpha_l \leq C_2 \quad \forall i,l \tag{5.42}$$

If we define new variables $\alpha_n^{'} = r_n\alpha_n$. Note that the following the index n and q enumerate both target and outlier objects.

The SVDD with negative examples is identical to the normal SVDD. The constraints given in (Eq. 5.38) and (Eq. 5.39) change into $\sum_n \left(\alpha_n^{'}\right)^* = 1$ and $a^* = \sum_n \left(\alpha_n^{'}\right)^* \mathbf{x}_n$ and again the testing function Eq.5.34 can be used.

Once a valid kernel $\mathcal{K}$ has been chosen, to find the coefficient α_i in the positive case (analogously in the positive and negative case) it is sufficient to:

maximize

$$W(\alpha) = \sum_i \alpha_i \mathcal{K}(\mathbf{x_i},\mathbf{x_i}) - \sum_{i,j}\alpha_i\alpha_j\mathcal{K}(\mathbf{x}_i,\mathbf{x}_j) \tag{5.43}$$

subject to:

$$0 \leq \alpha_i \leq C \tag{5.44}$$

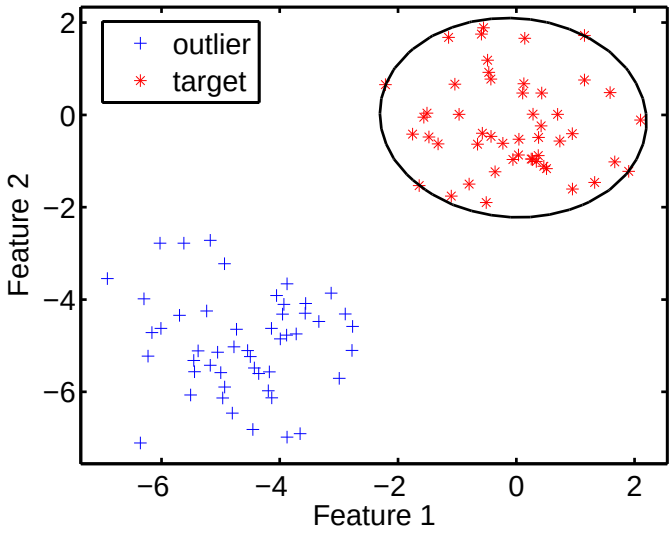


Figure 5.3: Linear classifier separation by an hypersphere.

The new testing function is expressed as:

$$f(\mathbf{z}) = \left(\mathcal{K}(\mathbf{z},\mathbf{z}) - 2\sum_i \alpha_i^* \mathcal{K}(\mathbf{z},\mathbf{x_i}) + \sum_{i,j} \alpha_i^* \alpha_j^* \mathcal{K}(\mathbf{x_i},\mathbf{x_j}) \right) \leq R^2 \tag{5.45}$$

For the case with negative examples, the testing function is expressed as:

$$f(\mathbf{z}) = \left(\mathcal{K}(\mathbf{z},\mathbf{z}) - 2\sum_n (\alpha_n^{'})^* \mathcal{K}(\mathbf{z},\mathbf{x_n}) + \sum_{n,q} (\alpha_n^{'})^* (\alpha_q^{'})^* \mathcal{K}(\mathbf{x_n},\mathbf{x_q}) \right) \leq R^2 \tag{5.46}$$

5.2 Support Vector Data Description including Dependency Hypothesis

The change-detection/classification problem is tackled in an unsupervised way using the kernel-based approach. However, the main bottleneck of kernel methods is the choice of the kernel function which depends strongly of the application [Scholkopf00]. Although the basic kernel functions are more or less successfully applied for change-detection, they do not exploit additional constraints often available, such as the dependency and the distribution of different features. We particularly show that the change-detection should be more robust, more accurate and more efficient if such information is integrated and correctly modeled within the change-detection method. In order to take into account these characteristics in our change-detection scheme, we propose the new kernel function which combines the old kernel functions with a new information about features distribution and dependency. The challenge is then to find the appropriate way to handle this dependency.

To this end, we have opted for the copula theory which has proved its effectiveness to handle dependency [Joe97]. Several studies show the effectiveness of the SVDD method to detect changes [Yang04, Camps-Valls08]. Indeed, in the case with few available labeled information, purely supervised approaches like SVMs yield poor solutions since there is no information on the change class. Contrarily, SVDD offers very good results since the method tries to model the 'unchange' class accurately rather than building a separating hyperplane 'change'/'unchange' [Camps-Valls06]. For this reason we opt for the SVDD method. Moreover, we show that the use of the new kernel function increases the performance of the change-detection compared to the basic kernel functions. The proposed method is denoted SV3DH (SV3DH is the acronym for Support Vector Data Description including Dependency Hypothesis).

5.2.1 Copula kernel function

The proposed kernel function

We remind that we seek to blindly classify the data into two classes: class of targets (*i.e.*; unchange) and the outliers using the SVDD method. In this part we define the proposed kernel function. Our aim is then to properly model and integrate both the dependency and the distribution of different features in the kernel function to reach a more accurate change-detection result. The new kernel function should combine the old kernel functions (in our case the RBF function which offers some freedom degree thanks to the hyperparameter σ) with a new information about correlated features distribution. To this end, we propose a simple, yet powerful, kernel function based on the copula theory.

Several studies show the effectiveness of the Gaussian copula c_G to handle dependency [Joe97] and we adopt this one: $\forall\, \mathbf{y} = (y_1, \cdots, y_L) \in \mathbb{R}^L$,

$$c_G(\mathbf{y}) = |\Gamma|^{-\frac{1}{2}} \exp\left[-\frac{\tilde{\mathbf{y}}^T(\Gamma^{-1} - I)\tilde{\mathbf{y}}}{2}\right] \tag{5.47}$$

where $\tilde{\mathbf{y}} = (\Phi^{-1}(y_1), \cdots, \Phi^{-1}(y_L))^T$ with $\Phi(.)$ the standard Gaussian cumulative distribution, Γ is the inter-data correlation matrix and I the $L \times L$ identity matrix.

The proposed kernel function is given by:

$$\mathcal{K}(\mathbf{x_i}, \mathbf{x_j}) = C_G(\mathbf{x_i}, \mathbf{x_j}).\, exp(-(\mathbf{x_i} - \mathbf{x_j})^T(\mathbf{x_i} - \mathbf{x_j})/2\sigma^2) \quad \sigma > 0 \tag{5.48}$$

where $C_G(\mathbf{x_i}, \mathbf{x_j}) = (\frac{1}{N}\sum_{k=1}^{N} c_G(x_i(k), x_j(k))$ and N is the length of the vector $\mathbf{x_i}$ and $\mathbf{x_j}$. Simply expressed, the more the couple $(\mathbf{x_i}, \mathbf{x_j})$ is dependent the more $C_G(\mathbf{x_i}, \mathbf{x_j})$ is close to 1. The hyperparameters of the copula function are estimated with the Maximum Likelihood (ML) procedure. Since the new kernel is the sum and the product of positive definite kernels, it is a positive definite kernel as well [Hofmann08].

5.2.2 The SV3DH algorithm

The proposed scheme is based on two steps: 1) an initialization step 2) the SV3DH core algorithm.

Fuzzy K-means initialization

The first step of the proposed change-detection scheme is to identify two classes: the class of targets and the class of outliers which are required to initialize the SVDD classifier. In order to address the gradual transition between both classes, we apply the fuzzy K-means method [Duda01] to extract classes. To estimate the membership function defining the membership degree of an element to the class of targets, we used an S-membership function.

Let μ be the estimated membership of an object to the target class. At the end of this K-means-based initialization step, we get two hard classes and two fuzzy classes: 1) Hard class of target population: $\mu = 1$ denoted C_{ht}, 2) Fuzzy class of target population: $\mu > 0.5$ denoted C_{ft}, 3) Fuzzy class of outlier population: $0 < \mu \leq 0.5$ denoted C_{fo} and 4) Hard class of outlier population: $\mu = 0$ denoted C_{ho}. This result will be used for initializing the SVDD algorithm.

The SV3DH core algorithm

The second step aims at describing the target class by exploiting the information present in the target and outlier sets defined in the initialization step (we use the SVDD with positive and negative patterns). For this, we replaced the kernel function in Eq. 5.46 by the proposed one. The leave-one-out cross-validation estimator was used to estimate our model hyperparameters [Cawley03]. This algorithm, often cited as being highly attractive for the purposes of model selection, provides an almost unbiased estimate.

5.3 Experiments

In this section, we present the experimental results obtained with the proposed method on synthetic on real dataset which have been introduced in [Ratsch01] as a benchmark collection. The advantage of this collection is that the ground truth is available. Unfortunately, since no ground truth is available for 2D HSQC spectra, only a comparison between results obtained with the proposed method on 2D HSQC spectra and results obtained on 1D spectra (both 1D and 2D spectra are obtained from the same biopsy) is presented in chapter 6.

Let us start with artificial toys problem to demonstrate the effects of different algorithm initialization strategies. For this, we have generated three artificial toys. Samples $(x,y) \in \mathbb{R}^{10} \times \pm 1$ are drown as follows: first we fix a label y with equal probability, then:

1. Database 1: we set $x_i = g_i + z_i$ for $i \in 1,...,5$ and $x_i = z_i$, where the $z_i \sim \mathcal{N}(0,1)$ are standard normal distribution and $g_i \sim \mathcal{G}(1+y_i/4,1)$ are the Gamma distribution

where its expression is given by Eq.5.49. For both z_i and g_i, the dependency of samples is 0.5.

$$\mathcal{G}(g_i,\alpha,\beta) = k^{(\alpha-1)}\frac{\beta^{\alpha}}{\Gamma(\alpha)}exp^{(-\beta g_i)} \quad g_i > 0 \tag{5.49}$$

2. Database 2: we set $x_i = g_i + z_i$ for $i \in 1,...,5$ and $x_i = z_i$, where the $z_i \sim \mathcal{N}(0,1)$ are standard normally distribution and $g_i \sim GG(\alpha,\sigma,\mu)$ ($\sigma = 1, \mu = y_i$) are the Generalized Gaussian distribution where its expression is given by Eq.5.50. For both z_i and g_i, the dependency of samples is 0.5.

$$GG(g_i;\alpha,\sigma,\mu) = \frac{\eta(\alpha)\alpha}{[2\Gamma(1/\alpha)]}\exp\left[-(\eta(\alpha)|g_i-\mu|)^{\alpha}\right] \tag{5.50}$$

where $\eta(\alpha) = \left[\frac{\Gamma(3/\alpha)}{\sigma^2\Gamma(1/\alpha)}\right]^{\frac{1}{2}}$, $\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1}\exp^{-t}dt$ and μ, σ, α the mean, standard deviation and shape parameter.

3. Database 3: we set $x_i = y_i/2 + z_i$ for $i \in 1,...,5$ and $x_i = z_i$, where the $z_i \sim \mathcal{N}(0,1)$ are standard normally distribution. For z_i, the dependency of samples is 0.5.

Thus, all coordinates are noisy, and only the first five coordinates carry task relevant information. We draw 5000 examples which were split into 50 partitions. In order to emphasize the benefit of the proposed initialization algorithm and particularly the use of the fuzzy k-means, three different methods were used to initialize the SV3DH: the proposed method, the k-means method and the Maximum likelihood method. The results of validation on synthetic databases are summarized in Table. 5.1. As one can see, our method performed the best. This means that our initialization algorithm is well adapted to the proposed change detection method.

initialization	database1	database 2	database 3
fuzzy k-means	**4.49 ± 0.45**	**4.52 ± 0.68**	**3.52 ± 0.48**
k-means	5.39 ± 0.88	5.98 ± 0.82	4.99 ± 0.57
ML	5.12 ± 0.82	5.69 ± 0.71	4.74 ± 0.51

Table 5.1: SV3DH averaged classification error in % and the standard deviation on synthetic data sets obtained with different initialization algorithms.

Moreover, and in order to evaluate the performance of the proposed algorithm on real datasets, we considered two multitemporal remote sensing image data sets acquired from a geographical area of Alaska and Philadelphia which are available from [lsiml]. The first database (Alaska image) contains a high resolution (1305 x 1520 pixels) set of multispectral images collected on a geographical area of Alaska. These images were acquired by Landsat-5 Thematic Mapper (TM) on July 22, 1985 and July 13, 2005, respectively. An area with 1024 x 1024 pixels is selected for experiments. The Landsat-5 TM provides optical imageries using seven spectral bands, Bands 1-7. The instrument's pixel resolution is

30 m. The ground truth of the change detection maps is available in [lsiml].

The second database (Philadelphia image) contains a high resolution (2000 x 2000 pixels) set of multispectral images collected on a geographical area of Philadelphia. These images were acquired by Landsat-5 Thematic Mapper (TM) on June 28, 1988 and a Landsat-7 Enhanced Thematic Mapper (ETM+) on September 23, 1999, respectively. As the Landsat-5, the Landsat-7 provides optical imageries using seven spectral bands. An area with 1024 x 1024 pixels is selected from Philadelphia image for experiments. Pixel size for all bands is 28.5 m. This includes the Landsat 7 ETM+ thermal band which has been resampled from its 57 m resolution and the Landsat 5 TM thermal band which has been resampled from its 114 m resolution. The ground truth of the change detection maps is available in [lsiml].

For multitemporal change detection, we consider the multispectral difference image $I_\delta = I2 - I1$ on 7 spectral bands. Therefore the high dimensional information present in the multispectral difference image is considered to improve the change detection accuracy. Fig. 5.4 (resp Fig. 5.5) displays the feature distribution of the unchanged class (gray) and changed class (dark) pixels in the 2-dimensional I_δ Alaska image (resp Philadelphia image) according to the available ground truth map. As one can seen from Fig. 5.5, the change detection problem on Philadelphia image is quite more complex than that on Alaska image, as the target and outlier classes are significantly overlapped.

In order to perform the change detection evaluation, we use the False Alarm *PFA*, the Miss Detection *PMD* and the Total Error *PTE* measurements computed in percentage and defined by:

$$PFA = \frac{FA}{N_F} \times 100\%; \quad PMD = \frac{MD}{N_M} \times 100\%; \quad PTE = \frac{MD+FA}{N_M+N_F} \times 100\%$$

where *FA* stands for the number of unchanged pixels that were incorrectly determined as changed ones, N_F the total number of unchanged pixels, *MD* the number of changed pixels that were mistakenly detected as unchanged ones, N_M the total number of changed pixels.

Tab. 5.2 presents the false detection, missed detection and total errors on both databases resulting from:

- The proposed SV3DH method initialized with the fuzzy K-means algorithm,
- The SVM method with the proposed copula kernel function (SVM with Dependance Handling SVMDH) initialized with the fuzzy K-means algorithm,
- The proposed method trained using only positive examples (SV3DH+) initialized with the fuzzy K-means algorithm,
- The proposed method initialized with the k-means algorithm (hard-SV3DH),
- The SVDD with the RBF kernel function trained using positive and negative examples (SVDD) and initialized with the fuzzy K-means algorithm,

Alaska Image	False detection	Missed detection	Total Errors
SV3DH	0.71 %	5.01 %	1.09 %
SVMDH	**0.70** %	**4.96** %	**1.07** %
SV3DH+	0.78 %	5.32 %	1.18 %
Hard-SV3DH	0.84 %	5.99 %	1.29 %
SVDD	1.87 %	6.81 %	2.01 %
SVDD+	1.89 %	7.03 %	2.11 %
SVM	1.04 %	6.31 %	1.75 %
Philadelphia Image	False detection	Missed detection	Total Errors
SV3DH	**3.82** %	**14.54** %	**8.34** %
SVMDH	4.27 %	16.07 %	9.37 %
SV3DH+	4.41 %	16.84 %	9.79 %
Hard-SV3DH	4.87 %	16.93 %	10.35 %
SVDD	5.09 %	17.79 %	11.09 %
SVDD+	6.17 %	18.38 %	12.21 %
SVM	5.31 %	17.91 %	11.39 %

Table 5.2: False detection, missed detection and total errors resulting from: the proposed method SV3DH, the SVM method with the proposed copula kernel function SVMDH, the proposed method trained using only positive examples SV3DH+, the proposed method initialized with the k-means algorithm, the SVDD trained using positive and negative examples (SVDD), the SVDD trained using only positive examples (SVDD+) and the SVM method.

- The SVDD with the RBF kernel function trained using only positive examples (SVDD+) and initialized with the fuzzy K-means algorithm,
- The SVM with the RBF kernel function methods initialized with the fuzzy K-means algorithm,.

As one can remark, the SV3DH and the SVMDH perform similar results. That means that the proposed kernel function improves the features discrimination for both standard SVDD and SVM methods. Moreover, the fuzzy k-means initialization allows us to obtain a better results than a k-means initialization particularly in the high uncertainty situation (Philadelphia image). Indeed, we obtained 8.34 % of total error with the fuzzy initialization while the k-means initialization lead to 10.35 % of total errors.

In order to emphasize the benefit of the Gaussian copula for features dependency handling, we have compared the feature fit goodness of the proposed copula with five other copula functions: the t-student copula [Demarta05], the Farlie-Gumbel-Morgenstern (FGM) copula [Cossette08], the Gumbel copula, the Frank copula and finally the Clayton copula functions [Rodriguez07]. To this end, we used the copula goodness-of-fit measurement approach proposed in [Genest08]. This approach consists in measuring the discrete

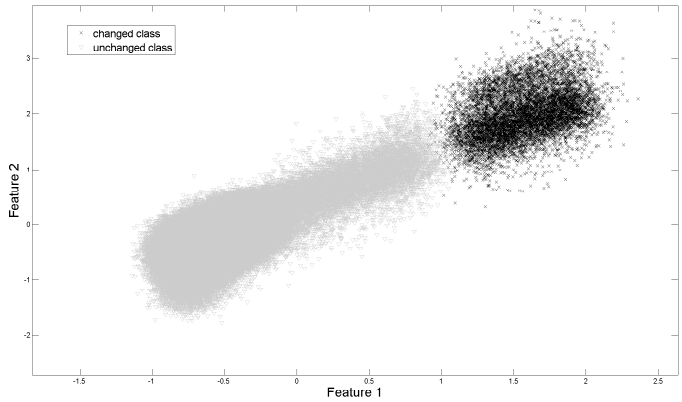


Figure 5.4: Distribution of the unchanged class (gray) and changed class (dark) pixels in the 2-dimensional I_δ Alaska image according to the available ground truth.

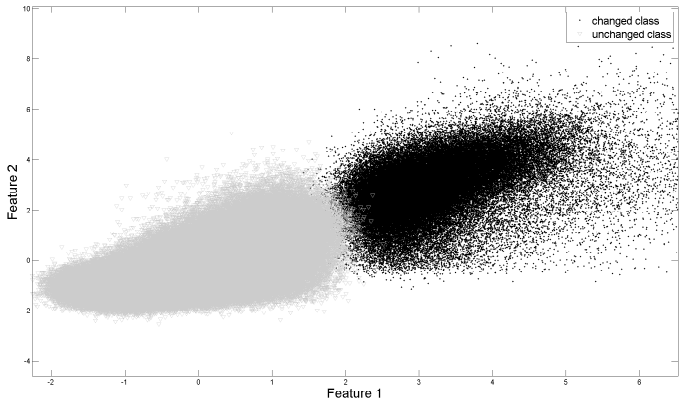


Figure 5.5: Distribution of the unchanged class (gray) and changed class (dark) pixels in the 2-dimensional I_δ Philadelphia image according to the available ground truth.

copula	Alaska image	Philadelphia image
Gaussian	**5.12** 10^{-3}	**4.02** 10^{-3}
t-student	5.48 10^{-3}	4.19 10^{-3}
FGM	5.91 10^{-3}	4.61 10^{-3}
Clayton	6.85 10^{-3}	5.17 10^{-3}
Frank	5.57 10^{-3}	4.28 10^{-3}
Gumbel	6.11 10^{-3}	4.97 10^{-3}

Table 5.3: The L^2 norm for different copula types with the empirical copula.

L^2 norm between a set of copulas and the empirical copula and then select the one with the minimum difference. We have applied this approach and the real databases. Results are presented in Tab. 5.3. As on can remark, Gaussian copula seems to be the one which better approximates the empirical copula.

Conclusion

In this chapter, the third step step in the indexing scheme is tackled. Indeed, the SV3DH method for unsupervised change-detection/classification based on SVDD has been proposed. This method could be used either for assigning a new query to the appropriate profile defined in the off-line phase (classification) or for detecting changes between two medical signal groups (change detection). We particularly focus on the formulation of change problem as a minimum enclosing ball problem with unchanged samples as target objects. The use of the dependency measurement for the first time to our knowledge in the SVDD framework increases the robustness of the proposed change-detection scheme with comparison to the classical SVM and SVDD methods. Of course, any performance gain depends on the quality of the prior samples distribution, which amounts to the quality of chosen distributions and consequently, the copula theory was used. Indeed, it provides tools to model samples dependency even if their distribution does not follow a gaussian law (HSQC spectrum peaks and fMRI object surface). Experimental results clearly indicate the benefit of the proposed method. Different applications of the proposed method on real HSQC spectra are presented in the next chapter.

CHAPTER 6

Results and discussion

Contents

Introduction

This chapter aims at experimentally validating the proposed treatment framework on real HSQC spectra and fMRI images. All comments, suggestions and conclusions presented in this chapter are the result of analysis conducted in conjunction with experts in the field of NMR: Pr. Karim Elbayed from *Institut de Chimie, University of strasbourg* and Pr Izzie Jacques Namer from *Department of Biophysics and Nuclear Medicine, University Hospitals of Strasbourg* and in the field of fMRI images: Dr Jacques Foucher from *Clinique psychiatrique, University Hospitals of Strasbourg*.

In section 6.1, we describe first the HSQC database used to perform the experiments. The second part details the complete scheme for HSQC spectrum processing. Finally, the experimental validation of the proposed approaches is presented and discussed.

After describing the real data involved in the fMRI validation and the way they have been reconstructed from the raw data, we details in the second part of section 6.2 the complete scheme for fMRI image analysis. Finally, the experimental results obtained with the proposed approaches is presented.

Finally, some conclusions are drawn based on the validation results in section 6.3.

6.1 HSQC spectrum Experiments

6.1.1 HSQC spectra data sets

Our data base contains two datasets: dataset1 and dataset2. The first one is dedicated to the Multiple Sclerosis (MS) pathology of central nerve system (c.f, Fig.6.1). Note that we have used the Experimental Autoimmune Encephalomyelitis (EAE) as a model of multiple sclerosis pathology. Dataset2 is dedicated to the colon cancer pathology. A detailed description of the database is given in Tab.6.1.

