Contributions à l'apprentissage automatique pour l'analyse d'images cérébrales anatomiques

Rémi Cuingnet

Faculté des sciences d'Orsay

UNIVERSITÉ PARIS-SUD 11

THÈSE DE DOCTORAT

SPÉCIALITÉ: PHYSIQUE

École Doctorale "Sciences et Technologies de l'Information des Télécommunications et des Systèmes"

présentée par

Rémi Cuingnet

Contributions à l'apprentissage automatique pour l'analyse d'images cérébrales anatomiques

Membres du jury:

Président :	Emmanuel DURAND
Rapporteurs :	Xavier PENNEC
	Jean-Philippe VERT
Examinateurs :	Jean-François MANGIN
	Pierre CELSIS
	Alain TROUVÉ
Directeurs de thèse :	Habib BENALI
	Olivier COLLIOT

Remerciements

Si je fais le bilan de la formation académique que j'ai eu la chance de recevoir, le doctorat et la « prépa » ont été indéniablement les années les plus exaltantes. Ces années n'auraient pu se faire sans un cadre humain solide. Je remercie chaleureusement toutes les personnes qui m'ont soutenu durant ces années.

Je remercie tout d'abord les membres du jury présidé par Emmanuel Durand. Je remercie tout particulièrement Xavier Pennec et Jean-Philippe Vert de m'avoir fait l'honneur d'être rapporteurs de ce travail. Je leur exprime toute ma reconnaissance pour le temps qu'ils ont consacré à la lecture approfondie de ce manuscrit ainsi que pour les retours si constructifs qu'ils m'ont adressés. Je remercie aussi chaleureusement Pierre Celsis, Jean-Francois Mangin et Alain Trouvé d'avoir accepté de faire partie de mon jury de thèse.

Je remercie chaleureusement mes deux directeurs de thèse Habib Benali et Olivier Colliot. Merci à toi Olivier de m'avoir proposé cette thèse. Merci d'avoir su m'épauler tout en me laissant cette exaltante liberté propre à la recherche. Merci d'avoir cru en mes idées et de m'avoir consacré tant de temps. Merci pour tes conseils avisés, ton écoute et ton sens des rapports humains. Il en va aussi pour Marie Chupin, pierre angulaire du laboratoire malgré elle. Sa connaissance, sa gentillesse, sa rigueur, sa disponibilité et son empathie ont joué un rôle clé dans ma thèse.

Je remercie également Sylvain Baillet et Line Garnero d'avoir accueilli au LENA le néophyte que j'étais en traitement d'images médicales.

Au sein de l'équipe CogImage, je remercie particulièrement Mario Chavez et Jacques Martinerie pour leur enthousiasme, leur gentillesse et leur disponibilité. Je remercie l'ensemble des masters, doctorants et postdoctorants du laboratoire pour la bonne ambiance qu'ils créent. Je pense plus particulièrement à mes aînés : Julien Lefèvre que je n'ai croisé que durant mon stage et avec qui j'aurai eu grand plaisir à travailler, Anaël Dossevi, à Benoit Cottereau et son souci de «classitude» dans les présentations, les deux niçois Guillaume Auzias avec son empathie sans limite et Yohan Attal avec sa droiture, son courage et son optimisme exemplaire,

Manik Battacharjee (Je suis persuadé que l'on aurait fait un très bon duo), Florence Gombert la confidente, Michel Besserve à qui je souhaite d'intégrer rapidement le CNRS et Sheraz. Je pense également aux étudiants qui sont arrivés en même temps ou après moi : l'ingénieur-chercheur Thomas Samaille et son esprit si pragmatique, la danseuse Émilie Gerardin, ma co-déléguée Lucile Gamond et ses talents culinaires, Thibault Dumas si souriant, Guillaume Dumas (l'esprit bouillonnant) et Claire Boutet. Sans oublier les anciens stagiaires du LENA dont Gaëtan Yvert qui a dû passer presque plus de temps à résoudre mes énigmes qu'à son stage et surtout Sarah Montagne et Jérôme Tessieras maintenant devenus de très bons amis. Je remercie également Ali et Chabha pour leur gentillesse kabyle. Au sein du LIF, je remercie particulièrement Mélanie, Guillaume, Vincent, Caroline, David, Arnaud et Alexandre. Plus généralement je remercie tous les membres de CogImage et du LIF pour leur accueil, leur disponibilité et la bonne ambiance qu'ils créent.

J'exprime toute ma reconnaissance au brillant mathématicien Joan Glaunès, qui m'a tant aidé dans les tous derniers mois de la thèse.

Parmi les chercheurs rencontrés au cours de ces trois ans et demi, je remercie Gaël Varoquaux, Alexandre Gramfort, Timothée Cour[1] et Bertrand Thirion pour les discussions si enrichissantes que nous avons eues.

Je remercie également les neuroradiologues, neurologues et médecins nucléaires du groupe hospitalier de la Pitié-Salpêtrière sans qui nos travaux seraient dépourvus de sens. Je pense plus particulièrement à Didier Dormont, Charlotte Rosso, Yves Samson, Stéphane Lehéricy, Lionel Thivard, Marie-Odile Habert, Marie Sarazin et Dominique Hasboun qui malgré leurs contraintes professionnelles restent toujours disponibles pour nous.

Je remercie chaleureusement Dominique Martin de l'École Doctorale STITS qui s'est tant dévouée pour aider les doctorants.

Je remercie mes amis très proches. Merci Elyès Jaillet non seulement d'avoir relu une partie du manuscrit mais d'avoir toujours était présent dès que j'avais besoin d'aide. Il en va aussi pour Thomas Guillet et Thierry. Thomas, merci aussi d'avoir répondu à toutes mes questions, quelque fussent le sujet et le niveau de difficulté et à n'importe quel moment de la journée. Tu m'impressionnes vraiment.

Un immense merci à ma famille pour leur soutien et leur amour inconditionnels. Je pense plus particulièrement à Papa, Maman, Thomas, Clémence, à mes Grands-Parents (Bon Papa, Papy et Many) et à Lily qui partage désormais ma vie et à qui je dois cette thèse.

Merci à toi Bonne Maman pour ta finesse, ta délicatesse, ta gentillesse, ta droiture et ton courage. Tu as été et resteras un modèle à suivre. J'aurais été si fier de te présenter ma thèse.

[1] et aussi pour m'avoir encadré à l'UPenn

Table des matières

Table des notations et symboles

Notations générales :

$\mathcal{H}$	espace de Hilbert	16
$\langle \cdot \vert \cdot \rangle$	produit scalaire	16
$\mathcal{L}$	ensemble des applications linéaires	91
$\mathcal{H}^*$	espace dual de $\mathcal{H}$	17
$\mathbf{h}^*$	dual d'un vecteur $\mathbf{h} \in \mathcal{H}$	91
$M^\dagger$	opérateur adjoint de l'opérateur M	90
M^T	transposée de la matrice M	88
$\mathrm{Sp}(M)$	spectre d'un opérateur M	107
κ_2	conditionnement relatif à la norme spectrale	102
I_d	matrice identité de taille $d \times d$	100
$\mathbf{1}_d$	vecteur colonne de taille d constitué que de 1	101
δ_{ij}	symbole de Kronecker	111
δ_x	distribution de Dirac	121
$\mathbf{P}\{\omega\}$	probabilité d'un événement ω	11
$\mathbf{E}[X]$	espérance de la variable aléatoire X	166
P	opérateur de régularisation	90
G	opérateur de Green	90

Notations pour les SVM :

$\mathcal{X}$	ensemble des observations ou *input space*	10
$\mathbf{x}$	observation (élément de $\mathcal{X}$)	10
$\mathcal{Y}$	ensemble des classes	10
y	classe d'une observation $\mathbf{x}$	10
$\mathcal{S}$	ensemble des sujets	86

s	indice d'un sujet	86
N	nombre de sujets de la population d'étude	11
f	fonction de classification	11
$\mathcal{F}$	ensemble d'hypothèses	13
ℓ	fonction de perte	12
ℓ_{0-1}	fonction de perte binaire	12
ℓ_{hinge}	fonction de perte « *hinge loss* »	20
$R_\ell[f]$	fonctionnelle de risque associée à la fonction de perte ℓ	12
R^*	risque de Bayes	11
$R_\ell^N[f]$	risque empirique pour un échantillon de taille N	13
f^*	classifieur de Bayes	11
$\hat{f}_N$	estimateur ERM	13
K	noyau (positif semi-défini)	15
ϕ	fonction qui envoie les observations dans l'espace des *features*	16
$\mathcal{R}_{N,X}$	complexité de Rademacher	18
$\mathbf{w}^{\text{opt}}$	vecteur de pondérations optimal dans l'espace des *f*eatures	20
α^{opt}	vecteur de pondérations optimal du problème dual	22
b	biais ou seuil	20
C	paramètre de régularisation du SVM	21
λ	paramètre de régularisation	19
ξ	*slack variable* ou variable ressort	21
m	marge du SVM	148
$\mathcal{V}$	domaines des images ou surfaces corticales	87
v	voxel ou nœud du maillage cortical	87
d	dimension de l'espace des données (ex : nombre de voxels)	86

Notations pour les graphes

A	matrice d'adjacence d'un graphe	99
L	Laplacien d'un graphe	91
$\tilde{L}$	Laplacien normalisé d'un graphe	100
$\mathcal{A}_r$	région r d'un atlas probabiliste	99
R	nombre de régions r d'un atlas probabiliste	99
$d^{(r)}$	nombre de voxels de la région r d'un atlas	101

Notations pour les variétés riemanniennes

$\mathcal{M}$	variété différentielle	111
g	tenseur métrique	111
h	inverse du tenseur métrique g	96

Δ opérateur de Laplace-Beltrami 95

Notations pour la diffusion

β paramètre de diffusion 91

$\mathbf{K}$ matrice de rigidité 113

$\mathbf{M}$ matrice de masse 113

ψ élément fini rectangulaire 113

Introduction

Les méthodes de neuroimagerie ont, depuis une vingtaine d'années, permis l'étude non invasive et in vivo du cerveau humain. En particulier, l'imagerie par résonance magnétique permet de visualiser sa structure avec une précision spatiale de l'ordre du millimètre.

L'analyse automatique des structures macroscopiques du cerveau a de nombreuses applications pour la compréhension et l'aide au diagnostic de pathologies neurologiques. Un des objectif majeurs de telles analyses est de permettre au neuroradiologue d'obtenir, à partir d'images structurelles, des informations souvent indécelables en routine clinique sur l'état d'un patient. Cela peut s'avérer crucial pour la détection précoce des pathologies évolutives telles que la maladie d'Alzheimer. En effet, le seul diagnostic de certitude de la maladie d'Alzheimer est post mortem. Quant au diagnostic de probabilité, basé essentiellement sur des tests neurospychologiques, il reste encore souvent tardif.

* *
*

Parmi les analyses d'images structurelles en neuroimagerie, on distingue les approches locales ou ciblées, des approches globales. Les approches locales ont un fort *a priori* sur les structures touchées par la pathologie. Elles réduisent le champ d'investigation à quelques structures prédéfinies. Les approches globales, plus prospectives, analysent le cerveau dans son ensemble. Elles ont, ces dernières années, principalement reposé sur les analyses univariées voxel-à-voxel. Dans de telles analyses, les images sont préalablement recalées dans un espace stéréotaxique commun à tous les sujets. On cherche à détecter les voxels en lesquels il y a des différences significatives entre les deux populations ; pour ce faire, des tests univariés sont effectués en chaque voxel. Cependant, la sensibilité de ces approches est limitée quand les différences entre les populations étudiées mettent en jeu des combinaisons complexes

de structures cérébrales. De plus, ces analyses sont généralement conçues pour l'analyse de groupe. Elles ne fournissent pas d'information au niveau individuel.

Récemment, il y a eu dans la communauté de la neuroimagerie un intérêt croissant pour les méthodes de classification comme les machines à vecteurs supports. De telles approches permettent de dépasser les limites de l'analyse univariée en prenant en compte des relations multivariées présentes dans les données.

* *
*

Dans cette thèse, nous nous intéressons principalement à l'apprentissage automatique par machines à vecteurs supports pour l'analyse de populations et la classification de patients en neuroimagerie. Plus particulièrement, il s'agit de contribuer au développement d'outils d'aide au diagnostic à partir d'images structurelles obtenues en imagerie par résonance magnétique.

Le développement d'une méthode de classification d'images cérébrales suppose d'effectuer différents choix méthodologiques. Il faut d'abord choisir les caractéristiques ou *features* utilisées pour la classification (par exemple le volume ou l'épaisseur de structures cérébrales). Différentes stratégies pour réduire ce nombre de *features* sont possibles. Il faut enfin définir un classifieur. Différentes approches ont été proposées et testées dans la littérature. Elles ont toutefois été évaluées sur des populations différentes, ce qui rend délicat toute comparaison de leurs performances. Nous avons donc comparé les performances de différentes stratégies, dans le contexte de la maladie d'Alzheimer, à partir de 509 images IRM anatomiques provenant d'une base de données commune, la base ADNI.

Les différentes stratégies que nous avons comparées tiennent insuffisamment compte de la distribution spatiale et anatomique des *features*. Cela se reflète en particulier dans les hyperplans séparateurs qui manquent de cohérence avec l'anatomie sous-jacente. Dans cette thèse, nous proposons un cadre de régularisation spatiale et anatomique des machines à vecteurs supports pour des données de neuroimagerie volumiques ou surfaciques, dans le formalisme de la régularisation laplacienne. Ce cadre pemet d'intégrer différents types d'informations *a priori* directement dans un classifieur SVM à l'aide d'opérateurs de régularisation via la définition de la notion de proximité.

Nous proposons différents modèles de proximité correspondant à diverses contraintes spatiales ou anatomiques. Le premier modèle de proximité est ce que nous appelons la proximité spatiale. Autrement dit, deux *features* sont proches si et seulement s'ils sont spatialement proches. Le deuxième modèle de proximité, que nous appelons proximité anatomique, est un peu plus complexe. Dans ce modèle, deux *features* sont considérés comme proches s'ils appartiennent à la même structure cérébrale ou s'ils sont connectés (par exemple par un faisceau de fibres). Nous proposons ensuite des approches pour combiner ces deux types de

proximités (spatiale et anatomique). En remplaçant la régularisation du SVM linéaire par une régularisation spectrale, nous pouvons intégrer dans le SVM diverses contraintes spatiales ou anatomiques. Une telle régularisation conduit à l'utilisation de noyaux de diffusion.

Nous avons enfin appliqué cette nouvelle approche à deux problématiques cliniques : la maladie d'Alzheimer et les accidents vasculaires cérébraux.

* *
*

Le manuscrit de thèse est organisé de la façon suivante.

Le **1**er **chapitre** présente une introduction sur l'apprentissage statistique en général, avant de décrire plus en détails une méthode fréquemment utilisée, les machines à vecteurs supports. Nous présentons ensuite une brève étude bibliographique des méthodes de classification d'images structurelles en neuroimagerie. L'étude de ces méthodes et de leurs limites nous conduira aux améliorations envisageables et ainsi aux principaux objectifs de la thèse.

Le **chapitre 2** est consacré à la comparaison de différentes stratégies de classification automatique dans le contexte de la maladie d'Alzheimer sur une grande base de données multicentrique (509 sujets, base ADNI).

Dans le **chapitre 3**, nous proposons un cadre de régularisation spatiale et anatomique des machines à vecteurs supports en neuroimagerie, dans le formalisme de la régularisation laplacienne. Nous proposons notamment des approches pour les données volumiques et pour les données surfaciques. Les cadres continus et discrets sont également considérés. Enfin, nous proposons des approches pour combiner différents types de régularisation.

Dans une dernière partie, **chapitre 4**, cette méthode de régularisation est appliquée à deux problématiques cliniques. La première concerne la classification automatique de patients atteints de la maladie d'Alzheimer. La seconde porte sur l'analyse des régions cérébrales asssociées au pronostic des patients victimes d'un accident vasculaire cérébral.

Chapitre 1

Problématique

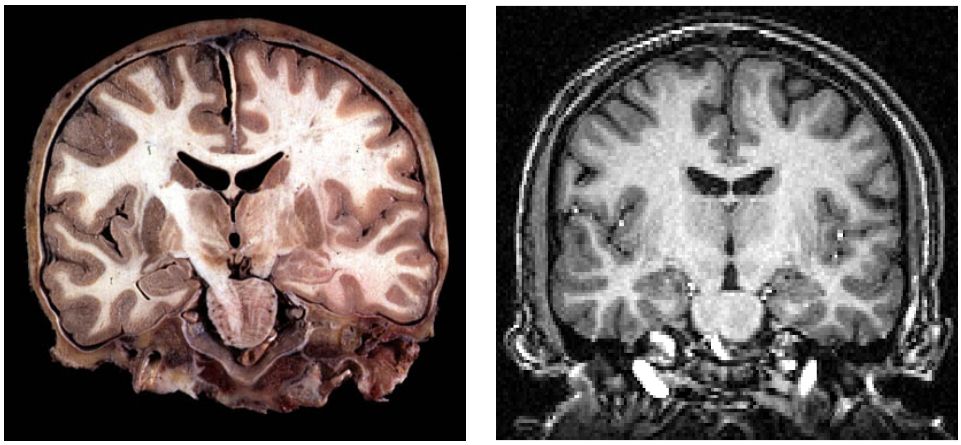
Le but de ce chapitre est de situer la problématique dans laquelle le travail de thèse s'inscrit et de présenter un bref état de l'art de la classification en imagerie médicale. Il est organisé de la façon suivante. Dans la section 1.1, nous verrons dans quelle mesure l'anatomie computationnelle, issue de l'émergence de nouvelles modalités non invasives d'imagerie et des avancées en traitement d'images et intelligence artificielle, a permis de réaliser des progrès dans l'étude des pathologies. La section 1.2 présente une introduction sur l'apprentissage statistique en général puis, plus particulièrement, sur les machines à vecteurs supports. La section 1.3 présente une courte bibliographie sur les méthodes de classification d'images structurelles en neuroimagerie. Nous définirons ensuite les principaux objectifs de la thèse dans la section 1.4.

1.1 L'apport de l'anatomie computationnelle à l'étude des pathologies cérébrales

Cette section a été en partie inspirée de la présentation de Xavier Pennec (colloque ETVC'08).

1.1.1 L'anatomie computationnelle

Depuis les années 1980, on assiste à l'essor de l'imagerie médicale avec l'émergence de nouvelles modalités permettant l'observation in vivo d'organismes, de leur structure et de leur fonctionnement. L'apparition de l'imagerie par résonance magnétique (IRM) a notamment eu un impact considérable. Elle permet de mesurer et de visualiser en trois dimensions et avec une précision spatiale de l'ordre du millimètre (figure 1.1), certaines caractéristiques sur les tissus observés comme notamment la concentration en eau dans la matière molle. Cette technique est basée sur le phénomène de *résonance magnétique nucléaire*.



(a) (b)

Figure 1.1 : (a) : coupe coronale d'un cerveau ; (b) : coupe coronale d'une image IRM anatomique (pondérée en T_1) d'un cerveau (source de l'image (a) : cours de D. Hasboun).

Étant le plus souvent non-invasives, ces méthodes d'acquisition ont permis la création de larges bases de données. Il est maintenant possible, grâce aux avancées dans les domaines du traitement d'images, de la vision par ordinateur et de l'intelligence artificielle, d'analyser ces données dans le but de décrire et/ou de modéliser l'anatomie d'un organisme vivant. C'est l'**anatomie computationnelle**. L'anatomie computationnelle consiste donc à modéliser et à analyser les observations structurelles.

1.1.2 Objectifs de l'anatomie computationnelle pour l'étude des pathologies cérébrales

1.1.2.1 L'étude des pathologies

En ce qui concerne les pathologies cérébrales, l'émergence de grandes bases de données d'imagerie telles que la base ADNI (voir chapitre 2) ont permis, en comparant différentes populations, de confirmer ou mettre en évidence des schémas d'altérations spécifiques dans de nombreuses pathologies. L'anatomie computationnelle a été appliquée à de nombreuses affections neurologiques et psychiatriques.

Les **affections neurologiques** souvent étudiées en neuroanatomie computationnelle sont :

- les **démences**, qu'elles soient **neurodégénératives**, comme la maladie d'Alzheimer [Lehéricy et al., 2007], la démence à corps de Lewy [McKeith et al., 2004 ; Watson et al., 2009] ou les démences fronto-temporales [Rosen et al., 2002 ; Du et al., 2007], ou d'**origines vasculaires** [Barber et al., 1999] ;

- les **affections motrices** comme la maladie de Parkinson [Brooks, 2010] ou la chorée de Huntington [Esmaeilzadeh et al., 2010] ;
- l'**épilepsie** [Richardson, 2010] ;
- la **sclérose en plaques** [Zivadinov & Cox, 2007] ;
- les **tumeurs cérébrales** [Gerstner et al., 2008] ;
- les **affections d'origines lésionnelles** telles que les accidents vasculaires cérébraux [Stinear, 2010] ou les traumatismes crâniens [Gallagher et al., 2007].

De nombreuses études se sont également intéressées aux **maladies psychiatriques** comme :

- la **schizophrénie** [Wright et al., 1995 ; Pearlson & Marsh, 1999] ;
- l'**autisme** [Verhoeven et al., 2010] ;
- la **dépression** Soares & Mann [1997].

L'anatomie computationnelle a permis d'étudier des pathologies in vivo de façon systématique sur de grandes populations, alors qu'autrefois les analyses étaient restreintes aux études post mortem sur quelques cas. Elle a ainsi permis de réaliser des progrès dans la compréhension de la physiopathologie des affections étudiées. Par exemple, des altérations structurelles ont été mises en évidence dans des affections psychiatriques telles que la schizophrénie. Elle permet également de définir de nouveaux marqueurs de substitution pour les essais thérapeutiques (développement de nouveaux médicaments) (ex : dans la maladie d'Alzheimer [Jack Jr et al., 2003]).

1.1.2.2 Aide au diagnostic

La compréhension des dysfonctionnements pathologiques, au niveau de la population, permet également de mettre en place, au niveau individuel, des stratégies de prévention, de détection/diagnostic et de suivi thérapeutique.

Toute la difficulté réside dans l'utilisation de l'information inférée à partir d'une population pour un problème au niveau individuel. Ce problème peut généralement s'écrire comme un problème de régression ou de classification. Jusqu'à aujourd'hui, ces outils sont principalement utilisés dans un but d'aide au diagnostic. L'objectif est d'aider le clinicien à détecter une maladie. On appelle cela le diagnostic assisté par ordinateur ou *computer-aided diagnosis*.

On rencontre ce genre d'approches par exemple pour la **maladie d'Alzheimer** [Freeborough & Fox, 1998 ; Teipel et al., 2007 ; Colliot et al., 2008a ; Duchesne et al., 2008 ; Magnin et al.,

2009 ; Vemuri et al., 2008 ; Gerardin et al., 2009 ; Querbes et al., 2009 ; Hinrichs et al., 2009 ; Chupin et al., 2009a ; Fan et al., 2008b,a ; Davatzikos et al., 2008a,b ; Misra et al., 2009 ; Desikan et al., 2009 ; Zhang et al., 2011], les **démences fronto-temporales** [Klöppel et al., 2008a ; Davatzikos et al., 2008b], la **schizophrénie** [Caan et al., 2006 ; Fan et al., 2005, 2007 ; Ingalhalikar et al., 2010 ; Golland et al., 2005], l'**épilepsie** [Duchesne et al., 2006] ou encore l'**autisme**[Ingalhalikar et al., 2010].

1.1.3 Les méthodes

1.1.3.1 Le principe général

Le principe général de l'anatomie computationnelle en neuroimagerie est le suivant :

1. **Images in vivo** – La première étape est l'acquisition d'images in vivo du cerveau.

2. **Extraction d'invariants** – Une fois ces images obtenues, la deuxième étape consiste à en extraire des invariants par rapport à la modalité. Autrement dit on cherche à obtenir, à partir des observations, des descripteurs du cerveau. Ces descripteurs ne doivent dépendre, dans l'idéal, que du cerveau, de sa forme ou de ses propriétés physiques, et non plus de la modalité. Ces invariants peuvent être un ensemble de points (points de repère anatomiques ou fonctionnels), de courbes (ex : sillons [Mangin et al., 2004a ; Fillard et al., 2007 ; Perrot et al., 2009]), de surfaces (ex : cortex [Fischl & Dale, 2000]) ou d'images (ex : *voxel-based morphometry* - VBM [Ashburner & Friston, 2000]) (figure 1.2).

3. **Analyse des invariants** – La dernière étape est l'analyse de ces invariants. Cela peut nécessiter la définition de certains objets mathématiques tels qu'une mesure d'affinité entre les invariants, ou d'un noyau défini positif si l'on veut faire de la classification de type machines à vecteurs supports.

   En général, pour faciliter ces définitions, cette étape est généralement précédée d'une mise en correspondance entre les invariants d'un couple ou d'un groupe de sujets. Cette étape d'appariement est appelée **recalage**. Le plus souvent, cette étape se fait en cherchant une transformation qui déforme les invariants d'un cerveau pour qu'ils soient le plus similaires possible, à une variance résiduelle près, à ceux d'un autre cerveau (ex : [Miller & Younes, 2001 ; Miller et al., 2006 ; Ashburner, 2007 ; Pennec et al., 2008 ; Auzias et al., 2009]). Notons qu'il existe d'autres moyens de mise en correspondance (ex : dans le cas du *primal sketch* [Coulon et al., 2000 ; Cachia et al., 2003]). L'analyse se fait ensuite en analysant directement les transformations (ex : [Durrleman et al., 2007, 2008]) ou en analysant les différences entre chaque couple d'invariants mis en correspondance (ex : VBM [Ashburner & Friston, 2000], cf. figures 1.4 et 1.5).

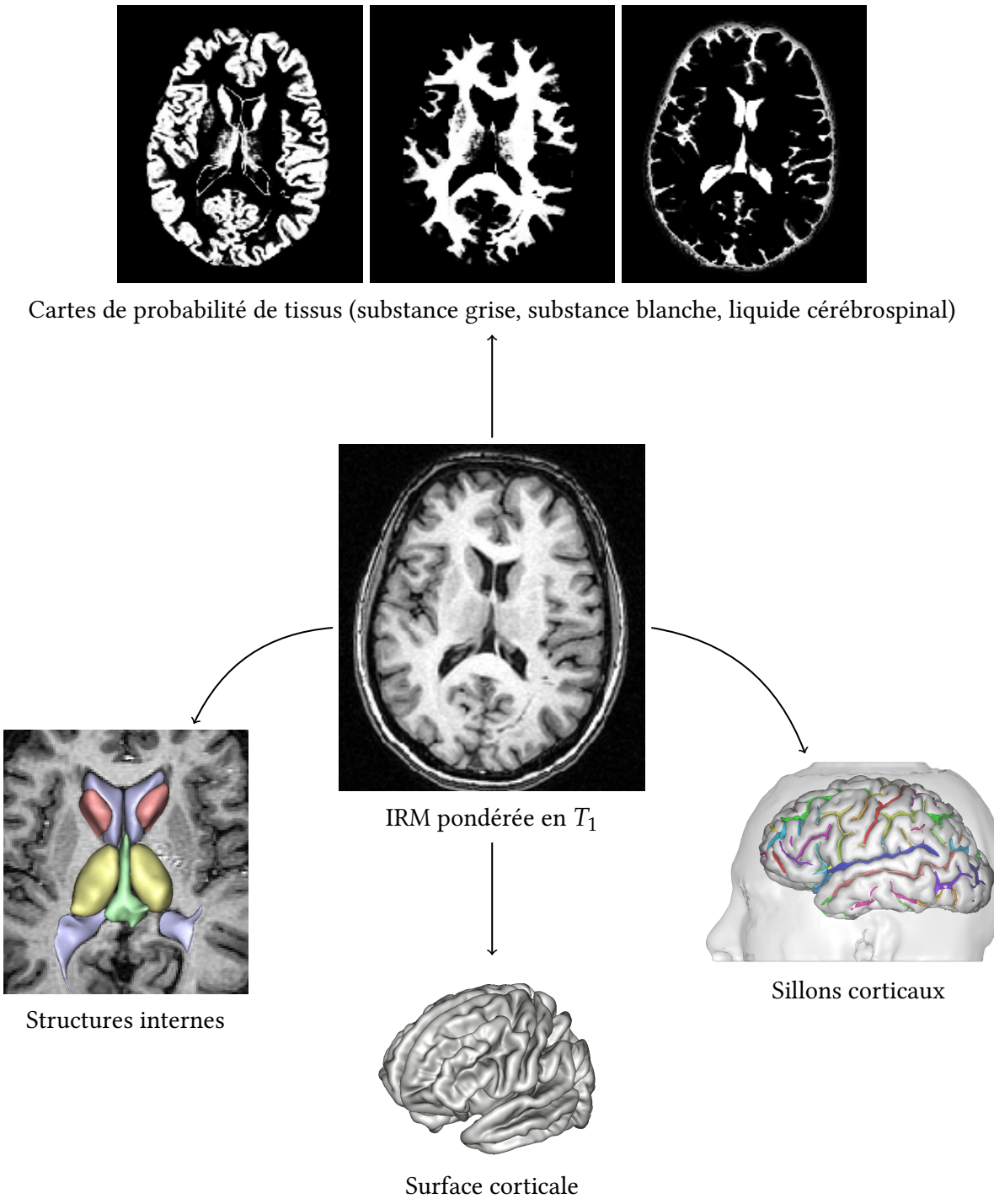


FIGURE 1.2 : Exemples d'invariants extraits à partir d'une image IRM pondérée en T_1. Les sillons corticaux ont été extraits avec BrainVisa (image fournie par G. Auzias). La surface corticale provient de FreeSurfer. Les tissus ont été segmentés avec SPM.

1.1.3.2 L'analyse voxel-à-voxel

L'approche la plus couramment utilisée en neuroimagerie computationnelle consiste à recaler toutes les images/cartes dans un espace commun à tous les sujets de telle sorte que le i-ème voxel d'une image d'un sujet corresponde au i-ème voxel des images des autres sujets.

Cette approche a l'avantage d'offrir un cadre d'analyse simple. Notamment, un des avantages de ce cadre de travail est que la combinaison linéaire d'images recalées a un sens. Cependant, il faut garder en mémoire que le problème de mise en correspondance de deux cerveaux n'est absolument pas résolu. Il n'y a donc aucune garantie que l'hypothèse de la correspondance voxel-à-voxel soit vérifiée [Mangin et al., 2004b].

Les approches voxel-à-voxel sont généralement utilisées pour des analyses de groupes de type VBM [Ashburner & Friston, 2000]. Des tests univariés sont alors effectués au niveau de chaque voxel afin de détecter des différences significatives entre les groupes étudiés. L'évaluation simultanée du pouvoir discriminant de chaque voxel d'une image à partir d'un unique échantillon de données est un problème connu en statistique sous le nom de **comparaisons multiples**. Différentes méthodes ont été proposées pour résoudre ce problème (*False Discovery Rate* [Benjamini & Yekutieli, 2001], *Random Field Theory* [Worsley et al., 1996], ...).

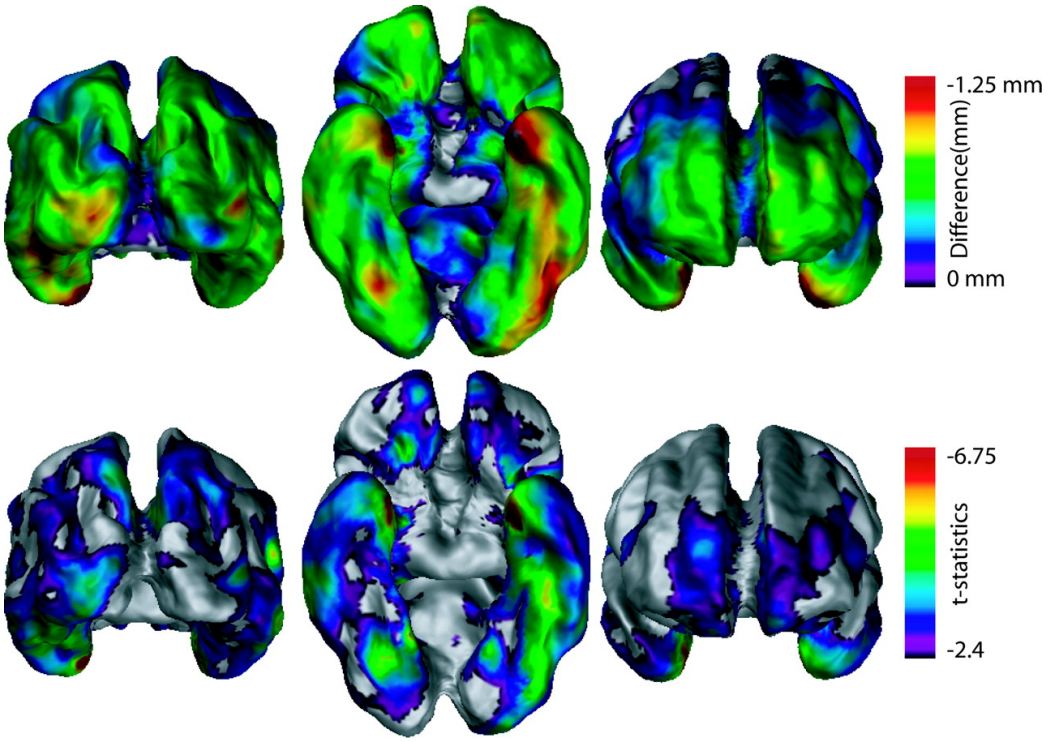


Figure 1.3 : Exemple de résultats d'analyses d'épaisseur corticale. Différences d'épaisseur corticale entre des sujets atteints de la maladie d'Alzheimer et des témoins (ligne du haut) et cartes statistiques t correspondantes (ligne du bas) (extrait de [Lerch et al., 2005]).

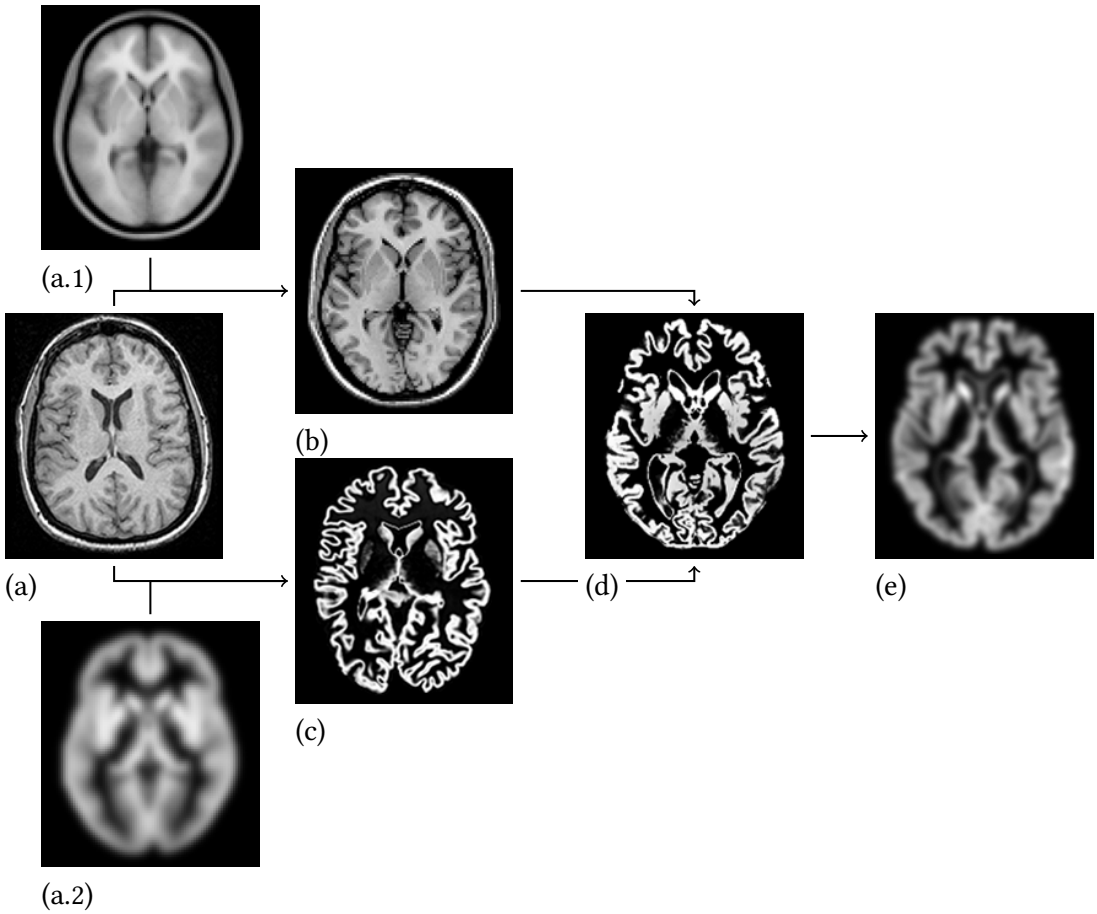


FIGURE 1.4 : Principe de la *voxel-based morphometry* (VBM). L'image IRM (a) est normalisée spatialement (b) à l'aide du template (a.1). Elle est également segmentée (c) à l'aide des *a priori* sur les tissus (a.2). La transformation est appliquée à l'image segmentée (c). Pour conserver la quantité de tissus, la carte de gris segmentée et normalisée est modulée par le jacobien de la transformation (d). Elle est ensuite lissée (e) avec un noyau gaussien avant les analyses voxel-à-voxel.

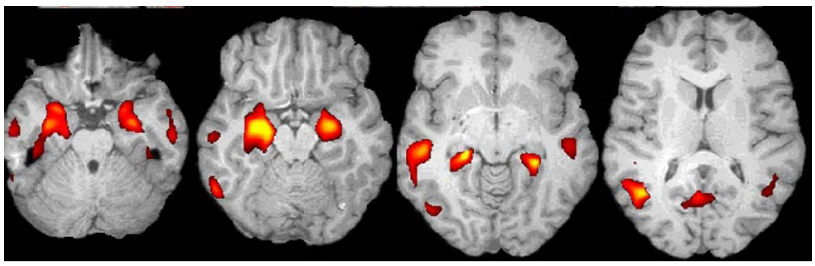


FIGURE 1.5 : Exemple de résultats d'analyse VBM : cartes de réduction de matière grise chez des patients atteints de la maladie d'Alzheimer par rapport à des sujets âgés sains (extrait de [Lehéricy et al., 2007]).

1.1.4 Situation du problème

Cependant, ces approches d'analyses univariées ont principalement deux défauts : (i) elles apportent peu d'information au niveau individuel et (ii) leur sensibilité est limitée quand les différences étudiées mettent en jeu des combinaisons de différentes structures du cerveau [Davatzikos, 2004]. Afin de dépasser ces limites de l'analyse univariée de masse, différents auteurs (ex : [Lao et al., 2004]) ont proposé d'utiliser des outils de classification multivariés tels que les **machines à vecteurs supports** (SVM) [Vapnik, 1995 ; Shawe-Taylor & Cristianini, 2000]. Historiquement, ces approches ont eu de nombreuses d'applications en bioinformatique [Mukherjee et al., 1999 ; Furey et al., 2000 ; Guyon et al., 2002 ; Brown et al., 2000 ; Hua & Sun, 2001 ; Ding & Dubchak, 2001 ; Jaakkola et al., 2000 ; Vert, 2002 ; Schölkopf et al., 2004]. Leur utilisation en neuroimagerie est plus récente et moins répandue.

Cette thèse a pour thème l'apprentissage automatique pour l'analyse de populations et la classification de patients en neuroimagerie. Nous nous placerons dans un cadre d'analyse voxel-à-voxel, tout en restant conscient des limites de cette approche [Mangin et al., 2004b]. Nous nous intéresserons principalement à l'utilisation des SVM pour ces analyses.

Le reste de ce chapitre est organisé de la façon suivante. Après une introduction sur l'apprentissage automatique, nous verrons quelles sont les méthodes d'apprentissage automatique principalement utilisées en neuroimagerie, leur limites ainsi que les améliorations possibles. Nous préciserons alors les objectifs de la thèse

1.2 Les méthodes d'apprentissage automatique

1.2.1 L'apprentissage par machines à vecteurs supports

Cette introduction sur la classification automatique est essentiellement basée sur les cinq livres suivants, [Devroye et al., 1996 ; Schölkopf & Smola, 2001 ; Hastie et al., 2005 ; Shawe-Taylor & Cristianini, 2000, 2004], sur les tutoriaux de Burges [1998], Vapnik [1999], Muller et al. [2001] et Bousquet et al. [2004] ainsi que sur le cours de Jean-Philippe Vert du Master mathématiques-vision-apprentissage (MVA) de l'ENS de Cachan (2007).

1.2.1.1 Le principe de la classification automatique

La **classification automatique** ou la **reconnaissance automatique de formes** a pour but d'identifier ou de prédire la « nature » d'un objet : noir ou blanc, un ou zéro, malade ou sain. Les objets, que l'on note $\mathbf{x}$, sont appelés des **observations** et l'ensemble des observations, $\mathcal{X}$, est appelé **domaine** ou *input space*. On appelle **classe** d'une observation $\mathbf{x}$ sa nature notée y. Soit $\mathcal{Y}$ l'ensemble des classes possibles. Nous supposerons dans la suite que $\mathcal{Y}$ est fini ; soit M son

cardinal. L'ensemble $\mathcal{Y}$ étant fini, une classe peut donc être représentée par un entier naturel : $\mathcal{Y} = [\![1, M]\!]$.

Lorsqu'il n'y a que deux classes, on parle alors de **classification binaire**. Notons que, dans le cas de problèmes de classification binaire, à la notation $\{1, 2\}$ est préférée[1] la notation $\{-1, +1\}$ pour les labels des classes. Une fonction $f \in \mathcal{Y}^{\mathcal{X}}$ qui à une observation $\mathbf{x}$ associe sa classe supposée $f(x)$ est appelée **classifieur** ou **fonction de classification** (également **fonction de prédiction** ou **hypothèse**). L'objectif de la classification automatique est donc de trouver la fonction f la plus « proche » de la réalité. Dans la suite, nous ne considérerons que le problème de classification binaire.

Il se peut que la classe ne soit pas une fonction déterministe de l'observation. Pour cette raison, on se place dans un cadre probabiliste. Les observations sont modélisées par une variable aléatoire X à valeur dans $\mathcal{X}$ et les classes comme une variable aléatoire Y à valeur dans $\mathcal{Y}$. Un classifieur f se trompe quand $f(X) \neq Y$; la **probabilité d'erreur d'un classifieur** f est donc :

$$R_{\ell_{0-1}}[f] = \mathbf{P}\{f(X) \neq Y\}$$

L'objectif de la classification automatique est donc de trouver un classifieur qui donne la probabilité d'erreur la plus petite possible, $R^*_{\ell_{0-1}}$, appelée **erreur de Bayes**. Supposons qu'un tel classifieur existe. Notons le f^*. Il est donc par définition donné par :

$$f^* = \underset{f \in \{\pm 1\}^{\mathcal{X}}}{\arg\min}\, R_{\ell_{0-1}}[f]$$

Ce classifieur est appelé **classifieur de Bayes** ou **règle de Bayes**. Le problème est que, la plupart du temps, la distribution de (X, Y) est inconnue et, par conséquent, f^* l'est aussi. Peut-on tout de même trouver un classifieur proche de f^* ?

Pour construire un tel classifieur, on suppose que l'on a accès à un échantillon, appelé **ensemble d'apprentissage**, composé de N couples de variables aléatoires $\{(X_i, Y_i)\}_{i \in [1,N]}$. Pour que la construction du classifieur ait une chance d'être correcte, l'échantillon utilisé doit être représentatif de la distribution de (X, Y). Pour cela, nous faisons donc l'hypothèse que ces variables aléatoires $\{(X_i, Y_i)\}_{i \in [1,N]}$ sont indépendantes et identiquement distribuées[2] selon la loi de (X, Y). On construit alors un classifieur à partir de la base de données d'apprentissage, $\{(X_1, Y_1), \cdots, (X_N, Y_N)\}$. Elle contient des observations déjà traitées et validées par un

[1] La résolution des problèmes de classification binaire passe généralement par une relaxation de l'ensemble des classes à l'espace $\mathbb{R}$ tout entier où les nombres réels positifs correspondent aux labels d'une classe et les nombres négatifs à ceux de l'autre classe.

[2] Notons que cette hypothèse est très forte et qu'elle n'est en pratique que rarement réalisée.

expert (*i.e.* leur classe est connue). Pour cette raison on parle d'**apprentissage supervisé**[3] (*supervised learning*) ou parfois même d'**apprentissage avec un professeur** (*learning with a teacher*). Nous noterons $f_N(\cdot; X_1, Y_1, \cdots, X_N, Y_N)$ un tel classifieur. L'erreur du classifieur f_N est donc :

$$R_{\ell_{0-1}}[f_N] = \mathbf{P}\left\{f_N(X; X_1, Y_1, \cdots, X_N, Y_N) \neq Y | X_1, Y_1, \cdots, X_N, Y_N\right\}$$

On appelle **règle de classification** toute suite de fonctions $(f_N)_{N\in\mathbb{N}}$. Une règle de classification est une bonne règle de classification si et seulement si elle est **consistante**[4], autrement dit si l'erreur $R_{\ell_{0-1}}[f_N]$ converge en probabilité vers l'erreur de Bayes et qu'elle **converge rapidement**.

Stone [1977] a montré qu'il existe des règles de classification qui sont consistantes indépendamment de la distribution de (X, Y) (chapitre 6 de [Devroye et al., 1996]). On dit qu'elles sont **universellement consistantes**. Malheureusement, la vitesse de convergence d'une règle de classification est, quant à elle, dépendante de la distribution du couple (X, Y) (chapitre 7 de [Devroye et al., 1996]) : il n'existe pas de borne universelle sur la vitesse de convergence. L'objectif de la classification automatique est donc de trouver, pour un problème donné, la meilleure règle de classification possible.

1.2.1.2 Minimisation du risque empirique

Le problème est maintenant de choisir la règle de classification. Dans ce qui précède, l'évaluation de la concordance entre une prédiction $f(x)$ et la réalité y est quantifiée par la fonction d'erreur binaire ℓ_{0-1} définie par :

$$\forall (x, y) \in \mathcal{X} \times \mathcal{Y},\ \ell_{0-1}(x, y, f(x)) = \begin{cases} 0 & \text{si } y = f(x) \\ 1 & \text{sinon} \end{cases}$$

Cette fonction pose certains problèmes. En particulier, n'étant pas convexe, elle conduit à des problèmes d'optimisation difficiles (NP-complets). D'autres fonctions sont donc utilisées pour évaluer la concordance entre la réalité et la prédiction. Ces fonctions sont appelées **fonctions de perte** (*loss function*). Une fonction de perte est une fonction $\ell \in {\mathbb{R}_+}^{\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}}$ (ou parfois $\ell \in {\mathbb{R}_+}^{\mathcal{X}\times\mathcal{Y}\times\mathbb{R}}$).

De manière générale, le problème d'apprentissage consiste donc à minimiser le **risque** $R_\ell[f]$ (ou **fonctionnelle de risque**) défini comme l'espérance de la fonction de perte. Afin de minimiser la fonctionnelle de risque pour une distribution de (X, Y) inconnue, le principe d'induction

[3] Par opposition à l'apprentissage non-supervisé, où la connaissance est limitée à un échantillon de l'ensemble des observations sans connaître leur classe.

[4] Pour plus de précision sur la notion de consistance, le lecteur pourra se référer à l'ouvrage de Vapnik [1995].

suivant est généralement utilisé. Le risque $R_\ell[f]$ est remplacé par le **risque empirique**, $R_\ell^N[f]$, construit à l'aide de l'échantillon :

$$R_\ell^N[f] = \sum_{i=1}^{N} \ell(X_i, Y_i, f(X_i))$$

Le principe consiste à approcher la fonction f^* qui minimise le risque R_ℓ par la fonction $\hat{f}_N$ qui minimise le risque empirique R_ℓ^N. Ce principe d'induction est appelé **minimisation du risque empirique** (ERM pour *empirical risk minimization*). Cette idée d'obtenir une règle de classification en minimisant le risque empirique a été principalement développée par les travaux de Vapnik et de Chervonenkis (ex : [Vapnik, 1995]).

Mais lorsque l'on minimise le risque empirique, minimise-t-on le risque réel ? En d'autres termes, apprend-on des données ? Sans autres conditions, une règle obtenue en suivant le principe de l'ERM n'est en générale pas consistante : une fonction qui minimise le risque empirique ne minimise pas forcément le risque réel. On est alors dans un cas de **surapprentissage**. L'idée de Vapnik et Chervonenkis est de ne pas chercher la fonction de classification dans l'ensemble des fonctions mais de restreindre la recherche à un sous ensemble $\mathcal{F}$. L'ensemble de recherche $\mathcal{F}$ est appelé **ensemble d'hypothèses**.

Soit $\hat{f}_N$ une fonction de $\mathcal{F}$ qui minimise le risque empirique R_ℓ^N. Une telle fonction est appelée **estimateur ERM**. Remarquons qu'en notant R_ℓ^* le risque de Bayes, on a :

$$R_\ell[\hat{f}_N] - R_\ell^* = \underbrace{\left(R_\ell[\hat{f}_N] - \inf_{f\in\mathcal{F}} R_\ell[f]\right)}_{\text{erreur d'approximation (variance)}} + \underbrace{\left(\inf_{f\in\mathcal{F}} R_\ell[f] - R_\ell^*\right)}_{\text{erreur d'estimation (biais)}}$$

Intuitivement, si l'ensemble d'hypothèses $\mathcal{F}$ est très large, il y aura de fortes chances qu'il contienne la fonction de classification optimale : l'erreur d'estimation sera faible. En revanche, il est peu probable qu'une fonction minimisant le risque empirique ait un faible risque réel. Si par exemple $\mathcal{F}$ est l'ensemble des fonctions mesurables, il est toujours possible d'annuler le risque empirique en prenant la fonction de classification f définie par : $\forall i \in [1, N],\ f(X_i) = Y_i$ et f vaut 1 ailleurs. Dans ce cas, la fonction trouvée en suivant le principe de l'ERM risque de mal se généraliser. L'erreur d'approximation peut en effet s'avérer très importante. C'est un cas de surapprentissage. Inversement, si $\mathcal{F}$ est trop restreint, l'erreur d'approximation sera très faible mais il peut être impossible de trouver une bonne solution : l'erreur d'estimation sera grande. Ce compromis sur la taille ou richesse de $\mathcal{F}$ est connu en statistique classique comme le **compromis entre le biais et la variance**. Trois questions se posent alors :

1. Dans quel cas a-t-on les convergences suivantes en probabilité ?

$$\lim_{N\to\infty} R_\ell^N\left[\hat{f}_N\right] = \lim_{N\to\infty} R_\ell\left[\hat{f}_N\right] = \inf_{f\in\mathcal{F}} R_\ell[f] \tag{1.1}$$

2. Quelle est la vitesse de convergence de $R_\ell^N\left[\hat{f}_N\right]$?

3. Sait-on quantifier $\left(\inf_{f\in\mathcal{F}} R_\ell[f] - R_\ell^*\right)$?

Vapnik et Chervonenkis furent les premiers à formaliser cette idée pour la classification automatique en introduisant le concept de dimension de Vapnik-Chervonenkis ou VC-dimension (exemple : [Vapnik, 1995, 1999]). La VC-dimension de $\mathcal{F}$, notée $\dim_{\mathrm{VC}}(\mathcal{F})$, quantifie la notion de complexité de l'ensemble d'hypothèses. Elle mesure en quelque sorte la capacité de l'ensemble $\mathcal{F}$ à contenir une fonction séparant deux ensembles de points. Ils ont montré qu'une condition nécessaire et suffisante pour que les équations (1.1) soient vraies est la finitude de la VC-dimension (question 1). Vapnik emploie le terme de consistance en parlant de ce problème de convergence. Notons que ce n'est pas la consistance de la règle de classification puisque l'erreur limite n'est pas forcément l'erreur de Bayes (si $f^* \notin \mathcal{F}$). Vapnik a également donné une borne sur la vitesse de convergence (question 2) sous la forme d'une borne universelle[5] du risque réel [Vapnik, 1999]. Cette borne dépend du risque empirique, de $\dim_{\mathrm{VC}}(\mathcal{F})$ et de la taille N de l'échantillon.

Comment obtient-on une borne de convergence universelle alors qu'il n'existe pas de borne universelle de convergence pour une règle de classification ? Cette universalité est obtenue au prix de la restriction de l'ensemble d'hypothèses à l'ensemble $\mathcal{F}$. Sait-on quantifier ce que l'on perd en se restreignant à l'ensemble $\mathcal{F}$ (question 3) ? De manière générale, cela peut être très important comme le montre le théorème suivant [Benedek & Itai, 1994 ; Devroye et al., 1996].