The 2D HSQC spectra were recorded on a Bruker Avance III 500 spectrometer operating at a proton frequency of 500.13 MHz. This instrument is installed at the Hautepierre University Hospital in Strasbourg and is dedicated to the analysis of biopsies by HR-MAS. It is operated by qualified scientific and medical personnel in the context of the CARMeN project. Indeed, the CARMeN project (Cancer RMN) is a consortium that gathers Strasbourg University Hospitals, Strasbourg University, CNRS, INSERM and Bruker BioSpin. It was labeled on December 20th 2006 by the Pole of Competitiveness «Therapeutic Innovations» of the Alsace region. This project aimed at creating a metabolic database in cancer research using the metabolic phenotype of tumors to identify high risk cancers and to develop personalized treatments.

The used *corpus* contains 45 referenced metabolites given by the physicians. Tab.6.2 displays different metabolites present in the *corpus* as well as the chemical shifts of their peaks in ppm. Note that for privacy concerns, the decimal digits of the chemical shift of hydrogen ^{1}H and the carbon ^{13}C are not provided.

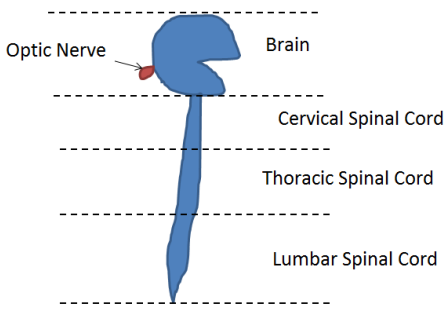


Figure 6.1: Central nerve system of the rat.

	biopsy	spectrum number
dataset1	health Optical Nerve (ON)	10
	EAE Optical Nerve (ON)	10
	healthy Cervical Spinal Cord (CSC)	10
	EAE Cervical Spinal Cord (CSC)	10
	healthy Thoracic Spinal Cord (TSC)	10
	EAE Thoracic Spinal Cord (TSC)	10
	healthy Lumber Spinal Cord (LSC)	10
	EAE Lumber Spinal Cord (LSC)	10
	biopsy	spectrum number
datatset2	healthy colon	28
	cancerous colon	28

Table 6.1: HSQC database description

Metabolite	**Abbreviation**	^{13}C	^{1}H	**Metabolite**	**Abbreviation**	^{13}C	^{1}H
Succinate	Succinate	36	2	Taurine	Tau	50	3
Theronine	Thr	63	3	Taurine	Tau	38	3
Theronine	Thr	68	4	Thrimethylamine	Thrimethylamine	47	2
Theronine	Thr	22	1	Tyrosine	Tyr	133	7
Uracil	Uracil	103	5	Tyrosine	Tyr	118	6
Uracil	Uracil	146	7	Valine	Val	31	2
Alpha-Glucose	alpha-Glc	94	5	Valine	Val	20	1
Alpha-Glucose	alpha-Glc	75	3	Valine	Val	19	0.9
Alpha-Glucose	alpha-Glc	72	3	Beta-Glucose	Beta-Glc	98	4
Alpha-Glucose	alpha-Glc	73	3	Beta-Glucose	Beta-Glc	72	3
Alpha-Glucose	alpha-Glc	63	3	Beta-Glucose	Beta-Glc	78	3
Myo-Inositole	mI	73	3	Beta-Glucose	Beta-Glc	63	3
Myo-Inositole	mI	74	4	GABA	GABA	36	2
Myo-Inositole	mI	75	3	GABA	GABA	26	1
Myo-Inositole	mI	77	3	Adrenaline	Adrenaline	35	2
Serine	Ser	59	3	Adrenaline	Adrenaline	118	6
Serine	Ser	62	3	Adrenaline	Adrenaline	121	6
Proline	Pro	26	1	Adrenaline	Adrenaline	116	6

Metabolite	**Abbreviation**	^{13}C	^{1}H	**Metabolite**	**Abbreviation**	^{13}C	^{1}H
Acetate	Ace	26	1	Alanine	Ala	53	3
Asparagine	Asn	53	3	Alanine	Ala	18	1
Asparagine	Asn	37	2	Arginine	Arg	30	1
Asparagine	Asn	37	2	Arginine	Arg	43	3
Aspartate	Asp	54	3	Arginine	Arg	26	1
Aspartate	Asp	39	2	Ascrobate	Ascrobate	81	4
Aspartate	Asp	39	2	Betaine	Betaine	56	3
Creatine	Cr	56	3	Choline	Cho	58	4
Creatine	Cr	39	3	Choline	Cho	96	3
Cysteine	Cys	27	3	Ethanol	ETHO	19	1
Cysteine	Cys	27	3	Ethanol	ETHO	60	3
Ethanolamine	Ethanolamine	60	3	Lipide (a)	FA (a)	25	1
Ethanolamine	Ethanolamine	32	2	Lipide (a)	FA (a)	34	1
Glucose 6-phosphate	G6P	73	3	Lipide (b)	FA (b)	130	5
Glucose 6-phosphate	G6P	71	3	Lipide (b)	FA (b)	132	5
Glucose 6-phosphate	G6P	65	3	Lipide (b)	FA (b)	27	2
Phosphatidylcholine	GPCho	37	3	Lipide (b)	FA (b)	28	2
Phosphatidylcholine	GPCho	62	4	Lipide (c)	FA (c)	27	1
Phosphatidylcholine	GPCho	68	3	Lipide (c)	FA (c)	36	2
Glutathione	GSH	28	2	Glutamine	Gln	33	2
Glutathione	GSH	58	4	Glutamine	Gln	29	2
Glutathione	GSH	56	3	Glutamic acid	Glu	36	2
Glutathione	GSH	46	3	Glycine	Gly	44	3
Glutathione	GSH	28	2	Glycerol	Glyc	65	3
Glutathione	GSH	33	2	Glycerol	Glyc	65	3
Hypothorine	Hypothorine	36	3	Glycerol	Glyc	74	3
Hypothorine	Hypothorine	13	0.9	Isoleucine	Ile	13	0.9
Isobutyrate	Isobutyrate	24	0.9	Isoleucine	Ile	17	1
Isovalerate	Isovalerate	24	0.9	Lactate	Lac	71	4
Isovalerate	Isovalerate	71	4	Lactate	**Lac**	**22.7**	**1.33**
Leucine	Leu	55	3	Lysine	Lys	32	1
Leucine	Leu	42	1	Lysine	Lys	29	1
Leucine	Leu	24	0.9	Lysine	Lys	41	3
Leucine	Leu	23	0.9	Lysine	Lys	24	1
Methionine	Meth	16	2	AcetylGlutamate	AcetylGlutamate	57	4
Methionine	Meth	31	2	AcetylGlutamate	AcetylGlutamate	24	2
N-Acetyl-Aspartate	NAA	24	2	Phosphocholine	Pcho	56	3
N-Acetyl-Aspartate	NAA	55	4	Phosphocholine	Pcho	60	4
Phenylalanine	Phe	131	7	Phosphocholine	Pcho	68	3
Phenylalanine	Phe	132	7	Proline	Pro	63	4
Phenylalanine	Phe	58	3	Proline	Pro	31	2
Phenylalanine	Phe	39	3	Proline	Pro	31	2
Scyllo-Inositol	Scyllo-Inositol	76	3	Proline	Pro	48	3

Table 6.2: Different metabolites present in the used *corpus* as well as the hydrogen chemical shifts ^{1}H and the carbon chemical shift ^{13}C of their peaks in ppm.

Before applying the treatment chain, a manual calibration of spectra using a peak reference (the lactate which is presented in bold in Tab. 6.2) with well known location (22.7ppm for carbon ^{13}C axis and 1.33ppm for proton ^{1}H axis) is performed. Then all the HSQC spectrum intensities are divided by their biopsies masses.

6.1.2 Treatment framework

We recall that our indexing scheme is composed of an off-line step which consists in establishing the profile of each diseases and then an on-line step which aims at assigning a new individual/group to a pathologic/healthy profile.

In the case of the HSQC spectrum analysis, the considered objects are the peaks within the spectrum. A summary of the treatment framework is presented in Fig. 6.2.

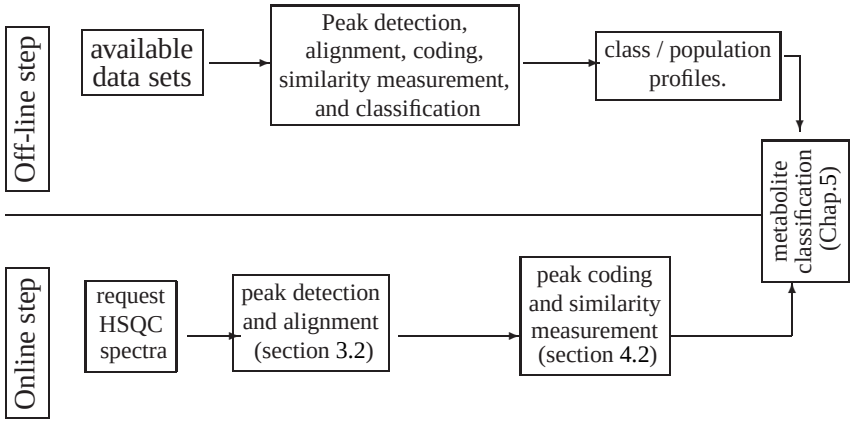


Figure 6.2: Overview diagram of the proposed classification framework for HSQC 2D NMR spectra.

The first step of the indexing framework consists in detecting and aligning different peaks within spectra (an overview of the proposed method is presented in Fig.6.3). This step requires up to 3h30 of computation time with Intel 2.66 GHz and a combination of C and matlab codes. Once this step achieved, we turn to the metabolite similarity measurement step. This step consists in identifying different metabolites in the spectra using the

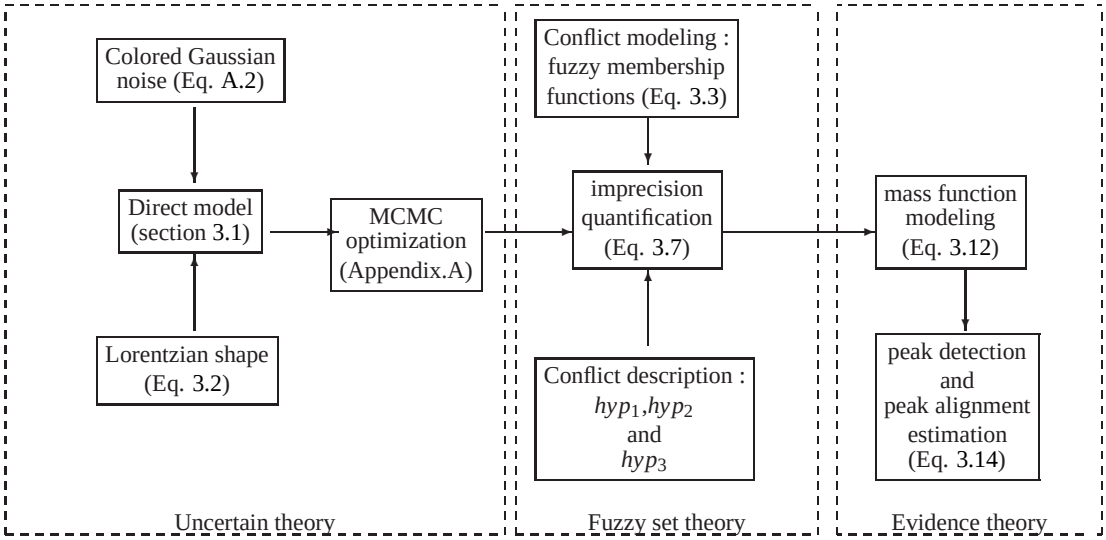


Figure 6.3: Overview diagram of the peak detection and alignment chain.

corpus (Fig.6.4). This step requires up to 15mn of computation time with Intel 2.66 GHz and matlab codes.

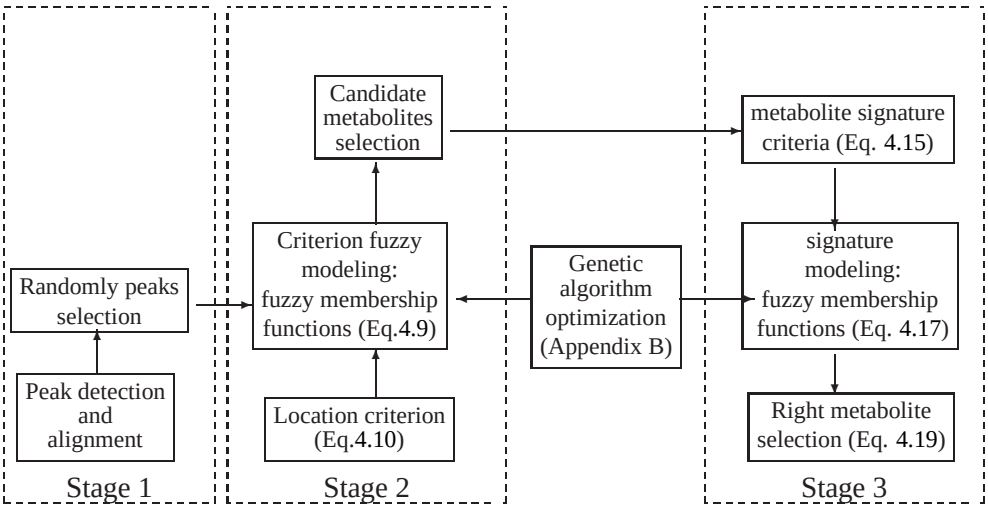


Figure 6.4: Overview diagram of the metabolite identification chain.

Once the metabolite is identified, we address the metabolite classification/change detection step. This step aims at comparing a new individual/population of spectra by identifying changed metabolites from unchanged ones. In order to achieve a more accurate metabolite change detection results, we introduce more *a priori* knowledge we have on the spectra into the change detection scheme. As a matter of fact, we assume that the residual spectrum image (the difference between the observed and the parameterized spectrum) is the same for all observations. Indeed, as the acquisition system is isolated from the outside environment, the characteristics of acquisition noise should be the same. Thus, any difference between two residual images is due either to errors in estimating the hyperparameters of the spectrum, to a deterioration of tissues or to a modification in biopsy features (*e.g*; the pH). To achieve a better change detection result, all these disturbances must be taken into account. To this end, we propose a method to compensate the estimation error within the change detection scheme (Alg.2). Finally, the overview diagram of the metabolite classification/change detection chain is described in Fig.6.5. Note that, this step requires up to 20mn of computation time with Intel 2.66 GHz and matlab codes.

Algorithm 2 Error compensation estimation

Input: a reference spectrum Y_{ref}, another spectrum Y as well as their parameterized (reconstructed) forms $Y^{rec,\,ref}$ and Y^{rec} (Eq.3.1).

1- For each two assigned peaks $x_{ref}(i,j)$ and $x(i,j)$ estimated from Y_{ref} and Y (Eq.3.1) :

- Extract the local spectrum areas containing $Y_{local,\,ref}$ and Y_{local}: $Y^{rec,ref}_{local,\,ref}$ and Y^{rec}_{local} respectively.

- Calculate the covariance matrix Γ^{ref}_{local} (resp Γ_{local}) of $\left(Y_{local,ref} - Y^{rec,ref}_{local,\,ref}\right)$ (resp $\left(Y_{local} - Y^{rec}_{local}\right)$).

- Perform the Principal Component Analysis PCA algorithm on both Γ^{ref}_{local} and Γ_{local}

- Let λ_{ref} (resp λ) be the largest eigenvalue of the PCA decomposition performed on Γ^{ref}_{local} (resp Γ_{local}): the estimated peak amplitude $x(i,j)$ is normalized as follows:

$$x(i,j) = \frac{\lambda}{\lambda_{ref}}\, x(i,j) \tag{6.1}$$

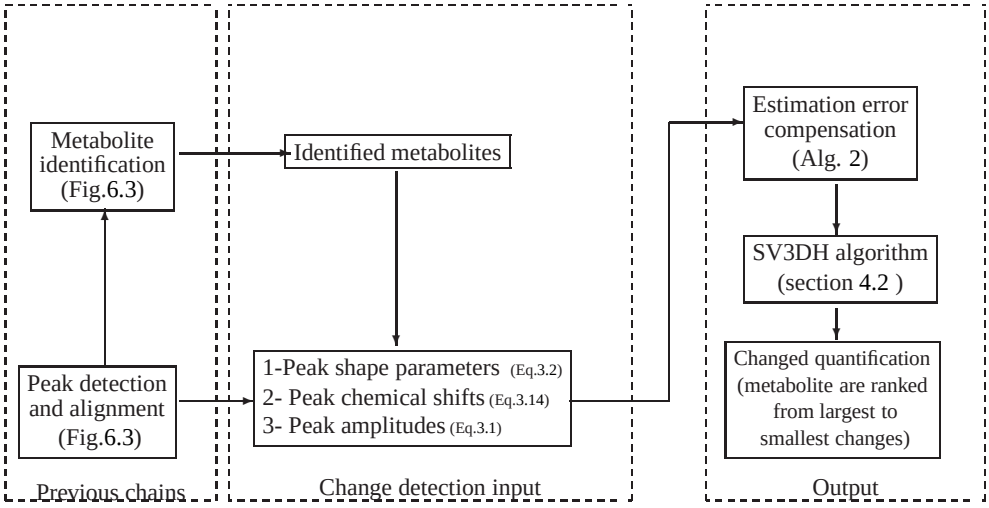


Figure 6.5: Overview diagram of the metabolite classification/change detection chain.

In the practice, the proposed metabolite classification/change detection (SV3DH) algorithm can be used for three different uses:

1. **Metabolite change detection**: given two spectrum populations, we aim at detecting the changed metabolites from unchanged ones. In this case, a feature vector x_{M_i} is associated to each metabolite M_i. The vector x_{M_i} consists of the peak amplitudes of the metabolite (Eq.3.1), the peak shape hyperparameters (Eq.3.2) and the peak chemical shifts (Eq.3.14). As one can remark, each metabolite is separately treated from the other ones. The first step consists in selecting a spectrum population and then estimating the hypersphere hyperparameters (a_{M_i}, R_{M_i}) that model the profile of each metabolite M_i using the SV3DH method. Then, the decision function f (Eq. 5.46) is applied to each spectrum belonging to the second population. In order to quantify the metabolite change degree denoted μ_{M_i}, we propose the use of an S-membership function f_1 (Eq.3.3). The expression of μ_{M_i} is given by:

$$\mu_{M_i} = f_1(f(x_{M_i})/R_{M_i}) \tag{6.2}$$

   where (a_{M_i}, R_{M_i}) are the hypershpere hyperparameters. Therefore, the more μ_{M_i} is close to 1 the more the metabolite M_i has changed. Thus the metabolites can be ranked according to their change degrees allowing the physicians to select the relevant changed metabolites by a simple threshold and then to control the results. In order to facilitate the threshold setting, we have classified the metabolites according to the change degree into two classes: "weak change" and "high change". To this end, we used a Maximum Likelihood (ML) classifier with respect to the statistical distribution law of metabolite change degrees. For example, the estimated threshold is about 0.4 for dataset1 et 0.35 for dataset2. Note that μ_{M_i} is assumed to follow a Generalized Gaussian distribution. Moreover, as we already remarked, μ_{M_i} mainly depends on the hypersphere radius R_{M_i} (the distance from the center of the sphere a_{M_i} to any of the support vectors on the boundary and the feature vector x_{M_i}. In order to quantify our uncertainty on the metabolite change degree estimation, we assign to each μ_{M_i} a confidence margin which is equal to:

$$\varepsilon_{M_i} = f_1((e_{M_i})/R_{M_i}) \tag{6.3}$$

   where e_{M_i} consists of the difference between the estimated metabolite peak amplitudes (Eq.3.1) and the normalized peak amplitude (Eq.6.1). Therefore, the more the difference is great the more the metabolite peaks amplitude estimation is unreliable and the more ε_{M_i} is great. In the practice, the f_1 hyperparameters (a_1, b_1, c_1) are set as follows:

   - $a_1 = 1$,
   - $b_1 = median(\{\mu_{M_i}\}_{i=1...N}/\mu_{M_i} > 1)$, where N is the number of metabolites within the spectra and $median(.)$ is the median function,
   - $c_1 = \max(\{\mu_{M_i}\}_{i=1...N}/\mu_{M_i} > 1)$.

Once all metabolite change degree for each spectrum belonging to the second population are estimated, a *p-value* is associated with each metabolite allowing to reject the null hypothesis H_0 (changed metabolite). Hence, computing the significance of the metabolite change over all the second population spectra amounts to compare $p(\mu_{M_i} > 0.1/H_0)$ with α (generally set to 10^{-3}). Thus, if $p(\mu_{M_i} > 0.1/H_0) < \alpha$ the null hypothesis H_0 cannot be rejected.

2. **Spectrum discrimination**: given a set of spectra, we aim at discriminating two groups. In this case, a feature vector x_{X_s} is associated to each spectrum X_s. The vector x_{X_s} consists of the peak amplitudes of the spectra (Eq.3.1). As one can remark, all metabolites are conjointly treated. Note that in spectrum discrimination case, the SV3DH method allows us to directly obtain the discrimination results and no statistical test is required.

3. **Spectrum classification**: given several spectrum populations $P_{kk>1}$, we aim at classifying a new spectrum X. As the spectrum discrimination case, the feature vector contains all the spectrum peaks. The first step consists in estimating the hypershpere hyperparameters (a_{P_k}, R_{P_k}) modeling the profile of each population P_k (or the **class profile**) using the SV3DH method. Then, the decision function is applied to the new spectrum X which will be assigned to the population with the lowest $(f(x_X)/R_{P_k})$ value.

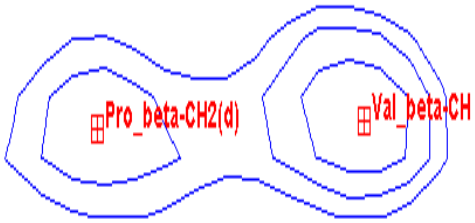
6.1.3 Results on real spectra

The results obtained by the proposed methods on real HSQC spectra are presented in two parts. In the first one, we focus on some case studies in order to emphasize the benefit of the proposed schemes and particularly the use of the deconvolution model for peaks detection and the use of fuzzy set theory to deal with the ambiguity which is in the heart of the metabolite identification task. In the second part, some metabolite change detection results and a comparison with results obtained with 1D spectra are presented.