Théorème 1.2.1 *Pour tout ensemble d'hypothèses $\mathcal{F}$ de VC-dimension finie, et pour tout $\epsilon > 0$, il existe une distribution de (X, Y) telle que :*

$$\inf_{f\in\mathcal{F}} R[f] - R^* > \frac{1}{2} - \epsilon \tag{1.2}$$

Il est possible d'augmenter la taille de $\mathcal{F}$ (VC-dimension infinie) tout en conservant la consistance et une borne universelle de convergence en suivant un autre principe d'induction, appelé minimisation du risque structurel ou SRM (*structural risk minimization* ou *complexity regularization*) [Vapnik, 1995, 1999]. Le SRM consiste, lorsque cela est possible, à considérer $\mathcal{F}$ comme la limite d'une suite croissante $(\mathcal{F}_i)$ de sous-ensembles de $\mathcal{F}$ de VC-dimension finie et à minimiser non pas l'erreur empirique mais une borne sur l'erreur réelle. Si la somme de la série de terme principal $e^{-\dim_{\mathrm{VC}}(\mathcal{F}_i)}$ est finie, le principe du SRM est consistant et il existe une borne universelle explicite de la vitesse de convergence [Vapnik, 1995, 1999 ; Devroye et al., 1996]. Ce principe permet de considérer des ensembles de fonctions beaucoup plus grands

[5] Universelle signifie indépendante de la distribution de (X, Y).

mais pas tous les ensembles de fonctions. En particulier, lorsque $\mathcal{X} = \mathbb{R}^d$, l'ensemble des fonctions de classification mesurables (Borel) ne peut pas s'écrire sous la forme d'une union dénombrable d'ensembles de VC-dimension finie [Benedek & Itai, 1994 ; Devroye et al., 1996].

Nous ne détaillerons pas dans ce manuscrit l'étude de la consistance des règles de classification ni leur vitesse de convergence. Le point clé à retenir des précédents paragraphes est le suivant : même sans connaître la distribution de (X, Y), **il est possible, à partir d'un échantillon de données, d'inférer une fonction de classification f qui se généralise correctement à l'ensemble des données $\mathcal{X}$, à condition que l'ensemble de recherche $\mathcal{F}$ soit adapté à la « taille » de $\mathcal{X}$ et de l'échantillon**.

Maintenant que nous avons vu ce qu'est l'apprentissage statistique et comment inférer ou plus simplement comment apprendre quelque chose à un classifieur à partir de données d'apprentissage, regardons les classifieurs les plus couramment utilisés obtenus en suivant ce principe : les classifieurs linéaires à large marge. Ces classifieurs séparent l'ensemble des données $\mathcal{X}$ en deux parties à l'aide d'une droite, d'un plan ou, de manière plus générale, d'un hyperplan ; ils sont appelés pour cette raison classifieurs linéaires. La séparatrice est choisie de telle sorte que la « distance » ou « marge » entre les données d'apprentissage de la classe 1 et celle de la classe -1 soit la plus grande possible. On dit pour cela qu'ils sont à large marge. Le problème est que, en général, l'ensemble $\mathcal{X}$ est quelconque et la notion d'hyperplan dans $\mathcal{X}$ n'existe pas. Pour cette raison, on passe par ce que l'on appelle des noyaux.

1.2.1.3 Les méthodes à noyaux

L'idée centrale des méthodes à noyaux est de représenter les données d'un ensemble $\mathcal{X}$ quelconque à l'aide d'une fonction $K \in \mathbb{R}^{\mathcal{X}\times\mathcal{X}}$ (cf. figure 1.6). Un élément $\mathbf{x} \in \mathcal{X}$ est alors représenté par la fonction $K(\mathbf{x}, \cdot)$; un ensemble de points $\{\mathbf{x}_i\}_i$ est représenté par la matrice $\big(K(\mathbf{x}_i, \mathbf{x}_j)\big)_{i,j}$ appelée **matrice de Gram**. La fonction K peut être considérée comme une « fonction de comparaison » ou une « mesure de similarité ».

Cette représentation facilite énormément les choses pour deux raisons : (i) lorsque K a certaines propriétés, cela revient à travailler dans un espace de Hilbert[6,7] (ce qui permet de considérer l'existence d'hyperplans...) [Aronszajn, 1950], et (ii) l'information « utile » d'un ensemble de données de N éléments est stockée dans une matrice de taille $N \times N$.

Les propriétés sur K permettant un lien avec les espaces de Hilbert font appel à la notion de noyau défini positif.

[6] *i.e.* Espace vectoriel muni d'un produit scalaire, complet pour la norme induite.
[7] Dans ce manuscrit, les espaces de Hilbert sont sur $\mathbb{R}$.

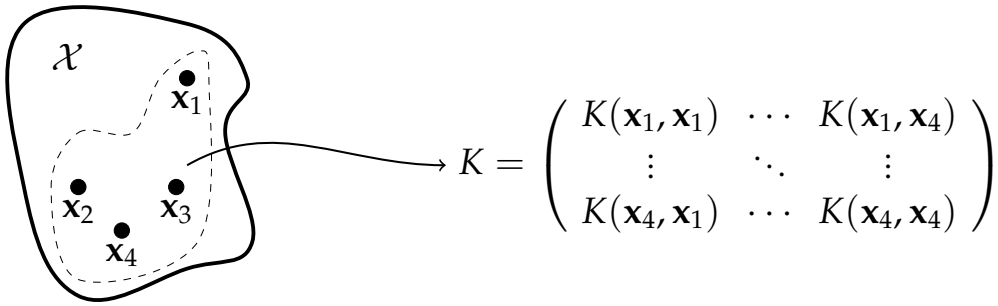


Figure 1.6 : Le noyau comme représentation des données (d'après de le cours de J.-P. Vert).

Définition 1.2.2 (Noyau défini positif) *Un **noyau défini positif** sur un ensemble $\mathcal{X}$ est une fonction $K \in \mathbb{R}^{\mathcal{X}\times\mathcal{X}}$ vérifiant les propriétés suivantes :*

(i) symétrie : $\forall(\mathbf{x}_1,\mathbf{x}_2) \in \mathcal{X}^2,\ K(\mathbf{x}_1,\mathbf{x}_2) = K(\mathbf{x}_2,\mathbf{x}_1)$

(ii) pour tout $N \in \mathbb{N}$ *et* $(\mathbf{x}_i)_{i\in[1,N]}$, *la matrice* $\left(K(\mathbf{x}_i,\mathbf{x}_j)\right)_{(i,j)\in[1,N]^2}$ *est semi-définie positive*

Aronszajn [1950] a montré que représenter les données avec un noyau défini positif revient à les envoyer dans un espace de Hilbert.

Théorème 1.2.3 (Moore-Aronszajn) *Si K est un noyau défini positif sur un ensemble $\mathcal{X}$, alors il existe un espace de Hilbert $\mathcal{H}$ et une fonction $\phi \in \mathcal{H}^{\mathcal{X}}$ tels que :*

$$\forall(\mathbf{x}_1,\mathbf{x}_2) \in \mathcal{X}^2,\ K(\mathbf{x}_1,\mathbf{x}_2) = \langle\phi(\mathbf{x}_1),\phi(\mathbf{x}_2)\rangle_{\mathcal{H}} \tag{1.3}$$

Réciproquement pour tout espace de Hilbert $\mathcal{H}$ et toute fonction $\phi \in \mathcal{H}^{\mathcal{X}}$, l'application $(\mathbf{x}_1,\mathbf{x}_2) \in \mathcal{X}^2 \mapsto \langle\phi(\mathbf{x}_1),\phi(\mathbf{x}_2)\rangle_{\mathcal{H}}$ est un noyau défini positif.

La preuve de ce théorème fait intervenir les espaces de Hilbert à noyau reproduisant (*reproducing kernel Hilbert space* - RKHS). Nous n'en avons pas besoin pour poursuivre. Le lecteur pourra se référer à [Aronszajn, 1950] pour plus de détails sur ces espaces.

Remarque : Un tel espace $\mathcal{H}$ est appelé **espace des *features***. Il n'y a unicité ni de ϕ, ni de $\mathcal{H}$; deux espaces de *features* peuvent même avoir des dimensions différentes. En revanche il y a unicité de la complétion de $\text{Vect}\left(\phi(\mathbf{x})\right)_{\mathbf{x}\in\mathcal{X}}$, l'espace vectoriel engendré par $\left(\phi(\mathbf{x})\right)_{\mathbf{x}\in\mathcal{X}}$, à une isométrie près.

Remarque : Choisir un noyau défini positif revient à envoyer les données dans un espace de Hilbert à l'aide d'une fonction ϕ. Lorsque l'on n'a besoin que du produit scalaire entre deux éléments de l'image de ϕ, comme :

$$\langle \phi(\mathbf{x}_1), \phi(\mathbf{x}_2) \rangle_{\mathcal{H}} = K(\mathbf{x}_1, \mathbf{x}_2),\ (\mathbf{x}_1, \mathbf{x}_2) \in \mathcal{X}^2$$

il est donc inutile de calculer $\phi(\mathbf{x}_i)$, ni même de connaître ϕ. On travaille alors dans l'espace des *features* de manière implicite ; c'est ce que l'on appelle l'**astuce du noyau** ou *kernel trick*.

Pour résumer, n'ayant pas forcément de « structure » sur l'espace $\mathcal{X}$, une possibilité est d'envoyer les observations de $\mathcal{X}$ dans un espace dans lequel il est facile de travailler, un espace de Hilbert $\mathcal{H}$, à l'aide d'une fonction ϕ. Pour faire cela il suffit de définir un noyau défini positif K sur $\mathcal{X}$.

Remarque : Si ϕ est injective, ce qui est le cas avec les noyaux classiques (linéaire, gaussien, polynomial [Schölkopf & Smola, 2001]), on ne perd pas d'information en changeant d'espace.

1.2.1.4 Application de l'ERM aux classifieurs à large marge

Regardons maintenant une catégorie de classifieurs appelés classifieurs à large marge. Ils seront entraînés en suivant le principe d'induction de l'ERM. Pour cela, on ne considère plus des fonctions de $\{\pm 1\}^{\mathcal{X}}$ mais des fonctions à valeurs dans $\mathbb{R}$. Par abus de langage, elles seront encore appelées fonctions de classification. Pour une fonction de classification $f \in \mathbb{R}^{\mathcal{X}}$ et une observation $\mathbf{x} \in \mathcal{X}$ données, la classe prédite est le signe de $f(\mathbf{x})$. La valeur absolue de $f(\mathbf{x})$ indique un degré de confiance dans le résultat. Soit y la classe de $\mathbf{x}$. On appelle **marge** de f pour le couple $(\mathbf{x}, y)$ le produit $yf(\mathbf{x})$. Un classifieur obtenu en minimisant le risque lorsque la fonction de perte est une fonction décroissante de la marge est appelé **classifieur à large marge**.

On suppose que $\mathcal{X}$ est doté d'un noyau défini positif K. Soient $\mathcal{H}$ un espace de Hilbert et $\phi \in \mathcal{H}^{\mathcal{X}}$ telle que : $\forall (\mathbf{x}_1, \mathbf{x}_2) \in \mathcal{X}^2, K(\mathbf{x}_1, \mathbf{x}_2) = \langle \phi(\mathbf{x}_1), \phi(\mathbf{x}_2) \rangle_{\mathcal{H}}$. L'existence de $\mathcal{H}$ et ϕ est donnée par le théorème 1.2.3. On se retrouve alors dans un espace dans lequel il est facile de travailler.

On cherche des fonctions de la forme $f = h \circ \phi$ avec $h \in \mathbb{R}^{\mathcal{H}}$. Plus particulièrement on se restreint à des fonctions h linéaires continues[8]. L'ensemble d'hypothèses est donc de la forme $\mathcal{F} = \mathcal{E} \circ \phi$ avec $\mathcal{E}$ un sous-ensemble de l'espace dual $\mathcal{H}^*$ de $\mathcal{H}$. Regardons ce que donne l'estimateur ERM $\hat{f}_N$. D'après ce que l'on a vu précédemment, cela dépend de la « capacité » de

[8] Donc aux fonctions affines de manière générale – si on remplace X par $(X, 1)$.

l'ensemble d'hypothèses $\mathcal{F}$. On a donc besoin d'une mesure de la « capacité » de l'ensemble d'hypothèses $\mathcal{F}$. Nous utiliserons la **complexité de Rademacher**, notée $\mathcal{R}$, [Bartlett et al., 2002]. C'est une mesure de la « capacité » d'un ensemble de fonctions par rapport à une distribution et à une taille d'échantillon. Si l'on suppose que la fonction de perte ℓ est k-lipschitzienne et que la perte est bornée de borne c, alors :

Proposition 1.2.4 *Pour tout $\delta > 0$, on a, avec probabilité au moins $1 - \delta$:*

$$R_\ell\left[\hat{f}_N\right] - R^* \leq \left(\inf_{f\in\mathcal{F}} R_\ell[f] - R^*_\ell\right) + 2k\mathcal{R}_{N,\phi(X)}(\mathcal{E}) + 2c\sqrt{\frac{\ln(1/\delta)}{2N}} \tag{1.4}$$

La démonstration est disponible à l'annexe A.

Par conséquent, si l'on suppose que $K(X, X)$ est borné[9] de borne κ^2 ($\kappa > 0$) et que

$$\mathcal{E} = \{h \in \mathcal{H}^* / \|h\|_{\mathcal{H}} \leq B\}$$

avec $B > 0$, on a $c \leq B\kappa^2$ et $\mathcal{R}_{N,\phi(X)}(\mathcal{E}) \leq \frac{2B\kappa}{\sqrt{N}}$. On a alors, à partir de l'inégalité (1.4) :

$$R_\ell\left[\hat{f}_N\right] - R^* \leq \underbrace{\frac{4kB\kappa}{\sqrt{N}} + 2B\kappa^2\sqrt{\frac{\ln(1/\delta)}{2N}}}_{\text{borne sur l'erreur d'approximation}} + \underbrace{\inf_{f\in\mathcal{F}} R_\ell[f] - R^*_\ell}_{\text{erreur d'estimation}} \tag{1.5}$$

Quand B augmente, l'erreur d'estimation $\left(\inf_{f\in\mathcal{F}} R[f] - R^*\right)$ diminue mais la borne sur l'erreur d'approximation augmente. Il faut donc trouver un compromis. Le meilleur compromis serait de choisir B de façon à minimiser la borne donnée par l'inégalité (1.5). Malheureusement, cela est généralement impossible. **En pratique, on estime donc la valeur optimale de B en testant plusieurs valeurs par validation croisée**.

Pour un ensemble d'apprentissage $\{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)\} \in (\mathcal{X}, \mathcal{Y})^N$, on résout donc le problème d'optimisation suivant :

$$\min_{h\in\mathcal{H}^*, \|h\|_{\mathcal{H}}\leq B} \frac{1}{N}\sum_{s=1}^{N} \ell\left(y_s h \circ \phi(\mathbf{x}_s)\right) \tag{1.6}$$

Lorsque la fonction ℓ est convexe, le problème d'estimation ERM (1.6) est équivalent à :

$$\min_{h\in\mathcal{H}^*} \underbrace{\frac{1}{N}\sum_{s=1}^{N} \ell\left(y_s h \circ \phi(\mathbf{x}_s)\right)}_{\text{erreur empirique}} + \underbrace{\lambda\|h\|^2_{\mathcal{H}}}_{\text{regularisation}} \tag{1.7}$$

[9] Ce sera toujours le cas pour nos données.

Le paramètre λ est le paramètre dual de B ; il est appelé **paramètre de régularisation**.

Pour le moment, nous avons fixé un espace $\mathcal{H}$ et une transformation ϕ. Nous allons voir que l'estimateur obtenu par la minimisation (1.6) ne dépend ni de ϕ, ni de $\mathcal{H}$.

Proposition 1.2.5 *Il existe* $(\alpha_1, \cdots, \alpha_N) \in \mathbb{R}^N$ *tel que la solution du problème d'optimisation 1.7 soit de la forme :*

$$\forall \mathbf{x} \in \mathcal{X},\ \hat{f}_N = \hat{h}_N \circ \phi(\mathbf{x}) = \sum_{s=1}^{N} \alpha_s K(\mathbf{x}_s, \mathbf{x}) \tag{1.8}$$

Démonstration À notre connaissance, ce théorème est en général démontré en utilisant le théorème du représentant. Or dans ce chapitre, nous travaillons dans un espace qui n'est pas nécessairement un RKHS. Nous en donnons donc la démonstration. Elle suit le même schéma que la démonstration du théorème du représentant (ex : [Schölkopf & Smola, 2001] ou le cours de Jean-Philippe Vert).

Soit $h \in \mathcal{H}^*$. D'après le théorème de représentation de Riesz, il existe $\mathbf{w} \in \mathcal{H}$ tel que : $\forall \mathbf{v} \in \mathcal{H},\ h(\mathbf{v}) = \langle \mathbf{w}, \mathbf{v} \rangle_{\mathcal{H}}$. L'espace Vect $(\{\phi(\mathbf{x}_s)\}_{s=1,\cdots,N})$ est un espace vectoriel fini. Il existe donc un unique couple $(\mathbf{w}_V, \mathbf{w}_\perp) \in$ Vect $(\{\phi(\mathbf{x}_s)\}_{s=1,\cdots,N}) \times$ Vect $(\{\phi(\mathbf{x}_s)\}_{s=1,\cdots,N})^\perp$ tel que : $\mathbf{w} = \mathbf{w}_V + \mathbf{w}_\perp$. Le paramètre λ est strictement positif donc $\mathbf{w}_\perp = 0$. Autrement dit, on a $\mathbf{w} \in$ Vect $(\{\phi(\mathbf{x}_s)\}_{s=1,\cdots,N})$. Soit $(\alpha_1, \cdots, \alpha_N) \in \mathbb{R}^N$ tel que : $\mathbf{w} = \sum_{s=1}^N \alpha_s \phi(\mathbf{x}_s)$. Par conséquent, on a : $\forall \mathbf{x} \in \mathcal{X},\ \hat{f}_N = \hat{h}_N \circ \phi(\mathbf{x}) = \sum_{s=1}^N \alpha_s \langle \phi(\mathbf{x}_s), \phi(\mathbf{x}) \rangle_{\mathcal{H}} = \sum_{s=1}^N \alpha_s K(\mathbf{x}_s, \mathbf{x})$.

■

Remarque : Si $\mathcal{H}$ est un espace de Hilbert à noyau reproduisant (RKHS), la proposition 1.2.5 est une conséquence directe du théorème du représentant [Kimeldorf & Wahba, 1971].

L'estimateur ERM obtenu ne dépend ni du choix de ϕ, ni de l'espace de Hilbert $\mathcal{H}$. **Le classifieur estimé dépend uniquement de la représentation des données d'apprentissage par le noyau K, c'est à dire de la matrice de Gram**. Le travail dans l'espace de Hilbert $\mathcal{H}$ est implicite. L'existence d'un espace des *features* sert uniquement à avoir un cadre théorique pour l'apprentissage.

1.2.2 Les machines à vecteur supports

Cette partie sur les SVM utilise le cours de Jean-Philippe Vert (MVA 2007) ainsi que [Schölkopf & Smola, 2001 ; Shawe-Taylor & Cristianini, 2000].

1.2.2.1 Un classifieur à large marge

Nous avons vu dans les paragraphes précédents que, lorsque les observations sont dans un espace $\mathcal{X}$ quelconque, représenter les données à l'aide d'un noyau défini positif revient à les envoyer dans un espace de Hilbert appelé espace des *features*. Il est alors possible d'utiliser

des classifieurs très simples : les classifieurs linéaires. Un point fort de cette approche est que, grâce à l'astuce du noyau, ce travail d'apprentissage dans l'espace des *features* reste implicite.

Pour que la fonction de classification obtenue se généralise bien, autrement dit pour que l'on apprenne des données, il faut que l'ensemble d'hypothèses soit restreint. Une manière de restreindre cet ensemble de recherche est de ne pas minimiser l'erreur empirique seule mais l'erreur empirique régularisée (cf. équation (1.7)).

Il nous reste maintenant à choisir la fonction de perte ℓ. Nous avons fait les hypothèses suivantes : la fonction de perte est une fonction décroissante de la marge, k-lipschitzienne et convexe. Une fonction d'erreur fréquemment utilisée est la fonction ***hinge loss*** ℓ_{hinge} définie par (cf. figure 1.7) :

$$\forall u \in \mathbb{R},\ \ell_{\text{hinge}}(u) = \max(0, 1-u)$$

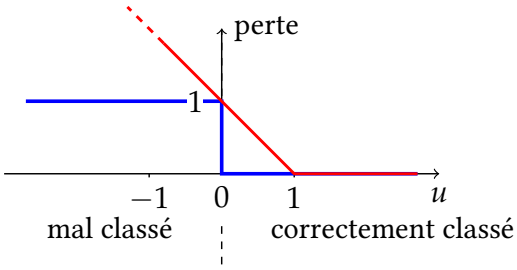


Figure 1.7 : En rouge, la fonction de perte ℓ_{hinge} appelée « *hinge loss* » et en bleu la fonction « erreur de classification » ℓ_{0-1}.

Les classifieurs obtenus avec cette fonction de perte sont appelés **machines à vecteurs supports** ou SVM (*support vector machine*). Les machines à vecteurs supports ont été introduites par Vapnik [1995] et Cortes & Vapnik [1995].

1.2.2.2 Différentes formulations du problème d'optimisation du SVM

Pour alléger les notations, nous supposerons que $\mathcal{X} = \mathcal{H}$ (espace de Hilbert). Le problème d'optimisation du SVM est donc :

$$\left(\mathbf{w}^{\text{opt}}, b^{\text{opt}}\right) = \underset{\mathbf{w}\in\mathcal{H}, b\in\mathbb{R}}{\arg\min} \frac{1}{N}\sum_{s=1}^{N} \ell_{\text{hinge}}\left(y_s\left[\langle \mathbf{w}, \mathbf{x}_s\rangle_{\mathcal{H}} + b\right]\right) + \lambda\|\mathbf{w}\|_{\mathcal{H}}^2 \tag{1.9}$$

Cette formulation permet de considérer les machines à vecteurs supports comme des machines qui minimisent l'erreur empirique régularisée. L'avantage de cette formulation est qu'elle distingue clairement la partie d'attache aux données de la partie régularisation ou autrement dit l'information *a priori*. C'est pour cela que nous utiliserons cette formulation (équation (1.9)) au chapitre 3.

Nous avons vu dans la section précédente que les SVM sont un cas particulier de classifieurs à large marge. Les classifieurs à large marge cherchent la séparatrice qui sépare au maximum les sujets de la classe de label 1 de l'ensemble d'apprentissage de ceux de la classe de label -1. En ce qui concerne le SVM, ceci vient du fait que la fonction de perte pénalise aussi les sujets bien classés lorsqu'ils se trouvent proches de l'hyperplan séparateur. En effet, pour un sujet s, même bien classé (*i.e.* tel que $y_s\left[\langle \mathbf{w}, \mathbf{x}_s\rangle_{\mathcal{H}} + b\right] > 0$), il y a pénalisation si $y_s\left[\langle \mathbf{w}, \mathbf{x}_s\rangle_{\mathcal{H}} + b\right] < 1$. On appelle **marge** du SVM la distance entre les hyperplans $\{\mathbf{x}|\langle \mathbf{w}, \mathbf{x}_s\rangle_{\mathcal{H}} + b = 1\}$ et $\{\mathbf{x}|\langle \mathbf{w}, \mathbf{x}_s\rangle_{\mathcal{H}} + b = -1\}$ (cf. figure 1.8).

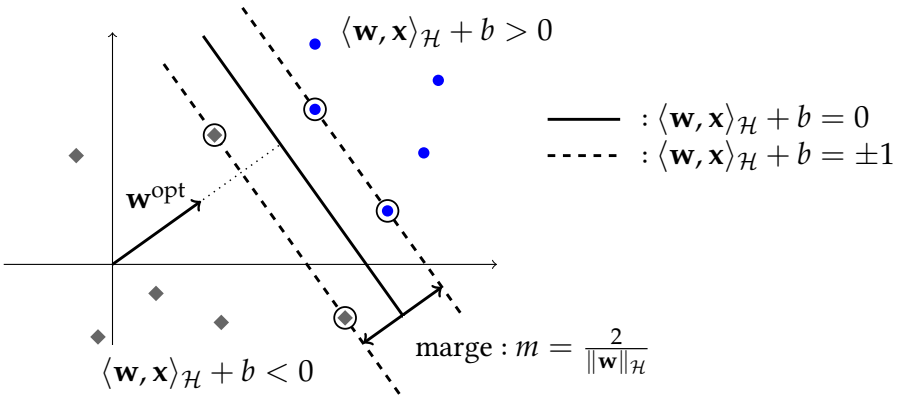


FIGURE 1.8 : Illustration d'un hyperplan séparateur obtenu avec un SVM linéaire. Les vecteurs supports sont entourés.

Une autre formulation du problème d'optimisation permet de mieux voir le SVM comme une machine que maximise la marge entre deux groupes. Le problème d'optimisation (1.9) est équivalent à :

$$\begin{aligned} \left(\mathbf{w}^{\text{opt}}, b^{\text{opt}}, \xi^{\text{opt}}\right) = \underset{\mathbf{w}\in\mathcal{H}, b\in\mathbb{R}, \xi\in\mathbb{R}^N}{\arg\min} \quad & \frac{1}{2}\|\mathbf{w}\|_{\mathcal{H}}^2 + C\sum_{s=1}^{N}\xi_s \\ \text{avec :} \quad & y_s[\langle \mathbf{w}, \mathbf{x}_s\rangle_{\mathcal{H}} + b] \geq 1 - \xi_s \\ & \xi_s \geq 0 \end{aligned} \tag{1.10}$$

avec $C = \frac{1}{2N\lambda}$. Le SVM maximise la marge ; or, comme la marge est proportionnelle à l'inverse de $\|\mathbf{w}\|_{\mathcal{H}}$, cela revient à minimiser $\|\mathbf{w}\|_{\mathcal{H}}^2$. Le rôle des variables ξ_s, appelées ***slack variables*** ou **variables ressort**, est de relâcher la contrainte $y_s[\langle \mathbf{w}, \mathbf{x}_s\rangle_{\mathcal{H}} + b] \geq 1$. Cela permet par exemple de gérer le cas d'un ensemble d'apprentissage non séparable. Cela permet également d'être plus robuste au bruit dans les données d'apprentissage (par exemple erreur de label). Le SVM cherche donc l'hyperplan qui sépare le mieux les données, autrement dit l'hyperplan de marge maximum, et pénalise les données mal classées à l'aide des variables ressort. La pénalisation par $\sum_{s=1}^{N}\xi_s$ revient à minimiser le nombre d'erreurs d'apprentissage. Plus précisément, minimiser le nombre d'erreurs d'apprentissage reviendrait à pénaliser par la

pseudo-norme ℓ_0 de ξ. L'inconvénient est qu'une telle pénalisation conduit à des problèmes d'optimisation NP-complets. On peut voir le terme $\sum_{s=1}^{N} \xi_s$ comme une relaxation convexe[10] à l'aide de la norme ℓ_1.

D'un point de vue algorithmique, le problème d'optimisation est une minimisation d'une fonction convexe quadratique sous contraintes linéaires. L'inconvénient majeur de ces formulations précédentes est qu'elles n'exploitent pas le fait que $\mathbf{w}^{\text{opt}}$ appartient à l'espace engendré par les données d'apprentissage. Pour cette raison, on utilise le formalisme lagrangien. Le problème d'optimisation dual est :

$$\begin{aligned} \alpha^{\text{opt}} = \underset{\alpha \in \mathbb{R}^N}{\arg\max} \quad & 2\alpha^T \mathbf{y} - \alpha^T K \alpha \\ \text{avec :} \quad & 0 \leq y_s \alpha_s \leq C \end{aligned} \tag{1.11}$$

avec $\mathbf{y} = (y_s)_s$ et K la matrice de Gram. On a alors :

$$\mathbf{w}^{\text{opt}} = \sum_{s=1}^{N} \alpha_s^{\text{opt}} \mathbf{x}_s$$

Il s'agit toujours d'un problème quadratique, mais la dimension est non plus la dimension de $\mathcal{X}$ mais le nombre de sujets de l'espace d'apprentissage. Dans nos études, la résolution du SVM est donc extrêmement rapide. Nous utiliserons libSVM [Chang & Lin, 2001] pour la résolution du SVM. La formulation (1.11) nous donne un autre point de vue plus géométrique du SVM. Les seules contraintes actives correspondent aux vecteurs $\mathbf{x}_s$ qui sont soit sur la marge, soit du mauvais côté de la marge (*i.e.* tel que $y_s \left[\langle \mathbf{w}, \mathbf{x}_s \rangle_{\mathcal{H}} + b\right] \leq 1$). Ces vecteurs sont appelés **vecteurs supports**. Leur poids dans la fonction de classification est compris entre 0 et C lorsqu'ils se situent sur la marge et est égal à C sinon. Quant aux autres vecteurs, leur poids est nul (cf. figure 1.9). Seules les observations situées à la frontière entre les deux classes interviennent dans la séparatrice optimale. Le SVM conduit donc à **une solution parcimonieuse**. Steinwart [2003] a montré que, lorsque λ est à la bonne échelle[11], le nombre de vecteurs supports est asymptotiquement équivalent à $2\eta N$ avec η la plus petite erreur d'apprentissage possible avec un classifieur linéaire.

1.2.2.3 Le paramètre de régularisation

En pratique, le paramètre de régularisation que l'on fixe n'est pas λ mais le paramètre $C = \dfrac{1}{2\lambda N}$. Pour cette raison, on parle parfois de C-SVM. L'interprétation de C n'est pas intuitive.

[10] Notons que la norme ℓ_1 permet également de prendre en compte la « quantité d'erreur », ce que ne fait pas la norme ℓ_0.

[11] $\lim_{n\to\infty} \lambda_N = 0$ et $\lim_{n\to\infty} N\lambda_N^3 = +\infty$.

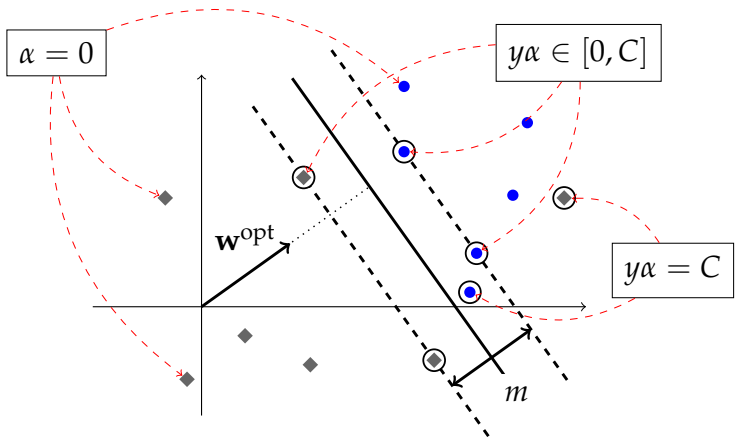


FIGURE 1.9 : Interprétation géométrique du SVM (d'après le cours de J.-P. Vert).

Remarquons tout de même que, plus C est important, plus le classifieur pénalise les erreurs d'apprentissage. En d'autres termes, plus C est grand, moins le classifieur est régularisé.

Ce paramètre est lié au rayon de la boule B qui intervient dans la borne (1.5) sur l'erreur d'approximation. On pourrait donc vouloir choisir la valeur de C qui minimise la borne (1.5). En pratique, cette minimisation n'est pas réalisable. On se contente donc de faire une recherche en grille et de choisir la valeur de C qui donne les meilleurs résultats (estimés par validation croisée).

1.2.2.4 SVM et séparatrices non linéaires

Les SVM sont historiquement les premières méthodes à noyaux utilisées pour la classification. Pour le moment nous n'avons parlé que de séparatrices linéaires puisqu'un SVM cherche une séparatrice linéaire dans l'espace des *features*. Cependant, la séparatrice trouvée n'est pas forcément linéaire dans l'espace $\mathcal{X}$. Grâce à l'astuce du noyau, le SVM permet de trouver des séparatrices non-linéaires dans l'espace $\mathcal{X}$ sans accroître la difficulté algorithmique.

Les noyaux non linéaires les plus couramment utilisés sont : le **noyau gaussien** et le **noyau polynomial**. Le noyau gaussien est défini par :

$$\forall (\mathbf{x}, \mathbf{x}') \in \mathcal{X}^2,\ K(\mathbf{x}, \mathbf{x}') = e^{-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma^2}}$$

avec σ un paramètre du noyau et $\|\cdot\|$ une norme (lorsqu'il en existe) sur $\mathcal{X}$. Quant au noyau polynomial (homogène) de degré d_0, si $\mathcal{X}$ admet un produit scalaire, il est défini par :

$$\forall (\mathbf{x}, \mathbf{x}') \in \mathcal{X}^2,\ K(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle_{\mathcal{X}}^{d_0}$$

1.2.2.5 Le choix du SVM linéaire

Dans la suite du manuscrit, nous travaillerons avec un SVM linéaire pour les raisons suivantes.

Gestion des grandes dimensions. Les problèmes de classification abordés dans cette thèse sont des problèmes de classification en grande dimension. Le SVM se comporte bien avec ce genre de problème en restreignant son espace de recherche à celui engendré par l'ensemble d'apprentissage.

Nous utiliserons des séparatrices linéaires pour deux raisons. L'utilisation d'un noyau non-linéaire augmente la taille de l'espace des *features* et augmente ainsi également le risque de surapprentissage. Les séparatrices linéaires sont parmi les séparatrices fréquemment utilisées les plus restrictives et donc les plus adaptées aux problèmes en grande dimension.

Parcimonie. Les problèmes de classification que nous abordons sont essentiellement des problèmes du type : patients *vs* témoins. Or un groupe de patients est souvent très hétérogène. En particulier il comprend généralement des individus qui sont à des stades différents de la maladie. La prise en compte de patients à des stades très avancés de la maladie apporte peu d'informations sur la discrimination entre les deux groupes. Elle risque principalement de décaler l'hyperplan séparateur du côté du demi-espace des patients et ainsi de baisser la sensibilité. Un des avantages du SVM est qu'il se focalise seulement sur les observations (sujets) situées à la « frontière »entre les deux classes.

1.3 Classification d'images IRM anatomiques

Dans cette thèse, nous nous intéressons au problème de classification d'IRM anatomiques. Il s'agit d'assigner à un sujet un diagnostic à partir de son image IRM. C'est un problème de classification en grande dimension : la dimension d des données est de l'ordre de $10^5 - 10^6$ alors que le nombre de sujets (ou d'observations) est de l'ordre de la centaine.

Une méthode de classification, que ce soit pour la classification d'images médicales ou non, peut se décomposer en trois étapes :

1. choix des *features* ou caractéristiques utilisées pour la classification
2. réduction de dimension de l'espace des *features*
3. choix du classifieur

Nous verrons ci-dessous les spécificités de ces trois étapes pour les méthodes de classification d'IRM anatomiques.

1.3.1 Choix des caractéristiques utilisées pour la classification

La grande majorité des méthodes de classification d'IRM anatomiques utilise des caractéristiques calculées au niveau du voxel (probabilité de matière grise, anisotropie fractionnelle,...). Les images sont ensuite recalées dans un espace commun à tous les sujets, de telle sorte que, dans l'idéal, le i-ème voxel d'un sujet corresponde au i-ème voxel des autres sujets [Caan et al., 2006 ; Lao et al., 2004 ; Teipel et al., 2007 ; Duchesne et al., 2006, 2008 ; Magnin et al., 2009 ; Vemuri et al., 2008 ; Klöppel et al., 2008b,a ; Hinrichs et al., 2009 ; Fan et al., 2005, 2007, 2008b,a ; Davatzikos et al., 2008a,b ; Misra et al., 2009].

D'autres méthodes, basées sur le même principe, utilisent comme caractéristiques des surfaces colorées (ex : épaisseur corticale). Ces surfaces sont également recalées dans un espace commun à tous les sujets [Querbes et al., 2009 ; Desikan et al., 2009]. Duchesnay et al. [2007] présentent une alternative originale à l'analyse de type voxel-à-voxel (ou équivalent cortical) permettant une analyse du cerveau entier. Elle est basée sur des descripteurs morphométriques des sillons corticaux.

Des méthodes n'utilisent pas l'ensemble du cerveau pour leur analyse mais se focalisent sur certaines structures du cerveau telles que les hippocampes. Elles étudient soit leur volume [Colliot et al., 2008a ; Chupin et al., 2009b,a], soit leur forme à l'aide de descripteurs de formes [Golland et al., 2005 ; Gerardin et al., 2009].

1.3.2 Réduction de dimension

Une fois l'ensemble des caractéristiques choisi, toutes les méthodes, mises à part celles de Klöppel et al. [2008b,a], utilisent au moins une étape de réduction de dimension. L'objectif de la réduction de dimension est de rendre la méthode plus robuste au surapprentissage. Il est en effet difficile de faire de l'inférence lorsque le nombre de variables est grand par rapport au nombre d'expériences, autrement dit de sujets. On appelle cela la malédiction de la dimension ou *curse of dimensionality*. Pour réduire la dimension, deux approches existent :

- l'**extraction de *features***;
- la **sélection de *features***.

L'objectif de l'**extraction de *features*** est de transformer les *features* en une représentation plus condensée contenant la même information discriminante. On travaille ici avec des images. L'approche la plus simple pour réduire la dimension est donc de **diminuer la résolution de l'image** en lissant et en sous-échantillonnant [Lao et al., 2004 ; Vemuri et al., 2008]. Une autre

approche utilisant la structure de l'image est de faire un changement de base à l'aide d'une **transformée en ondelettes** et de coupler ce changement avec une étape de sélection de *features* [Lao et al., 2004]. Sur le même principe de changement de base, une technique fréquemment utilisée en traitement d'images est l'**analyse en composantes principales** [Turk & Pentland, 2002]. Elle est également fréquemment utilisée à cet effet en neuroimagerie [Caan et al., 2006 ; Teipel et al., 2007 ; Duchesne et al., 2006, 2008]. Une autre méthode consiste à regrouper les voxels en régions anatomiques à l'aide d'un **atlas**. Les images ou surfaces colorées étant normalisées dans un espace standard, il est possible de les parcelliser en R régions à l'aide d'un atlas préalablement défini dans le même espace. Pour chaque sujet, son vecteur des *features* est alors remplacé par un vecteur de taille R où la r-ème composante est une fonction (en général la somme ou la moyenne) des intensités des voxels de la région r. Cette technique est utilisée dans [Lao et al., 2004 ; Magnin et al., 2009 ; Querbes et al., 2009 ; Desikan et al., 2009 ; Ye et al., 2008]. L'inconvénient majeur de cette approche est que l'atlas utilisé n'est en général pas spécifique à la pathologie étudiée : les frontières des régions atteintes ne correspondent pas forcément à celles de l'atlas. Pour cette raison, Fan et al. [2007] ont proposé d'utiliser une **parcellisation du cerveau adaptée à la pathologie** au lieu d'un atlas prédéfini. Pour cela, ils effectuent une étude de groupe puis un *clustering* sur les cartes de corrélations ainsi obtenues afin d'obtenir des régions homogènes en termes de discrimination. Cette dernière approche a été utilisée dans de nombreuses études : [Fan et al., 2008b,a ; Davatzikos et al., 2008a,b ; Misra et al., 2009].

La **sélection de *features*** réduit la dimension en sélectionnant le sous-ensemble des *features* le plus discriminant possible. Il existe de nombreuses techniques de sélection de *features*. La grande majorité des méthodes proposées en neuroimagerie utilise une **méthode univariée** [Lao et al., 2004 ; Gerardin et al., 2009 ; Fan et al., 2005, 2007, 2008b,a ; Davatzikos et al., 2008a,b ; Misra et al., 2009 ; Hinrichs et al., 2009]. Le principe est le suivant. A l'aide d'un test univarié tel que le test de Student, on obtient en chaque voxel un **score de « corrélation »** avec la variable étudiée (la pathologie). Les *features* sont alors sélectionnés par seuillage (score supérieur au seuil) ou en sélectionnant les n *features* les plus corrélés à la pathologie. Les deux inconvénients majeurs de cette approche sont (i) les *features* sont testés indépendamment les uns des autres et (ii) cette approche n'est en général pas spécifique au classifieur utilisé. La conséquence de la première critique est que l'ensemble des *features* sélectionnés peut contenir des *features* fortement redondants. Pour des classifieurs tels que le SVM qui travaille dans l'espace engendré par l'ensemble d'apprentissage cette redondance importe peu en pratique. En revanche pour des classifieurs plus classiques, il pourrait être plus adapté de faire une première étape d'extraction des données avec une analyse en composantes indépendantes (ICA) afin d'avoir des *features* indépendants. Une autre possibilité est d'utiliser des méthodes de *ranking* qui évaluent tous les *features* en même temps. Ingalhalikar et al. [2010] utilisent le *minimum redundancy maximum*

relevance (mRMR). Cependant, la deuxième critique reste valable : ce genre d'approches n'est pas spécifique au classifieur. Une deuxième catégorie de méthodes de sélection de *features* (souvent appelées *wrapper*) **utilise le classifieur pour sélectionner les *features***. Suivant la même idée d'attribution de score et de seuillage, [Vemuri et al., 2008] utilisent un SVM linéaire et prennent comme scores les coefficients de l'hyperplan séparateur optimal. Quant à Desikan et al. [2009], ils effectuent pour chaque *feature* un test de classification (à l'aide de cette seule variable) et lui attribuent comme score l'aire sous la courbe ROC. Mais l'idéal en sélection de *features* est d'obtenir le sous-ensemble le plus discriminant. C'est ce que cherchent à faire plus directement les autres méthodes. [Querbes et al., 2009] évaluent par validation croisée le pouvoir discriminant de tous les sous-ensembles possibles de *features* et prennent celui qui donne le meilleur résultat. En général, le nombre de *features* est trop grand pour pouvoir évaluer tous les sous-ensembles possibles. Des approches gloutonnes (greedy) appelées ***stepwise*** sont souvent utilisées. L'ensemble des *features* sélectionnés est obtenu par élimination ou agrégation récursive des *features* à partir de l'ensemble des *features* (élimination) ou de l'ensemble vide (agrégation). Cette approche est utilisée non seulement avec des classifieurs classiques [Freeborough & Fox, 1998] mais également avec des classifieurs tels que le SVM à l'aide du **SVM-RFE** (*recursive features elimination*) de Guyon et al. [2002] [Fan et al., 2005, 2007, 2008b,a ; Davatzikos et al., 2008a,b ; Misra et al., 2009 ; Ingalhalikar et al., 2010]. Pour de plus amples détails sur la réduction de dimension, le lecteur pourra se référer à [Guyon & Elisseeff, 2003].

Cette approche de réduction de dimension est parfois incorporée au classifieur. Nous n'avons pas rencontré ce cas en neuroimagerie. Nous en reparlerons dans la discussion du chapitre 3.

1.3.3 Choix du classifieur

Une fois l'ensemble des caractéristiques utilisées pour la classification fixé, la dernière étape est le choix du classifieur.

Lorsque le nombre de *features* est beaucoup plus faible que le nombre de sujets (au moins un rapport 10), il est possible d'utiliser des **méthodes de classification issues des statistiques classiques**. Dans le cas multivarié, on rencontre généralement l'emploi de l'**analyse linéaire discriminante** (LDA) [Caan et al., 2006 ; Duchesne et al., 2006, 2008 ; Querbes et al., 2009] ou son équivalent quadratique, la **QDA** [Duchesne et al., 2008]. La **régression logistique** est également souvent utilisée [Teipel et al., 2007 ; Desikan et al., 2009].

Quand le nombre de variables est au moins de l'ordre du nombre de sujets, ce qui est le plus souvent le cas, il est nécessaire d'utiliser des méthodes issues des statistiques modernes capables de gérer les problèmes de grande dimension. La grande différence avec les méthodes évoquées dans le paragraphe précédent est qu'elles minimisent non plus une erreur empirique

mais une erreur empirique régularisée. Notons que la plupart des méthodes présentées dans le paragraphe précédent ont des équivalents régularisés. La grande majorité des approches utilise des **SVM** [Lao et al., 2004; Magnin et al., 2009; Gerardin et al., 2009; Vemuri et al., 2008; Klöppel et al., 2008b,a; Fan et al., 2005, 2007, 2008b,a; Davatzikos et al., 2008b,a; Misra et al., 2009; Golland et al., 2005; Ingalhalikar et al., 2010]. Hinrichs et al. [2009] utilisent quant à eux une approche de type ***boosting***. Plus précisément, ils utilisent le *LPBoost* de Demiriz et al. [2002]. L'idée du *boosting* est de créer un classifieur sous la forme d'une combinaison linéaire de H classifieurs $(h_i)_i$ appelés *weak classifiers* qui, pris individuellement, peuvent avoir de faibles pouvoirs prédictifs. Cela revient à minimiser :

$$\begin{array}{rl} \min\limits_{\alpha\in\mathbb{R}^N,\xi\in\mathbb{R}^N} & \sum\limits_{i=1}^{H}\alpha_i + C\sum\limits_{s=1}^{N}\xi_s \\ \text{avec :} & y_s\sum_{i=1}^{H}\alpha_i h_i(\mathbf{x}_s) \geq 1-\xi_s \\ & \xi_s \geq 0 \end{array} \tag{1.12}$$

Hinrichs et al. [2009] ont choisi comme *weak classifier* h_i un seuil sur l'intensité du voxel i. Notons que cette méthode correspond également à un classifieur à large marge. Le choix des *weak classifiers* d'Hinrichs et al. [2009] est tel que la fonction de classification trouvée peut s'écrire comme le signe d'une combinaison linéaire de l'intensité des voxels. La seule différence avec le SVM est que la régularisation est sous la forme d'une pénalisation ℓ_1 et non quadratique. Une telle régularisation entraîne de la parcimonie. En réalité Hinrichs et al. [2009] n'utilisent pas le *LPBoost* tel quel, ils introduisent une pénalisation qui force les classifieurs correspondant à des voxels voisins à avoir le même poids.

1.3.4 Limites et améliorations possibles

La question centrale qui se pose est la suivante. Ces méthodes de classification sont-elles adaptées à notre problème ? Nous abordons ici à un problème de classification en grande dimension d'images cérébrales pour l'aide au diagnostic de pathologies.

Grandes dimensions. La dimension de l'espace des *features* est très grande puisqu'elle correspond au nombre de voxels d'une image anatomique ou au nombre de nœuds d'un maillage cortical ($\sim 10^5 - 10^6$). L'approche classique consiste à réduire la dimension des données par une ou des étapes de sélection de *features* ou d'extraction de *features*. C'est l'approche utilisée dans l'ensemble des méthodes présentées.

La limite majeure de cette approche est qu'elle n'est en général pas spécifique au classifieur utilisé, en particulier pour les méthodes de type extraction de *features*. Certaines méthodes de sélection de *features* essaient de surmonter cette limite en cherchant le sous-ensemble le plus discriminant pour un classifieur donné [Querbes et al., 2009]. Ce type de méthodes présente

principalement deux inconvénients. Premièrement, dès que le nombre de *features* devient important, cette procédure est trop lourde. Des stratégies gloutonnes et donc généralement sous-optimales sont alors employées. Deuxièmement, le nombre d'observations en imagerie médicale étant faible, la sélection de *features* est souvent effectuée sur l'ensemble d'apprentissage. La conséquence est que leurs résultats sont plus sensibles à l'ensemble d'apprentissage choisi. En effet, la sélection de *features* peut être considérée comme une étape d'apprentissage. Cette étape augmente la taille de l'ensemble des fonctions de classification possibles et ainsi également le risque de surapprentissage. Une manière de réduire cet effet est d'utiliser une procédure de type *bootstrap* [Hastie et al., 2005].

L'alternative que nous proposons dans cette thèse est d'intégrer directement cette étape de réduction de dimensions dans un classifieur (capable de gérer les grandes dimensions tel que le SVM) sous forme de contraintes, en le régularisant.

Images cérébrales. Les observations que nous cherchons à classer sont des images anatomiques de cerveaux (ou équivalents surfaciques). Les méthodes de classification doivent donc exploiter la structure de l'image et l'anatomie du cerveau. En d'autres termes, il existe une régularité entre les *features* qui peut-être vue comme une covariance. Elle est due au fait que les observations sont des images de cerveaux pour une pathologie donnée. Cette régularité peut être de plusieurs natures. Au niveau le plus simple, elle peut être liée à la proximité spatiale des *features*, mais elle peut être aussi liée à une connectivité sous-jacente de nature anatomique (ex : architecture des réseaux de fibres) ou de nature fonctionnelle (ex : synchronies cérébrales ou corrélation du signal IRMf). Exploiter ces structures cérébrales sous-jacentes revient donc à définir une notion de proximité entre les *features* et à forcer le classifieur à considérer les *features* « proches » comme similaires.

Dans les approches vues précédemment, la structure de l'image est généralement exploitée indirectement via une étape préliminaire de lissage ou de décomposition en ondelettes. L'inconvénient majeur est qu'elle ne prend pas en compte l'anatomie du cerveau. Une autre approche fréquemment utilisée, qui prend en compte la structure de l'image et aussi l'anatomie du cerveau, est la parcellisation du cerveau à l'aide d'un atlas préalablement défini dans un espace standard. Néanmoins lorsque les voxels sont agrégés, l'information individuelle de chaque voxel est perdue. Il semblerait donc plus adapté d'introduire directement une notion de proximité spatiale et anatomique dans le classifieur plutôt que de procéder à une étape de réduction de dimension.

À notre connaissance, seuls Hinrichs et al. [2009] ont proposé ce genre d'approches, mais ils n'utilisent que la structure de l'image et non l'anatomie. Ils considèrent que deux voxels sont voisins si et seulement si ils sont voisins dans l'image. De plus, comme nous l'avons vu, leur approche utilise une pénalisation ℓ_1. Or, une telle pénalisation est très lourde computationnellement et ils sont obligés de faire une étape de sélection de *features* pour ne

garder que 1% des voxels. De plus une telle pénalisation entraîne de la parcimonie ; autrement dit, cela fait de la sélection de *features*. Or, la principale pathologie à laquelle nous nous sommes intéressés durant cette thèse est la maladie d'Alzheimer. Dans cette maladie, les lésions sont distribuées sur plusieurs régions. Nous préférons donc la pénalisation quadratique qui agit comme un lissage des données et non comme une sélection de variables. **Nous utiliserons donc un SVM qui, grâce au noyau, ne pose pas de problème computationnel et qui, de plus, utilise une régularisation quadratique**.

1.4 Objectifs de la thèse

L'analyse computationnelle en neuroimagerie a pour but une meilleure compréhension des pathologies cérébrales afin de permettre la mise en place, au niveau individuel, de stratégies de prévention, de détection et de suivi thérapeutique.

La grande majorité des approches en neuroimagerie computationnelle sont des études de groupes basées sur des analyses univariées de masse, voxel-à-voxel. Cependant, ces approches univariées (i) apportent peu d'information au niveau individuel et (ii) ont une sensibilité limitée quand les différences étudiées mettent en jeu différentes structures du cerveau. Les approches de classification en haute dimension tels que les SVM sont une alternative pour dépasser les limites de ces analyses.

Dans une méthode de classification, différentes stratégies sont possibles pour la définition des *features*, la réduction de la dimension ou encore la classification proprement dite. Certaines stratégies ont été proposées et testées dans la littérature. Elles ont toutefois été évaluées sur des populations d'étude différentes et/ou des problèmes de classification différents, ce qui rend difficile toute comparaison directe. Pour cette raison, une première partie (chapitre 2) sera consacrée à la **comparaison de différentes stratégies de classification** automatique de patients Alzheimer à partir de caractéristiques anatomiques (substance grise, épaisseur du cortex, hippocampe) sur une grande base de données, la base ADNI.

Les problèmes de classification abordés en neuroimagerie sont des problèmes de classification d'images en grande dimension. Les méthodes proposées abordent généralement ce problème à l'aide d'une étape de réduction de dimension préliminaire à la classification. Cependant, cette étape n'est pas toujours spécifique au classifieur utilisé. Il nous semble plus adapté de remplacer cette étape par des **contraintes supplémentaires directement introduites dans le classifieur**. Nous proposerons donc au chapitre 3 un cadre général pour l'introduction de

telles contraintes dans un SVM. Nous nous attacherons également à **définir des contraintes qui prennent en compte la structure spatiale et anatomique du cerveau**.

Les objectifs de la thèse sont donc :

1. Évaluer différentes stratégies de classification sur un ensemble de données communes.
2. Proposer un cadre général d'introduction de contraintes spatiales et anatomiques dans un SVM.
3. Définir des contraintes adaptées respectant la structure spatiale et anatomique des IRM.
4. Évaluer cette méthode sur des applications cliniques : la maladie d'Alzheimer et les accidents vasculaires cérébraux.

Chapitre 2

Évaluation de stratégies de classification sur une grande base d'images cérébrales

Différents choix président à la construction d'une méthode de classification : définition des caractéristiques (volumiques, surfaciques...), réduction de la dimension, choix d'un classifieur. Récemment, plusieurs méthodes de classification ont été proposées pour détecter automatiquement les patients atteints de la maladie d'Alzheimer ou atteints de troubles cognitifs léger à partir d'IRM pondérées en T_1. Cependant, ces méthodes ont été évaluées sur différentes populations, ce qui rend leur comparaison difficile. Dans ce chapitre, nous évaluons les performances de différentes stratégies pour le problème de classification automatique dans la maladie d'Alzheimer. Trois de types de *features* sont utilisés : la concentration de tissus dans des approches voxel-à-voxel, l'épaisseur corticale dans des approches surfaciques et le volume ou la forme de l'hippocampe dans des approches locales. Différentes méthodes de sélection ou de réduction de *features* sont évaluées. L'évaluation est faite à partir d'images IRM anatomiques de 509 sujets de la base de données ADNI.