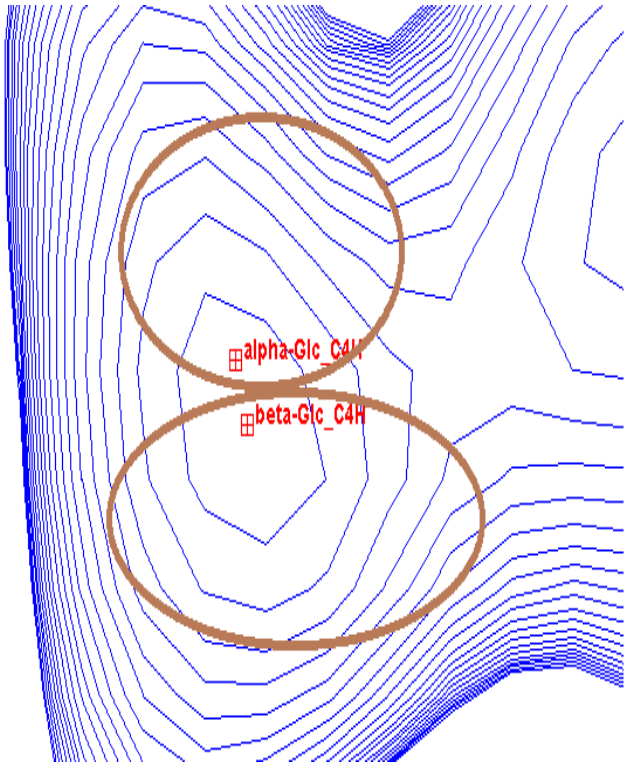
Metabolite identification results

The first step in the HSQC spectrum analysis is the peak detection. This task is very important since all the rest of the processing chain depends on it. Indeed, a poor peak detection can cause the fail of the framework. For this reason we have paid an important attention to this task. We recall that the proposed peak detection algorithm relies on the deconvolution model to achieve a better fit of the HSQC spectrum (Eq.3.1) which allows us to overcome the problem of peak overlap. For example, Fig.6.6.(a) shows two peaks that can be easily detected without a deconvolution step. However, in some cases (*e.g* Fig.6.6.(b)) two peaks could be overlapped and then a manual peak extraction seems to be a difficult task. Thus, the peak deconvolution allows us to overcome such problem.

However, the problem of peak overlap is unfortunately not the single one to be addressed to automatic the peak analysis process. In fact, the high complexity of this type of



(*a*)



(*b*)

Figure 6.6: Examples of (a) two separated peaks, (b) two overlapped peaks. In both cases, the proposed evidential peak detection method has correctly identified peaks.

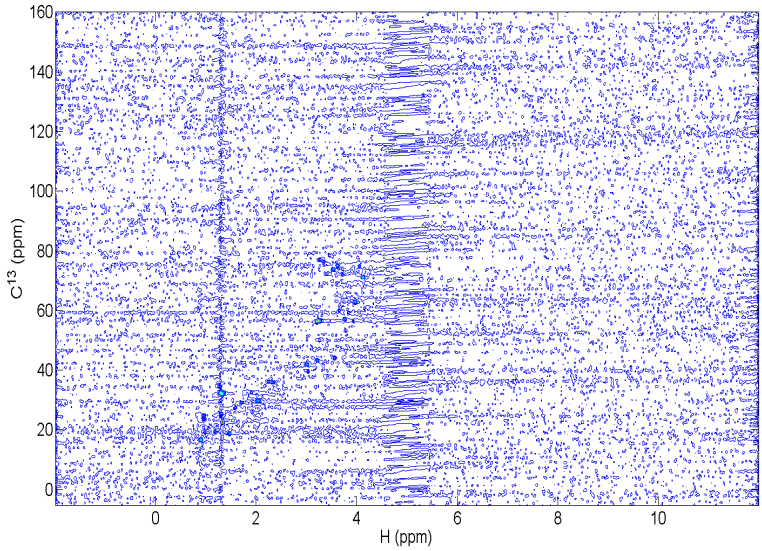


Figure 6.7: A real HSQC spectrum

spectra brought by the noise and the artifacts make the detection task more delicate and difficult [Becker00]. Indeed, due to experimental conditions, correlated vertical lines appear sometimes in the observed spectrum [Becker00]. In order to overcome this problem, we proposed the use of the multivariate Gaussian distribution to model the noise correlation. We recall that the synthetic validation of the proposed approach has shown its robustness to the high level of noise (Tab.4.1). Unfortunately, the validation task on real spectra is not trivial since no ground/absolute truth is available. A manual validation by NMR experts is then required in order to assess the performances of the proposed scheme on real spectra. Note that the only available ground truth is the biological nature of the biopsies (a metabolite exists or does not exist) and hence a peak presence or absence does not have a biological sense. Therefore, our NMR experts have validated the metabolite identification results and thus explicitly the peak detection and alignment results. The validation results show that most metabolites belonging to the *corpus* are properly detected and identified. For example, only 3/45 metabolites were wrongly identified in the dataset2.

Metabolite change detection results

We are now faced with the problem of metabolites change detection on real HSQC HR-MAS 2D spectra. We recall that our first objective is the determination of discriminant parameters between two states of the same biological system (*e.g.*, two evolution stages of the same type of tumor). Indeed, the identification of metabolic fingerprint associated with specific biological states could reveal metabolic differences related to different considered groups of spectra (*e.g*, healthy and pathological groups of spectra). In a second step, these metabolic fingerprints (discriminants or biomarkers) should allow us to under-

stand the metabolic processes associated with each treated group of spectra and then to establish the group profiles.

Metabolite change detection results on dataset1

In the following, we detail the change detection results on the dataset1 (EAE disease). Since no ground truth is available, we will compare our results with those obtained with the 1D spectrum experiments to study the result coherence (concordance). To this end, we distinguish three different experiments: the 1D experiment and the 2D HSQC experiment on the same sample. In the last one, we repeat the 1D experiment on the sample taken from the spectrometer after the 2D experiment. Indeed, due to the high time sample spinning (rotation) during the 2D HSQC experiment, the concentrations of some metabolites particularly the Ace, Ala, Asp, Cr, Cho, GABA, Glu Gly Lac, mI, NAA, Pcho and Tau may be changed [Detour11].

To facilitate the result interpretations, we used a table for each configurations of EAE/Healty spectrum group comparisons (c.f, Tab 6.1). In this table, we distinguish four typefaces:

1. Normal font (*e.g*, Asp). In this case, the metabolite was detected as a changed metabolite only with the 2D experiment.
2. Italic font (*e.g*, *Asp*). In this case, the metabolite was detected as a changed metabolite with the 2D experiment and the first 1D experiment.
3. Bold font (*e.g*, **Asp**). In this case, the metabolite was detected as a changed metabolite with the 2D experiment and the second 1D experiment.
4. Italic bold font (*e.g*, ***Asp***). In this case, the metabolite was detected as a changed metabolite with the 2D experiment, the first 1D experiment and the second 1D experiment.

Note that the metabolites detected as changed only with the first and second 1D experiment are presented in the table caption. We present likewise in the same table the metabolite concentration changes ("inc" for concentration increase and "dec" for concentration decrease) as well as the metabolite change detection degree μ_M (Eq.6.2).

Metabolite	μ_M	concentration
Arg	0.519	inc
Ser	0.461	inc
Asp	0.390	inc
Cho	0.331	dec
Pcho	0.327	dec
Glu	0.321	dec
Ace	0.298	dec
Ala	0.214	dec
Ile	0.214	dec
Gln	0.267	dec
Ascorbate	0.218	inc
mI	0.165	dec
Cr	0.166	dec
Gly	0.168	inc
NAA	0.150	inc
Lys	0.144	inc
Lac	0.110	dec

Table 6.3: Identified biomarkers for healthy ON biopsy vs. EAE ON biopsy with 2D spectra. The changed metabolites identified with the first 1D experiment are: GPCho and N-acetylglutamate. The changed metabolites identified with the second 1D experiment are: Thr, Succinate, GABA and Tau. As one can see, we identify with the 2D experiment most metabolites identified with the first and the second 1D experiment. Moreover, some metabolite like the "NAA", the "Ala" and the "Lys" were identified only with the first 1D experiment and the 2D experiment. We recall that the concentrations of these metabolites change during the 2D experiment and for this reasons they were not detected in the second 1D experiment. However, we still able to identify them as changed metabolites. This could be explained by the fact that we used in our SV3DH algorithm beside the metabolite concentration other features like the peak chemical shifts and the peak shapes for change detection. In addition, we remark that the Arg, Ile, Gly and Ser were detected as changed metabolites only with the 2D experiment.

Metabolite	μ_M	concentration
alpha-Glc	0.824	inc
Hypotaurine	0.834	inc
beta-Glc	0.724	inc
mI	0.684	dec
Gln	0.642	inc
Cho	0.627	dec
Thr	0.546	inc
GPCHO	0.510	inc
Ace	0.464	dec
Tau	0.446	inc
Asp	0.394	dec
Glu/Gln	0.389	inc
Lys	0.347	inc
Ser	0.341	inc
Ala	0.260	inc
Arg	0.257	inc
PCho	0.239	dec
Ethanolamine	0.234	dec
Gly	0.179	dec
Lac	0.124	dec
Val	0.116	dec

Table 6.4: Identified biomarkers for healthy CSC biopsy vs. EAE CSC biopsy with 2D spectra. The changed metabolites identified only with the first 1D experiment are: Succ, Ile and NAA. The changed metabolite identified only with the second 1D experiment is the Cr. As one can remark, we identify with the 2D experiment most metabolites identified with the first and the second 1D experiment. Moreover, some metabolite like the "Ser", the "Gly" and the "Gln" were identified only with the first 1D experiment and the 2D experiment.

Metabolite	μ_M	concentration
beta-Glc	0.807	inc
Ace	0.747	dec
Cho	0.688	dec
Ser	0.659	inc
Alpha-Glc	0.654	inc
Gln	0.646	inc
Thr	0.634	inc
mI	0.604	dec
Lys	0.590	inc
Asp	0.516	dec
Ethanolamine	0.462	dec
Tau	0.371	inc
Glu/Gln	0.263	inc
Tyr	0.256	inc
Ile	0.250	inc
Gly	0.221	dec
PCho	0.176	dec
GPCho	0.161	dec

Table 6.5: Identified biomarkers for healthy TSC biopsy vs. EAE TSC biopsy with 2D spectra. The changed metabolites identified only with the first 1D experiment are: Ala, Succ, Hypotaurine, Val and NAA. The changed metabolites identified only with the second 1D experiment are: Ala, Hypotaurine. As one can remark, we identify with the 2D experiment most metabolites identified with the first and the second 1D experiment. However, the Ala and the Hypotaurine were detected only with the first and the second 1D experiments.

Metabolite	μ_M	concentration
mI	0.978	dec
beta-Glc	0.942	inc
Thr	0.904	inc
Glu/Gln	0.669	inc
Gln	0.619	inc
Gly	0.606	inc
Ser	0.494	inc
Ace	0.479	dec
Cho	0.429	dec
Arg	0.409	inc
Ile	0.361	inc
Asp	0.382	dec
Ala	0.370	inc
Cr	0.320	dec
PCHo	0.255	dec
Ethanolamine	0.213	dec
Val	0.194	inc
Thr	0.131	inc

Table 6.6: Identified biomarkers for healthy LSC biopsy vs. EAE LSC biopsy with 2D spectra. The changed metabolites identified only with the first 1D experiment are: Lys, Hypotau, Succ, Tau, PCho, GPCho and NAA. The changed metabolites identified only with the second 1D experiment are: Lys, Hypotau, Tau and GPCho. As one can remark, we identify with the 2D experiment most metabolites identified with the first and the second 1D experiment. However, as one can observe, some metabolites like the Lys and Hypotau were detected only with the first and the second 1D experiments while the Val was detected only with the 2D and the second 1D experiments.

Metabolite	μ_M	concentration
Arg	0.452	inc
Tau	0.363	dec
Val	0.291	inc
Gln	0.256	inc
NAA	0.198	inc
Thr	0.116	inc

Table 6.7: Identified biomarker for healthy CSC biopsy vs. healthy LSC biopsy with 2D spectra. In this experiment, no changed metabolite was detected with the first and second 1D experiments while 6 metabolites were be detected with the 2D experiment. This means that our method is able to discriminate the healthy CSC biopsy from the healthy LSC biopsy.

Metabolite	μ_M	concentration
mI	0.341	inc
Arg	0.338	inc
Tau	0.320	dec
Gln	0.311	inc
Glu/Gln	0.224	inc
NAA	0.211	inc
Tyr	0.184	dec
Cho	0.161	inc
Cr	0.157	inc
PCho	0.149	inc

Table 6.8: Identified biomarkers for healthy CTC biopsy vs. healthy LTC biopsy with 2D spectra. In this experiment, no changed metabolite was detected with the first and second 1D experiments while 10 metabolites were be detected with the 2D experiment.

Metabolite change detection results on dataset2

We address now the metabolite change detection on colon cancer HSQC spectra. Fig. 6.8 (a) shows a healthy colon biopsy spectrum whereas Fig. 6.8 (b) displays a cancerous colon biopsy spectrum. The mean image of the 25 reconstructed healthy spectra (after peak detection with the MCMC procedure) is presented in Fig. 6.8 (c). Fig. 6.8 (d) displays the mean image of the 25 reconstructed cancerous spectra. As one can remark, all the spectrum noise was removed and only the relevant peaks (belonging to the corpus) were preserved allowing an easier interpretation of the spectra by physicians. Fig. 6.9 presents some metabolites change detection results on a cancerous spectrum (drawn in red arrows). As one can see, it is difficult to manually detect changed metabolites. The metabolite change detection results on dataset2 are presented in Tab.6.9.

Metabolite	μ_M	concentration
Tau	0.895	inc
mI	0.771	dec
beta-Glc	0.638	dec
Asp	0.621	inc
Glu	0.502	inc
Lac	0.445	inc
Pcho	0.394	inc

Table 6.9: Identified biomarkers for healthy colon biopsy vs. cancerous colon biopsy with 2D spectra. As one can remark, we identify the same metabolite as in first 1D experiment. Nota that no second 1D experiment was performed in the case of the colon biopsy spectra.

All these results for both datasets were examined and validated by NMR experts. The first conclusion that 2D experiment results confirm those obtained with 1D experiments. Secondly, the 2D experiment can be used to discriminate some spectrum populations (c.f, Tab.6.7 and Tab.6.8).

Spectrum classification results

We address now the spectrum classification validation. To this end, we used the Leave-One-Out Cross-Validation (LOOCV). Indeed, this validation involves to use a single observation from a spectrum group as the validation data (the spectrum query), and the remaining spectra as the training data (the spectrum group profiles) (c.f, Fig. 6.2). In other words, we select a spectrum form the database and we try to assign it to one of the available spectrum groups. In our case, we dispose of ten spectrum groups: the ON healthy group, the ON EAE group, the CSC healthy group, the CSC EAE group, the TSC healthy group, the TSC EAE group, the LSC healthy group, the LSC EAE group, the colon healthy group and finally the colon cancerous group (c.f, 6.1). In order to emphasize the benefits of the proposed classification method and particularly the use of the copula kernel function, we use two methods for spectrum classification: the proposed SV3DH method and the classical SVDD method with a RBF kernel function. Note that we trained both methods only with target class (c.f, Chap.5). Tab 6.10 shows the spectrum classification results. As one can remark, most of the spectra were correctly classified except the ON group. This can be explained by the high biopsy degradation during the HSQC experiments. Moreover, the proposed method performs the best comparing to the classical SVDD method.

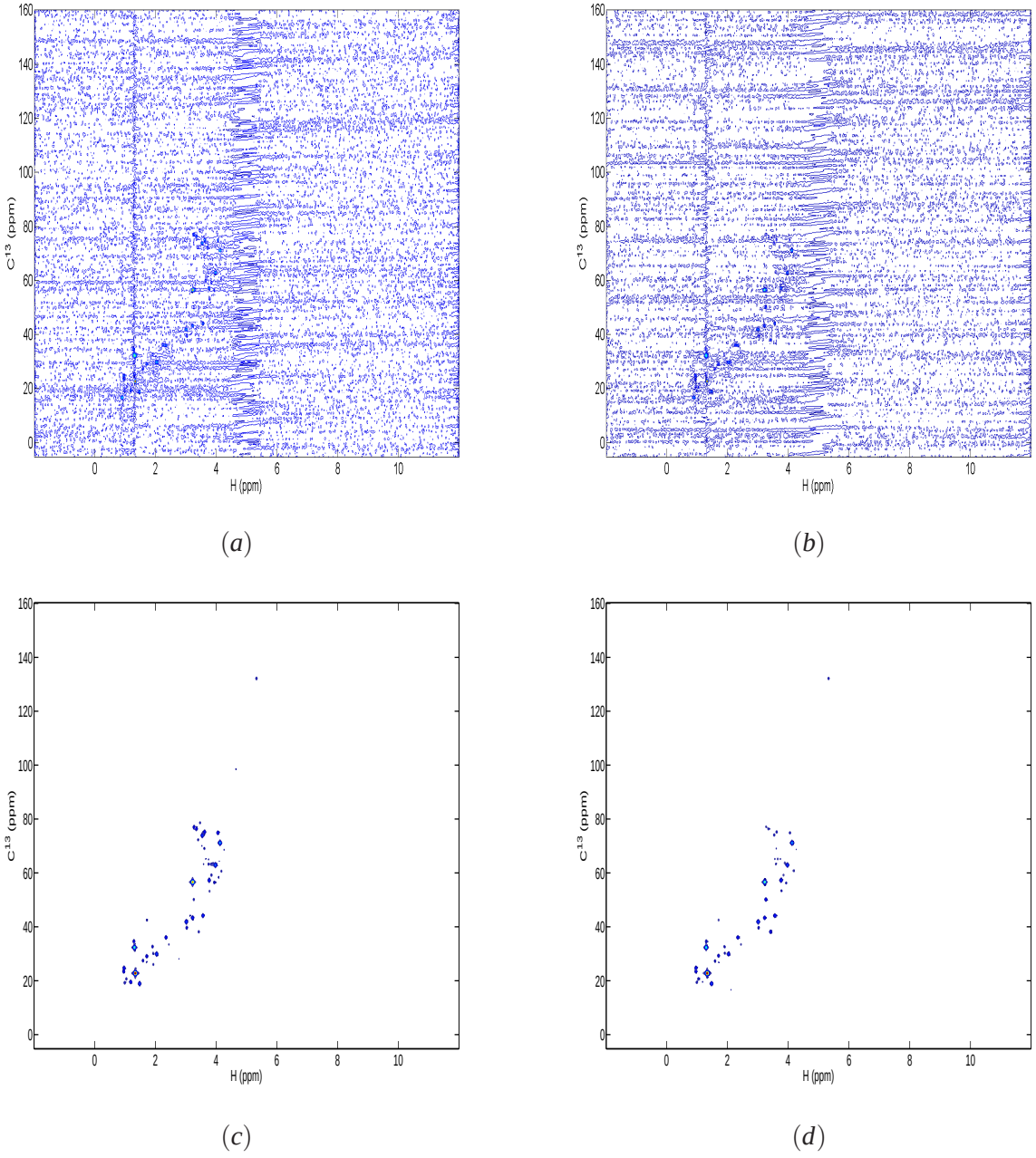


Figure 6.8: Example of (a) a healthy colon biopsy spectrum, (b) a cancerous colon biopsy spectrum, (c) The mean image of the 28 reconstructed healthy spectra, (d) The mean image of the 28 reconstructed cancerous spectra.

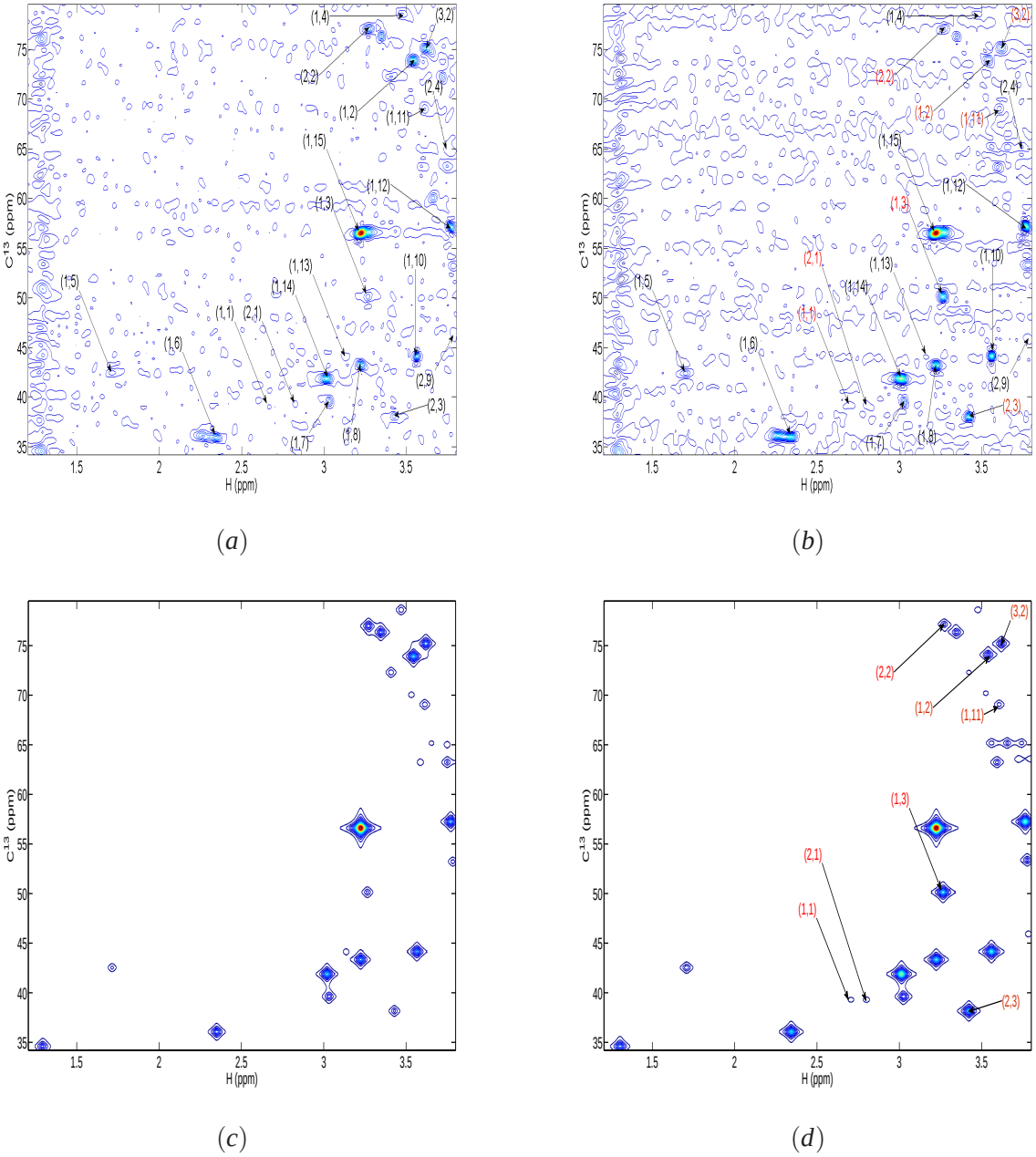


Figure 6.9: Identified metabolites in the same area of (a) a healthy colon biopsy spectrum, (b) a cancerous colon biopsy spectrum. The changed metabolite is presented in red arrows where every peak p belonging to a given metabolite M is labeled with (p,M). (c) The mean image of the 28 reconstructed healthy, (d) The mean image of the 28 reconstructed cancerous of the same area.