Les résultats présentés dans ce chapitres ont été obtenus en collaboration avec Marie Chupin (segmentation de l'hippocampe), Émilie Gerardin (morphométrie de l'hippocampe) et Jérôme Tessieras (épaisseur corticale).

Ce chapitre est organisé de la façon suivante. Après une introduction sur la maladie d'Alzheimer et l'apport de l'imagerie anatomique (section 2.1), nous décrirons la base de données utilisée pour la comparaison (section 2.2) puis les méthodes de classification évaluées (section 2.3). Les sections 2.4 et 2.5 présentent respectivement les expériences de classification et les résultats obtenus. Ces résultats sont ensuite discutés dans la section 2.6.

2.1 Contexte : aide au diagnostic de la maladie d'Alzheimer

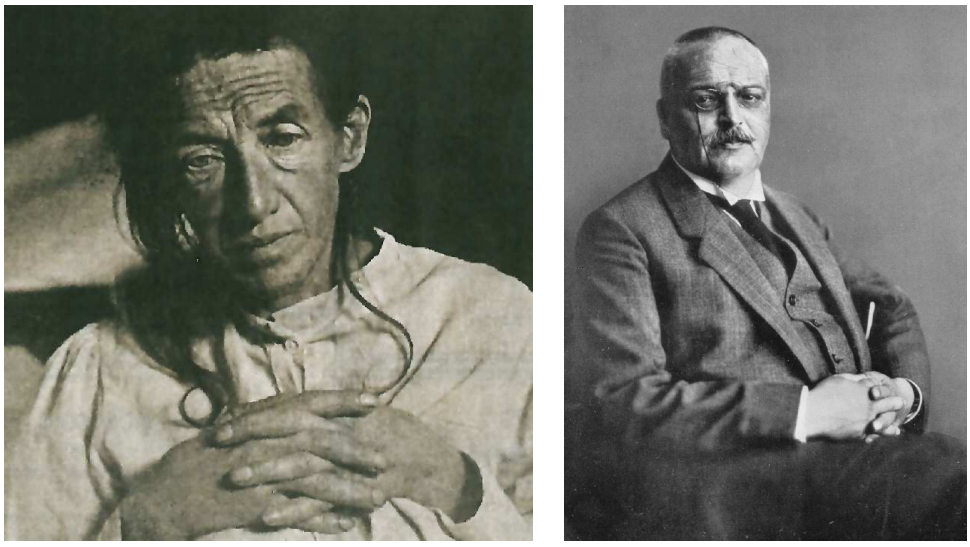
2.1.1 La maladie d'Alzheimer

2.1.1.1 Une maladie découverte depuis plus d'un siècle.

Ce résumé de l'historique de la maladie d'Alzheimer est basé sur l'article de Maurer et al. [1997].

Auguste D. Le 26 novembre 1901, Auguste D. (figure 2.1) fut admise à l'hôpital de Francfort. Elle avait de nombreux symptômes : compréhension diminuée, mémoire diminuée, aphasie, perte du sens de l'orientation... Elle fut prise en charge dès son arrivée par un neuropathologiste allemand, Alois Alzheimer. Elle mourût d'une septicémie le 8 avril 1903.

Alois Alzheimer. Alois Alzheimer (figure 2.1) débuta sa carrière de médecin à l'hôpital des malades mentaux et épileptiques de Francfort. Franck Nissl le rejoignit à Francfort et collabora avec lui sur la neuropathologie des troubles de la démence. Franck Nissl mis notamment au point des techniques de coloration particulièrement adaptées à l'étude histologique des pathologies nerveuses.



(a) *(b)*

Figure 2.1 : *(a)* Auguste D. première malade diagnostiquée ; *(b)* Alois Alzheimer.

Alois Alzheimer quitta Francfort en 1903 et alla s'installer à la clinique psychiatrique royale de Munich sous la direction d'Emil Kraeplin, le fondateur de ce qui deviendra par la suite le *Max-Planck Institut für Psychiatrie*. Franck Nissl avait déjà rejoint Kraeplin en 1895. Il continua cependant à suivre sa patiente. À la mort d'Auguste D., Alois Alzheimer analysa les aspects histopathologiques de la maladie dont souffrait Auguste D. L'autopsie révéla des **plaques séniles** et une forme de dégénérescence neuronale particulière : des **enchevêtrements neurofibrillaires** (figure 2.2). Elles restent les seuls éléments permettant le diagnostic de certitude [Amieva et al., 2007].

FIGURE 2.2 : Dessins originaux d'Alois Alzheimer : enchevêtrements neurofibrillaires (image extraite de [Maurer et al., 1997]).

La maladie d'Alzheimer. Lors de la 37-ème conférence des psychiatres allemands à Tübingen, Alois Alzheimer exposa ses observations d'un nouveau type de démence. Ses résultats furent publiés, en 1907, dans un article intitulé *Une maladie caractéristique grave du cortex cérébral*. Le nom de maladie d'Alzheimer fut mentionné pour la première fois par Emil Kraeplin dans la huitième édition du *Handbook of Psychiatry* (1910).

D'autres psychiatres comme F. Bonfiglio, O. Fisher et Perusini étudièrent au début des années 1900 des patients présentant des symptômes semblables à Auguste D. Quant à Perusini, son premier cas d'étude fut également Auguste D.

2.1.1.2 Qui touche de plus en plus de personnes

Les démences neurodégénératives représentent les formes de démences les plus fréquentes. Plus particulièrement, de 60% à 80% [Ott et al., 1995 ; Ramaroson et al., 2003 ; Kalaria et al., 2008] des démences dans le monde seraient relatives à la maladie d'Alzheimer.

Les démences. De nombreuses études (ADNI, PAQUID, 3 Cités, EURODEM, Rotterdam, Nun,...) ont permis d'obtenir des informations sur l'épidémiologie des démences, notamment sur leur prévalence. La **prévalence** désigne la proportion d'individus atteints d'une affection dans une population à un moment donné. En 2005, Ferri et al. [2006] ont réalisé une estimation de la prévalence des démences à travers le monde, pour Alzheimer Disease International (ADI, http ://www.alz.co.uk), une fédération de 71 associations en relation avec l'Organisation Mondiale de la Santé OMS. Les résultats sont rapportés dans le tableau *T*ab. 2.1 par secteur géographique et par tranche d'âge. Au total, Ferri et al. [2006] estiment que 3.9% des plus de 60 ans sont atteints de démences. Cette proportion correspond à 24 millions d'individus (en 2001).

Il faut toutefois considérer ces résultats avec une grande prudence. Il y a un très grand déséquilibre dans la répartition géographique des études épidémiologiques (*F*ig. 2.3) et donc une très grande variabilité dans la fiabilité des résultats. L'Amérique du Nord, l'Europe, le Japon et l'Australie sont les seules régions où les estimations sont fiables. Néanmoins, des études plus récentes dans des pays en voie de développement (et en Chine) rapportent des estimations du même ordre de grandeur [Kalaria et al., 2008].

En ce qui concerne la France, la prévalence serait de 17.8% chez les sujets âgés de plus de 75 ans pour la période 1998/1999 [Ramaroson et al., 2003] (étude PAQUID). Cela représenterait 769 000 individus.

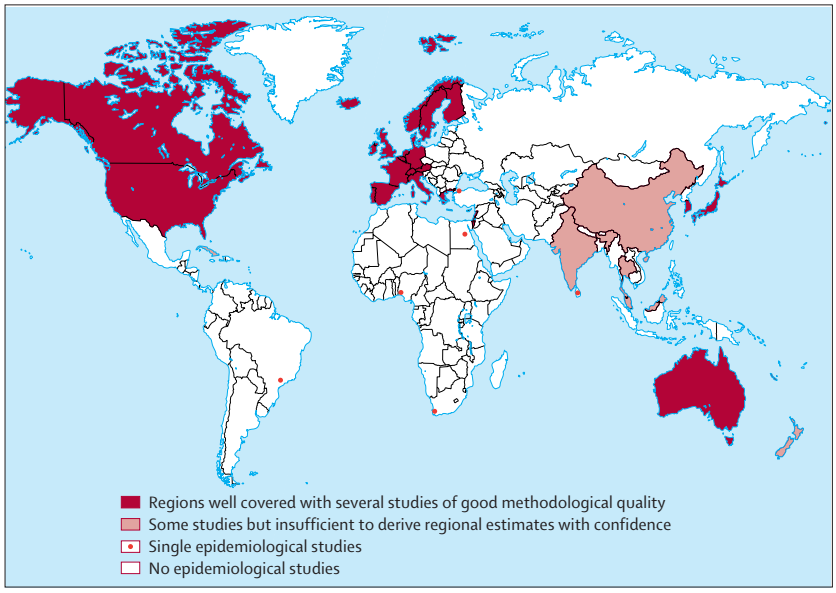


Figure 2.3 : Répartition des études épidémiologiques permettant une estimation de la prévalence de la maladie d'Alzheimer. Carte extraite de [Ferri et al., 2006].

TABLE 2.1 : Prévalence (moyenne et écart type en %) des démences dans le monde en 2005 selon ADI [Ferri et al., 2006] par groupe d'âge (en année). Les régions sont séparées selon le taux de mortalité noté de A (plus faible taux) à E (plus fort taux).

		60-64	65-69	70-74	75-79	80-84	≥85
Europe	A	0.9 (0.1)	1.5 (0.2)	3.6 (0.2)	6.0 (0.2)	12.2 (0.8)	24.8 (1.0)
	B	0.9 (0.1)	1.3 (0.1)	3.2 (0.3)	5.8 (0.3)	12.2 (0.3)	24.7 (2.3)
	C	0.9 (0.1)	1.3 (0.1)	3.2 (0.2)	5.8 (0.2)	11.8 (0.5)	24.5 (1.8)
les Amériques	A	0.8 (0.1)	1.7 (0.1)	3.3 (0.3)	6.5 (0.5)	12.8 (0.5)	30.1 (1.1)
	B	0.8 (0.1)	1.7 (0.1)	3.4 (0.2)	7.6 (0.4)	14.8 (0.6)	33.2 (3.5)
	D	0.7 (0.1)	1.5 (0.3)	2.8 (0.4)	6.2 (1.1)	11.1 (2.0)	28.1 (5.2)
Afrique du Nord et Moyen Orient	B	0.9 (0.3)	1.8 (0.1)	3.5 (0.3)	6.6 (0.2)	13.6 (0.8)	25.5 (2.3)
	D	1.2 (0.3)	1.9 (0.2)	3.9 (0.3)	6.6 (0.4)	13.9 (1.3)	23.5 (2.3)
Pacifique	A	0.6 (0.1)	1.4 (0.1)	2.6 (0.3)	4.7 (0.6)	10.4 (1.2)	22.1 (3.5)
	B	0.6 (0.1)	1.8 (0.2)	3.7 (0.4)	7.0 (0.9)	14.4 (1.9)	26.2 (3.9)
Asie du Sud	B	1.0 (0.1)	1.7 (0.2)	3.4 (0.2)	5.7 (0.5)	10.8 (1.2)	17.6 (2.7)
	D	0.4 (0.1)	0.9 (0.1)	1.8 (0.2)	3.7 (0.4)	7.2 (1.2)	14.4 (2.7)
Afrique	D	0.3 (0.1)	0.6 (0.1)	1.3 (0.2)	2.3 (0.5)	4.3 (1.0)	9.7 (1.9)
	E	0.5 (0.3)	1.0 (0.4)	1.9 (0.9)	3.8 (1.7)	7.0 (3.6)	14.9 (7.2)

La maladie d'Alzheimer. Les données concernant la prévalence de la maladie d'Alzheimer seule sont plus rares. Ramaroson et al. [2003] estiment à 14% la prévalence chez les plus de 75 ans. Ott et al. [1995] sont les seuls à notre connaissance à rapporter leur estimation de la prévalence de la maladie d'Alzheimer par tranche d'âge (*F*ig. 2.4). Ils estiment la prévalence de la maladie d'Alzheimer à 13.2% chez les personnes de plus de 75 ans.

2.1.1.3 Facteurs de risque

La connaissance des facteurs de risque de la maladie d'Alzheimer est nécessaire à toute approche préventive de la maladie. Les facteurs de risque de la maladie d'Alzheimer connus actuellement peuvent être classés en trois catégories : *(i)* les facteurs de risque directs, *(ii)* les altérations de la réserve cognitive, *(iii)* les facteurs confondants.

Avant d'aller plus loin, il paraît nécessaire de préciser qu'un facteur de risque est une variable corrélée à la maladie. Par conséquent, il n'y a aucune notion de causalité. De plus, comme le rapporte Claudine Berr dans l'ouvrage collectif [Dubois et al., 2003], les facteurs mis en

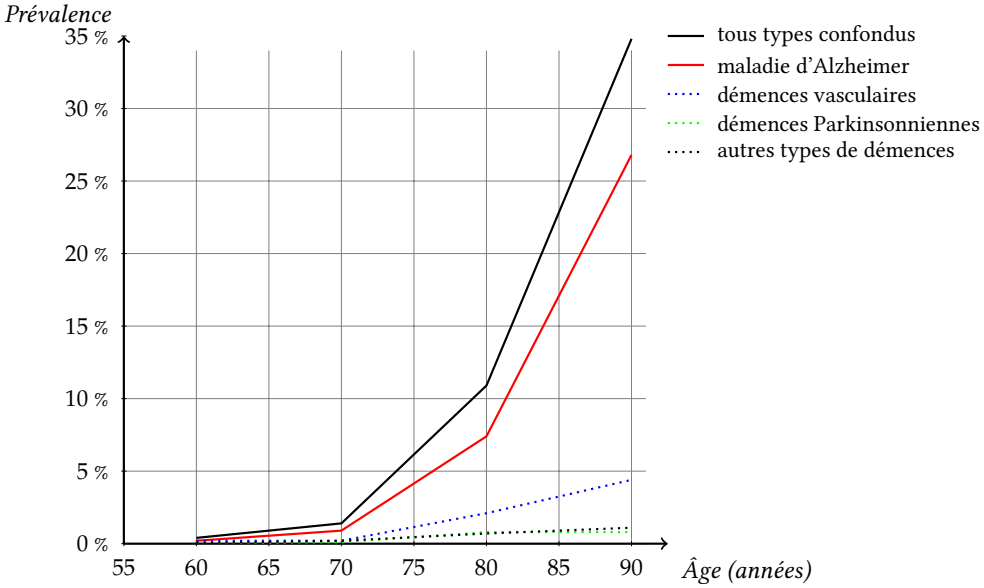


FIGURE 2.4 : Prévalence en fonction de l'âge (extrait de [Ott et al., 1995])

évidence par une étude de la prévalence peuvent s'avérer être plus liés à la survie des malades qu'au risque de la maladie elle-même.

Facteurs de risque directs. Dans ce paragraphe nous allons voir les principaux facteurs de risque à proprement parler de la maladie d'Alzheimer.

L'**âge** est le facteur de risque principal des démences et en particulier des maladies neurodégénératives telles que la maladie d'Alzheimer [Ott et al., 1995 ; Jorm & Jolley, 1998 ; Fratiglioni et al., 2000 ; Ramaroson et al., 2003 ; de Pedro-Cuesta et al., 2009]. L'incidence de la maladie, c'est à dire le nombre de nouveaux cas pour une population dans un intervalle de temps défini, croît exponentiellement avec l'âge : elle double approximativement tous les cinq ans.

Le deuxième facteur de risque incontesté après l'âge est un **facteur** génétique : le polymorphisme du gène de l'alipoprotéine E (**ApoE**). Les trois formes les plus fréquentes de l'ApoE sont chez l'homme : l'ApoE2, l'ApoE3 et l'ApoE4. Elles sont codées respectivement par les allèles $\epsilon 2$, $\epsilon 3$ et $\epsilon 4$. Le risque en fonction des allèles est rapporté dans le tableau 2.2. La présence de l'allèle $\epsilon 4$ du gène augmente fortement le risque de la maladie d'Alzheimer [Raber et al., 2004]. Le gène de l'ApoE est donc un gène de susceptibilité par opposition aux gènes déterministes : sa présence corrèle avec le risque de démence. Des d'informations supplémentaires sur les associations entre la maladie d'Alzheimer et la génétique sont disponibles sur le site : `http://alzgene.org`.

D'autres facteurs de risque plus controversés ont été mis en évidence ou proposés dans

TABLE 2.2 : Risque en fonction des allèles du gène de l'ApoE (d'après Raber et al. [2004]).

Allèles	Population (%)	Risque (%)
$\epsilon 2/\epsilon 2$	1	0.08
$\epsilon 2/\epsilon 3$	12	3.2
$\epsilon 3/\epsilon 3$	60	5.1
$\epsilon 3/\epsilon 4$	21	18
$\epsilon 4/\epsilon 4$	2	67

différentes études.

Le **genre** est un facteur de risque controversé. L'étude PAQUID [Ramaroson et al., 2003] et l'étude EURODEM [Letenneur et al., 2000] ont mis en évidence une prévalence plus élevée chez les femmes. Cependant, il est difficile de déterminer la cause de cette différence. Elle peut être liée au fait que la survie est deux fois plus longue chez les femmes une fois la maladie débutée [Helmer et al., 2001] (étude PAQUID).

Un autre facteur de risque également très discuté est l'**aluminium** [Rondeau et al., 2009]. L'origine de cette hypothèse est due à la présence d'aluminium dans les plaques séniles et dans les dégénérescences neurofibrillaires. Cependant, aucune étude, à notre connaissance, n'a mis clairement en évidence l'aluminium comme facteur de risque. Cela peut être dû aux difficultés méthodologiques d'une telle étude.

Les **facteurs nutritionnels** sont également mal connus. D'après Claudine Berr [Dubois et al., 2003], quelques pistes émergent cependant : les antioxydants et acides gras ainsi qu'une consommation modérée d'alcool seraient des facteurs protecteur. De nouvelles études telles que l'étude MAPT [Vellas et al., 2008] dont les résultats sont prévus pour 2014, sont en cours.

Enfin quelques études ont été réalisées sur l'impact du tabagisme [Wang et al., 1999 ; Fratiglioni & Wang, 2000 ; Reitz et al., 2007]. Ces études sont peu concluantes ou très critiquables dans la mesure où il y a un biais de survie très important. D'autres facteurs sont étudiés tels que les traitements hormonaux de la ménopause et la prise d'anti-inflammatoires [Dubois et al., 2003].

Altérations des réserves cérébrales et cognitives. Les concepts de réserve cérébrale et de réserve cognitive sont apparus au début des années 1990. Cette idée de réserve correspond à la capacité du cerveau à compenser des altérations, qu'elles soient liées à l'âge ou pathologiques. Cette capacité du cerveau serait liée au nombre de neurones et à la densité synaptique [Katzman et al., 1988 ; Katzman, 1993 ; Orrell & Sahakian, 1995] ; on parle alors de **réserve neuronale** ou **réserve cérébrale**.

Une autre hypothèse, celle de la **réserve cognitive** [Stern, 2002 ; Stern et al., 2005 ; Stern, 2006], explique cette capacité du cerveau à supporter les altérations par un mécanisme de compensation. Le cerveau serait capable de compenser la perte neuronale par exemple en étant plus actif ou en utilisant des réseaux plus efficaces ou moins atteints pas la pathologie. Cette capacité serait variable d'un individu à l'autre et dépendrait de divers facteurs tels que l'éducation [Mortimer & B., 1993 ; Letenneur et al., 2000], les activités sociales [Fratiglioni et al., 2004], ...

Ces réserves neuronales ou cognitives n'influent pas *a priori* sur l'apparition des lésions dues à la maladie mais sur le délai d'apparition des symptômes. En particulier, si la réserve cérébrale et/ou cognitive est suffisante, un individu atteint de la maladie d'Alzheimer peut décéder avant l'apparition des symptômes cliniques.

Certains facteurs (par exemple environnementaux) influeraient sur ces réserves et donc sur la vitesse d'apparition des symptômes. Puisque cela n'influe pas sur la maladie mais sur les symptômes de celle-ci, il est donc délicat de parler de facteurs de risque à proprement parler. On retrouve par exemple les traumatismes crâniens [O'Meara et al., 1997 ; Nemetz et al., 1999].

2.1.2 Le diagnostic de la maladie d'Alzheimer

Le diagnostic avec certitude requiert une confirmation histopathologique de la présence de plaques amyloïdes et de cas de dégénérescence neurofibrillaire ; il est donc post mortem. Cependant, un diagnostic sûr dès les premiers stades de la maladie est nécessaire dans la perspective de nouvelles thérapies. Le diagnostic clinique actuel est quant à lui basé sur une batterie d'examens cliniques et neuropsychologiques [Blennow et al., 2006]. Le rôle traditionnel de l'imagerie est principalement le diagnostic négatif : elle se limite essentiellement à éliminer d'autres causes symptomatiques (tumeur, hématome, hydrocéphalie, démence vasculaire). Dans certains centres, d'autres examens sont effectués en cas de doute, comme l'imagerie PIB[1] ou la recherche de biomarqueurs dans le liquide cérébro-spinal.

Les patients atteints de la maladie d'Alzheimer à des stades prodromaux sont, pour la plupart, atteints de troubles cognitifs légers [Petersen et al., 1999 ; Dubois & Albert, 2004]. On les appellera MCI pour *mild cognitive impairment* dans la suite du manuscrit. Tous les patients MCI ne sont pas des patients atteints de la maladie d'Alzheimer ; à ce stade, par définition, leurs symptômes ne sont pas suffisants pour un diagnostic clinique. Ils peuvent également être atteints d'autres pathologies.

Récemment, de nouveaux critères ont été proposés pour le diagnostic de la maladie d'Alzheimer à des stades plus précoces [Dubois et al., 2007, 2010]. Ces critères utilisent la combinaison d'un score clinique évaluant la mémoire épisodique avec des biomarqueurs issus

[1] Le PIB (*Pittsburgh compound B* est un marqueur des plaques amyloïdes utilisé en tomographie par émission de positons (TEP).

de l'imagerie IRM et TEP ou de biomarqueurs du liquide cérébro-spinal obtenus par ponction lombaire. La neuro-imagerie permettrait ainsi d'aider au diagnostic précoce de la maladie en fournissant par exemple des mesures telles que l'atrophie du lobe temporal à l'aide d'IRM anatomique ou des mesures provenant de la tomographie par émission de positrons soit avec du fluorodésoxyglucose (FDG) soit avec des marqueurs des plaques amyloïdes [Fox & Schott, 2004 ; Jagust, 2006].

2.1.3 L'apport de l'imagerie anatomique

Cette introduction sur le rôle de l'imagerie anatomique dans la maladie d'Alzheimer se base essentiellement sur [Colliot et al., 2008b].

Dans le cadre de la maladie d'Alzheimer, l'imagerie anatomique est principalement utilisée pour rechercher les signes positifs d'atrophie qui sont le reflet de la perte neuronale [Bobinski et al., 1999]. L'IRM anatomique (pondérée en T_1) fournit un contraste important entre la substance grise et la substance blanche, avec une taille de voxel de l'ordre de 1 mm^3 ; elle représente un bon compromis pour étudier l'atrophie cérébrale. Différentes études visant à localiser et à quantifier cette atrophie ont été réalisées. Elles ont été réalisées à partir de :

- mesures de régions d'intérêts [Convit et al., 1997, 2000 ; Jack Jr. et al., 1997, 1998 ; Juottonen et al., 1998 ; Laakso et al., 1998, 2000 ; Busatto et al., 2003 ; Xu et al., 2000 ; Good et al., 2002 ; Chételat & Baron, 2003 ; Rusinek et al., 2004 ; Tapiola et al., 2008] ;
- d'études morphométriques voxel-à-voxel [Good et al., 2002 ; Busatto et al., 2003 ; Karas et al., 2003, 2004 ; Chételat et al., 2005 ; Whitwell et al., 2007, 2008] ;
- de comparaison de cartes d'épaisseurs corticales [Thompson et al., 2001, 2003, 2004 ; Lerch et al., 2005, 2008 ; Bakkour et al., 2009 ; Dickerson et al., 2009 ; Hua et al., 2009 ; McDonald et al., 2009].

Ces études ont montré que l'atrophie cérébrale chez les sujets atteints de la maladie d'Alzheimer (AD) et chez les AD prodromaux touche de nombreuses régions. Les régions principalement touchées par l'atrophie sont : le cortex entorhinal, les hippocampes, les structures temporales latérales et inférieures ainsi que le gyrus cingulaire antérieur et postérieur. Cependant, la plupart de ces études met en évidence des différences au niveau de la population et non au niveau individuel. Leur intérêt pour le diagnostic individuel reste donc limité.

Les avancées en apprentissage statistique avec notamment le développement d'algorithmes capables en pratique de traiter des problèmes de classification dans des espaces de très grandes dimensions tels que les machines à vecteurs supports (SVM) [Vapnik, 1995 ; Shawe-Taylor &

Cristianini, 2004 ; Schölkopf & Smola, 2001] ont permis le développement de nouveaux outils de diagnostic à partir de l'imagerie par résonance magnétique anatomique[2] (cf. chapitre 1). Récemment, différentes approches de classification automatique de patients AD et/ou MCI à partir d'IRM anatomique ont vu le jour [Fan et al., 2008b,a ; Davatzikos et al., 2008a,c ; Klöppel et al., 2008b ; Vemuri et al., 2008 ; Chupin et al., 2009b,a ; Desikan et al., 2009 ; Gerardin et al., 2009 ; Hinrichs et al., 2009 ; Magnin et al., 2009 ; Misra et al., 2009 ; Querbes et al., 2009]. Ces approches peuvent potentiellement aider au diagnostic précoce de la maladie d'Alzheimer. On peut les regrouper en trois catégories selon les caractéristiques utilisées pour la classification :

1. les méthodes voxelliques ;
2. les méthodes utilisant l'épaisseur corticale ;
3. les méthodes utilisant l'hippocampe.

Dans la première catégorie de méthodes, les méthodes voxelliques, les caractéristiques sont définies au niveau des voxels de l'IRM. Plus précisément, les caractéristiques sont l'intensité en chaque voxel des cartes de probabilité de substance grise, de substance blanche et de liquide cérébrospinal [Fan et al., 2008b,a ; Davatzikos et al., 2008a,c ; Klöppel et al., 2008b ; Vemuri et al., 2008 ; Hinrichs et al., 2009 ; Magnin et al., 2009]. Klöppel et al. [2008b] utilisent directement ces caractéristiques comme données pour la classification (SVM). Toutes les autres approches réduisent la dimension de l'espace des données à l'aide de différentes étapes d'extraction et/ou de sélection de *features*. Vemuri et al. [2008] sous-échantillonnent les cartes puis effectuent une étape de sélection de *features*. Une autre possibilité est de grouper les voxels en régions anatomiques à l'aide d'un atlas [Ye et al., 2008 ; Magnin et al., 2009]. Cependant, une telle parcellisation du cerveau n'est pas forcément adaptée à la maladie d'Alzheimer : la frontière des régions touchées ne correspond pas forcément à celles de l'atlas. Pour cette raison, Fan et al. [2007] ont proposé de faire une parcellisation adaptative du cerveau à l'aide d'une étude de groupe, afin d'obtenir un ensemble de régions homogènes en terme de pouvoir discriminant. Cette méthode a été utilisée dans de nombreuses études telles que : [Davatzikos et al., 2008a,c ; Fan et al., 2008b,a ; Misra et al., 2009].