Spectrum group	number of correct classification	number of wrong classification
ON healthy group	SV3DH 8 SVDD 6	SV3DH 2 SVDD 4
ON EAE group	SV3DH 9 SVDD 8	SV3DH 1 SVDD 2
CSC healthy group	SV3DH 10 SVDD 9	SV3DH 0 SVDD 1
CSC EAE group	SV3DH 10 SVDD 10	SV3DH 0 SVDD 0
TSC healthy group	SV3DH 10 SVDD 8	SV3DH 0 SVDD 1
TSC EAE group	SV3DH 9 SVDD 8	SV3DH 1 SVDD 2
LSC healthy group	SV3DH 10 SVDD 9	SV3DH 0 SVDD 1
LSC EAE group	SV3DH 10 SVDD 10	SV3DH 0 SVDD 0
colon healthy group	SV3DH 27 SVDD 25	SV3DH 1 SVDD 3
colon cancerous group	SV3DH 28 SVDD 26	SV3DH 0 SVDD 2

Table 6.10: Spectrum classification results with the SV3DH method and the classical SVDD method.

Spectrum discrimination results

Let's turn now to the spectrum discrimination validation. In this experiment, we try to discriminate between two groups of the same biopsy: healthy group and pathological group. We compared likewise our method to the classical SVDD and SVM methods. Tab 6.11 shows the spectrum discrimination results. As one can remark, most of the spectra were correctly discriminated. Moreover, the proposed method performs the best in all tested configurations compared to the classical SVM and SVDD methods.

Spectrum group	number of wrong discriminated spectra	
ON healthy *vs*. ON EAE groups	SV3DH	3/20
	SVM	4/20
	SVDD	6/20
CSC healthy *vs*. CSC EAE groups	SV3DH	0/20
	SVM	3/20
	SVDD	3/20
TSC healthy *vs*. TSC EAE groups	SV3DH	1/20
	SVM	1/20
	SVDD	2/20
LSC healthy *vs*. LSC EAE groups	SV3DH	0/20
	SVM	2/20
	SVDD	4/20
Healthy *vs*. cancerous colon groups	SV3DH	2/56
	SVM	4/56
	SVDD	4/56

Table 6.11: Spectrum discrimination with the SV3DH, the SVDD and the SVM methods.

6.2 fMRI experiments

6.2.1 Material and Database

Participants

After giving written informed consent, 62 right-handed healthy participants (age 37.3 $\pm$ 7.9 years) with no history of neurological or psychiatric disorders underwent a resting-state fMRI session. This study was part of a protocol approved by the local Ethics Committee. Participants were instructed to lie down with their eyes closed without falling asleep.

Data acquisition

Four hundred and five whole-brain T2*-weighted echo planar images were acquired interleaved on a 2T Bruker scanner (Ettlingen, Germany) (session parameters: TR = 3 s; flip angle = 90°; TE = 43ms; FOV = 256 mm x 256 x mm x 128mm; Imaging matrix = $64 \times 64 \times 32$; 4 mm isotropic voxels, with fat saturation preparation) during 20 minutes.

Preprocessing

After conversion to Analyze format, images were preprocessed using Statistical Parametric Mapping toolbox v99 (Welcome Department of Cognitive Neurology, London, UK) working on Matlab R2009b (The MathWorks, Inc., Sherborn, MA, USA).
For each participant, the first 5 images were removed to account for T1 partial saturation. The 400 remaining images were then motion corrected, and all the volumes were realigned on the 200th volume (sinc interpolation).

Statistical analyzes of fMRI data

For each participant, Independent Component Analysis (ICA) [Te-Won98] was performed using FMRLAB toolbox 2.3 (Swartz Center for Computational Neuroscience, University of San Diego, CA, USA) with an implementation of INFOMAX algorithm [Theis03]. Since we planned to capture even small spontaneous activities for medical application, the dimension of the data was only reduced from 400 to 250 using a principal component approach for each participant. This procedure allowed maintaining the computational time for the algorithm to converge in acceptable limits. For display purpose, the components were superimposed on the (EPI) (Echo-Planar Imaging) mean image at a threshold of ± 1.5 standard deviation [Weiskopf05]. Each ICA component is called a Spontaneous Activity Map (SAM). At the end of the ICA algorithm, we obtain 250 SAM.

Selection of relevant SAM

In order to only select the relevant SAM, several criteria should be respected. Indeed, these criteria suppose that the whole brain volume is displayed with positive and negative parts of the spatial components overlapped on the mean EPI (z-score is above or below ± 1.5). The time course has to be evaluated on the component time course more than on the average region of interest time course. *Positive selection criteria*: a plausible BOLD signal is expected to fit with every following criteria for the whole cerebral volume or time course:

1. Spatially coherent positive or negative 3D blobs, i.e. within and between slices. The component can be followed on slice series and its parts look like a 3D coherent map. In the case of an interleaved acquisition, a signal occurring every two slices is not considered as spatially coherent.
2. The spatial distribution of the blobs overlaps on grey matter only.
3. The signal time course is in the appropriate frequency range, i.e., most of the power is below the frequency of the hemodynamic response (< 0.08 Hz), but oscillating at a higher frequency than 0.004 Hz (max. period of 2 min.).

Examples of typical SAMs are shown in Fig. 6.10.(1) to (5) (radiological convention) (1: Default Mode Network DMN; 2: verbal working memory network; 3: visuospatial working memory/attentional network; 4-5: visual networks).

Rejection criteria. To avoid artifact or noisy components, none of the following criteria should be present anywhere in the volume or the time course or represent a negligible aspect of it:

1. A spatial alternating aspect, *i.e.*, a juxtaposition of significantly correlated and anti-correlated voxels, alternating in space and sometimes appearing like a reticule (Fig. 6.10.(6)).
2. A spatial noisy aspect, i.e. the voxels are mildly significant and disseminated (Fig. 6.10.(7)).
3. Brutal crash or slow drift on the temporal time course.
4. No aspect of any known artifacts:

    - Head motion artifacts. Translation or rotation movements are surrounding high spatial contrasts (Cerebrospinal fluid (CSF)/brain, CSF or brain/skull etc) sometimes with a symmetrical aspect (positive correlation on one side and negative on the other side). Temporal course comes with slow drift or brutal crashes. Fig. 6.10.(8) shows the aspect of a z-translation residuum after registration.
    - Ocular movements artifacts: signal in the eyeball with more or less trails in the encoding phase axis, Dirac spike on the time course (Fig. 6.10.(9a) and (9b))
    - CSF-pulse artifacts. Arterial pulse and respiration induce CSF flux and this T1 partial saturation effect leads to signal fluctuation in sensitive regions, i.e., the temporal pole, Sylvian sulcus, skull base around the circle of Willis, aqueduct of Sylvius, foramen of Monro or ventricles (Fig. 6.10.(10)). The temporal course is mostly at high frequency.
    - Scanner artifacts, *i.e.*, radio frequency (trails of alternating significant voxels) or Analogic-Digital converter artifacts (signal drop or instability in one slice).

The pretreatment chain is presented in Fig. 6.11.

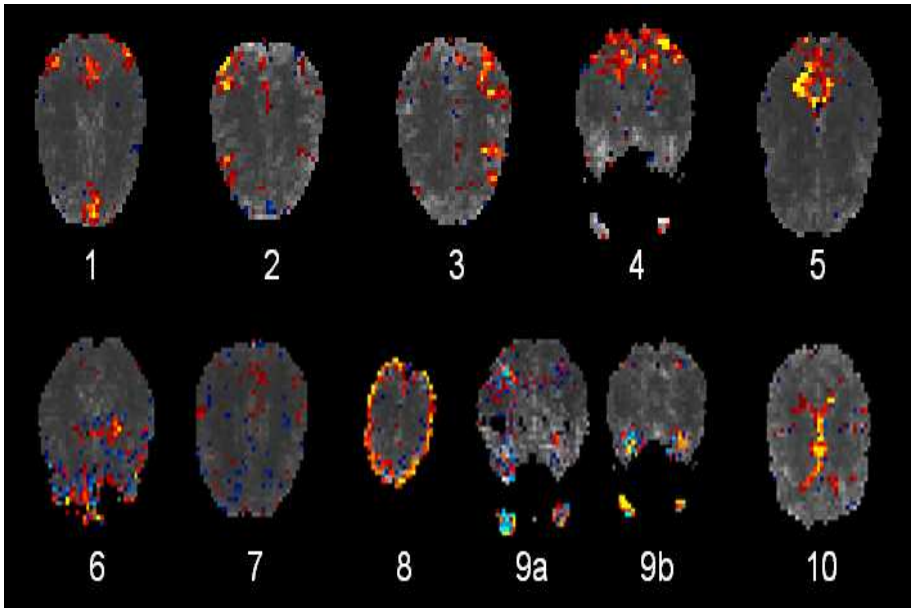


Figure 6.10: Spontaneous Activity Map (SAM) selection criteria

Figure 6.11: Overview diagram of the fMRI pretreatment chain.

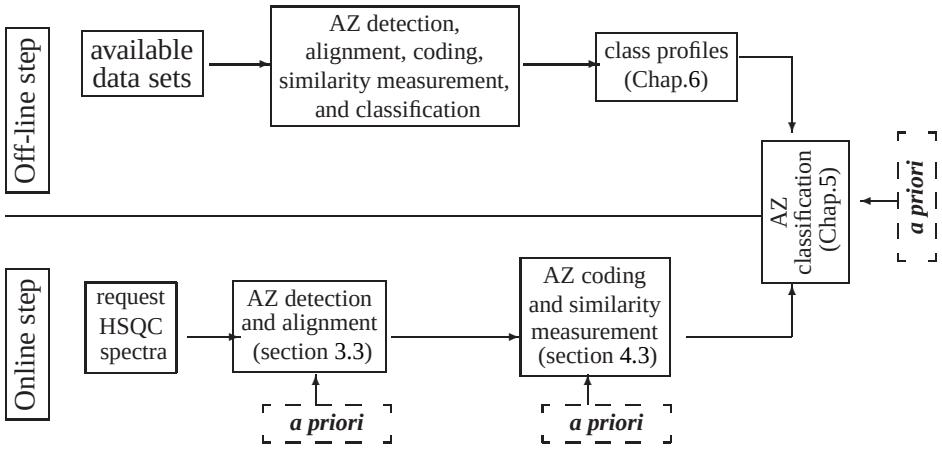


Figure 6.12: Overview diagram of the proposed classification framework for fMRI images.

6.2.2 fMRI treatment framework

In the case of the fMRI image analysis, the considered objects are the active zones (AZ). A summary of the treatment framework is presented in Fig. 6.12.

The first step of the indexing framework consists in detecting and aligning different active zones within fMRI images. To this end, we firstly perform an active zone detection step using a Hidden Markov Chain (HMC) segmentation algorithm [Bricq08]. However, when only weak differences occurred between samples of the considered classes (*i.e*; active zone with positive intensity class, active zone class with negative intensity class and no active zone), the probability density function (pdf) of one calss may be confused with the pdf of another class (*e.g.*, the Hidden Markov Model method generally tries to regularize bad segmentation results due to this ill-posed problem and the presence of outliers in the data [Belghith09]). To overcome this issue, we applied a fuzzy contrast enhancement (FCE) method on each detected object with the HMC algorithm. This method aims at segmenting an unknown sample based on the intensity of its neighbors by considering a fuzzy class transition (*i.e* the edge intensities of two adjacent classes follow a S-membership function). Once the active zones are detected we apply first the proposed active zone alignment algorithm (Fig.6.13). This step requires up to 3mn of computation time with Intel 2.66 GHz and matlab codes. Then the AZ coding and similarity measurement (Fig.6.14) to cluster similar objets (about 1mn of computation time with a matlab codes). However, two objects with two similar shapes can not be affected together if they are not in the same brain area.

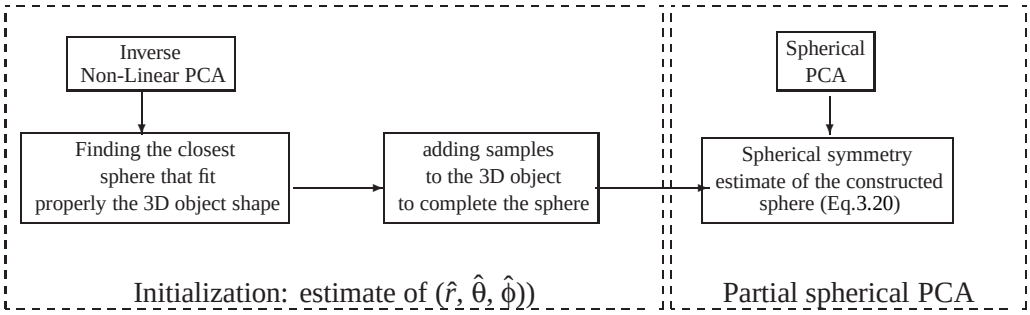


Figure 6.13: Overview diagram of the proposed active zones alignment chain.

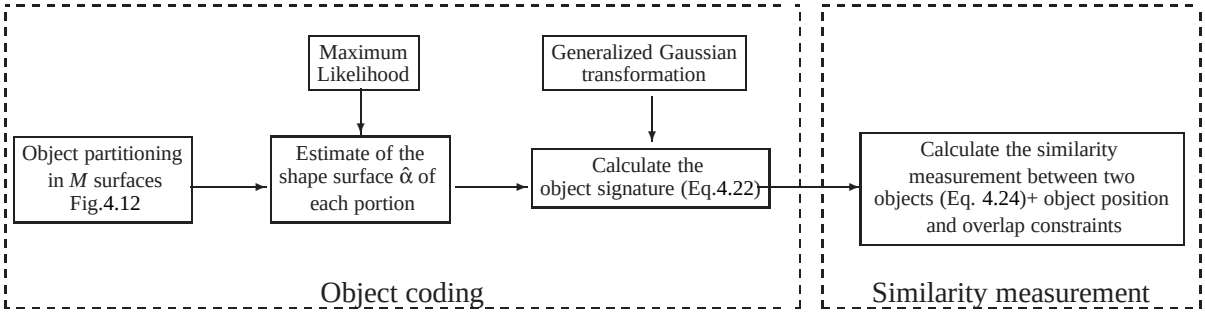


Figure 6.14: Overview diagram of the proposed active zones coding and similarity measurement chain.

For this, we add two new constraints to the objects clustering step: the object position and overlap. Thus, the more their positions are close, the more they can be assigned together.

6.2.3 Real results

Since the fMRI images of subjects with psychologic pathologies are not yet available, we only focus on the two first steps of the indexing scheme: the active zone alignment step and the active zone clustering step (active zone coding and similarity measurement). We have applied the proposed methods on our fMRI database. At the end of the second step, we obtained 72 fMRI object clusters.

Let us start with the object alignment validation. In order to emphasize the benefit of the alignment method and particularly the use of the partial-spherical PCA, we have compared our algorithm with the continuous PCA method [Vranić01a] on some objects belongings to three different clusters (C1, C2 and C3). Fig.6.15.(a) (resp Fig.6.16.(a) and Fig.6.17.(a)) displays the active zone alignment results on three objects belonging to C1 (resp C2 and C3) obtained with the proposed method. Fig.6.15.(b) (resp Fig.6.16.(b) and

Fig.6.17.(b)) displays the active zone alignment results on three objects belonging to C1 (resp C2 and C3) obtained with the continuous PCA method. As one can see, our method performs the best alignment result. Moreover, even with no pronounced partial spherical shape of the active zones (cf. Fig.6.16), the proposed method still works well and provides better results than the continues PCA. This can be explain by the fact that the initialization step properly fit the object shape thanks to the sphere hyperparameters ($\hat{r}$, $\hat{\theta}$, $\hat{\phi}$). For example, the more the object shape seems to a plan, the more $\hat{r}$ is great.

We address now the object clustering validation. Since no ground truth is available, the clustering results were examined by an expert. The preliminary results are promising and we are now working on a comprehensive validation of the results. Fig.6.18 and Fig.6.19 display four active zone clusters obtained by the proposed methods.

6.3 Conclusion

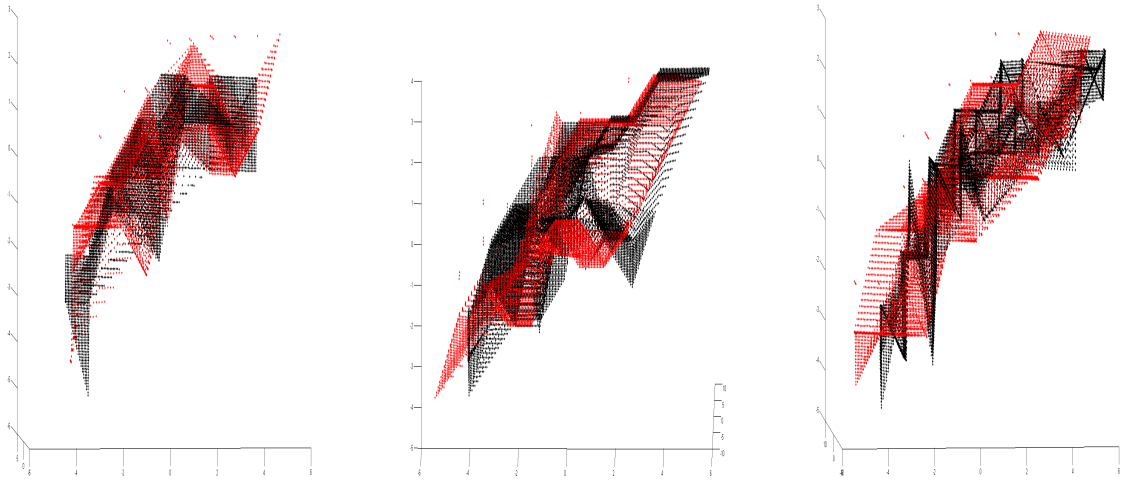
The evaluation of real cases and statistical comparisons with the results obtained by experts were used to validate the behavior of our methods on large scale datasets. We were able to obtain good or satisfactory arrangements for all considered characteristics. This study and critical analysis carried out with the experts allowed us to identify the main limitations of the algorithms on a significant number of objects and then determine a number of changes to solve most encountered problems. Below a summary of physician suggestions:

For the HSQC spectra:

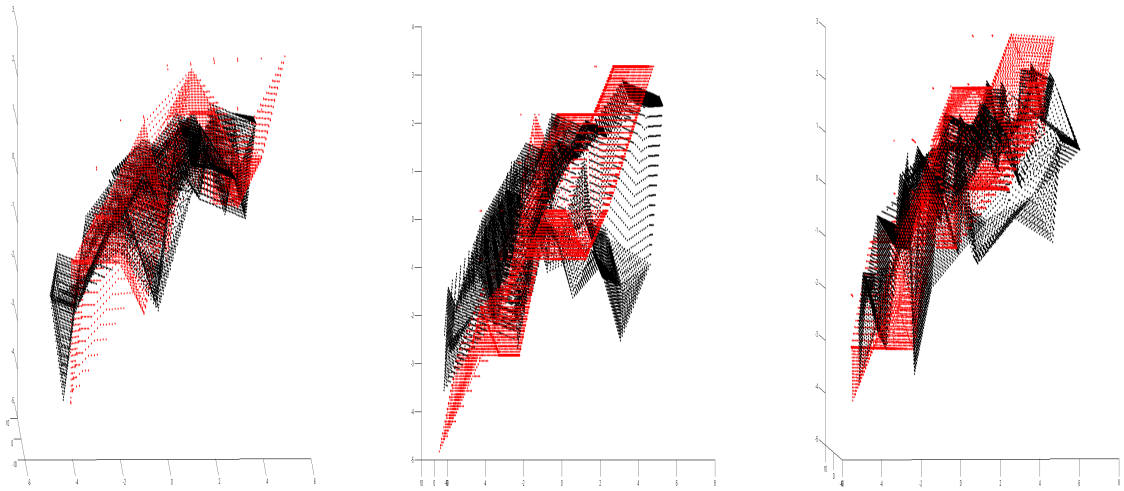
- Consideration of the peak with the largest amplitude as reference to calculate the peak ratios of a given metabolite in the annotation scheme;
- Normalization of the spectra with their biopsy masses;
- Introduction of a metabolite confidence degree for each identified metabolite;
- Introduction of a change threshold to select the relevant changed metabolites.

For the fMRI images:

- Separation between fMRI objects with positive and negative intensities;
- Tightening of the fMRI object location and object overlap constraints in the clustering step;



(*a*)



(*b*)

Figure 6.15: Alignment results of three active zones belonging to the same cluster with the (a) the proposed method and (b) the Continues PCA method.

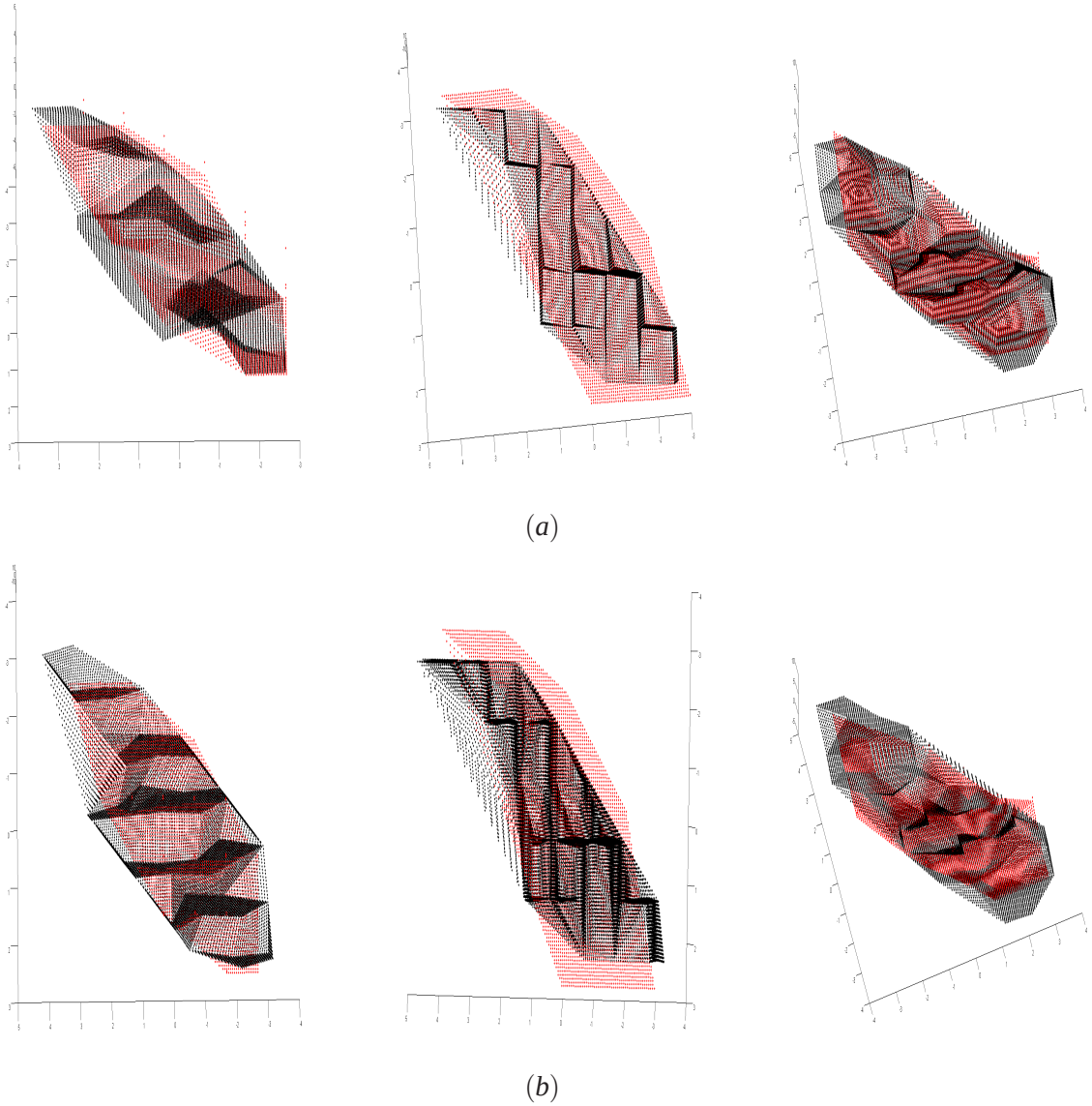
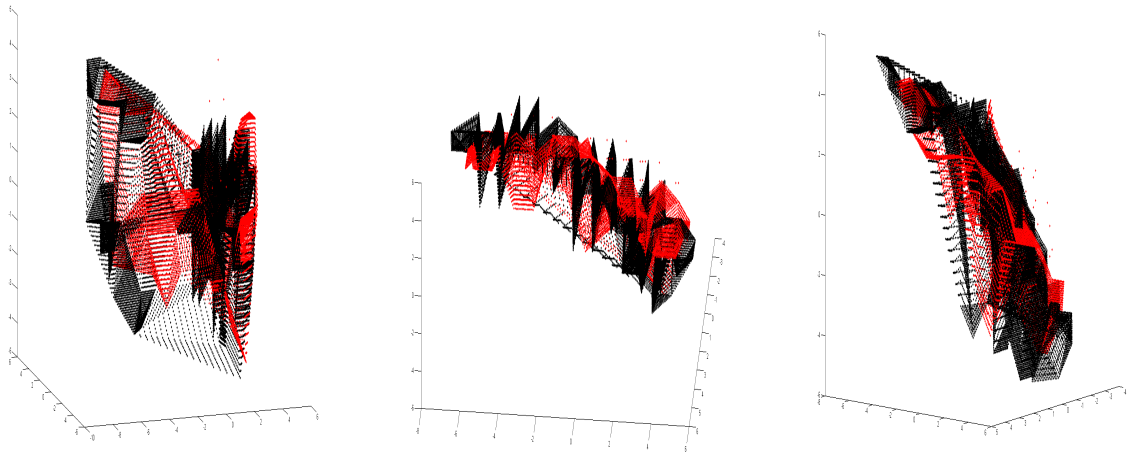
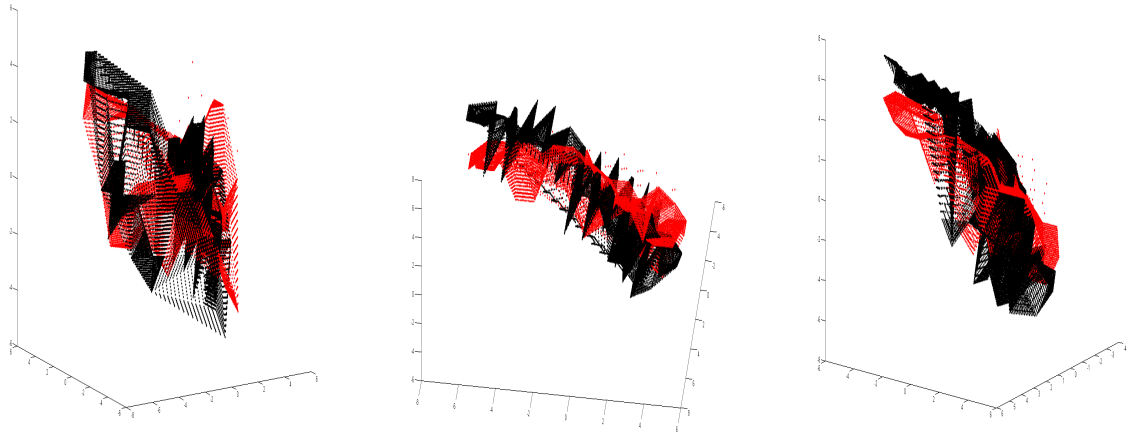


Figure 6.16: Alignment results of three active zones with no pronounced partial spherical shape belonging to the same cluster with the (a) the proposed method and (b) the Continues PCA method. As one can see, even with no pronounced partial spherical shape, the proposed method performs better than the Continues PCA method.



(*a*)



(*b*)

Figure 6.17: Alignment results of three active zones belonging to the same cluster with the (a) the proposed method and (b) the Continues PCA method.

Figure 6.18: Two active zone clusters obtained by the proposed method. A reference brain is displayed to indicate the position of the active zone clusters.

Figure 6.19: Two active zone clusters obtained by the proposed method. A reference brain is displayed to indicate the position of the active zone clusters.

Conclusion and perspectives

This research work aimed at developing an approach for biomarker identification from medical signals and particularly the HSQC HR-MAS 2D NMR spectra and fMRI images. Our approach should be based on characterizations having a clear and proven physician interpretation. We have answered this problem in two steps. On the one hand, the development of a hierarchical model providing a high semantics description level of medical signals, and an access to the relevant information of the medical signal using a new global content-based object indexing and retrieval scheme. On the other hand, we are interested to properly model and integrate the *a priori* knowledge we have on the biological signal allowing us to propose thereafter appropriate methods to each indexing scheme step and each type of treated signals. The performances obtained by the combination of these two aspects were then evaluated on a consistent medical databases. The results were discussed in depth with physicians and we were able to show the relevance and robustness of the proposed methods.

The main contributions of this PhD are the followings:

1. **A global content-based object indexing and retrieval scheme**. We have imagined a strategy to adapt classical indexing scheme to biomarker identification problem, thus providing a global application framework valid for most types of medical signals. To this end, we proposed the add of a classification step to the classical indexing scheme. Indeed, the biomarker identification consists in classifying for example a group of medical signals into the healthy and the pathologic classes (*e.g* cancer or psychological diseases) and to detect then the differences (changes) between them.
   One perspective of this development involves the introduction of an interaction module between the system and the user (a real time feedback) allowing the results control according to the physicians expectations.

2. **An evidential peaks detection and alignment method**. This method combined the modeling of the knowledge by means of the evidence theory and integrates the fuzzy theory to quantify the imprecision degree presented in the spectra. The handling of both imprecision and uncertainty by the evidence theory increased the robustness of the proposed alignment scheme with comparison to the Bayesian method. In addition, we have used the deconvolution model to achieve a better fit of the HSQC spectrum and the multivariate Gaussian distribution to model the noise correlation enabling method robustness to a high level of noise, one of the most delicate issues in HSQC spectra. All these developments resulted in a detection procedure and fully automatic alignment of peaks to a fully parametric representation of the observed spectra.
   In this work, we used a deconvolution method for spectrum peak detection. The optimization of the problem is addressed by a classic MCMC procedure where the Gibbs algorithm was used to model hyperparameters sampling. This allowed us to

take advantage of the sampler computing time rapidity. However, more sophisticated samplers can potentially be used to achieve better results like the Reversible Jump MCMC procedure [Larocque02] allowing a variant peak number handling. Moreover, it would be particularly interesting to replace the evidence theory by the possibility theory. Indeed, possibility theory is one of the current uncertainty theories devoted to the handling of incomplete information. Basically, this theory is similar to the probability theory because it is based on set-functions but differs from the latter by the use of a pair of dual set functions called possibility and necessity measures [Dubois06].

3. **An active zone alignment method**. Based on the reflection symmetry, this method allows us to find the most object natural pose and align visually similar objects in the same manner. Indeed, in order to integrate our *a priori* knowledge and particularly the partial spherical active zone shape, we proposed a new method for spherical symmetry estimation based on the non-linear PCA to model the reflection symmetry of the cortical active zones. To calculate the spherical symmetries of the fMRI active zones, we develop previous works proposed in [Kirby96] where authors adapt the network to the case of circular data (2D data). In a first step, we extend this work to the 3D data case (entire spherical shape) and then to the partial spherical shape which is well suited to fit the active zone shape due the human cortex shape.
   In neuroimaging, it is well known that the brain is made up of three main components: white matter, gray matter and cerebral spinal fluid. Many efficient segmentation algorithms are now available and allow to precisely extract these three components. Based on such segmentation, the active zones detection algorithm may be extended to take into account the differences between these three tissue types inside the brain. Moreover, extending the proposed active zone alignment method by combining the detection and alignment steps in a joint framework would also lead to a fully automatic method exploiting the brain tissue varieties.

4. **A metabolite similarity measurement method**. We have proposed the use of the fuzzy set theory to deal with the ambiguity which is in the heart of such identification task. The use of the metabolite likelihood measure as a metabolite signature has increased the robustness of the proposed identification scheme with comparison to the SVM method which does not take into account the *a priori* knowledge. Genetic Algorithm (GA) was employed to train the comparison model in order to calibrate the parameters in an unsupervised way.
   One perspective of this development involves the use of the residual image (the image difference between the observation and the parametric spectrum representation) to compensate the estimation error within the similarity measurement scheme.

5. **A fMRI active zone coding and similarity measurement method**. Inspired from the Gaussian transformation , the proposed active zone coding method is improved by using of the Generalized Gaussian transformation which is well suited to describe the surface topology of the fMRI active zones. In particular, we showed that the proposed method not only provides a compact representation of the object in its

space but also a signature faithfully attached to its surface topology (flat surface or surface with reliefs). Moreover, in order to be less sensitive to small displacements or minor geometric variations, we introduced a new similarity measure.
It would be interesting to extend the method to other complex surface topology by combining the GG transformation with other transformations (e.g mixture Gaussian and mixture GG transformations).

6. **A object classification/change detection method**. In this method, we proposed a new SV3DH kernel function which combines the characteristics of basic kernel functions with new information about features distribution and then dependency between samples. The dependency between samples was handled based on copulas theory that is be used for the first time to our knowledge in the SVDD framework.
The proposed classification method is limited to two classes. As a matter of fact, new individual/group is classified into two classes: changed and unchanged class. This method can potentially be extended to the multi-class case allowing a multi-diseases discrimination.

Finally, the different proposed methods were validated in a first part on simulated data to demonstrate their behavior compared to existing methods. In a second part, all the results obtained on real data have been examined by experts in each domain (HR-MAS NMR spectroscopy and fMRI). This validation shows the good performance of our algorithms leading to similar results to those obtained by physicians in a short time both for the HSQC spectra (HR-MAS NMR) and fMRI images. The discussions we had with physicians convinced us of the relevance of the proposed approaches, the proximity of the parametric model with the biological model allowing a dialogue and an easy feedback between both communities.

APPENDIX A

MCMC

In this Appendix, we develop a Monte Carlo Markov Chain (MCMC) procedure to estimate the peak locations, amplitudes and shapes required for the proposed peak detection and alignment scheme. The principle of MCMC method is to generate samples drawn from the posterior densities and then to be able to achieve hyperparameter estimation using the Marginal Mean (MPM) estimator [Gilks96]:

$$\hat{X} = \mathrm{E}[X/Y] \tag{A.1}$$

where X is the variable to estimate (theoretical spectrum) from the observation Y (observed spectrum).

We use a Gibbs sampler [Smith93] based on a stationary ergodic Markov chain allowing to draw samples whose distribution asymptotically follows the *a posteriori* density $p(X,\theta/Y)$, $\theta = \{\theta_{\gamma^Y}, \theta_b, \theta_x\}$) where:

- θ_{γ^Y} stands for the pdf hyperparameters of the Lorentzian shape filter γ^Y (Eq.3.2),
- θ_b stands for the hyperparameters of additive noise $B = b(i,j)_{i=1...M, j=1...N}$ (Eq.3.1),
- θ_x represents the hyperparameters of the theoretical 2D spectrum image X (location, shape, amplitude (Eq.3.1)).

Noise Model

MCMC method requires the definition of a noise model that can be based in a Bayesian framework as additive, white and Gaussian. However, the hypothesis of a white gaussian noise is not always entirely justified [Bodenhausen80]. Thus, we propose here to keep the Gaussian behavior but to take into account the correlation of the additive noise. Indeed, vertical lines appear sometimes in the observed spectrum and which are due to experimental condition [Becker00]: it leads us to introduce a correlated noise modeling this kind of artefact. Then, we adopt a multivariate Gaussian distribution with covariance matrix Γ_b and mean μ_b. The expression of the noise density is given by:

$$\mathcal{N}(B;\mu_b,\Gamma_b) = \prod_{j=1}^{N} \frac{1}{(2\pi)^{M/2}|\Gamma_b|^{1/2}} exp\left(-\frac{1}{2} \left(\begin{bmatrix} b(1,j) \\ b(2,j) \\ . \\ . \\ . \\ b(M,j) \end{bmatrix} - \mu_b \right)^T \Gamma_b^{-1} \left(\begin{bmatrix} b(1,j) \\ b(2,j) \\ . \\ . \\ . \\ b(M,j) \end{bmatrix} - \mu_b \right) \right) \tag{A.2}$$

where $(M \times N)$ is the spectrum size. The likelihood corresponding to the data-driven term is expressed as:

$$p(Y/X,\gamma^Y,\theta_b) = \prod_{j=1}^{N} \mathcal{N}\left(y_j - x_{\gamma^Y,j}, \theta_b\right) \tag{A.3}$$

where $y_j = [y(1,j), y(2,j), ..., y(M,j)]^T$, $x_j = \left[x_{\gamma^Y,j}(1), ..., x_{\gamma^Y,j}(i), ..., x_{\gamma^Y,j}(M)\right]^T$, $x_{\gamma^Y,j}(i) = \sum\sum_{k1,k_2} x(k_1,k_2)\, h(i-k_1, j-k_2)$ and $\theta_b = \{\Gamma_b, \mu_b\}$.

Simulation scheme

We present in this paragraph the implementation of the iterative Gibbs algorithm we used for sampling. We introduce the variables θ_x^{prior} and θ_x^{vs} which represent the parameter of the *a priori* probability $p(X/\theta_x^{prior})$ and the likelihood probability $P(Y/X,\theta_x^{vs})$ respectively. To sample $p(X,\gamma,\theta)$, at every iteration l, the main steps consist in:

1. sampling the theoretical 2D spectrum image $X^{[l+1]}$ from

$$p(X/Y,\theta_x^{vs},\theta_x^{prior}) \tag{A.4}$$

with

$$\begin{cases} \theta_x^{vs} = \left\{X^{[l]}, \Gamma_b^{[l]}, \gamma^{Y\,[l]}\right\} \\ \theta_x^{prior} = \left\{\alpha^{[l]}, \beta^{[l]}\right\} \end{cases} \tag{A.5}$$

where $\left\{\alpha^{[l]}, \beta^{[l]}\right\}$ are the hyperparameters of the Gamma distribution modeling the *a priori* on X. The gamma distribution is an exponential family distribution which is used for fitting non-negative data [Hsiao03]. Indeed, the shape parameters ($\left\{\alpha^{[l]}, \beta^{[l]}\right\}$) of the Gamma distribution allow to fit spectral data that may present some sparsity and possibly a background [Dobigeon09]. The gamma density $\mathcal{G}$ is expressed as:

$$\begin{aligned} p(X;\alpha,\beta) &= \prod_{i=1}^{M}\prod_{j=1}^{N} \mathcal{G}\left(x(i,j),\alpha,\beta\right) \quad x(i,j) > 0 \\ &= \prod_{i=1}^{M}\prod_{j=1}^{N} x(i,j)^{(\alpha-1)} \frac{\beta^\alpha}{\Gamma(\alpha)} exp^{(-\beta x(i,j))} \end{aligned} \tag{A.6}$$

2. sampling the hyperparameters of the Lorentzian shape filter $\gamma^{[l+1]}$ from

$$p(\gamma^Y/Y,\theta_{\gamma^Y}^{vs},\theta_{\gamma^Y}^{prior}) \tag{A.7}$$

where

$$\begin{cases} \theta_{\gamma^Y}^{vs} = \left\{X^{[l+1]}, \Gamma_b^{[l]}, \gamma^{Y\,[l]}\right\} \\ \theta_{\gamma^Y}^{prior} = \left\{\sigma^{[l]}\right\} \end{cases} \tag{A.8}$$

where $\{\sigma^{[l]}\}$ is the variance of the gaussian distribution modeling our *a priori* on $\{\gamma^Y\}$

$$p(\gamma^Y;\sigma) = \prod_{c=1,2} \frac{1}{\sqrt{2\pi\sigma^{[l]}}} exp\left(-\left(\frac{(\gamma_c^Y)^2}{2\sigma^{[l]}}\right)\right) \tag{A.9}$$

3. sampling the covariance matrix of the noise $\Gamma_b^{[l+1]}$ from

$$p(\Gamma_b/Y, X^{[l+1]}, \gamma^{Y\,[l+1]}) \tag{A.10}$$

4. sampling the hyperparameter $\alpha^{[l+1]}$ from

$$p(\alpha/Y, X^{[l+1]}, \gamma^{Y\,[l+1]}, \beta^{[l]}) \tag{A.11}$$

5. sampling the hyperparameter $\beta^{[l+1]}$ from

$$p(\beta/Y, X^{[l+1]}, \gamma^{Y\,[l+1]}, \alpha^{[l]}) \tag{A.12}$$

6. sampling the hyperparameter $\sigma^{[l+1]}$from

$$p(\frac{1}{\sigma}/Y, X^{[l+1]}, \gamma^{Y\,[l+1]}) \tag{A.13}$$

After l_{max} iterations, $\hat{X}$ and $\hat{\gamma^Y}$ are given by

$$\begin{cases} \hat{X} = \frac{1}{l_{max}-l_{min}} \sum_{l=l_{min}+1}^{l_{max}} X^{[l]} \\ \hat{\gamma^Y} = \frac{1}{l_{max}-l_{min}} \sum_{l=l_{min}+1}^{l_{max}} \gamma^{Y\,[l]} \end{cases} \tag{A.14}$$

where l_{min} stands for the number of iterations corresponding to the burn-in time of the Markov chain [Cowles96]. In our case l_{min} is equal to 200 iterations whereas l_{max} is equal to 500. Concerning the computation time, the MCMC algorithm requires up to 3h30 to converge with a 2.66 GHz Intel processor and a combination of matlab and C code.

APPENDIX B

Genetic algorithm

Genetic Algorithms (GA) are adaptive heuristic search algorithm premised on the evolutionary ideas of natural selection and genetic [Goldberg89]. The basic concept of GA is designed to simulate processes in natural system necessary for evolution, specifically those that follow the principles first laid down by Charles Darwin of survival of the fittest. As such they represent an intelligent exploitation of a random search within a defined search space to solve a problem. The aim of genetic algorithm is to use simple representations to encode complex structures and simple operations to improve theses structures.

We describe in this part how to use GA to ascertain the hyperparameters of different membership functions. As a matter of fact, we used the GA algorithm for the peak detection-alignment and the metabolite identification schemes. To this end, we use a training data base containing 10 spectra where each one consists of 30 metabolites with known characteristics (the number of metabolite peaks, the locations and the shape od each peaks).

Concerning the fuzzy membership functions hyperparameters used in the peak detection and alignment scheme, let us denote by $(a_1, b_1, c_1, a_2, b_2, c_2, a_3, b_3, c_3)$ the population representing the hyperparameters of these functions where (a_1, b_1, c_1) are the hyperparameters of the function f_{hyp1} (Eq.3.4), (a_2, b_2, c_2) the hyperparameters of f_{hyp2} (Eq. 3.5) and (a_3, b_3, c_3) the hyperparameters of f_{hyp3} (Eq. 3.6). We apply real coding to encode the chromosomes of the population. Each individual is represented by a vector of $\{0, 1\}$. The different steps of GA are:

1. Generating an initial population:

$$(a_1^{[0]}, b_1^{[0]}, c_1^{[0]}, a_2^{[0]}, b_2^{[0]}, c_2^{[0]}, a_3^{[0]}, b_3^{[0]}, c_3^{[0]})$$

   The initialization is done in an experimental way.

2. Define individual fitness function to indicate the fitness of every chromosome. The proposed function is expressed as:

$$f_{opt} = (\hat{\mu}^{[l]}(i, j) - \mu_e(i, j))^2 \tag{B.1}$$

   where $\hat{\mu}^{[l]}(i, j)$ is given by Eq.3.7 with respect to the estimated hyperparameters at iteration l and μ_e is the expected solution.

3. Generating offspring by selection and crossover: 20% of the population which has best fitness is copied directly to next generation to keep the best gene. The other

80% of the population is obtained by crossover operation with a probability $P_c = 0.8$. We select randomly one cut-point and then we exchange the right part of two chromosomes.

4. Mutation operation: only the chromosomes having undergone crossover can be affected by the mutation. It consists in modifying a gene with a probability $Pm = 0.008$.