Dans la deuxième catégorie regroupant les méthodes utilisant l'épaisseur corticale, les caractéristiques utilisées pour la classification sont les mesures d'épaisseur corticale en chaque nœud du maillage cortical [Desikan et al., 2009 ; Querbes et al., 2009].

Enfin les méthodes de la troisième catégorie se focalisent sur l'étude de l'hippocampe : son volume [Colliot et al., 2008a ; Chupin et al., 2009b,a] ou sa forme [Gerardin et al., 2009].

[2] pondérée en T_1

Toutes ces approches obtiennent de bonnes performances avec des taux de classifications supérieurs à 84% pour la classification entre les témoins et les patients. Cependant, leur évaluation a été réalisée sur des populations d'étude différentes, ce qui rend difficile toute comparaison directe. En effet, de nombreux facteurs tels que le stade de la maladie, l'âge, le genre, le génotype, le niveau d'éducation ou encore la qualité des IRM peuvent modifier l'estimation des performances de classification. Outre ces différents facteurs, le faible nombre de sujets dans les différentes études réalisées rend toute méta-analyse difficile.

Pour cette raison, nous proposons une étude de comparaison de différentes méthodes de classification de sujets atteints de la maladie d'Alzheimer à partir d'IRM anatomiques T_1 en utilisant une même population d'étude provenant de la base de données ADNI. Nous comparons dix méthodes.

Nous évaluons tout d'abord les performances de cinq approches voxelliques : une approche directe [Klöppel et al., 2008b], une approche utilisant un volume d'intérêt [Klöppel et al., 2008b], une approche utilisant une parcellisation du cerveau à l'aide d'un atlas [Lao et al., 2004 ; Magnin et al., 2009] ainsi que les approches proposées respectivement par Vemuri et al. [2008] et Fan et al. [2007]. Afin d'évaluer l'influence de l'étape de recalage et du choix des cartes de probabilité de tissu utilisées, nous testons deux algorithmes de recalage (SPM5 [Ashburner & Friston, 2005] et DARTEL [Ashburner, 2007]) et l'utilisation uniquement des cartes de substance grise ou de tous les tissus.

Trois approches basées sur l'épaisseur corticale sont également évaluées : une approche directe similaire à celle de Klöppel et al. [2008b] dans le cas voxellique, une approche utilisant un atlas anatomique ainsi que l'approche proposée par Desikan et al. [2009].

Enfin deux méthodes utilisant uniquement l'hippocampe sont également testées : l'une ne se basant que sur le volume [Colliot et al., 2008a ; Chupin et al., 2009b,a] et l'autre analysant la forme de l'hippocampe [Gerardin et al., 2009].

2.2 La base de données ADNI

Les données utilisées pour la comparaison de méthodes de classification d'images cérébrales sont issues de la base de données ADNI (*Alzheimer's Disease Neuroimaging Initiative*). ADNI est une étude multicentrique financée par un partenariat public/privé.

L'objectif primordial de cette étude est le développement et la validation de biomarqueurs de la maladie d'Alzheimer. Cela concerne principalement les biomarqueurs issus de l'imagerie IRM et TEP ainsi que les biomarqueurs du liquide cérébro-spinal et du sang.

Près de 800 sujets ont été recrutés pour ADNI. Parmi ces sujets, il y a approximativement 200 témoins âgés, 400 patients souffrant de troubles cognitifs légers (MCI - *mild cognitive*

Table 2.3 : Caractéristiques cliniques et démographiques de la population d'étude. Les valeurs sont indiquées comme : moyenne± écart-type [intervalle].

Ensemble	Diag.	Nb.	Age	Genre	MMS
appr.	CN	81	76.1 ± 5.6 [60 – 89]	38 M / 43 F	29.2 ± 1.0 [25 – 30]
	AD	69	75.8 ± 7.5 [55 – 89]	34 M / 35 F	23.3 ± 1.9 [18 – 26]
	MCI_c	39	74.7 ± 7.8 [55 – 88]	22 M / 17 F	26.0 ± 1.8 [23 – 30]
	MCI_{nc}	67	74.3 ± 7.3 [58 – 87]	42 M / 25 F	27.1 ± 1.8 [24 – 30]
test	CN	81	76.5 ± 5.2 [63 – 90]	38 M / 43 F	29.2 ± 0.9 [26 – 30]
	AD	68	76.2 ± 7.2 [57 – 91]	33 M / 35 F	23.2 ± 2.1 [20 – 27]
	MCI_c	37	74.9 ± 7.0 [57 – 87]	21 M / 16 F	26.9 ± 1.8 [24 – 30]
	MCI_{nc}	67	74.7 ± 7.3 [58 – 88]	42 M / 25 F	27.3 ± 1.7 [24 – 30]
total	CN	162	76.3 ± 5.4 [60 – 90]	76 M / 86 F	29.2 ± 1.0 [25 – 30]
	AD	137	76.0 ± 7.3 [55 – 91]	67 M / 70 F	23.2 ± 2.0 [18 – 27]
	MCI_c	76	74.8 ± 7.4 [55 – 88]	43 M / 33 F	26.5 ± 1.9 [23 – 30]
	MCI_{nc}	134	74.5 ± 7.2 [58 – 88]	84 M / 50 F	27.2 ± 1.7 [24 – 30]

impairment) et 200 patients atteints de la maladie d'Alzheimer. Les critères d'inclusion des participants à cette étude sont par disponibles en ligne[3]. Le suivi des sujets dure de deux à trois ans avec un examen tous les 6 à 12 mois. La fréquence des examens dépend de leur nature.

2.2.1 Participants

Pour effectuer la comparaison de méthodes de classification d'images cérébrales, nous avons considéré l'ensemble des sujets de la base ADNI dont les images prétraitées étaient disponibles au moment où nous avons commencé notre travail. Au total, 509 sujets ont ainsi été sélectionnés. Parmi ces sujets, on dénombre 162 témoins (CN - *cognitively normal elderly controls*), 137 patients atteints de la maladie d'Alzheimer (AD - *Alzheimer's Disease*), 76 patients souffrants d'un déclin cognitif léger qui ont converti vers la maladie d'Alzheimer dans les 18 mois qui suivirent l'IRM initiale (MCI_c) ainsi que 134 autres MCI qui n'ont pas converti vers la maladie d'Alzheimer durant ces 18 premiers mois (MCI_{nc}). Les caractéristiques cliniques et démographiques de la population d'étude sont résumées dans le tableau 2.3.

[3] `http://www.adni-info.org/Scientists/AboutADNI.aspx#`

Afin de s'assurer de l'absence de différences entre les groupes autres que le diagnostic clinique, des tests statistiques sont réalisés : un test de Student pour l'âge et un test du χ^2 de Pearson pour le genre. En fixant le seuil de significativité à $p = 0.05$, aucune différence significative n'est alors trouvée entre les différents groupes.

Pour pouvoir effectuer une estimation non biaisée des performances des différentes méthodes, nous avons divisé aléatoirement la population d'étude en deux sous-ensembles de même taille. L'un des deux sous-ensembles est utilisé pour l'apprentissage des classifieurs, l'autre pour l'évaluation de leur sensibilité et spécificité. Le processus de division respecte les distributions de l'âge et du genre de chaque groupe clinique (tableau 2.3).

2.2.2 Acquisitions IRM

Les images sont les IRM anatomiques pondérées en T_1 acquises à 1.5T lors de la visite de *baseline* quand celles-ci sont disponibles. Lorsque ce n'est pas le cas, nous prenons celles de la visite de *screening*. Le protocole d'acquisition des IRM est détaillé dans [Jack et al., 2008].

Dans le protocole d'acquisition des données d'ADNI, les sujets sont scannés deux fois à chaque visite. Pour chaque sujet, nous considérons donc uniquement celle des deux IRM qui a la meilleure qualité selon la cellule de contrôle qualité d'ADNI.

ADNI met non seulement à disposition les images originales, mais aussi les images ayant subi certains prétraitements. Nous utilisons les images ayant subi les prétraitements suivants :

- *grad-warp* : correction des distorsions dues à la non-linéarité du gradient ;
- correction des inhomogénéités de champ B_1.

Ces prétraitements peuvent tous les deux être obtenus en routine clinique.

Les 509 images proviennent de 41 centres différents. Il n'y a aucun autre critère d'exclusion.

2.3 Les méthodes évaluées

Les différentes méthodes que nous avons comparées peuvent être groupées en 3 catégories selon les caractéristiques (*features*) utilisées pour la classification, comme nous l'avons fait dans la section 2.1.3. Pour les classifieurs de la première catégorie (paragraphe 2.3.1), les caractéristiques utilisées pour la classification sont définies au niveau du voxel de l'IRM. Plus précisément, ce sont les probabilités définies en chaque voxel d'avoir de la substance grise (GM - *gray matter*), de la substance blanche (WM - *white matter*) ou du liquide cérébrospinal (CSF - *cerebrospinal fluid*). La deuxième catégorie (paragraphe 2.3.2) regroupe les classifieurs pour

lesquels les caractéristiques sont l'épaisseur corticale définie à chaque nœud du maillage du cortex cérébral. Les méthodes de la dernière catégorie (paragraphe 2.3.3) utilisent uniquement l'hippocampe pour la classification.

2.3.1 Méthodes voxelliques

La première catégorie de méthode regroupe les méthodes qui utilisent comme caractéristiques pour la classification les cartes de probabilités de tissus (GM, WM et CSF). Ces cartes sont obtenues en suivant la procédure suivante. Les IRM anatomiques sont simultanément segmentées et recalées à l'aide de la segmentation unifiée de SPM5 (Statistical Parametric Mapping, London, UK) [Ashburner & Friston, 2005]. La segmentation et le recalage sont réalisés avec les paramètres par défaut de SPM5. À la fin de cette étape, nous avons pour chaque individu trois cartes de probabilités, SPM5_GM, SPM5_WM et SPM5_CSF qui correspondent respectivement aux cartes de probabilités de substance grise, de substance blanche et de liquide cérébrospinal.

Afin d'évaluer l'influence de l'étape de recalage sur la classification, nous avons également recalé les cartes de substance grise et de substance blanche segmentées par SPM5 en utilisant l'algorithme de recalage DARTEL [Ashburner, 2007]. Plus précisément, DARTEL utilise les cartes de probabilité de substance grise et de substance blanche dans l'espace natifs des sujets. DARTEL itère ensuite les deux étapes suivantes :

1. recalage des cartes de probabilités sur le *template* commun

2. construction d'un nouveau *template* commun à partir des cartes recalées obtenues à l'étape 1

Au final, les cartes de probabilités de substance grise et de substance blanche sont recalées sur un *template* commun généré à partir de la population d'étude. Nous les appellerons dans la suite DARTEL_GM et DARTEL_WM. Les transformations obtenues sont également appliquées aux cartes de CSF. Nous appellerons ces cartes DARTEL_CSF. Les paramètres par défaut de DARTEL ont été utilisés pour le recalage. En particulier l'étape recalage/construction de *template* a été itérée six fois. Toutes les cartes recalées, que ce soit par la segmentation unifiée de SPM5 ou par DARTEL, ont été modulées par le jacobien de leur transformation. Cela permet de préserver les quantités de tissus (le jacobien correspond aux déformations locales du volume). Aucun lissage spatial n'a été réalisé au préalable.

Certaines méthodes, dans leur version originale, n'utilisent que les cartes de substance grise tandis que d'autres utilisent l'ensemble des cartes de probabilités. Nous avons évalué chacune des méthodes de manière systématique avec les cartes de substance grise uniquement et également avec l'ensemble des cartes de probabilités.

Les différentes méthodes de cette catégorie diffèrent dans leur manière d'extraire et/ou de sélectionner les caractéristiques utilisées pour la classification à partir des cartes de probabilités. Nous les décrivons succinctement dans les paragraphes suivants. Elles sont également résumées dans le tableau 2.4.

2.3.1.1 Directe

L'approche la plus simple consiste à utiliser directement les cartes de probabilités comme caractéristiques pour la classification. Nous appellerons cette approche *Voxel-Direct* dans la suite du manuscrit. Ce type d'approche a été proposé par Klöppel et al. [2008b]. Ils ont proposé deux versions de cette approche. La première version utilise les cartes de probabilités de tout le cerveau. Quant à la deuxième, elle restreint son analyse à un volume d'intérêt (VOI - *volume of interest*) situé dans la partie antérieure moyenne du lobe temporal. Cette région d'intérêt englobe en partie l'hippocampe. Plus précisément, cette région d'intérêt est définie par deux parallélépipèdes rectangles centrés respectivement en $(-17, -8, -18)$ et $(16, -9, -18)$ dans le repère MNI (les valeurs sont données en mm). Leurs dimensions sont les suivantes : 12 mm, 16 mm et 12 mm suivant l'axe transversal, l'axe sagittal (ou antéropostérieur) et l'axe vertical respectivement. Cette approche sera appelée *Voxel-Direct_VOI* dans la suite du manuscrit.

Dans leur version originale, ces méthodes utilisent seulement DARTEL_GM ; nous les avons testé avec les ensembles suivants :

- SPM5_GM uniquement,
- SPM5_GM, SPM5_WM et SPM5_CSF,
- DARTEL_GM uniquement,
- DARTEL_GM, DARTEL_WM et DARTEL_CSF.

2.3.1.2 STAND-score

Vemuri et al. [2008] ont proposé une approche appelée *STAND-score*. Dans cette approche, la dimensionnalité de l'espace des caractéristiques est réduite à l'aide d'une succession d'étapes d'agrégation et de sélection de variables. Plus précisément, la méthode est la suivante.

1. Tout d'abord, les voxels du cervelet sont retirés à l'aide d'un masque dans l'espace MNI.
2. Les images sont ensuite lissées avec un filtre gaussien puis sous-échantillonnées après un filtrage par moyenne glissante.

Table 2.4 : Résumé des méthodes voxelliques comparées.

Caract.	Recalage	Tissus	Classifieur	Méthode	
Direct	DARTEL	GM	SVM lin.	*1.1.1 a*	*Voxel-Direct-D-gm*
		tout	SVM lin.	*1.1.1 b*	*Voxel-Direct-D-all*
	SPM5	GM	SVM lin.	*1.1.2 a*	*Voxel-Direct-S-gm*
		tout	SVM lin.	*1.1.2 b*	*Voxel-Direct-S-all*
Direct VOI	DARTEL	GM	SVM lin.	*1.2.1 a*	*Voxel-Direct_VOI-D-gm*
		tout	SVM lin.	*1.2.1 b*	*Voxel-Direct_VOI-D-all*
	SPM5	GM	SVM lin.	*1.2.2 a*	*Voxel-Direct_VOI-S-gm*
		tout	SVM lin.	*1.2.2 b*	*Voxel-Direct_VOI-S-all*
STAND-score	DARTEL	GM	SVM lin.	*1.3.1 a*	*Voxel-STAND-D-gm*
		tout	SVM lin.	*1.3.1 b*	*Voxel-STAND-D-all*
	SPM5	GM	SVM lin.	*1.3.2 a*	*Voxel-STAND-S-gm*
		tout	SVM lin.	*1.3.2 b*	*Voxel-STAND-S-all*
	SPM5 custom template	GM	SVM lin.	*1.3.3 a*	*Voxel-STAND-Sc-gm*
		tout	SVM lin.	*1.3.3 b*	*Voxel-STAND-Sc-all*
Atlas	DARTEL	GM	SVM lin.	*1.4.1 a*	*Voxel-Atlas-D-gm*
		tout	SVM lin.	*1.4.1 b*	*Voxel-Atlas-D-all*
	SPM5	GM	SVM lin.	*1.4.2 a*	*Voxel-Atlas-S-gm*
		tout	SVM lin.	*1.4.2 b*	*Voxel-Atlas-S-all*
COMPARE	DARTEL	GM	SVM gauss.	*1.5.1 a*	*Voxel-COMPARE-D-gm*
		tout	SVM gauss.	*1.5.1 b*	*Voxel-COMPARE-D-all*
	SPM5	GM	SVM gauss.	*1.5.2 a*	*Voxel-COMPARE-S-gm*
		tout	SVM gauss.	*1.5.2 b*	*Voxel-COMPARE-S-all*

3. Ne sont gardés pour l'analyse que les voxels qui contiennent plus de 10% de tissus (GM, WM ou CSF) chez au moins la moitié des sujets. Ceci est une manière de ne garder que les voxels du cerveau.

4. Une première étape de sélection de variables est réalisée. Cette étape est réalisée indépendamment pour la substance grise, pour la substance blanche et pour le liquide cérébrospinal. Les données sont entrées dans un SVM linéaire. Le SVM linéaire donne un poids à chaque voxel (ce sont les coefficients de l'hyperplan séparateur). Ne sont gardés que les voxels qui ont un poids donné par le SVM indiquant une atrophie (cf. paragraphe 2.5.6).

5. La deuxième étape de sélection est également appliquée séparément aux différents tissus. Les voxels voisins des voxels sélectionnés sont également sélectionnés.

6. L'ensemble des caractéristiques issues des différents tissus sont alors concaténées pour ne former qu'un seul vecteur. C'est ce vecteur qui est utilisé pour la classification.

Cette méthode sera appelée *Voxel-STAND* dans la suite du manuscrit. Dans sa version originale, cette méthode utilise les cartes de substance grise, de substance blanche et de CSF segmentées et recalées par la segmentation unifiée de SPM5 avec un *template* généré à partir de la population d'étude. Nous avons donc testé cette méthode non seulement avec les cartes décrites précédemment mais également avec celles obtenues de cette manière.

2.3.1.3 Atlas

Une autre méthode consiste à regrouper les voxels en régions anatomiques à l'aide d'un atlas. Ce type d'approche a été utilisé notamment par [Lao et al., 2004 ; Magnin et al., 2009].

Dans cette approche, les cartes de probabilités sont chacune découpées en 116 régions d'intérêt à l'aide de l'atlas AAL (*Automatic Anatomical Labeling*) [Tzourio-Mazoyer et al., 2002]. On utilise ensuite les moyennes d'intensités de chaque région comme caractéristiques pour la classification. Nous appellerons cette approche *Voxel-Atlas* dans la suite du manuscrit.

L'atlas AAL est un atlas anatomo-fonctionnel. Il n'a pas été spécifiquement construit pour l'étude de la maladie d'Alzheimer. Par conséquence, ses régions ne représentent pas nécessairement des régions homogènes sur le plan de la physiopathologie ou de l'atrophie.

2.3.1.4 COMPARE

Fan et al. [2007, 2008b,a] proposent d'utiliser une parcellisation du cerveau adaptée à la pathologie au lieu d'un atlas prédéfini (2.3.1.3). Cette méthode est décrite de manière détaillée dans [Fan et al., 2007]. Nous n'en donnons ici que les grandes lignes :

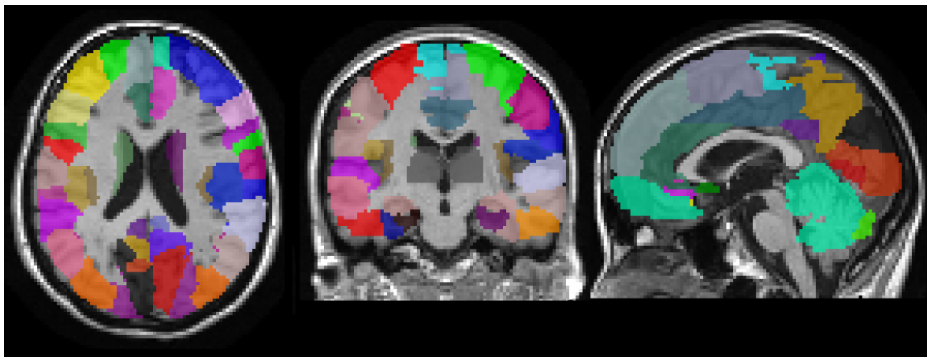


Figure 2.5 : Atlas AAL [Tzourio-Mazoyer et al., 2002] en coupe axiale, coronale et sagittale.

1. La première étape de COMPARE consiste à créer des régions homogènes en terme de discrimination.

2. Une fois ces régions déterminées, les sommes des intensités dans chacune des régions sont utilisées comme caractéristiques pour la classification.

3. Des étapes de sélection de variables sont ensuite réalisées (test T de Student et SVM-RFE) [Guyon et al., 2002].

Cette méthode sera appelée *Voxel-COMPARE* dans la suite du document. Nous avons utilisé l'implémentation de COMPARE disponible en ligne[4].

2.3.2 Méthodes utilisant l'épaisseur corticale

Les méthodes regroupées dans cette catégorie utilisent comme caractéristiques l'épaisseur corticale des sujets calculée en chaque point du maillage du cerveau. L'épaisseur corticale caractérise directement l'atrophie du cortex. Elle est donc potentiellement un biomarqueur de choix pour l'aide au diagnostic de la maladie d'Alzheimer [Thompson et al., 2001, 2003, 2004 ; Lerch et al., 2005 ; Bakkour et al., 2009 ; Dickerson et al., 2009 ; Hua et al., 2009 ; McDonald et al., 2009].

Les mesures d'épaisseur corticale sont réalisées avec le logiciel d'analyse FreeSurfer (Massachusetts General Hospital, Boston, MA). Ce logiciel est disponible gratuitement en ligne à l'adresse suivante : `http://surfer.nmr.mgh.havard.edu`. Les détails techniques de ce logiciel sont décrits principalement dans les papiers suivants : [Sled et al., 1998 ; Dale et al., 1999 ; Fischl et al., 1999a,b ; Fischl & Dale, 2000]. Toutes les cartes d'épaisseur corticale sont recalées sur le *template* par défaut de FreeSurfer. Notons que les sujets étant recalés dans

[4] `https://www.rad.upenn.edu/sbia/software/index.html`

TABLE 2.5 : Résumé des méthodes comparées utilisant l'épaisseur corticale.

Caract.	Recalage	Classifieur	Méthode
Direct	Freesurfer	SVM lin.	*2.1 Thickness-Direct*
Atlas	Freesurfer	SVM lin.	*2.2 Thickness-Atlas*
ROI	Freesurfer	reg. log	*2.3 Thickness-ROI*

un espace commun, le maillage du cerveau est le même pour tous les sujets. Quatre sujets n'ont pas pu être analysés par FreeSurfer. Ces sujets sont indiqués par un astérisque dans les tableaux C.1 à C.8. Ces sujets ne pouvaient donc pas être analysés par le SVM. Ils ont donc été éliminés de l'ensemble d'apprentissage. Quant à ceux de l'ensemble test, ils ont été considérés comme classés correctement avec une probabilité $\frac{1}{2}$.

Les différentes méthodes de cette catégorie diffèrent dans leur manière d'extraire et/ou de sélectionner les caractéristiques utilisées pour la classification à partir des cartes de probabilités. Nous les décrivons succinctement dans les paragraphes suivants. Elles sont également résumées dans le tableau 2.5.

2.3.2.1 Directe

De la même manière que pour les méthodes voxelliques présentées dans la section précédente (2.3.1), l'approche la plus simple consiste à considérer les mesures d'épaisseur corticale en chaque point du maillage directement comme les caractéristiques utilisées par le classifieur sans aucune autre étape de prétraitement des données. Cette approche sera appelée *Thickness-Direct*.

2.3.2.2 Atlas

Nous avons également testé l'approche qui consiste à regrouper les nœuds du maillage du cortex en régions anatomiques à l'aide d'un atlas. Ce type d'approche a été utilisé dans [Querbes et al., 2009 ; Desikan et al., 2009]. La parcellisation du cerveau en régions anatomiques est réalisée avec l'atlas de Desikan et al. [2006]. Cet atlas est constitué de 68 régions. Ces régions sont définies à partir des principaux gyri du cerveau. Dans chaque région, l'épaisseur corticale est moyennée. Ce sont ces valeurs qui sont utilisées pour la classification. Cette approche sera appelée *Thickness-Atlas*.

2.3.2.3 Régions d'intérêt

Desikan et al. [2009] ont divisé le cerveau en régions anatomiques à l'aide de l'atlas de Desikan et al. [2006] comme décrit dans le paragraphe précédent. Ils ont étudié le pouvoir discriminant de la moyenne de l'épaisseur corticale de chaque région ainsi que leur volume. Dans leur analyse, ils ont sommé les deux hémisphères. Ils ont également corrigé les volumes à l'aide du volume intracrânien total.

Leur étude a été réalisée sur un groupe de 97 sujets faisant partie de la base de données OASIS (*Open Access Series of Imaging Studies*) [Marcus et al., 2007]. En réalisant une régression logistique, ils ont trouvé que l'ensemble de caractéristiques le plus discriminant était la combinaison des trois caractéristiques suivantes :

- l'épaisseur moyenne du cortex entorhinal
- l'épaisseur moyenne du gyrus supramarginal
- le volume de l'hippocampe

Desikan et al. [2009] ont utilisé une régression logistique pour les classifications *CN vs AD* et *CN vs* MCI_c. Dans cette approche, nous avons donc utilisé uniquement ces trois caractéristiques pour la classification. Le classifieur utilisé est la régression logistique. Cette approche sera appelée *Thickness-ROI* dans la suite du manuscrit.

2.3.3 Méthodes utilisant l'hippocampe

La dernière catégorie regroupe les méthodes qui n'utilisent que l'hippocampe et non le cerveau dans son ensemble ou le cortex entier pour la classification. L'hippocampe est une structure du cerveau atteinte dès les premiers stades de la maladie. Il a été utilisé comme biomarqueur de la maladie dans un grand nombre d'études [Frisoni et al., 1999 ; Convit et al., 2000 ; Chupin et al., 2009b].

Pour cette étude, nous avons réalisé la segmentation de l'hippocampe à l'aide de la méthode de segmentation automatique SACHA inventée et développée par Chupin et al. [2007, 2009b].

Les différentes méthodes de cette catégorie sont décrites succinctement dans les paragraphes suivants. Elles sont également résumées dans le tableau 2.6.

2.3.3.1 Volume

Nous avons tout d'abord évalué les performances obtenues lorsque la seule caractéristique utilisée pour la classification est le volume hippocampique total. Nous avons donc calculé pour chaque sujet le volume des hippocampes. Ces volumes ont été normalisés par le volume

TABLE 2.6 : Résumé des méthodes comparées utilisant uniquement l'hippocampe

Caract.	Segmentation	Classifieur	Méthode
Volume	Freesurfer	parzen	*3.1.1 Hippo-Volume-F*
Volume	SACHA	parzen	*3.1.2 Hippo-Volume-S*
Shape	SACHA	SVM lin.	*3.2 Thickness-ROI*

intracrânien total (TIV - *total intracranial volume*) estimé en sommant les cartes de substance grise, de substance blanche et de liquide cérébrospinal obtenues avec SPM5. Pour plus de robustesse dans l'estimation des volumes hippocampiques, les volumes des hippocampes droits et gauches ont été sommés. Nous appellerons cette méthode *Hippo-Volume-S*. Nous avons également évalué cette approche en utilisant les volumes hippocampiques estimés avec FreeSurfer et normalisés par le TIV obtenu également avec Freesurfer. Cette approche sera appelée *Hippo-Volume-F*.

2.3.3.2 Analyse de forme

Nous avons ensuite testé une méthode dont les caractéristiques utilisées pour la classification sont des descripteurs de la forme de l'hippocampe [Gerardin et al., 2009]. Chaque hippocampe, segmenté automatiquement avec SACHA, est décomposé par une série d'harmoniques sphériques (SPHARM). Les coefficients des harmoniques sphériques encodent ainsi la forme des hippocampes. Ce sont ces coefficients qui sont utilisés pour la classification.

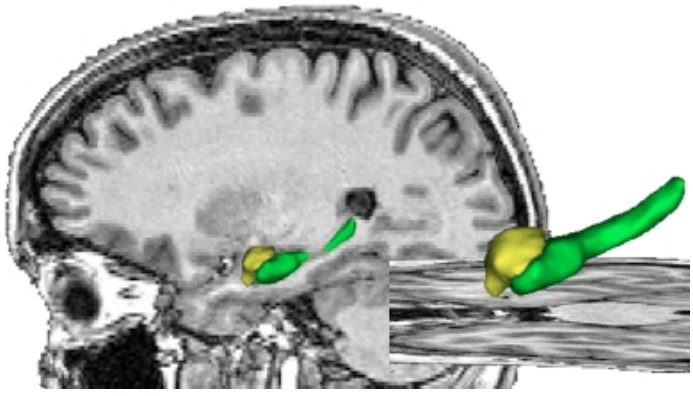


FIGURE 2.6 : Hippocampe (en vert) et amygdale (en jaune) segmentés par SACHA (image fournie par Marie Chupin). Sur l'image de gauche, l'hippocampe et l'amygdale sont supperposés à une coupe sagittale. Sur l'image de droite, l'hippocampe et l'amygdale sont représentés au dessus d'une coupe axiale.

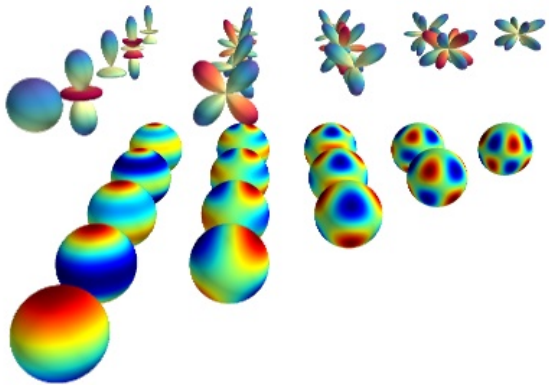


Figure 2.7 : Harmoniques sphériques.

Les coefficients des SPHARM sont calculés à l'aide du logiciel SPHARM-PDM [Styner et al., 2006] (*Spherical Harmonics-Point Distribution Model*) développé par l'Université of North Carolina et la National Alliance for Medical Image Computing[5]. Nous avons utilisé le degré 4 de la décomposition en SPHARM. Sa version originale [Gerardin et al., 2009] utilise une étape de sélection de variable. La raison principale était le petit nombre de sujets (moins de 30 sujets par groupe). Quand le nombre de sujets utilisés pour la classification est faible, le classifieur est sensible aux caractéristiques non discriminantes. Dans cette étude, nous avons éliminé cette étape. Le nombre de sujets étant plus important, cette étape devient moins nécessaire. En outre, une sélection de *features* augmente le risque de surapprentissage. Nous avons également testé avec la sélection de *features*; cela n'améliore pas les résultats de classification (*CN vs AD* : sensibilité 60%, spécificité 80%, *CN vs* MCI_c : sensibilité 50%, spécificité 85%, MCI_{nc} *vs* MCI_c : sensibilité 32%, spécificité 60%).

Quatre sujets n'ont pas pu être analysés par SPHARM-PDM. Ces sujets sont indiqués par une croix dans les tableaux C.1 à C.8. Ces sujets ne pouvaient donc pas être analysés par le SVM. Ils ont donc été éliminés de l'ensemble d'apprentissage. Quant à ceux de l'ensemble test, ils ont été considérés comme classés correctement avec une probabilité $\frac{1}{2}$.

Cette approche sera appelée *Hippo-Shape* dans la suite du manuscrit.

2.4 Classifications

2.4.1 Les expériences

Afin de comparer les différentes approches présentées dans la section précédente (2.3), nous avons effectué pour chaque méthode trois classifications. La première est la classification entre

[5] `http://www.na-mic.org/Wiki/index.php/Algorithm:UNC:Shape_Analysis`

les témoins (CN) et les patients AD probables. Nous l'appellerons *CN vs AD*. La deuxième expérience est la classification entre les témoins et les sujets MCI qui convertissent vers la maladie d'Alzheimer durant les 18 premiers mois suivant l'inclusion. Cette expérience de classification, appelée *CN vs* MCI_c, correspond à la détection de patients AD prodromaux selon la définition de [Dubois & Albert, 2004]. La troisième expérience est la classification entre les sujets MCI_{nc} et les sujets MCI_c (*MCI*$_{nc}$ *vs* MCI_c). Cela correspond à la prédiction à 18 mois de la conversion chez les patients MCI.

2.4.2 Les classifieurs

Les classifieurs utilisés dans cette étude sont des C-SVM linéaires exceptés pour *Voxel-COMPARE*, *Thickness-ROI* et *Hippo-Volume*. Afin de respecter leur version originale, nous avons utilisé un SVM non-linéaire avec un noyau gaussien pour *Voxel-COMPARE* et une régression logistique pour *Thickness-ROI*.

En ce qui concerne *Hippo-Volume*, la dimension de l'espace des caractéristiques est seulement un. Un classifieur beaucoup plus simple et sans hyperparamètre peut alors être utilisé ; chaque sujet est assigné au groupe le plus proche. Plus précisément, soient S_1 et S_2 deux groupes de sujets de moyennes de volumes hippocampiques respectivement m_1 et m_2. Considérons un nouvel individu avec un volume hippocampique x. Ce sujet va être classé comme appartenant au groupe S_1 si et seulement si $|m_1 - x| \leq |m_2 - x|$. On peut voir ce classifieur comme un cas particulier d'un classifieur à fenêtre de Parzen avec un noyau linéaire sous l'hypothèse d'une prévalence de 50% [Shawe-Taylor & Cristianini, 2004]. Nous avons également testé la classification avec un SVM linéaire et cela ne change pas les résultats de la classification.

L'implémentation utilisée pour le SVM est celle de LibSVM [Chang & Lin, 2001].

2.4.3 Évaluation des performances de classification

Pour obtenir des évaluations non-biaisées des performances de classification, la population d'étude a été divisée en deux sous-ensembles de même taille : un ensemble d'apprentissage et un ensemble test. La division est faite de manière à préserver les distributions de l'âge et du genre de chaque groupe clinique. L'ensemble d'apprentissage sert à déterminer les valeurs optimales des hyperparamètres pour chaque méthode et à l'apprentissage à proprement parler des classifieurs. Quant à l'ensemble de test, il est uniquement utilisé pour l'estimation des performances des méthodes. Les ensembles d'apprentissages et de tests sont les mêmes pour toutes les méthodes exceptées celles pour lesquelles le calcul des caractéristiques a échoué (voir plus haut : épaisseur corticale et SPHARM). Pour ces méthodes là, les sujets pour

lesquels l'analyse a échoué ont été soit retirés de l'analyse s'ils appartenaient à l'ensemble d'apprentissage, soit classés correctement avec probabilité $\frac{1}{2}$ s'ils faisaient partie de l'ensemble de test. Nous n'avons pas retiré complètement les sujets pour lesquels l'analyse avait échoué. En effet, nous considérons que c'est une faiblesse de la méthode correspondante et qu'il faut en tenir compte. Nous détaillons dans les paragraphes suivants la procédure d'évaluation des méthodes. Elle est également illustrée par le diagramme figure 2.8.

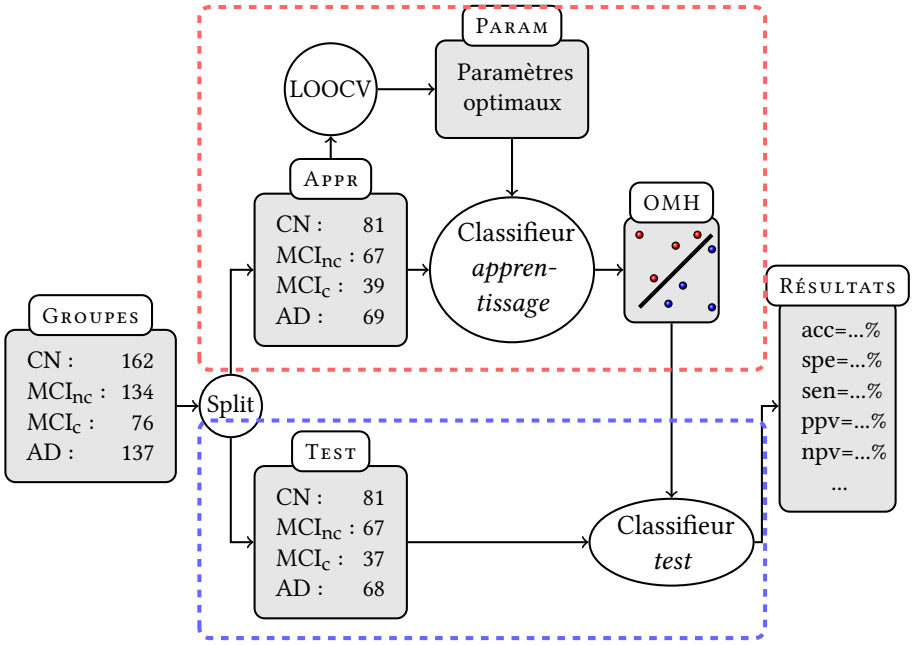


Figure 2.8 : Processus de comparaison des méthodes de classification. La population d'étude (Group) est divisée en deux ensembles. Le premier ensemble, l'ensemble d'apprentissage (Train), est utilisé pour l'optimisation des hyperparamètres et l'apprentissage du classifieur. L'optimisation des hyperparamètres se fait par recherche en grille avec une évaluation de type LOOCV. On choisit les hyperparamètres qui donnent le meilleur taux de classification avec cette évaluation. Une fois ces hyperparamètres fixés, on apprend le classifieur et on teste sur l'ensemble test (Test). Les performances sont ainsi estimées par validation croisée.

L'ensemble d'apprentissage est utilisé pour estimer par validation croisée (CV - *cross validation*) les valeurs optimales des hyperparamètres. Dans la plupart des méthodes, il n'y a qu'un seul hyperparamètre à estimer, c'est le paramètre C du C-SVM. Dans la méthode *Voxel-STAND*, il y a un second hyperparamètre à estimer : le seuil t sur les poids du SVM pour la sélection de variables. La méthode *Voxel-COMPARE* requiert deux autres hyperparamètres : σ qui détermine la taille du noyau gaussien du SVM et n_f qui correspond au nombre de *features* sélectionnés. Quant à la méthode *Hippo-Volume* elle n'a aucun hyperparamètre. L'estimation des valeurs optimales des hyperparamètres s'effectue par une recherche en

grille (grid-search). Pour chaque n-uplet de valeurs de la grille où n correspond au nombre d'hyperparamètres, le taux de classification est estimé par une procédure de *leave-one-out cross validation* (LOOCV) sur l'ensemble d'apprentissage. On choisit alors le n-uplet qui donne le meilleur taux de classification. La recherche en grille est effectuée sur les intervalles/ensembles suivants : $C = 10^{-5.0}, 10^{-4.5}, \cdots, 10^{3.0}$, $t = 0.06, 0.08, \cdots, 0.98$, $\sigma = 100, 200, \cdots, 1000$ et $n_f = 1, 2, \cdots, 150$ (sauf pour *Voxel-COMPARE* où $C = 10^{0.0}, 10^{1.5}, \cdots, 10^{2.5}$).

Pour chaque méthode, une fois les valeurs optimales des hyperparamètres estimées, l'ensemble d'apprentissage est utilisé pour l'apprentissage à proprement parler des classifieurs. Les performances (tableau 2.7) des méthodes sont ensuite évaluées sur l'ensemble test. De cette manière là, nous obtenons des estimations non biaisées des performances. Pour chaque méthode, on calcule le nombre de :

- vrais positifs (TP - *true positive*) – les sujets malades classés comme malades
- vrais négatifs (TN - *true negative*) – les sujets témoins classés comme témoins
- faux positifs (FP - *false positive*) – les sujets témoins classés comme malades
- faux négatifs (FN - *false negative*) – les sujets malades classés comme témoins

Ceci permet de calculer (tableau 2.7) :

- la sensibilité – taux de sujets malades correctement classés, à savoir : $\frac{TP}{TP+FN}$
- la specificité – taux de sujets témoins correctement classés, à savoir : $\frac{TN}{FP+TN}$
- la valeur prédictive positive (PPV) – $PPV = \frac{TP}{TP+FP}$
- la valeur prédictive négative (NPV) – $NPV = \frac{TN}{TN+FN}$

Remarque : Une autre manière de procéder pour l'évaluation des performances auraient été de ne pas diviser la population en un ensemble d'apprentissage et un ensemble de test mais d'estimer les hyperparamètres à l'aide de deux validations croisées de type LOOCV imbriquées. Nous n'avons pas choisi cette option principalement pour des raisons de temps de calculs (notamment de la méthode *Voxel-COMPARE*).

Pour savoir si les différentes méthodes obtiennent des résultats significativement meilleurs que le hasard, nous utilisons le test de χ^2 de McNemar. Le seuil de significativité a été fixé à $p = 0.05$. Nous avons également utilisé ce test du χ^2 de McNemar pour évaluer les différences

Table 2.7 : Récapitulatif des différentes mesures de performances en termes de vrais positifs (TP), faux positifs (FP), vrais négatifs (TN), faux négatifs (FN), sensibilité, spécificité, valeur prédictive positive (PPV) et valeur prédictive négative (NPV).

	Malades	**Sains**	
Classés comme malades	TP	FP erreur de type I	$\rightarrow PPV = \frac{TP}{TP+FP}$
Classés comme sains	FN erreur de type II	TN	$\rightarrow NPV = \frac{TN}{TN+FN}$
	$\downarrow$ sensibilité $= \frac{TP}{TP+FN}$	$\downarrow$ spécificité $= \frac{TN}{TN+FP}$	

en terme de performance de classification entre les classifications utilisant les cartes recalées avec DARTEL et celles utilisant les cartes recalées avec SPM5. Nous avons suivi la même approche pour évaluer l'apport des cartes de substance blanche et de CSF. Le test du χ^2 de McNemar quantifie les différences entre la proportion de sujets correctement classifiés par un classifieur et celle obtenue avec un autre. En d'autres termes, ce test quantifie les différences de taux de classification. La table de contingence correspondant à ce test est présentée au tableau 2.8.

Table 2.8 : Table de contingence utilisée pour le test de McNemar. a : nombre de sujets correctement classés par les deux classifieurs ; b : nombre de sujets correctement classés par le classifieur 1 mais mal classés par le classifieur 2 ; c : nombre de sujets mal classés par le classifieur 1 mais correctement classé par le classifieur 2 ; d : nombre de sujets mal classés par les deux classifieurs.

	Classifieur 2 : sujets correctement classés	Classifieur 2 : sujets mal classés
Classifieur 1 : sujets correctement classés	a	b
Classifieur 1 : sujets mal classés	c	d

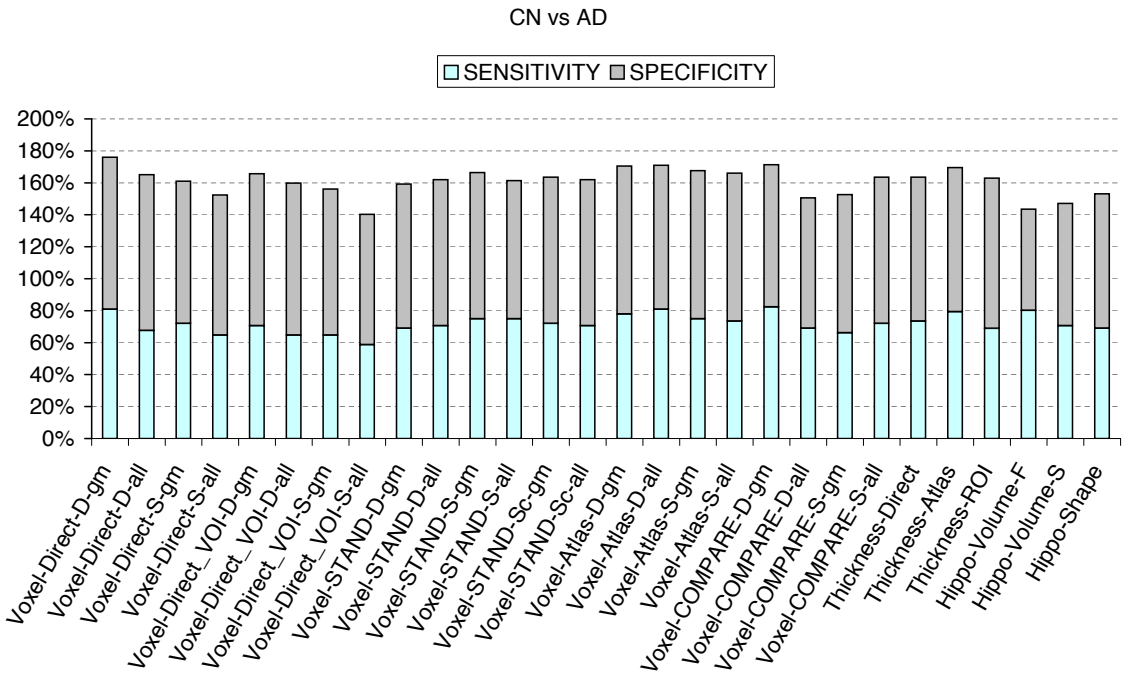


FIGURE 2.9 : Résultats des différentes méthodes pour la comparaison *CN vs AD*

2.5 Résultats

2.5.1 Performance des méthodes

Les résultats des différentes expériences de classifications présentées au paragraphe 2.4.1 sont résumés dans les tableaux 2.9, 2.10 et 2.11 respectivement pour les comparaisons *CN vs AD*, *CN vs* MCI_c et MCI_{nc} *vs* MCI_c. Ces résultats sont également résumés par les figures 2.9 et 2.10. Les méthodes sont appelées par les noms que nous leur avons donnés dans la section 2.3. Ces noms sont récapitulés dans les tableaux 2.4, 2.5 et 2.6.

2.5.1.1 *CN vs AD*

Les résultats de classification pour *CN vs AD* sont donnés dans le tableau 2.9 et illustrés par la figure 2.9. Toutes les méthodes ont des résultats meilleurs que le hasard ($p < 0.05$). Les quatre méthodes *Voxel* (*Voxel-Direct*, *Voxel-STAND*, *Voxel-Atlas* et *Voxel-COMPARE*) classent les AD des sujets témoins avec une très bonne spécificité (plus de 89%). La sensibilité obtenue est de 75% pour *Voxel-STAND* et est supérieure à 81% avec les trois autres méthodes.

Les méthodes utilisant l'épaisseur corticale comme caractéristiques pour la classification obtiennent des résultats similaires avec une spécificité supérieure à 90% et une sensibilité de

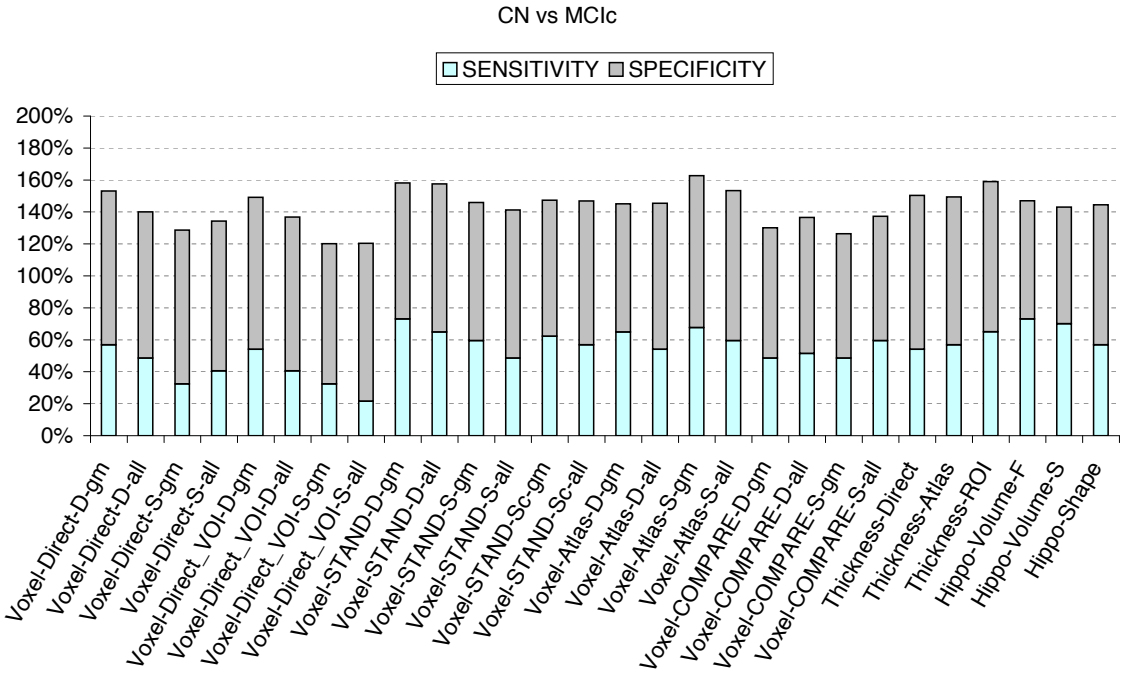


FIGURE 2.10 : Résultats des différentes méthodes pour la comparaison *CN vs MCI$_c$*

69%, 74% et 79% respectivement pour *Thickness-ROI*, *Thickness-Direct* et pour *Thickness-Atlas*.

En ce qui concerne les méthodes basées sur l'hippocampe, elles sont tout aussi sensibles mais moins spécifiques. Leur spécificité est de 63% pour *Hippo-Volume* et de 84% pour *Hippo-Shape*.

2.5.1.2 *CN vs MCI$_c$*

Les résultats de classification *CN vs MCI$_c$* sont résumés dans le tableau 2.10 et la figure 2.10. La majorité des méthodes est moins sensible que pour la classification *CN vs AD*. Toutes les méthodes exceptées *Voxel-COMPARE* et *Hippo* obtiennent des résultats significativement meilleurs que le hasard ($p < 0.05$). Il n'y a pas de grandes différences entre les résultats obtenus par les méthodes *Voxel-Direct*, *Voxel-Atlas* et ceux obtenus avec *Voxel-STAND*. Toutes ces méthodes ont une spécificité supérieure à 85%. La sensibilité est plus faible ; elle est comprise entre 51% (*Voxel-COMPARE*) et 73% (*Voxel-STAND*).

Les méthodes utilisant l'épaisseur corticale obtiennent des résultats similaires aux méthodes *Voxel*. La méthode *Hippo-Volume*, quant à elle, perd en spécificité mais reste constante en sensibilité par rapport à la comparaison *CN vs AD*.

2.5.1.3 MCI_{nc} *vs* MCI_c

Les résultats de la classification MCI_{nc} *vs* MCI_c sont présentés dans le tableau 2.11. Aucune des méthodes comparées n'obtient de résultats significativement meilleurs que le hasard. Quatre méthodes obtiennent toutefois des résultats légèrement meilleurs que le hasard (mais pas de manière significative). *Thickness-Direct* a une sensibilité de 32% et une spécificité de 91%.*Voxel-STAND* a une sensibilité de 57% et une spécificité de 78%. *Voxel-COMPARE* a une sensibilité de 62% et une spécificité de 67%. *Hippo-Volume* a une senbilité de 62% et une spécificité de 69%.

2.5.2 Complémentarité des méthodes

Les différentes méthodes testées dans cette étude abordent le problème de la classification de sujets atteints de la maladie d'Alzheimer sous des angles différents. Elles sont donc peut-être complémentaires. Afin de quantifier leur similarité, nous avons comparé leurs résultats à l'aide de l'indice de similarité de Jaccard. Dans notre cas, l'indice de Jaccard entre deux méthodes correspond au ratio entre le nombre de sujets classés correctement par les deux classifieurs et le nombre de sujets classés correctement par au moins un des deux classifieurs. Les résultats sont présentés dans la figures 2.11. Toutes les méthodes ont des résultats fortement similaires (indice de Jaccard supérieur à 0.6) et la plupart d'entre elles sont très fortement similaires. Les méthodes utilisant l'hippocampe uniquement sont celles qui se différencient le plus des autres méthodes.

Nous avons alors considéré une combinaison de trois approches, une de chaque catégorie : *Voxel-Direct-D-gm*, *Thickness-Atlas* et *Hippo-Volume-S*. Une manière de combiner ces différentes approches est d'utiliser comme noyau une combinaison linéaire convexe des noyaux issus de chacune des approches. La difficulté est de trouver la combinaison linéaire optimale des trois noyaux. Ce problème d'apprentissage à la fois de la combinaison linéaire convexe optimale des noyaux et pour cette combinaison de l'hyperplan séparateur optimal (*optimal margin hyperplane* - OMH) est appelé *Multiple Kernel Learning* (MKL) [Lanckriet et al., 2004 ; Bach et al., 2004 ; Sonnenburg et al., 2006]. Nous avons utilisé le logiciel SimpleMKL de Rakotomamonjy et al. [2008] pour le MKL. Nous avons testé les quatre combinaisons possibles. Pour pouvoir interpréter les coefficients des combinaisons linéaires obtenues, les noyaux sont préalablement normalisés par la trace de la matrice de Gram de l'ensemble d'apprentissage.

Remarque : Pour la méthode *Hippo-Volume-S*, le classifieur à fenêtre de Parzen a été remplacé par un SVM linéaire pour cette analyse.

Aucune de ces quatre combinaisons n'améliore le taux de classification dans la comparaison *CN vs AD*. Seulement la combinaison de *Hippo-Volume-S* et de *Thickness-Atlas* améliore

Table 2.9 : Résultats de la classification *CN vs AD*.