5. Ending condition: if the maximal evolutionary epoch (maximal number of iterations 500) is reached, the GA end.

For the metabolite identification problem, let us denote by:

$$(a_1, b_1, a_2, b_2, c_2, a_3, b_3, c_3, a_4, b_4, c_4, d_4)$$

the population representing the parameters of the used fuzzy membership functions. Indeed, (a_1, b_1) are the hyperparameters of Eq. 4.16, (a_2, b_2, c_2) the hyperparameters of Eq. 4.12, (a_3, b_3, c_3) the hyperparameters of Eq. 4.15 and (a_4, b_4, c_4, d_4) the hyperparameters of Eq. 4.18. We apply real coding to encode the chromosomes of the population. Therefore, each individual is represented by a vector of $\{0, 1\}$. The different steps of GA are:

1. Generating in an experimental way the initial population:

$$(a_1^{[0]}, b_1^{[0]}, a_2^{[0]}, b_2^{[0]}, c_2^{[0]}, a_3^{[0]}, b_3^{[0]}, c_3^{[0]}, a_4^{[0]}, b_4^{[0]}, c_4^{[0]}, d_4^{[0]})$$

2. Define individual fitness function to indicate the fitness of every chromosome. The proposed function is expressed as:

$$f_{opt} = (g - 1)^2 \tag{B.2}$$

where g is given by Eq. 4.19 and 1 is the expected result ($c.f$ Eq. 4.19).

3. Generating offspring by selection and crossover: 20% of the population which has best fitness is copied directly to next generation to keep the best gene. The other 80% of the population is obtained by crossover operation with a probability $P_c = 0.8$. We select randomly one cut-point and then we exchange the right part of two chromosomes.

4. Mutation operation: only the Chromosomes having undergone crossover can be affected by the mutation. It consists in modifying a gene with a probability $Pm = 0.008$.

5. Ending condition: if the maximal evolutionary epoch (maximal number of iterations 500) is reached, the GA end.

Concerning the computation time, the GA algorithm requires up to 6h45 to converge with a 2.66 GHz Intel processor and a matlab codes.

APPENDIX C

Evidence theory

The Dempster Shafer (DS) is a mathematical theory of evidence. In a finite discrete space, DS theory may be red as a generalization of probability theory. Indeed, the probabilities are assigned to sets as against to mutually exclusive singletons. In probability theory, evidence is only related to one hypothesis [Shafer76]. In DS theory, evidence is related to sets of events. As a matter of fact, the DS theory is designed to cope with varying levels of precision regarding the information. To this end, DS theory provides tools to represent the uncertainty of data where an imprecise may be characterized by a set or an interval and the resulting output is a set or an interval.

The mass function

Let us denote Θ the *frame of discernment*, which is defined as:

$$\Theta = \{H_1, H_2, ..., H_N\}$$

It is composed of N exhaustive and exclusive hypotheses $H_j, j = 1..N$. From the frame of discernment, let Ω be the power set composed with the 2^N propositions A of Θ:

$$\Omega = \{\emptyset, \{H_1\}, \{H_2\}, ..., \{H_N\}, \{H_1, H_2\}, ..., \Theta\}$$

The DS evidence theory provides a representation of both imprecision and uncertainty through the definition of two functions: plausibility *Pls* and belief *Bel*, which are both derived from a mass function *m*. This mass function *m* allows us to quantify the reliability degree of a proposition. *m* is defined for every element A of Ω and observation Y, such that the mass value $m(A;Y)$ belongs to the [0, 1] interval with respect to:

$$m : \begin{cases} m(\emptyset;Y) = 0 \\ \sum\limits_{A \subset \Omega} m(A;Y) = 1 \end{cases}$$

where $\emptyset$ is the empty set. In the Bayesian theory, the uncertainty about an hypothesis is calculated by the probability and imprecision associated with uncertainty is assumed to be null. In the evidence theory, the plausibility value may be explained as the maximum uncertainty value about A whereas the belief value of hypothesis A may be explained as the minimum uncertainty value about A. Therefore, this theory, which allows to represent both imprecision and uncertainty, appears as a more flexible and general approach than the Bayesian one. Indeed, when the mass affected to a compound hypothesis $\{H_1, H_2\}$ is nonzero, it means that we have an option not to make the decision between $\{H_1\}$ or $\{H_2\}$

as the Bayesian theory but rather leave the sample in the $\{H_1, H_2\}$ class. Applications were developed in medical signals [Chaabane09, Yazdani09], object detection [Aeberhard11], image segmentation [Pieczynski07, Ben Chaabane09], and remote sensing classification [Malpica07].

The belief and plausibility functions, derived from m, are respectively defined from Ω to [0, 1]:

$$Bel(A/Y) = \sum_{A \subset \Omega, B \subseteq A} m(B;Y) \tag{C.1}$$

$$Pls(A/Y) = \sum_{A \subset \Omega, B \cap A \neq \emptyset} m(B;Y) \tag{C.2}$$

DS Combination

In the case of problems taking into account both uncertain and imprecise data, it should be useful to combined the information obtained from several sources in order to get more relevant information. DS theory offers tools to combine the knowledge given by different sources. The orthogonal rule also called Dempster's rule of combination is the first combination defined within the framework of evidence theory. Let us denote $m(Y_1)$,...,$m(Y_L)$, L masses of belief coming from L distinct sources Y_l, $l = 1...L$. The belief function m resulting from the combination of the L sources by means of Dempster's combination rule is defined by:

$$m(A) = m(A;Y_1) \oplus m(A;Y_2) \oplus ... \oplus m(A;Y_L) \tag{C.3}$$

where $\oplus$ is defined by:

$$m(A;Y_1) \oplus m(A;Y_2) = \frac{1}{1-K} \sum_{B \bigcap C = A} m(B;Y_1).m(C;Y_2) \tag{C.4}$$

and

$$K = \sum_{B \bigcap C = \emptyset} m_1(B).m_2(C) \tag{C.5}$$

K is often interpreted as a measure of conflict between the different sources and is introduced as a normalization factor. The larger is K, the more the sources are conflicting and the less sense has their combination. The K factor indicates the amount of evidential conflict. If $K = 0$, this shows a complete compatibility and if $0 < K < 1$, it shows partial compatibility. Finally, the orthogonal sum does not exist when $K = 1$. In this case, the sources are totally contradictory, and it is no longer possible to combine them.

Bibliography

[Aeberhard11] M. Aeberhard, S. Paul, N. Kaempchen and T. Bertram. *Object existence probability fusion using dempster-shafer theory in a high-level sensor data fusion architecture*. In Intelligent Vehicles Symposium (IV), 2011 IEEE, pages 770–775. IEEE, 2011. (Cited on page 156.)

[Alum08] M.F. Alum, P.A. Shaw, B.C. Sweatman, B.K. Ubhi, J.N. Haselden and S.C. Connor. *4, 4-Dimethyl-4-silapentane-1-ammonium trifluoroacetate (DSA), a promising universal internal standard for NMR-based metabolic profiling studies of biofluids, including blood plasma and serum*. Metabolomics, vol. 4, no. 2, pages 122–127, 2008. (Cited on page 8.)

[Amberg09] B. Amberg, A. Blake and T. Vetter. *On compositional image alignment, with an application to active appearance models*. In conference of Computer Vision and Pattern recognition, pages 1714–1721, 2009. (Cited on page 32.)

[Ankerst99] M. Ankerst, G. Kastenmuller, H.P. Kriegel and T. Seidl. *3D shape histograms for similarity search and classification in spatial databases*. In Advances in Spatial Databases, pages 207–226. Springer, 1999. (Cited on pages xiv, 65 and 86.)

[Apostolova07] L.G. Apostolova and P.M. Thompson. *Brain mapping as a tool to study neurodegeneration*. Neurotherapeutics, vol. 4, no. 3, pages 387–400, 2007. (Cited on page 8.)

[Aster05] R.C. Aster, B. Borchers, C.H. Thurber and Ebooks Corporation. Parameter estimation and inverse problems. Elsevier Academic Press Amsterdam, The Netherlands, 2005. (Cited on page 18.)

[Baker04] S. Baker and I. Matthews. *Lucas-kanade 20 years on: A unifying framework*. International Journal of Computer Vision, vol. 56, no. 3, pages 221–255, 2004. (Cited on page 32.)

[Bales84] J.R. Bales, D.P. Higham, I. Howe, J.K. Nicholson and P.J. Sadler. *Use of high-resolution proton nuclear magnetic resonance spectroscopy for rapid multi-component analysis of urine*. Clinical chemistry, vol. 30, no. 3, page 426, 1984. (Cited on page 8.)

[Barba07] I. Barba, E. Jaimez-Auguets, A. Rodriguez-Sinovas and D. Garcia-Dorado. *1 H NMR-based metabolomic identification of at-risk areas after myocardial infarction in swine*. Magnetic Resonance Materials in Physics, Biology and Medicine, vol. 20, no. 5, pages 265–271, 2007. (Cited on page 8.)

[Bate98] A. Bate, M. Lindquist, IR Edwards, S. Olsson, R. Orre, A. Lansner and RM De Freitas. *A Bayesian neural network method for adverse drug reaction signal generation*. European Journal of Clinical Pharmacology, vol. 54, no. 4, pages 315–321, 1998. (Cited on page 70.)

[Bayro-Corrochano07] E. Bayro-Corrochano and J. Ortegon-Aguilar. *Lie algebra approach for tracking and 3D motion estimation using monocular vision*. Image and Vision Computing, vol. 25, no. 6, pages 907–921, 2007. (Cited on page 32.)

[Bazaraa06] M.S. Bazaraa, H.D. Sherali and C.M. Shetty. Nonlinear programming: theory and algorithms. John Wiley and Sons, 2006. (Cited on page 97.)

[Beaudet78] P.R. Beaudet. *Rotationally invariant image operators*. In Proceedings of the International Joint Conference on Pattern Recognition, pages 579–583, 1978. (Cited on page 31.)

[Becker00] E.D. Becker. High resolution NMR: theory and chemical applications. Academic Pr, 2000. (Cited on pages 122 and 149.)

[Beckonert03] O. Beckonert, J. Monnerjahn, U. Bonk and D. Leibfritz. *Visualizing metabolic changes in breast-cancer tissue using 1H-NMR spectroscopy and self-organizing maps*. NMR in Biomedicine, vol. 16, no. 1, pages 1–11, 2003. (Cited on page 8.)

[Beckonert07] O. Beckonert, H.C. Keun, T.M.D. Ebbels, J. Bundy, E. Holmes, J.C. Lindon and J.K. Nicholson. *Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts*. Nature protocols, vol. 2, no. 11, pages 2692–2703, 2007. (Cited on page 8.)

[Belghith09] A. Belghith and C. Collet. *Segmentation of respiratory signals by evidence theory*. In Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE, pages 1905–1908. IEEE, 2009. (Cited on pages 94 and 137.)

[Belongie02] S. Belongie, J. Malik and J. Puzicha. *Shape matching and object recognition using shape contexts*. IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 509–522, 2002. (Cited on page 32.)

[Ben Chaabane09] S. Ben Chaabane, F. Fnaiech, M. Sayadi and E. Brassart. *Relevance of the Dempster-Shafer evidence theory for image segmentation*. In Signals, Circuits and Systems (SCS), 2009 3rd International Conference on, pages 1–4. IEEE, 2009. (Cited on page 156.)

[Ben-Hur02] A. Ben-Hur, D. Horn, H.T. Siegelmann and V. Vapnik. *Support vector clustering*. The Journal of Machine Learning Research, vol. 2, pages 125–137, 2002. (Cited on page 94.)

[Berg05] A.C. Berg, T.L. Berg and J. Malik. *Shape matching and object recognition using low distortion correspondences*. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, volume 1, pages 26–33. IEEE, 2005. (Cited on page 31.)

[Berger04] S. Berger and S. Braun. 200 and more nmr experiments: A practical course. Wiley-Vch Weinheim, Germany, 2004. (Cited on page 17.)

[Bertram07] H.C. Bertram, C. Hoppe, B.O. Petersen, J.Ø. Duus, C. Mølgaard and K.F. Michaelsen. *An NMR-based metabonomic investigation on effects of milk and meat protein diets given to 8-year-old boys*. British Journal of Nutrition, vol. 97, no. 04, pages 758–763, 2007. (Cited on page 8.)

[Bezdek99] J.C. Bezdek, J. Keller, R. Krisnapuram and N.R. Pal. Fuzzy models and algorithms for pattern recognition and image processing. Kluwer Academic Publishers, 1999. (Cited on page 35.)

[Binder97] J.R. Binder, J.A. Frost, T.A. Hammeke, R.W. Cox, S.M. Rao and T. Prieto. *Human brain language areas identified by functional magnetic resonance imaging*. The Journal of Neuroscience, vol. 17, no. 1, page 353, 1997. (Cited on page 20.)

[Bishop95] C.M. Bishop. Neural networks for pattern recognition. Oxford university press, 1995. (Cited on pages ix and 50.)

[Bodenhausen80] G. Bodenhausen and PH Bolton. *Elimination of flip angle effects in two-dimensional NMR spectroscopy. Application to cyclic nucleotides.* J. Magn. Reson, vol. 39, page 399, 1980. (Cited on page 149.)

[Boser92] B.E. Boser, I.M. Guyon and V.N. Vapnik. *A training algorithm for optimal margin classifiers*. In Proceedings of the fifth annual workshop on Computational learning theory, pages 144–152. ACM, 1992. (Cited on page 98.)

[Bovey69] F.A. Bovey, L. Jelinski, P.A. Mirau, American Telephone and Telegraph Company. Nuclear magnetic resonance spectroscopy. Academic Press New York, 1969. (Cited on page 12.)

[Brenner93] RE Brenner, PMG Munro, SCR Williams, JD Bell, GJ Barker, CP Hawkins, DN Landon and WI McDonald. *The proton NMR spectrum in acute EAE: the significance of the change in the Cho: Cr ratio*. Magnetic resonance in medicine, vol. 29, no. 6, pages 737–745, 1993. (Cited on page 8.)

[Bricq08] S. Bricq, C. Collet and JP Armspach. *Unifying framework for multimodal brain MRI segmentation based on hidden Markov chains*. Medical image analysis, vol. 12, no. 6, pages 639–652, 2008. (Cited on pages ix, 49 and 137.)

[Brindle02] J.T. Brindle, H. Antti, E. Holmes, G. Tranter, J.K. Nicholson, H.W.L. Bethell, S. Clarke, P.M. Schofield, E. McKilligin, D.E. Mosedale*et al. Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using 1 H-NMR-based metabonomics*. Nature Medicine, vol. 8, no. 12, pages 1439–1445, 2002. (Cited on page x.)

[Bruzzone06] L. Bruzzone, M. Chi and M. Marconcini. *A novel transductive SVM for semisupervised classification of remote-sensing images*. Geoscience and Remote Sensing, IEEE Transactions on, vol. 44, no. 11, pages 3363–3373, 2006. (Cited on pages xii, xv and 70.)

[Bustos04] B. Bustos, D. Keim, D. Saupe, T. Schreck and D. Vranic. *An experimental comparison of feature-based 3D retrieval methods*. 2004. (Cited on page 49.)

[Calabrese08] C. Calabrese, A. Pisi, G. Di Febo, G. Liguori, G. Filippini, M. Cervellera, V. Righi, P. Lucchi, A. Mucci, L. Schenetti*et al. Biochemical alterations from normal mucosa to gastric cancer by ex vivo magnetic resonance spectroscopy*. Cancer Epidemiology Biomarkers & Prevention, vol. 17, no. 6, page 1386, 2008. (Cited on page 8.)

[Camicioli07] R.M. Camicioli, C.C. Hanstock, T.P. Bouchard, M. Gee, N.J. Fisher and WR Martin. *Magnetic resonance spectroscopic evidence for presupplementary motor area neuronal dysfunction in Parkinson's disease*. Movement disorders, vol. 22, no. 3, pages 382–386, 2007. (Cited on page 8.)

[Camps-Valls05] G. Camps-Valls and L. Bruzzone. *Kernel-based methods for hyperspectral image classification*. IEEE Transactions on Geoscience and Remote Sensing, vol. 43, no. 6, pages 1351–1362, 2005. (Cited on pages xiii and 78.)

[Camps-Valls06] G. Camps-Valls, L. Gomez-Chova, J. Muñoz-Mari, L. Alonso, J. Calpe-Maravilla and J. Moreno. *Multitemporal image classification and change detection with kernels*. In SPIE International Symposium Remote Sensing XII, volume 6365, page 63650H, 2006. (Cited on page 104.)

[Camps-Valls08] G. Camps-Valls, L. Gómez-Chova, J. Muñoz-Marí, J.L. Rojo-Alvarez and M. Martínez-Ramón. *Kernel-based framework for multitemporal and multisource remote sensing data classification and change detection*. Geoscience and Remote Sensing, IEEE Transactions on, vol. 46, no. 6, pages 1822–1835, 2008. (Cited on page 104.)

[Cawley03] G.C. Cawley and N.L.C. Talbot. *Efficient leave-one-out cross-validation of kernel Fisher discriminant classifiers*. Pattern Recognition, vol. 36, no. 11, pages 2585–2592, 2003. (Cited on page 105.)

[Chaabane09] S.B. Chaabane, M. Sayadi, F. Fnaiech and E. Brassart. *Dempster-Shafer evidence theory for image segmentation: application in cells images*. International Journal of Signal Processing, vol. 5, no. 1, 2009. (Cited on page 156.)

[Chaouch08] M. Chaouch and A. Verroust-Blondet. *A novel method for alignment of 3D models*. In Shape Modeling and Applications, 2008. SMI 2008. IEEE International Conference on, pages 187–195. IEEE, 2008. (Cited on pages 49 and 50.)

[Chaouch09] M. Chaouch and A. Verroust-Blondet. *3D Gaussian descriptor for 3D shape retrieval*. In Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on, pages 834–837. IEEE, 2009. (Cited on pages x, xiv, 24, 67, 69, 70, 71, 83 and 86.)

[Cohn09] B.R. Cohn, B.N. Joe, S. Zhao, J. Kornak, V.Y. Zhang, R. Iman, J. Kurhanewicz, K. Vahidi, J. Yu, A.B. Caughey*et al.* *Quantitative metabolic profiles of 2nd and 3rd trimester human amniotic fluid using 1 H HR-MAS spectroscopy*. Magnetic Resonance Materials in Physics, Biology and Medicine, vol. 22, no. 6, pages 343–352, 2009. (Cited on page 8.)

[Comaniciu03] D. Comaniciu, V. Ramesh and P. Meer. *Kernel-based object tracking*. IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 564–575, 2003. (Cited on page 32.)

[Corr06] P. Corr. *Imaging of neuro-AIDS*. Journal of psychosomatic research, vol. 61, no. 3, pages 295–299, 2006. (Cited on page 8.)

[Cortes95] C. Cortes and V. Vapnik. *Support-vector networks*. Machine learning, vol. 20, no. 3, pages 273–297, 1995. (Cited on page 97.)

[Cossette08] H. Cossette, E. Marceau and F. Marri. *On the compound Poisson risk model with dependence based on a generalized Farlie-Gumbel-Morgenstern copula*. Insurance: Mathematics and Economics, vol. 43, no. 3, pages 444–455, 2008. (Cited on page 108.)

[Cowles96] M.K. Cowles and B.P. Carlin. *Markov Chain Monte Carlo convergence diagnostics: a comparative review.* Journal of the American Statistical Association, vol. 91, no. 434, 1996. (Cited on page 151.)

[Dame09] A. Dame and E. Marchand. *Entropy-based visual servoing*. In Robotics and Automation, ICRA'09, IEEE International Conference on, pages 707–713. IEEE, 2009. (Cited on page 32.)

[Davies95] S.E.C. Davies, J. Newcombe, S.R. Williams, W.I. McDonald and J.B. Clark. *High resolution proton NMR spectroscopy of multiple sclerosis lesions*. Journal of neurochemistry, vol. 64, no. 2, pages 742–748, 1995. (Cited on page 8.)

[De Luca06] M. De Luca, CF Beckmann, N. De Stefano, PM Matthews and SM Smith. *fMRI resting state networks define distinct modes of long-distance interactions in the human brain*. Neuroimage, vol. 29, no. 4, pages 1359–1367, 2006. (Cited on page 22.)

[Demarta05] S. Demarta and A.J. McNeil. *The t copula and related copulas*. International Statistical Review, vol. 73, no. 1, pages 111–129, 2005. (Cited on page 108.)

[Derrode03] S. Derrode, G. Mercier and W. Pieczynski. *Unsupervised Change Detection in SAR Images Using Multicomponent HMC Models*. In in MultiTemp, pages 16–18, 2003. (Cited on page xi.)

[Detour11] J. Detour, K. Elbayed, M. Piotto, FM Moussallieh, A. Nehlig and IJ Namer. *Ultra fast in vivo microwave irradiation for enhanced metabolic stability of brain biopsy samples during HRMAS NMR analysis*. Journal of neuroscience methods, 2011. (Cited on page 123.)

[Dobigeon09] N. Dobigeon, S. Moussaoui, J.Y. Tourneret and C. Carteret. *Bayesian separation of spectral sources under non-negativity and full additivity constraints*. Signal Processing, vol. 89, no. 12, pages 2657–2669, 2009. (Cited on pages 73 and 150.)

[Dobrosotskaya08] J.A. Dobrosotskaya and A.L. Bertozzi. *A wavelet-laplace variational technique for image deconvolution and inpainting*. IEEE Transactions on Image Processing, vol. 17, no. 5, page 657, 2008. (Cited on page 36.)

[Downar01] J. Downar, A.P. Crawley, D.J. Mikulis and K.D. Davis. *The effect of task relevance on the cortical response to changes in visual and auditory stimuli: an event-related fMRI study*. Neuroimage, vol. 14, no. 6, pages 1256–1267, 2001. (Cited on page 20.)

[Dowson08] N. Dowson and R. Bowden. *Mutual information for lucas-kanade tracking (milk): An inverse compositional formulation*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 30, no. 1, pages 180–185, 2008. (Cited on page 32.)

[Dubois06] D. Dubois. *Possibility theory and statistical reasoning*. Computational statistics and data analysis, vol. 51, no. 1, pages 47–69, 2006. (Cited on page 146.)

[Duda01] R.O. Duda, P.E. Hart, D.G. Stork*et al.* Pattern classification, volume 2. wiley New York, 2001. (Cited on page 105.)

[Dunne05] V.G. Dunne, S. Bhattachayya, M. Besser, C. Rae and J.L. Griffin. *Metabolites from cerebrospinal fluid in aneurysmal subarachnoid haemorrhage correlate with vasospasm and clinical outcome: a pattern-recognition 1H NMR study*. NMR in Biomedicine, vol. 18, no. 1, pages 24–33, 2005. (Cited on page 8.)

[Eakins96] J.P. Eakins. *Automatic image content retrieval-are we getting anywhere?* In ELVIRA-PROCEEDINGS-, pages 121–134. Citeseer, 1996. (Cited on pages iv and 23.)

[Ekstrand77] KE Ekstrand, RL Dixon, M. Raben and CR Ferree. *Proton NMR relaxation times in the peripheral blood or cancer patients*. Physics in Medicine and Biology, vol. 22, page 925, 1977. (Cited on page 8.)

[Engel97] S.A. Engel, G.H. Glover and B.A. Wandell. *Retinotopic organization in human visual cortex and the spatial precision of functional MRI.* Cerebral cortex, vol. 7, no. 2, page 181, 1997. (Cited on pages 2 and 18.)

[Fahlman88] S E Fahlman. *An empirical study of learning speed in back-propagation networks*. Training, vol. 6, no. 4976, pages 1–19, 1988. (Cited on page 52.)

[Feliz06] M. Feliz, J. García, E. Aragón and M. Pons. *Fast 2D NMR ligand screening using Hadamard spectroscopy*. Journal of the American Chemical Society, vol. 128, no. 22, pages 7146–7147, 2006. (Cited on page 36.)

[Ferrandez03] AM Ferrandez, L. Hugueville, S. Lehericy, JB Poline, C. Marsault and V. Pouthas. *Basal ganglia and supplementary motor area subtend duration perception: an fMRI study*. Neuroimage, vol. 19, no. 4, pages 1532–1544, 2003. (Cited on page 20.)

[Fiehn02] O. Fiehn. *Metabolomics–the link between genotypes and phenotypes*. Plant Molecular Biology, vol. 48, no. 1, pages 155–171, 2002. (Cited on page 8.)

[Fjortoft03] R. Fjortoft, Y. Delignon, W. Pieczynski, M. Sigelle and F. Tupin. *Unsupervised classification of radar images using hidden Markov chains and hidden Markov random fields*. Geoscience and Remote Sensing, IEEE Transactions on, vol. 41, no. 3, pages 675–686, 2003. (Cited on page 30.)

[Fox86] P.T. Fox and M.E. Raichle. *Focal physiological uncoupling of cerebral blood flow and oxidative metabolism during somatosensory stimulation in human subjects*. Proceedings of the National Academy of Sciences of the United States of America, vol. 83, no. 4, page 1140, 1986. (Cited on page 20.)

[Foxall95] PJ Foxall, S. Bewley, GH Neild, CH Rodeck and JK Nicholson. *Analysis of fetal and neonatal urine using proton nuclear magnetic resonance spectroscopy.* Archives of Disease in Childhood. Fetal and Neonatal Edition, vol. 73, no. 3, page F153, 1995. (Cited on page 8.)

[Friebolin91] H. Friebolin and J.K. Becconsall. Basic one-and two-dimensional nmr spectroscopy. VCH Weinheim, 1991. (Cited on pages 9, 10 and 14.)

[Fukushima81] E. Fukushima and S.B.W. Roeder. Experimental pulse nmr: a nuts and bolts approach. Addison-Wesley Pub. Co., Advanced Book Program, 1981. (Cited on page 10.)

[Fumera00] G. Fumera, F. Roli and G. Giacinto. *Reject option with multiple thresholds*. Pattern Recognition, vol. 33, no. 12, pages 2099–2101, 2000. (Cited on pages xi and 94.)

[Furey00] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer and D. Haussler. *Support vector machine classification and validation of cancer tissue samples using microarray expression data*. Bioinformatics, vol. 16, no. 10, page 906, 2000. (Cited on page xii.)

[Ganis04] G. Ganis, W.L. Thompson and S.M. Kosslyn. *Brain areas underlying visual mental imagery and visual perception: an fMRI study*. Cognitive Brain Research, vol. 20, no. 2, pages 226–241, 2004. (Cited on page 20.)

[Garrod01] S. Garrod, E. Humpher, SC Connor, JC Connelly, M. Spraul, JK Nicholson and E. Holmes. *High-resolution 1H NMR and magic angle spinning NMR spectroscopic investigation of the biochemical effects of 2-bromoethanamine in intact renal and hepatic tissue*. Magnetic resonance in medicine, vol. 45, no. 5, pages 781–790, 2001. (Cited on page 8.)

[Ge09] Z. Ge, C. Yang and Z. Song. *Improved kernel PCA-based monitoring approach for nonlinear processes*. Chemical Engineering Science, vol. 64, no. 9, pages 2245–2255, 2009. (Cited on page 50.)

[Genest08] C. Genest and B. Rémillard. *Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models*. In Annales de l'Institut Henri Poincaré, Probabilités et Statistiques, volume 44, pages 1096–1127. Institut Henri Poincaré, 2008. (Cited on page 108.)

[Gibson97] S.F. Gibson. *3D chainmail: a fast algorithm for deforming volumetric objects*. In Proceedings of the 1997 symposium on Interactive 3D graphics, pages 149–ff. ACM, 1997. (Cited on page 33.)

[Gilks96] W.R. Gilks and DJ Spiegelhalter. Markov chain Monte Carlo in practice. Chapman & Hall/CRC, 1996. (Cited on pages 35 and 149.)

[Goldberg89] D.E. Goldberg. Genetic Algorithms in Search and Optimization. Addison-wesley, 1989. (Cited on page 153.)

[Graca09] G. Graca, I.F. Duarte, A.S. Barros, B.J. Goodfellow, S. Diaz, I.M. Carreira, A.B. Couceiro, E. Galhano and A.M. Gil. *1H NMR Based Metabonomics of Human Amniotic Fluid for the Metabolic Characterization of Fetus Malformations*. Journal of Proteome Research, vol. 8, no. 8, pages 4144–4150, 2009. (Cited on page 8.)

[Griffin00] JL Griffin, LA Walker, S. Garrod, E. Holmes, RF Shore and JK Nicholson. *NMR spectroscopy based metabonomic studies on the comparative biochemistry of the kidney and urine of the bank vole (Clethrionomys glareolus), wood mouse (Apodemus sylvaticus), white toothed shrew (Crocidura*

suaveolens) and the laboratory rat. Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology, vol. 127, no. 3, pages 357–367, 2000. (Cited on page 8.)

[Griffin04] J.L. Griffin and J.P. Shockcor. *Metabolic profiles of cancer cells*. Nature reviews cancer, vol. 4, no. 7, pages 551–561, 2004. (Cited on pages viii and 16.)

[Grootveld05] M. Grootveld and C.J.L. Silwood. *1H NMR analysis as a diagnostic probe for human saliva*. Biochemical and biophysical research communications, vol. 329, no. 1, pages 1–5, 2005. (Cited on page 8.)

[Grootveld06] M. Grootveld, D. Algeo, C.J.L. Silwood, J.C. Blackburn and A.D. Clark. *Determination of the illicit drug gamma-hydroxybutyrate (GHB) in human saliva and beverages by 1 H NMR analysis*. Biofactors, vol. 27, no. 1, pages 121–136, 2006. (Cited on page 8.)

[Hager98] G.D. Hager and P.N. Belhumeur. *Efficient region tracking with parametric models of geometry and illumination*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 20, no. 10, pages 1025–1039, 1998. (Cited on page 32.)

[Hager04] Gregory Hager, , Gregory D. Hager, Maneesh Dewan and Charles V. Stewart. *Multiple Kernel Tracking with SSD*. In In CVPR'04, pages 790–797, 2004. (Cited on page 32.)

[Halmos82] P.R. Halmos. A hilbert space problem book, volume 19. Springer, 1982. (Cited on page 98.)

[Hammen07] T. Hammen, M. Schwarz, M. Doelken, F. Kerling, T. Engelhorn, A. Stadlbauer, O. Ganslandt, C. Nimsky, A. Doerfler and H. Stefan. *1H-MR Spectroscopy Indicates Severity Markers in Temporal Lobe Epilepsy: Correlations between Metabolic Alterations, Seizures, and Epileptic Discharges in EEG*. Epilepsia, vol. 48, no. 2, pages 263–269, 2007. (Cited on page 8.)

[Harris88] C. Harris and M. Stephens. *A combined corner and edge detector*. In Alvey vision conference, volume 15, page 50. Manchester, UK, 1988. (Cited on page 31.)

[Hofmann08] T. Hofmann, B. Scholkopf and A.J. Smola. *Kernel methods in machine learning*. The annals of statistics, vol. 36, no. 3, pages 1171–1220, 2008. (Cited on pages 98 and 104.)

[Holmes00] HC Holmes, GJA I. Snodgrass and RA Iles. *Changes in the choline content of human breast milk in the first 3 weeks after birth*. European journal of pediatrics, vol. 159, no. 3, pages 198–204, 2000. (Cited on page 8.)

[Holzgrabe99] U. Holzgrabe, I. Wawer and B. Diehl. NMR spectroscopy in drug development and analysis. Wiley-VCH Verlag GmbH, 1999. (Cited on page 8.)

[Howe93] FA Howe, RJ Maxwell, DE Saunders, MM Brown and JR Griffiths. *Proton spectroscopy in vivo*. Magnetic resonance quarterly, vol. 9, no. 1, page 31, 1993. (Cited on page 8.)

[Howe03] F.A. Howe and K.S. Opstad. *1H MR spectroscopy of brain tumours and masses*. NMR in Biomedicine, vol. 16, no. 3, pages 123–131, 2003. (Cited on page 8.)

[Hsiao03] T. Hsiao, A. Rangarajan and G. Gindi. *Bayesian image reconstruction for transmission tomography using deterministic annealing*. Journal of Electronic Imaging, vol. 12, page 7, 2003. (Cited on page 150.)

[Hsieh98] W.W. Hsieh and B. Tang. *Applying neural network models to prediction and data analysis in meteorology and oceanography*. BULLETIN-AMERICAN METEOROLOGICAL SOCIETY, vol. 79, pages 1855–1870, 1998. (Cited on pages ix and 50.)

[ICID10] ICID. Institut Canadien d'Information sur la Santé, Tendances des dépenses nationales de santé, 1975 à 2010. Master's thesis, Décembre, 2010. (Cited on page 7.)

[Joe97] H. Joe. Multivariate models and dependence concepts. Chapman & Hall/CRC, 1997. (Cited on pages 75 and 104.)

[Joe08] B.N. Joe, K. Vahidi, A. Zektzer, M.H. Chen, M.S. Clifton, T. Butler, K. Keshari, J. Kurhanewicz, F. Coakley and M.G. Swanson. *1H HR-MAS spectroscopy for quantitative measurement of choline concentration in amniotic fluid as a marker of fetal lung maturity: Inter-and intraobserver reproducibility study*. Journal of Magnetic Resonance Imaging, vol. 28, no. 6, pages 1540–1545, 2008. (Cited on page 8.)

[Jolliffe86] I.T. Jolliffe. Principal component analysis. Springer Verlag, 1986. (Cited on page 33.)

[Jukarainen08] N.M. Jukarainen, S.P. Korhonen, M.P. Laakso, M.A. Korolainen, M. Niemitz, P.P. Soininen, K. Tuppurainen, J. Vepsalainen, T. Pirttila and R. Laatikainen. *Quantification of 1 H NMR spectra of human cerebrospinal fluid: a protocol based on constrained total-line-shape analysis*. Metabolomics, vol. 4, no. 2, pages 150–160, 2008. (Cited on page 8.)

[Jurie02] F. Jurie and M. Dhome. *Hyperplane approximation for template matching*. IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 996–1000, 2002. (Cited on page 32.)

[Kang01] K. Kang, J.P. Tarel, R. Fishman and D. Cooper. *A linear dual-space approach to 3D surface reconstruction from occluding contours using algebraic surfaces*. In Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on, volume 1, pages 198–204. IEEE, 2001. (Cited on page 33.)

[Kaplan99] A.M. Kaplan, D.J. Bandy, K.H. Manwaring, K. Chen, M.A. Lawson, S.D. Moss, J.D. Duncan, D.L. Wodrich, J.A. Schnur and E.M. Reiman. *Functional brain mapping using positron emission tomography scanning in preoperative neurosurgical planning for pediatric brain tumors.* Journal of neurosurgery, vol. 91, no. 5, pages 797–803, 1999. (Cited on page 18.)

[Kass88] M. Kass, A. Witkin and D. Terzopoulos. *Snakes: Active contour models*. International journal of computer vision, vol. 1, no. 4, pages 321–331, 1988. (Cited on pages 30 and 31.)

[Keller04] Y. Keller and A. Averbuch. *Fast motion estimation using bidirectional gradient methods*. Image Processing, IEEE Transactions on, vol. 13, no. 8, pages 1042–1054, 2004. (Cited on page 32.)

[Keun02] H.C. Keun, O. Beckonert, J.L. Griffin, C. Richter, D. Moskau, J.C. Lindon and J.K. Nicholson. *Cryogenic probe 13C NMR spectroscopy of urine for metabonomic studies*. Analytical chemistry, vol. 74, no. 17, pages 4588–4593, 2002. (Cited on page 8.)

[Kirby96] M.J. Kirby and R. Miranda. *Circular nodes in neural networks*. Neural Computation, vol. 8, no. 2, pages 390–402, 1996. (Cited on pages 52 and 146.)

[Kos01] G. Kos. *An algorithm to triangulate surfaces in 3D using unorganised point clouds*. Computing. Supplementum, vol. 14, pages 219–232, 2001. (Cited on page 33.)

[Kramer91] M.A. Kramer. *Nonlinear principal component analysis using autoassociative neural networks*. AIChE journal, vol. 37, no. 2, pages 233–243, 1991. (Cited on page 50.)

[Kruskal74] J.B. Kruskal and R.N. Shepard. *A nonmetric variety of linear factor analysis*. Psychometrika, vol. 39, no. 2, pages 123–157, 1974. (Cited on page 50.)

[Kwock06] L. Kwock, J.K. Smith, M. Castillo, M.G. Ewend, F. Collichio, D.E. Morris, T.W. Bouldin and S. Cush. *Clinical role of proton magnetic resonance spectroscopy in oncology: brain, breast, and prostate cancer*. The lancet oncology, vol. 7, no. 10, pages 859–868, 2006. (Cited on page 8.)

[La Cascia00] M. La Cascia, S. Sclaroff and V. Athitsos. *Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3D models*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 22, no. 4, pages 322–336, 2000. (Cited on page 32.)

[Larocque02] J.R. Larocque and J.P. Reilly. *Reversible jump MCMC for joint detection and estimation of sources in colored noise*. Signal Processing, IEEE Transactions on, vol. 50, no. 2, pages 231–240, 2002. (Cited on page 146.)

[Lawrence05] N. Lawrence. *Probabilistic non-linear principal component analysis with Gaussian process latent variable models*. The Journal of Machine Learning Research, vol. 6, pages 1783–1816, 2005. (Cited on page 50.)

[Le Bihan91] D. Le Bihan. *Molecular diffusion nuclear magnetic resonance imaging*. Magnetic resonance quarterly, vol. 7, no. 1, page 1, 1991. (Cited on page 18.)

[Lee95] A.T. Lee, G.H. Glover and C.H. Meyer. *Discrimination of Large Venous Vessels in Time-Course Spiral Blood-Oxygen-Level-Dependent Magnetic-Resonance Functional Neuroimaging*. Magnetic Resonance in Medicine, vol. 33, no. 6, pages 745–754, 1995. (Cited on page 20.)

[Leymarie93] F. Leymarie and M.D. Levine. *Tracking deformable objects in the plane using an active contour model*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 15, no. 6, pages 617–634, 1993. (Cited on page 31.)

[Li06] H. Li, Z. Lu and Z. Yue. *Support vector machine for structural reliability analysis*. Applied Mathematics and Mechanics, vol. 27, no. 10, pages 1295–1303, 2006. (Cited on page 94.)

[Liang06] M. Liang, Y. Zhou, T. Jiang, Z. Liu, L. Tian, H. Liu and Y. Hao. *Widespread functional disconnectivity in schizophrenia with resting-state functional magnetic resonance imaging*. Neuroreport, vol. 17, no. 2, page 209, 2006. (Cited on page 22.)

[Liboff98] R.L. Liboff. Introductory quantum mechanics. Addison-Wesley, 1998. (Cited on page 9.)

[Lindeberg98] T. Lindeberg. *Feature detection with automatic scale selection*. International Journal of Computer Vision, vol. 30, no. 2, pages 79–116, 1998. (Cited on page 31.)

[Lingoes67] J.C. Lingoes and L. Guttman. *Nonmetric factor analysis: a rank reducing alternative to linear factor analysis*. Multivariate Behavioral Research, 1967. (Cited on page 50.)

[Lof97] K. Lof, J. Hovinen, P. Reinikainen, LM Vilpo, E. Seppala and JA Vilpoa. *Kinetics of chlorambucil in vitro: effects of fluid matrix, human gastric juice, plasma proteins and red cells*. Chemico-biological interactions, vol. 103, no. 3, pages 187–198, 1997. (Cited on page 8.)

[Lowry08] D.F. Lowry, A. Stancik, R.M. Shrestha and G.W. Daughdrill. *Modeling the accessible conformations of the intrinsically unstructured transactivation domain of p53*. Proteins: Structure, Function, and Bioinformatics, vol. 71, no. 2, pages 587–598, 2008. (Cited on page 36.)

[lsiml] *Retrieved on December 2011 From the World Wide Web*, http://change.gsfc.nasa.gov/data_index.html. (Cited on pages xv, 106 and 107.)

[Lucas81] B.D. Lucas, T. Kanade*et al.* *An iterative image registration technique with an application to stereo vision*. In International joint conference on artificial intelligence, volume 3, pages 674–679. Citeseer, 1981. (Cited on page 32.)

[Lynch94] MJ Lynch, J. Masters, JP Pryor, JC Lindon, M. Spraul, PJD Foxall and JK Nicholson. *Ultra high field NMR spectroscopic studies on human seminal fluid, seminal vesicle and prostatic secretions*. Journal of pharmaceutical and biomedical analysis, vol. 12, no. 1, pages 5–19, 1994. (Cited on page 8.)

[MacQueen67] J. MacQueen*et al.* *Some methods for classification and analysis of multivariate observations*. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, volume 1, page 14. California, USA, 1967. (Cited on page 30.)

[Mahon04] M.M. Mahon, A.D. Williams, W.P. Soutter, I.J. Cox, G.A. McIndoe, G.A. Coutts, R. Dina and N.M. desouza. *1H magnetic resonance spectroscopy of invasive cervical cancer: an in vivo study with ex vivo corroboration*. NMR in Biomedicine, vol. 17, no. 1, pages 1–9, 2004. (Cited on page 8.)

[Malpica07] JA Malpica, MC Alonso and MA Sanz. *Dempster-shafer theory in geographic information systems: A survey*. Expert Systems with Applications, vol. 32, no. 1, pages 47–55, 2007. (Cited on page 156.)

[Martin-Pastor00] M. Martin-Pastor and C.A. Bush. *Conformational Studies of Human Milk Oligosaccharides Using 1H-13C One-Bond NMR Residual Dipolar Couplings*. Biochemistry, vol. 39, no. 16, pages 4674–4683, 2000. (Cited on page 8.)

[McInerney96] T. McInerney and D. Terzopoulos. *Deformable models in medical image analysis: a survey*. Medical image analysis, vol. 1, no. 2, pages 91–108, 1996. (Cited on page 31.)

[Melendez01] H.V. Melendez, D. Ahmadi, H.G. Parkes, M. Rela, G. Murphy and N. Heaton. *Proton nuclear magnetic resonance analysis of hepatic bile from donors and recipients in human liver transplantation*. Transplantation, vol. 72, no. 5, page 855, 2001. (Cited on page 8.)

[Minh06] H. Minh, P. Niyogi and Y. Yao. *Mercer's theorem, feature maps, and smoothing*. Learning theory, pages 154–168, 2006. (Cited on page 98.)

[Mintorovitch91] J. Mintorovitch, ME Moseley, L. Chileuitt, H. Shimizu, Y. Cohen and PR Weinstein. *Comparison of diffusion-and T2-weighted MRI for the early detection of cerebral ischemia and reperfusion in rats*. Magnetic resonance in medicine, vol. 18, no. 1, pages 39–50, 1991. (Cited on page 20.)

[Mitra06] N.J. Mitra, L.J. Guibas and M. Pauly. *Partial and approximate symmetry detection for 3D geometry*. ACM Transactions on Graphics (TOG), vol. 25, no. 3, pages 560–568, 2006. (Cited on page 49.)

[Mountford82] C.E. Mountford, G. Grossman, G. Reid and R.M. Fox. *Characterization of transformed cells and tumors by proton nuclear magnetic resonance spectroscopy*. Cancer Research, vol. 42, no. 6, page 2270, 1982. (Cited on page 8.)

[Nakata08] H. Nakata, K. Sakamoto, A. Ferretti, M. Gianni Perrucci, C. Del Gratta, R. Kakigi and G. Luca Romani. *Somato-motor inhibitory processing in humans: an event-related functional MRI study*. Neuroimage, vol. 39, no. 4, pages 1858–1866, 2008. (Cited on page 20.)

[Nicholson95] J.K. Nicholson, P.J.D. Foxall, M. Spraul, R.D. Farrant and J.C. Lindon. *750 MHz 1H and 1H-13C NMR spectroscopy of human blood plasma*. Analytical chemistry, vol. 67, no. 5, pages 793–811, 1995. (Cited on page 8.)

[Nocedal99] J. Nocedal and S.J. Wright. Numerical optimization. Springer verlag, 1999. (Cited on page 52.)

[Odunsi05] K. Odunsi, R.M. Wollman, C.B. Ambrosone, A. Hutson, S.E. McCann, J. Tammela, J.P. Geisler, G. Miller, T. Sellers, W. Cliby*et al. Detection of epithelial ovarian cancer using 1H-NMR-based metabonomics*. International journal of cancer, vol. 113, no. 5, pages 782–788, 2005. (Cited on page 8.)

[Ohbuchi03] R. Ohbuchi, T. Minamitani and T. Takei. *Shape-similarity search of 3D models by using enhanced shape functions*. In Theory and Practice of Computer Graphics, 2003. Proceedings, pages 97–104. IEEE, 2003. (Cited on page 65.)

[Opstad08] K.S. Opstad, B.A. Bell, J.R. Griffiths and F.A. Howe. *Toward accurate quantification of metabolites, lipids, and macromolecules in HRMAS spectra of human brain tumor biopsies using LCModel*. Magnetic Resonance in Medicine, vol. 60, no. 5, pages 1237–1242, 2008. (Cited on page 8.)

[Osada02] R. Osada, T. Funkhouser, B. Chazelle and D. Dobkin. *Shape distributions*. ACM Transactions on Graphics (TOG), vol. 21, no. 4, pages 807–832, 2002. (Cited on page 64.)

[Paczkowska03] A. Paczkowska, B. Toczylowska, P. Nyckowski, W. Patkowski, A. Kanski, M. Krawczyk and U. Oldakowska-Jedynak. *High-resolution 1H nuclear magnetic resonance spectroscopy analysis of bile samples obtained from a patient after orthotopic liver transplantation: new perspectives*. In Transplantation proceedings, volume 35, pages 2278–2280. Elsevier, 2003. (Cited on page 8.)

[Pappas92] T.N. Pappas. *An adaptive clustering algorithm for image segmentation*. Signal Processing, IEEE Transactions on, vol. 40, no. 4, pages 901–914, 1992. (Cited on page 30.)

[Paquet99] E. Paquet and M. Rioux. *The MPEG-7 standard and the content-based management of three-dimensional data: A case study*. In Multimedia Computing and Systems, 1999. IEEE International Conference on, volume 1, pages 375–380. IEEE, 1999. (Cited on page 65.)

[Paquet00] E. Paquet, M. Rioux, A. Murching and T. Naveen. *Description of shape information for 2-D and 3-D objects*. Signal Processing: Image Communication, vol. 16, no. 1-2, pages 103–122, 2000. (Cited on page 33.)

[Pieczynski07] W. Pieczynski. *Multisensor triplet Markov chains and theory of evidence*. International Journal of Approximate Reasoning, vol. 45, no. 1, pages 1–16, 2007. (Cited on page 156.)

[Piotto09] M. Piotto, F.M. Moussallieh, B. Dillmann, A. Imperiale, A. Neuville, C. Brigand, J.P. Bellocq, K. Elbayed and IJ Namer. *Metabolic characterization of primary human colorectal cancers using high resolution magic angle spinning 1 H magnetic resonance spectroscopy*. Metabolomics, vol. 5, no. 3, pages 292–301, 2009. (Cited on page 17.)

[Podolak06] J. Podolak, P. Shilane, A. Golovinskiy, S. Rusinkiewicz and T. Funkhouser. *A planar-reflective symmetry transform for 3D shapes*. In ACM Transactions on Graphics (TOG), volume 25, pages 549–559. ACM, 2006. (Cited on pages 34 and 49.)

[Ratsch01] G. Ratsch, T. Onoda and K.R. Muller. *Soft margins for AdaBoost*. Machine Learning, vol. 42, no. 3, pages 287–320, 2001. (Cited on page 105.)

[Reddy02] H. Reddy, S. Narayanan, M. Woolrich, T. Mitsumori, Y. Lapierre, DL Arnold and PM Matthews. *Functional brain reorganization for hand movement in patients with multiple sclerosis: defining distinct effects of injury and disability*. Brain, vol. 125, no. 12, page 2646, 2002. (Cited on page 20.)

[Righi07] V. Righi, A. Mucci, L. Shenetti, M.R. Tosi*et al.* *Ex vivo HR-MAS magnetic resonance spectroscopy of normal and malignant human renal tissues*. Anticancer research, vol. 27, no. 5A, page 3195, 2007. (Cited on page 8.)

[Rocca02] M.A. Rocca, A. Falini, B. Colombo, G. Scotti, G. Comi and M. Filippi. *Adaptive functional changes in the cerebral cortex of patients with nondisabling multiple sclerosis correlate with the extent of brain structural damage*. Annals of neurology, vol. 51, no. 3, pages 330–339, 2002. (Cited on page 20.)

[Roche01] A. Roche, X. Pennec, G. Malandain and N. Ayache. *Rigid registration of 3-D ultrasound with MR images: a new approach combining intensity and gradient information*. Medical Imaging, IEEE Transactions on, vol. 20, no. 10, pages 1038–1049, 2001. (Cited on page 32.)

[Rodriguez07] J.C. Rodriguez. *Measuring financial contagion: A copula approach*. Journal of Empirical Finance, vol. 14, no. 3, pages 401–423, 2007. (Cited on page 108.)

[Rombouts05] S.A.R.B. Rombouts, F. Barkhof, R. Goekoop, C.J. Stam and P. Scheltens. *Altered resting state networks in mild cognitive impairment and mild Alzheimer's disease: an fMRI study*. Human Brain Mapping, vol. 26, no. 4, pages 231–239, 2005. (Cited on page 22.)

[Rombouts09] S.A.R.B. Rombouts, J.S. Damoiseaux, R. Goekoop, F. Barkhof, P. Scheltens, S.M. Smith and C.F. Beckmann. *Model-free group analysis shows altered BOLD FMRI networks in dementia*. Human brain mapping, vol. 30, no. 1, pages 256–266, 2009. (Cited on page 18.)

[Rosen90] B.R. Rosen, J.W. Belliveau, J.M. Vevea and T.J. Brady. *Perfusion imaging with NMR contrast agents*. Magnetic resonance in medicine, vol. 14, no. 2, pages 249–265, 1990. (Cited on page 18.)

[Ross83] BD Ross and GK Radda. *Application of 31P nmr to inborn errors of muscle metabolism.* Biochemical Society transactions, vol. 11, no. 6, page 627, 1983. (Cited on page 8.)

[Ross94] B. Ross and T. Michaelis. *Clinical applications of magnetic resonance spectroscopy.* Magnetic Resonance Quarterly, vol. 10, no. 4, page 191, 1994. (Cited on page 8.)

[Rousseau03] D. Rousseau, G. Anand and F. Chapeau-Blondeau. *Nonlinear estimation from quantized signals : quantizer optimization and stochastic resonance*. Physics in Signal and Image Processing, PSIP03, Grenoble, France, pages 89–92, 2003. (Cited on page 98.)

[Roy90] C.S. Roy and CS Sherrington. *On the regulation of the blood-supply of the brain*. The Journal of physiology, vol. 11, no. 1-2, page 85, 1890. (Cited on page 18.)

[Sanchez-Hernandez07] C. Sanchez-Hernandez, D.S. Boyd and G.M. Foody. *One-class classification for mapping a specific land-cover class: SVDD classification of fenland*. Geoscience and Remote Sensing, IEEE Transactions on, vol. 45, no. 4, pages 1061–1073, 2007. (Cited on page 94.)

[Schanda05] P. Schanda, Ē. Kupče and B. Brutscher. *SOFAST-HMQC experiments for recording two-dimensional deteronuclear correlation spectra of proteins within a few seconds*. Journal of Biomolecular NMR, vol. 33, no. 4, pages 199–211, 2005. (Cited on page 36.)

[Schmidt-Rohr94] K. Schmidt-Rohr and H.W. Spiess. Multidimensional solid-state nmr and polymers. Academic Press London, 1994. (Cited on pages 2 and 17.)

[Schneiderman98] H. Schneiderman and T. Kanade. *Probabilistic modeling of local appearance and spatial relationships for object recognition*. In Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on, pages 45–51. IEEE, 1998. (Cited on page 31.)

[Scholkopf00] B. Scholkopf, R.C. Williamson, A.J. Smola, J. Shawe-Taylor and J. Platt. *Support vector method for novelty detection*. Advances in neural information processing systems, vol. 12, no. 3, pages 582–588, 2000. (Cited on pages xii and 103.)

[Scholvinck10] M.L. Scholvinck, A. Maier, F.Q. Ye, J.H. Duyn and D.A. Leopold. *Neural basis of global resting-state fMRI activity*. Proceedings of the National Academy of Sciences, vol. 107, no. 22, page 10238, 2010. (Cited on page 22.)

[Scholz05] M. Scholz, F. Kaplan, C.L. Guy, J. Kopka and J. Selbig. *Non-linear PCA: a missing data approach*. Bioinformatics, vol. 21, no. 20, page 3887, 2005. (Cited on pages ix, 50 and 53.)

[Scholz07] M. Scholz. *Analysing periodic phenomena by circular PCA*. Bioinformatics Research and Development, pages 38–47, 2007. (Cited on page 50.)

[Seo05] J.K. Seo, G.C. Sharp and S.W. Lee. *Range data registration using photometric features*. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, volume 2, pages 1140–1145. IEEE, 2005. (Cited on page 31.)

[Shafer76] Glenn Shafer. A mathematical theory of evidence. Princeton University Press, Princeton, 1976. (Cited on pages 35 and 155.)

[Shawe-Taylor04] J. Shawe-Taylor and N. Cristianini. Kernel methods for pattern analysis. Cambridge Univ Pr, 2004. (Cited on pages xii and 95.)

[Shilane04] Philip Shilane, Patrick Min, Michael M. Kazhdan and Thomas A. Funkhouser. *The Princeton Shape Benchmark*. In Social Media Impact Conference, pages 167–178. IEEE Computer Society, 2004. (Cited on page 65.)

[Shum02] H.Y. Shum and R. Szeliski. *Construction of Panoramic Image Mosaics with Global and Local Alignment*. International Journal of Computer Vision, vol. 48, no. 2, pages 151–152, 2002. (Cited on page 32.)

[Silwood99] C.J.L. Silwood, E.J. Lynch, S. Seddon, A. Sheerin, A.W.D. Claxson and M.C. Grootveld. *1H-NMR analysis of microbial-derived organic acids in primary root carious lesions and saliva*. NMR in Biomedicine, vol. 12, no. 6, pages 345–356, 1999. (Cited on page 8.)

[Simari06] P. Simari, E. Kalogerakis and K. Singh. *Folding meshes: hierarchical mesh segmentation based on planar symmetry*. In Proceedings of the fourth Eurographics symposium on Geometry processing, pages 111–119. Eurographics Association, 2006. (Cited on page 49.)

[Sinclair10] A.J. Sinclair, M.R. Viant, A.K. Ball, M.A. Burdon, E.A. Walker, P.M. Stewart, S. Rauz and S.P. Young. *NMR-based metabolomic analysis of cerebrospinal fluid and serum in neurological diseases–a diagnostic tool?* NMR in Biomedicine, vol. 23, no. 2, pages 123–132, 2010. (Cited on page 8.)

[Sitter04] B. Sitter, T. Bathen, B. Hagen, C. Arentz, F.E. Skjeldestad and I.S. Gribbestad. *Cervical cancer tissue characterized by high-resolution magic angle spinning MR spectroscopy*. Magnetic Resonance Materials in Physics, Biology and Medicine, vol. 16, no. 4, pages 174–181, 2004. (Cited on page 8.)

[Smith93] AFM Smith and GO Roberts. *Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods*. Journal of the Royal Statistical Society. Series B (Methodological), vol. 55, no. 1, pages 3–23, 1993. (Cited on pages 75 and 149.)

[Stamkopoulos98] T. Stamkopoulos, K. Diamantaras, N. Maglaveras and M. Strintzis. *ECG analysis using nonlinear PCA neural networks for ischemia detection*. Signal Processing, IEEE Transactions on, vol. 46, no. 11, pages 3058–3067, 1998. (Cited on pages ix and 50.)

[Surguladze10] S.A. Surguladze, E.M. Chu, N. Marshall, A. Evans, A.P.P. Anilkumar, C. Timehin, C. McDonald, C. Ecker, M.L. Phillips and A.S. David. *Emotion processing in schizophrenia: fMRI study of patients treated with risperidone long-acting injections or conventional depot medication*. Journal of Psychopharmacology, 2010. (Cited on page 20.)

[Swanson08] M.G. Swanson, K.R. Keshari, Z.L. Tabatabai, J.P. Simko, K. Shinohara, P.R. Carroll, A.S. Zektzer and J. Kurhanewicz. *Quantification of choline- and ethanolamine-containing metabolites in human prostate tissues using 1H HR-MAS total correlation spectroscopy*. Magnetic Resonance in Medicine, vol. 60, no. 1, pages 33–40, 2008. (Cited on page 8.)

[Tao07] W. Tao, H. Jin and Y. Zhang. *Color image segmentation based on mean shift and normalized cuts*. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, vol. 37, no. 5, pages 1382–1389, 2007. (Cited on page 30.)

[Tate00] A.R. Tate, P.J.D. Foxall, E. Holmes, D. Moka, M. Spraul, J.K. Nicholson and J.C. Lindon. *Distinction between normal and renal cell carcinoma kidney cortical biopsy samples using pattern recognition of 1H magic angle spinning (MAS) NMR spectra*. NMR in Biomedicine, vol. 13, no. 2, pages 64–71, 2000. (Cited on page 8.)

[Tax04] D.M.J. Tax and R.P.W. Duin. *Support vector data description*. Machine Learning, vol. 54, no. 1, pages 45–66, 2004. (Cited on pages xii, xv, 100, 101 and 102.)

[Te-Won98] L. Te-Won. Independent component analysis, theory and applications. Boston: Kluwer Academic Publishers, 1998. (Cited on page 134.)

[Theis03] F.J. Theis, A. Jung, C.G. Puntonet and E.W. Lang. *Linear geometric ICA: Fundamentals and algorithms*. Neural Computation, vol. 15, no. 2, pages 419–439, 2003. (Cited on page 134.)

[Thompson07] P.M. Thompson, K.M. Hayashi, R.A. Dutton, M.C. CHIANG, A.D. Leow, E.R. Sowell, G. De Zubicaray, J.T. Becker, O.L. Lopez, H.J. Aizenstein*et al.* *Tracking Alzheimer's disease*. Annals of the New York Academy of Sciences, vol. 1097, no. 1, pages 183–214, 2007. (Cited on page 8.)

[Tiberio06] M. Tiberio, D.T. Chard, D.R. Altmann, G. Davies, C.M. Griffin, M.A. McLean, W. Rashid, J. Sastre-Garriga, A.J. Thompson and D.H. Miller. *Metabolite changes in early relapsing–remitting multiple sclerosis*. Journal of neurology, vol. 253, no. 2, pages 224–230, 2006. (Cited on page 17.)

[Toews05] M. Toews, D.L. Collins and T. Arbel. *Maximum a posteriori local histogram estimation for image registration*. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2005, pages 163–170, 2005. (Cited on pages xiii and 43.)

[Tomlins98] A.M. Tomlins, P.J.D. Foxall, M.J. Lynch, J. Parkinson, J.R. Everett and J.K. Nicholson. *High resolution NMR spectroscopic studies on dynamic biochemical processes in incubated human seminal fluid samples*. Biochimica et Biophysica Acta (BBA)-General Subjects, vol. 1379, no. 3, pages 367–380, 1998. (Cited on page 8.)

[Tugnoli04] V. Tugnoli, A. Mucci, L. Schenetti, C. Calabrese, G. Di Febo, MC Rossi and MR Tosi. *Molecular characterization of human gastric mucosa by HR-MAS magnetic resonance spectroscopy*. International journal of molecular medicine, vol. 14, pages 1065–1072, 2004. (Cited on page 8.)

[Tugnoli06] V. Tugnoli, A. Mucci, L. Schenetti, V. Righi, C. Calabrese, A. Fabbri, F. DI*et al.* *Ex vivo HR-MAS magnetic resonance spectroscopy of human gastric adenocarcinomas: a comparison with healthy gastric mucosa*. Oncology reports, vol. 16, no. 3, pages 543–553, 2006. (Cited on page 8.)

[Van de Ven95] F.J.M. Van de Ven. Multidimensional NMR in liquids: basic principles and experimental methods. Vch, 1995. (Cited on pages 10 and 14.)

[Vapnik98] V.N. Vapnik. *Statistical learning theory*. 1998. (Cited on page 96.)

[Vapnik00] V.N. Vapnik. The nature of statistical learning theory. Springer Verlag, 2000. (Cited on pages 95 and 96.)

[Vazquez98] A.L. Vazquez and D.C. Noll. *Nonlinear Aspects of the BOLD Response in Functional MRI* 1*. Neuroimage, vol. 7, no. 2, pages 108–118, 1998. (Cited on page 18.)

[Villringer88] A. Villringer, B.R. Rosen, J.W. Belliveau, J.L. Ackerman, R.B. Lauffer, R.B. Buxton, Y.S. Chao, V.J. Wedeenand and T.J. Brady. *Dynamic imaging with lanthanide chelates in normal brain: contrast due to magnetic*

susceptibility effects. Magnetic resonance in medicine, vol. 6, no. 2, pages 164–174, 1988. (Cited on page 18.)

[Viola95] P. Viola and W.M. Wells III. *Alignment by maximization of mutual information*. In iccv, page 16. Published by the IEEE Computer Society, 1995. (Cited on page 32.)

[Vranic00] DV Vranic and D. Saupe. *3D model retrieval*. Proc. SCCG 2000, pages 3–6, 2000. (Cited on page 65.)

[Vranić01a] D.V. Vranić and D. Saupe. *3D shape descriptor based on 3D Fourier transform*. In Proceedings of the EURASIP Conference on Digital Signal Processing for Multimedia Communications and Services (ECMCS 2001), Budapest, Hungary. Citeseer, 2001. (Cited on pages ix, xiii, 33, 49, 50, 55 and 138.)

[Vranic01b] D.V. Vranic, D. Saupe and J. Richter. *Tools for 3D-object retrieval: Karhunen-Loeve transform and spherical harmonics*. In Multimedia Signal Processing, 2001 IEEE Fourth Workshop on, pages 293–298. IEEE, 2001. (Cited on pages ix, 33 and 49.)

[Waltz90] E. Waltz and J. Llinas. Multisensor data fusion. Artech House Boston, 1990. (Cited on page 71.)

[Weiskopf05] N. Weiskopf, U. Klose, N. Birbaumer and K. Mathiak. *Single-shot compensation of image distortions and BOLD contrast optimization using multi-echo EPI for real-time fMRI*. Neuroimage, vol. 24, no. 4, pages 1068–1079, 2005. (Cited on page 134.)

[Weljie06] A.M. Weljie, J. Newton, P. Mercier, E. Carlson, C.M. Slupsky*et al. Targeted profiling: quantitative analysis of 1H NMR metabolomics data*. Anal Chem, vol. 78, no. 13, pages 4430–4442, 2006. (Cited on page x.)

[Wevers94] R.A. Wevers, U. Engelke and A. Heerschap. *High-resolution 1H-NMR spectroscopy of blood plasma for metabolic studies*. Clinical chemistry, vol. 40, no. 7, page 1245, 1994. (Cited on page 8.)

[Wright10] A.J. Wright, G.A. Fellows, J.R. Griffiths, M. Wilson, B.A. Bell and F.A. Howe. *Ex-vivo HRMAS of adult brain tumours: metabolite quantification and assignment of tumour biomarkers*. Molecular Cancer, vol. 9, no. 1, page 66, 2010. (Cited on page 8.)

[Xia08] J. Xia, T.C. Bjorndahl, P. Tang and D.S. Wishart. *MetaboMiner – semi-automated identification of metabolites from 2 D NMR spectra of complex biofluids*. BMC bioinformatics, vol. 9, no. 1, page 507, 2008. (Cited on pages xiii, 70, 71 and 78.)

[Xiao08] L. Xiao and P. Li. *Improvement on mean shift based tracking using second-order information*. In Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, pages 1–4. IEEE, 2008. (Cited on page 32.)

[Yang04] M. Yang, H.G. Zhang, J.M. Fu and F. Yan. *A framework for adaptive anomaly detection based on support vector data description*. Network and Parallel Computing, pages 443–450, 2004. (Cited on page 104.)

[Yazdani09] A. Yazdani, T. Ebrahimi and U. Hoffmann. *Classification of EEG signals using Dempster Shafer theory and a k-nearest neighbor classifier*. In Neural Engineering, 2009. NER'09. 4th International IEEE/EMBS Conference on, pages 327–330. IEEE, 2009. (Cited on page 156.)

[Zhang08] X. Zhang, L. Jiao, F. Liu, L. Bo and M. Gong. *Spectral clustering ensemble applied to SAR image segmentation*. Geoscience and Remote Sensing, IEEE Transactions on, vol. 46, no. 7, pages 2126–2136, 2008. (Cited on page 30.)

[Zhang09] D. Zhang, J.M. Johnston, M.D. Fox, E.C. Leuthardt, R.L. Grubb, M.R. Chicoine, M.D. Smyth, A.Z. Snyder, M.E. Raichle and J.S. Shimony. *Preoperative sensorimotor mapping in brain tumor patients using spontaneous fluctuations in neuronal activity imaged with fMRI: initial experience*. Neurosurgery, vol. 65, no. 6 Suppl, page 226, 2009. (Cited on page 20.)

[Zheng07] M. Zheng, P. Lu, Y. Liu, J. Pease, J. Usuka, G. Liao and G. Peltz. *2D NMR metabonomic analysis: a novel method for automated peak alignment*. Bioinformatics, vol. 23, no. 21, page 2926, 2007. (Cited on pages 70 and 71.)

[Zimmermann09] K. Zimmermann, J. Matas and T. Svoboda. *Tracking by an optimal sequence of linear predictors*. IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 677–692, 2009. (Cited on page 32.)

Résumé:
Les techniques d'acquisition des signaux médicaux sont en constante évolution et fournissent une quantité croissante de données hétérogènes qui doivent être analysées par le médecin. Dans ce contexte, des méthodes automatiques de traitement des signaux médicaux sont régulièrement proposées pour aider l'expert dans l'analyse qualitative et quantitative en facilitant leur interprétation. Ces méthodes doivent tenir compte de la physique de l'acquisition, de l'*a priori* que nous avons sur ces signaux et de la quantité de données à analyser pour une interprétation plus précise et plus fiable. Dans cette thèse, l'analyse des tissus biologique par spectroscopie RMN et la recherche des activités fonctionnelles cérébrales et leurs connectivités par IRMf sont explorées pour la recherche de nouveaux bio-marqueurs. Chaque information médicale sera caractérisée par un ensemble d'objets que nous cherchons à extraire, à aligner, et à coder. Le regroupement de ces objets par la mesure de leur similitude permettra leur classification et l'identification de bio-marqueurs. C'est ce schéma global d'indexation et de recherche par le contenu d'objets pour la détection des bio-marqueurs que nous proposons. Pour cela, nous nous sommes intéressés dans cette thèse à modéliser et intégrer les connaissances *a priori* que nous avons sur ces signaux biologiques permettant ainsi de proposer des méthodes appropriées à chaque étape d'indexation et à chaque type de signal.

Mots clès: Identification de bio-marqueurs, spectres HSQC, images fMRI, indexation, détection et alignement d'objet, codage et mesure de similarité, détection de changement.

Abstract:
The medical signal acquisition techniques are constantly evolving in recent years and providing an increasing amount of data which should be then analyzed. In this context, automatic signal processing methods are regularly proposed to assist the expert in the qualitative and quantitative analysis of these images in order to facilitate their interpretation. These methods should take into account the physics of signal acquisition, the *a priori* we have on the signal formation and the amount of data to analyze for a more accurate and reliable interpretation. In this thesis, we focus on the two-dimensional 2D Heteronuclear Single Quantum Coherence HSQC spectra obtained by High-Resolution Magic Angle Spinning HR-MAS NMR for biological tissue analysis and the functional Magnetic Resonance Imaging fMRI images for functional brain activities analysis. Each processed medical information will be characterized by a set of objects that we seek to extract, align, and code. The clustering of these objects by measuring their similarity will allow their classification and then the identification of biomarkers. It is this global content-based object indexing and retrieval scheme that we propose. We are interested in this thesis to properly model and integrate the *a priori* knowledge we have on these biological signal allowing us to propose thereafter appropriate methods to each indexing step and each type of signal.

Keywords: Biomarker identification, HSQC spectra, fMRI images, indexing, object detection and alignment, object coding and similarity measurement, change detection.