	Méthode	Sens	Spec	VPP	VPN	McNemar
1.1.1 a	*Voxel-Direct-D-gm*	81%	95%	93%	86%	$p < 0.0001$
1.1.1 b	*Voxel-Direct-D-all*	68%	98%	96%	78%	$p < 0.0001$
1.1.2 a	*Voxel-Direct-S-gm*	72%	89%	84%	79%	$p < 0.0001$
1.1.2 b	*Voxel-Direct-S-all*	65%	88%	81%	75%	$p < 0.0001$
1.2.1 a	*Voxel-Direct_VOI-D-gm*	71%	95%	92%	79%	$p < 0.0001$
1.2.1 b	*Voxel-Direct_VOI-D-all*	65%	95%	92%	76%	$p < 0.0001$
1.2.2 a	*Voxel-Direct_VOI-S-gm*	65%	91%	86%	76%	$p < 0.0001$
1.2.2 b	*Voxel-Direct_VOI-S-all*	59%	81%	73%	70%	$p = 0.0012$
1.3.1 a	*Voxel-STAND-D-gm*	69%	90%	85%	78%	$p < 0.0001$
1.3.1 b	*Voxel-STAND-D-all*	71%	91%	87%	79%	$p < 0.0001$
1.3.2 a	*Voxel-STAND-S-gm*	75%	91%	88%	81%	$p < 0.0001$
1.3.2 b	*Voxel-STAND-S-all*	75%	86%	82%	80%	$p < 0.0001$
1.3.3 a	*Voxel-STAND-Sc-gm*	72%	91%	88%	80%	$p < 0.0001$
1.3.3 b	*Voxel-STAND-Sc-all*	71%	91%	87%	79%	$p < 0.0001$
1.4.1 a	*Voxel-Atlas-D-gm*	78%	93%	90%	83%	$p < 0.0001$
1.4.1 b	*Voxel-Atlas-D-all*	81%	90%	87%	85%	$p < 0.0001$
1.4.2 a	*Voxel-Atlas-S-gm*	75%	93%	89%	82%	$p < 0.0001$
1.4.2 b	*Voxel-Atlas-S-all*	74%	93%	89%	81%	$p < 0.0001$
1.5.1 a	*Voxel-COMPARE-D-gm*	82%	89%	86%	86%	$p < 0.0001$
1.5.1 b	*Voxel-COMPARE-D-all*	69%	81%	76%	76%	$p < 0.0001$
1.5.2 a	*Voxel-COMPARE-S-gm*	66%	86%	80%	75%	$p < 0.0001$
1.5.2 b	*Voxel-COMPARE-S-all*	72%	91%	88%	80%	$p < 0.0001$
2.1	*Thickness-Direct*	74%	90%	86%	80%	$p < 0.0001$
2.2	*Thickness-Atlas*	79%	90%	87%	84%	$p < 0.0001$
2.3	*Thickness-ROI*	69%	94%	90%	78%	$p < 0.0001$
3.1.1	*Hippo-Volume-F*	63%	80%	73%	72%	$p = 0.0007$
3.1.2	*Hippo-Volume-S*	71%	77%	72%	76%	$p = 0.0006$
3.2	*Hippo-Shape*	69%	84%	78%	76%	$p < 0.0001$

TABLE 2.10 : Résultats de la classification *CN vs* MCI_c.

	Méthode	Sens	Spec	VPP	VPN	McNemar
1.1.1 a	*Voxel-Direct-D-gm*	57%	96%	88%	83%	$p = 0.00052$
1.1.1 b	*Voxel-Direct-D-all*	49%	91%	72%	80%	$p = 0.046$
1.1.2 a	*Voxel-Direct-S-gm*	32%	96%	80%	76%	$p = 0.039$
1.1.2 b	*Voxel-Direct-S-all*	41%	94%	75%	78%	$p = 0.044$
1.2.1 a	*Voxel-Direct_VOI-D-gm*	54%	95%	83%	82%	$p = 0.0022$
1.2.1 b	*Voxel-Direct_VOI-D-all*	41%	96%	83%	78%	$p = 0.0095$
1.2.2 a	*Voxel-Direct_VOI-S-gm*	32%	88%	55%	74%	$p = 0.83$
1.2.2 b	*Voxel-Direct_VOI-S-all*	22%	99%	89%	73%	$p = 0.046$
1.3.1 a	*Voxel-STAND-D-gm*	73%	85%	69%	87%	$p = 0.025$
1.3.1 b	*Voxel-STAND-D-all*	65%	93%	80%	85%	$p = 0.0019$
1.3.2 a	*Voxel-STAND-S-gm*	59%	86%	67%	82%	$p = 0.082$
1.3.2 b	*Voxel-STAND-S-all*	49%	93%	75%	80%	$p = 0.025$
1.3.3 a	*Voxel-STAND-Sc-gm*	62%	85%	66%	83%	$p = 0.091$
1.3.3 b	*Voxel-STAND-Sc-all*	57%	90%	72%	82%	$p = 0.026$
1.4.1 a	*Voxel-Atlas-D-gm*	65%	80%	60%	83%	$p = 0.27$
1.4.1 b	*Voxel-Atlas-D-all*	54%	91%	74%	81%	$p = 0.021$
1.4.2 a	*Voxel-Atlas-S-gm*	68%	95%	86%	87%	$p = 0.00020$
1.4.2 b	*Voxel-Atlas-S-all*	59%	94%	81%	84%	$p = 0.0021$
1.5.1 a	*Voxel-COMPARE-D-gm*	49%	81%	55%	78%	$p = 0.73$
1.5.1 b	*Voxel-COMPARE-D-all*	51%	85%	61%	79%	$p = 0.28$
1.5.2 a	*Voxel-COMPARE-S-gm*	49%	78%	50%	77%	$p = 0.87$
1.5.2 b	*Voxel-COMPARE-S-all*	59%	78%	55%	81%	$p = 0.64$
2.1	*Thickness-Direct*	54%	96%	87%	82%	$p = 0.00084$
2.2	*Thickness-Atlas*	57%	93%	78%	82%	$p = 0.0071$
2.3	*Thickness-ROI*	65%	94%	83%	85%	$p = 0.00083$
3.1.1	*Hippo-Volume-F*	73%	74%	56%	86%	$p = 0.47$
3.1.2	*Hippo-Volume-S*	70%	73%	54%	84%	$p = 0.67$
3.2	*Hippo-Shape*	57%	88%	68%	82%	$p = 0.072$

Table 2.11 : Classification MCInc vs MCIc.

	Méthode	Sens	Spec	VPP	VPN	McNemar
1.1.1 a	*Voxel-Direct-D-gm*	0%	100%	–	64%	$p = 1.0$
1.1.1 b	*Voxel-Direct-D-all*	0%	100%	–	64%	$p = 1.0$
1.1.2 a	*Voxel-Direct-S-gm*	0%	100%	–	64%	$p = 1.0$
1.1.2 b	*Voxel-Direct-S-all*	0%	100%	–	64%	$p = 1.0$
1.2.1 a	*Voxel-Direct_VOI-D-gm*	43%	70%	44%	69%	$p = 0.62$
1.2.1 b	*Voxel-Direct_VOI-D-all*	0%	100%	–	64%	$p = 1.0$
1.2.2 a	*Voxel-Direct_VOI-S-gm*	0%	100%	–	64%	$p = 1.0$
1.2.2 b	*Voxel-Direct_VOI-S-all*	0%	100%	–	64%	$p = 1.0$
1.3.1 a	*Voxel-STAND-D-gm*	57%	78%	58%	76%	$p = 0.40$
1.3.1 b	*Voxel-STAND-D-all*	0%	100%	–	64%	$p = 1.0$
1.3.2 a	*Voxel-STAND-S-gm*	22%	91%	57%	68%	$p = 0.79$
1.3.2 b	*Voxel-STAND-S-all*	51%	79%	58%	75%	$p = 0.49$
1.3.3 a	*Voxel-STAND-Sc-gm*	35%	70%	39%	66%	$p = 0.30$
1.3.3 b	*Voxel-STAND-Sc-all*	41%	72%	44%	69%	$p = 0.61$
1.4.1 a	*Voxel-Atlas-D-gm*	0%	100%	–	64%	$p = 1.0$
1.4.1 b	*Voxel-Atlas-D-all*	0%	100%	–	64%	$p = 1.0$
1.4.2 a	*Voxel-Atlas-S-gm*	0%	100%	–	64%	$p = 1.0$
1.4.2 b	*Voxel-Atlas-S-all*	0%	100%	–	64%	$p = 1.0$
1.5.1 a	*Voxel-COMPARE-D-gm*	62%	67%	51%	76%	$p = 1.0$
1.5.1 b	*Voxel-COMPARE-D-all*	54%	78%	57%	75%	$p = 0.50$
1.5.2 a	*Voxel-COMPARE-S-gm*	32%	82%	50%	69%	$p = 0.84$
1.5.2 b	*Voxel-COMPARE-S-all*	51%	72%	50%	73%	$p = 0.87$
2.1	*Thickness-Direct*	32%	91%	67%	71%	$p = 0.24$
2.2	*Thickness-Atlas*	27%	85%	50%	68%	$p = 0.82$
2.3	*Thickness-ROI*	24%	82%	43%	66%	$p = 0.66$
3.1.1	*Hippo-Volume-F*	70%	61%	50%	79%	$p = 0.89$
3.1.2	*Hippo-Volume-S*	62%	69%	52%	77%	$p = 0.88$
3.2	*Hippo-Shape*	0%	100%	–	64%	$p = 1.0$

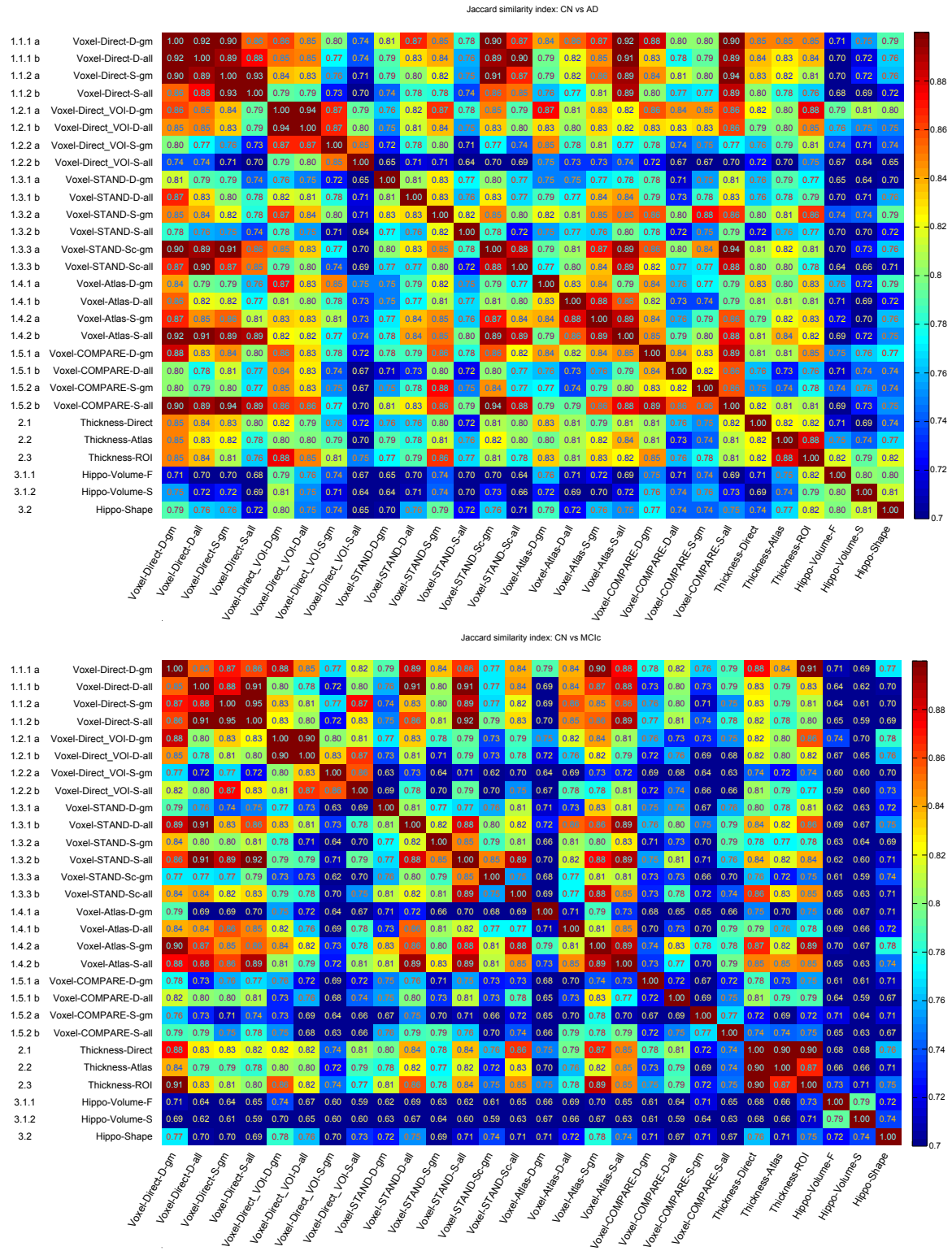


FIGURE 2.11 : Coefficient de similarité de Jaccard entre les différentes méthodes de classification pour les comparaisons *CN vs AD* et *CN vs MCI$_c$*. Le coefficient similarité de Jaccard entre deux méthodes est le nombre de sujets classés correctement par les deux méthodes divisé par le nombre de sujets classés correctement par au moins l'une des deux.

légèrement les performances de classification pour la comparaison *CN vs* MCI_c. La sensibilité est alors de 76% et la spécificité de 85%. Les coefficients de la combinaison linéaire optimale sont alors 0.057 et 0.943 pour les noyaux respectivement de *Hippo-Volume-S* et de *Thickness-Atlas*. Cette combinaison de classifieurs atteint une sensibilité de 43% et une spécificité de 83% pour la comparaison MCI_{nc} *vs* MCI_c. Les coefficients de la combinaison linéaire optimale correspondante sont 0.030 et 0.970.

2.5.3 Influence des prétraitements

2.5.3.1 Le recalage

Afin d'évaluer l'impact de l'étape de recalage sur les performances de classification, nous avons testé les différentes méthodes *Voxel* présentées dans la section 2.3.1 avec le recalage de la segmentation unifiée de SPM5 et également avec le recalage DARTEL décrit précédemment (2.3.1). L'influence de l'étape de recalage sur les performances de classification est illustrée par la figure 2.12. Les taux de classification obtenus lors de la comparaison MCI_{nc} *vs* MCI_c sont trop faibles pour pouvoir évaluer l'impact de l'étape de recalage. Pour cette raison nous n'avons pas considéré cette comparaison dans les analyses. L'utilisation du recalage difféomorphique DARTEL améliore de façon significative les résultats dans six cas sur 20 ($p < 0.05$). Cela détériore en revanche les résultats dans deux cas. D'après les tableaux 2.9, 2.10 et 2.11, l'utilisation d'un *template* généré à partir de la population d'étude n'améliore pas les performances de classification de *Voxel-STAND*.

2.5.3.2 Les cartes de WM et de CSF

Nous avons également comparé les performances de classifications obtenues en utilisant uniquement les cartes de substance grise avec celles obtenues avec l'ensemble des cartes de GM, WM et CSF. Les résultats de cette comparaison sont présentés par la figure 2.12. L'utilisation des trois cartes mène à des résultats significativement moins bons dans deux cas sur 20 (*Voxel-Direct_VOI-S* et *Voxel-COMPARE-D*). On n'obtient jamais de résultats significativement meilleurs.

2.5.4 Influence de l'âge et du genre sur la classification

Nous avons également regardé si l'âge des sujets influence les résultats. Pour cela, nous avons calculé l'âge moyen des vrais positifs (TP), des faux positifs (FP), des vrais négatifs (TN) et celui des faux négatifs (FN). Nous avons trouvé que les faux positifs, étaient souvent plus âgés que les vrais négatifs. En d'autres termes les témoins âgés étaient plus souvent mal classés que les

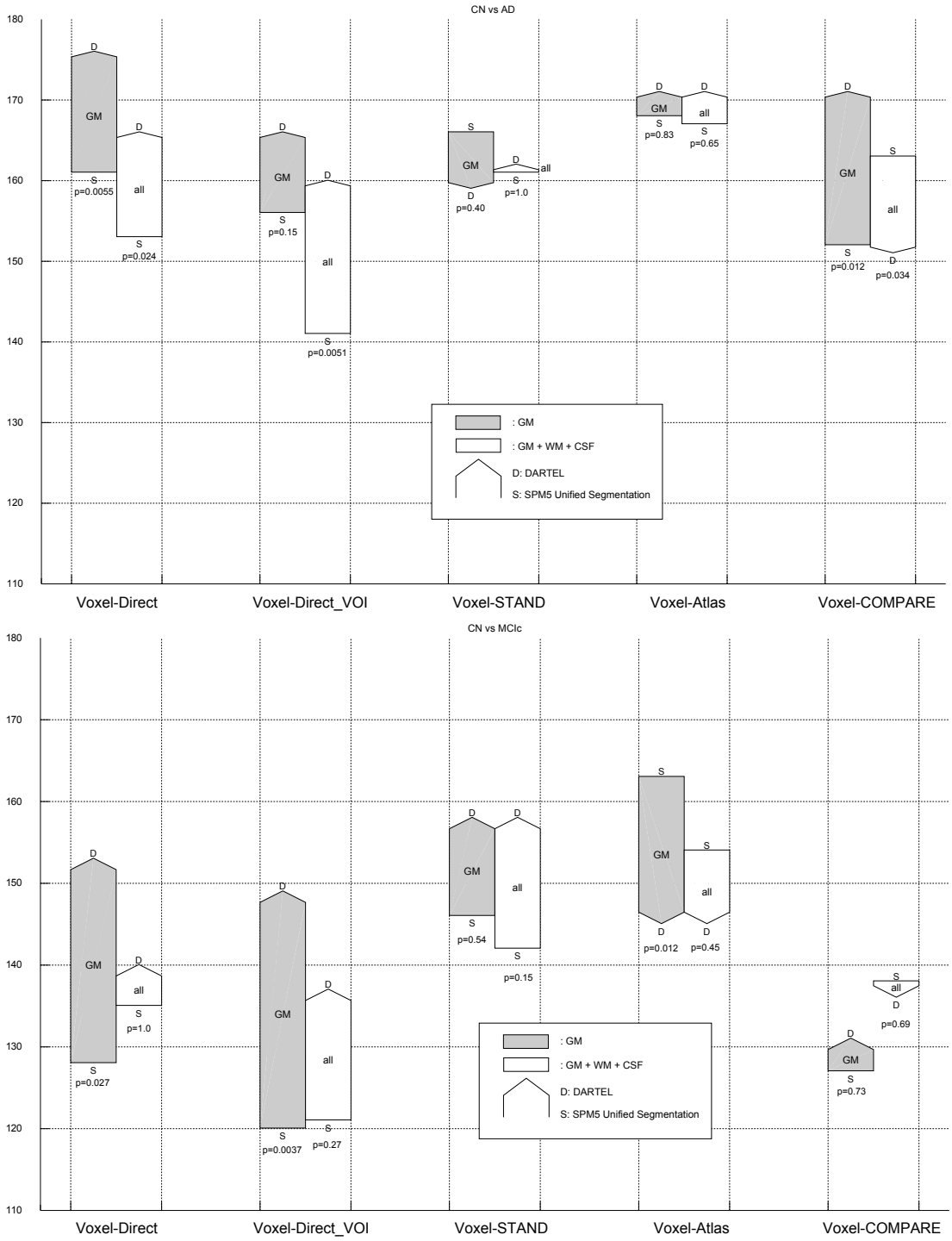


FIGURE 2.12 : Influence des étapes de prétraitements pour les comparaison *CN vs AD* et *CN vs MCI$_c$*. Sont représentées les sommes des sensibilités et spécificités. La pointe des flèches correspond aux résultats obtenus avec DARTEL tandis que l'autre extrémité correspond aux résultats obtenus avec la segmentation unifiée de SPM5. La couleur des flèches indique les cartes utilisées : les flèches sont grises lorsque seulement les cartes de substance grise sont utilisées et blanches lorsque les cartes de substance blanche et de CSF sont également utilisées. Les p-values indiquées sont celles obtenues avec le test du χ^2 de McNemar. Elles quantifient les différences entre les performances de classification entre SPM5 et DARTEL.

témoins jeunes. Plus précisément, ce fut le cas pour 25 des 28 approches dans la comparaison *CN vs AD* et pour 24 approches sur 28 pour la comparaison *CN vs* MCI_c. Inversement, les faux négatifs étaient souvent plus jeunes que les vrais positifs ; cela signifie que les jeunes patients sont plus souvent mal classés. Ce fut le cas pour 26 approches sur 28 pour la comparaison *CN vs AD* et pour toutes les approches pour la comparaison *CN vs* MCI_c. Le nombre de sujets mal classés étaient trop faible pour pouvoir avoir des différences statistiquement significatives. Cependant, le fait que ces différences soient présentes dans la plus grande majorité des comparaisons laisse penser que ce n'est pas dû au hasard. Nous avons également analysé l'influence du genre sur les résultats de la classification. Nous n'avons trouvé aucune différence.

2.5.5 Temps de calcul

Les calculs ont été effectués avec un processeur de 3.6 GHz avec 2 Gb de RAM. Le tableau 2.12 donne, pour chaque méthode, l'ordre de grandeur du temps de calcul. Nous avons séparé le processus complet en trois grandes étapes :

1. le calcul des caractéristiques (segmentation et recalage),
2. la construction du classifieur (optimisation des hyperparamètres et apprentissage)
3. la classification d'un nouveau sujet.

L'ordre de grandeur du temps de calcul de l'étape de segmentation et de recalage est de 10 minutes par sujet pour la segmentation unifiée de SPM5 et d'une heure pour le recalage utilisant DARTEL. Le calcul de l'épaisseur corticale et le recalage des cartes d'épaisseur corticale dans un espace commun avec FreeSurfer prennent à peu près une journée par sujet. La segmentation des hippocampes d'un sujet dure quelques minutes et la décomposition en SPHARM dure approximativement une heure.

L'estimation des hyperparamètres et la phase d'apprentissage durent de quelques minutes à plusieurs semaines pour des méthodes telles que *Voxel-STAND* et *Voxel-COMPARE*. Une fois ces étapes réalisées, la classification d'un nouveau sujet dure au plus quelques minutes.

2.5.6 Description des hyperplans séparateurs

La fonction de classification obtenue avec un SVM linéaire est, à une constante additive près, le signe du produit scalaire du vecteur des caractéristiques $\mathbf{x}_s$ du sujet à classer s avec $\mathbf{w}$, un vecteur orthogonal à l'hyperplan séparateur optimal (cf section 1.2.1). Par conséquent, si la i^{ieme} composante w_i du vecteur $\mathbf{w}$ est petite, la i^{ieme} caractéristique va avoir une faible

TABLE 2.12 : Ordre de grandeur du temps de calcul (en minutes, heures, jours et semaines) pour chacune des différentes méthodes comparées dans ce chapitre. Pour chaque méthode, on distingue trois phases : (i) le calcul des caractéristiques (segmentation, recalage, décomposition), (ii) l'optimisation des hyperparamètres et l'apprentissage du classifieur , (iii) la classification d'un nouveau sujet. Les calculs sont effectués avec un processeur de 3.6 GHz avec 2 Gb de RAM.

	Méthode	**Segmentation Recalage**	**Apprentissage**	**Test**
1.1.1 a	*Voxel-Direct-D-gm*	Heure(s) per subject	Minute(s)	Minute(s)
1.1.1 b	*Voxel-Direct-D-all*	Heure(s) per subject	Minute(s)	Minute(s)
1.1.2 a	*Voxel-Direct-S-gm*	10 Minutes per subject	Minute(s)	Minute(s)
1.1.2 b	*Voxel-Direct-S-all*	10 Minutes per subject	Minute(s)	Minute(s)
1.2.1 a	*Voxel-Direct_VOI-D-gm*	Heure(s) per subject	Minute(s)	Minute(s)
1.2.1 b	*Voxel-Direct_VOI-D-all*	Heure(s) per subject	Minute(s)	Minute(s)
1.2.2 a	*Voxel-Direct_VOI-S-gm*	10 Minutes per subject	Minute(s)	Minute(s)
1.2.2 b	*Voxel-Direct_VOI-S-all*	10 Minutes per subject	Minute(s)	Minute(s)
1.3.1 a	*Voxel-STAND-D-gm*	Heure(s) per subject	Jour(s)	Heure(s)
1.3.1 b	*Voxel-STAND-D-all*	Heure(s) per subject	Week(s)	Heure(s)
1.3.2 a	*Voxel-STAND-S-gm*	10 Minutes per subject	Jour(s)	Heure(s)
1.3.2 b	*Voxel-STAND-S-all*	10 Minutes per subject	Week(s)	Heure(s)
1.3.3 a	*Voxel-STAND-Sc-gm*	20 Minutes per subject	Jour(s)	Heure(s)
1.3.3 b	*Voxel-STAND-Sc-all*	20 Minutes per subject	Week(s)	Heure(s)
1.4.1 a	*Voxel-Atlas-D-gm*	Heure(s) per subject	Minute(s)	Minute(s)
1.4.1 b	*Voxel-Atlas-D-all*	Heure(s) per subject	Minute(s)	Minute(s)
1.4.2 a	*Voxel-Atlas-S-gm*	10 Minutes per subject	Minute(s)	Minute(s)
1.4.2 b	*Voxel-Atlas-S-all*	10 Minutes per subject	Minute(s)	Minute(s)
1.5.1 a	*Voxel-COMPARE-D-gm*	Heure(s) per subject	Week(s)	Heure(s)
1.5.1 b	*Voxel-COMPARE-D-all*	Heure(s) per subject	Week(s)	Heure(s)
1.5.2 a	*Voxel-COMPARE-S-gm*	10 Minutes per subject	Week(s)	Heure(s)
1.5.2 b	*Voxel-COMPARE-S-all*	10 Minutes per subject	Week(s)	Heure(s)
2.1	*Thickness-Direct*	Jour(s) per subject	Minute(s)	Minute(s)
2.2	*Thickness-Atlas*	Jour(s) per subject	Minute(s)	Minute(s)
2.3	*Thickness-ROI*	Jour(s) per subject	Minute(s)	Seconds
3.1.1	*Hippo-Volume-F*	Jour(s) per subject	Minute(s)	Seconds
3.1.2	*Hippo-Volume-S*	10 Minutes per subject	Minute(s)	Seconds
3.2	*Hippo-Shape*	Heure(s) per subject	Minute(s)	Seconds

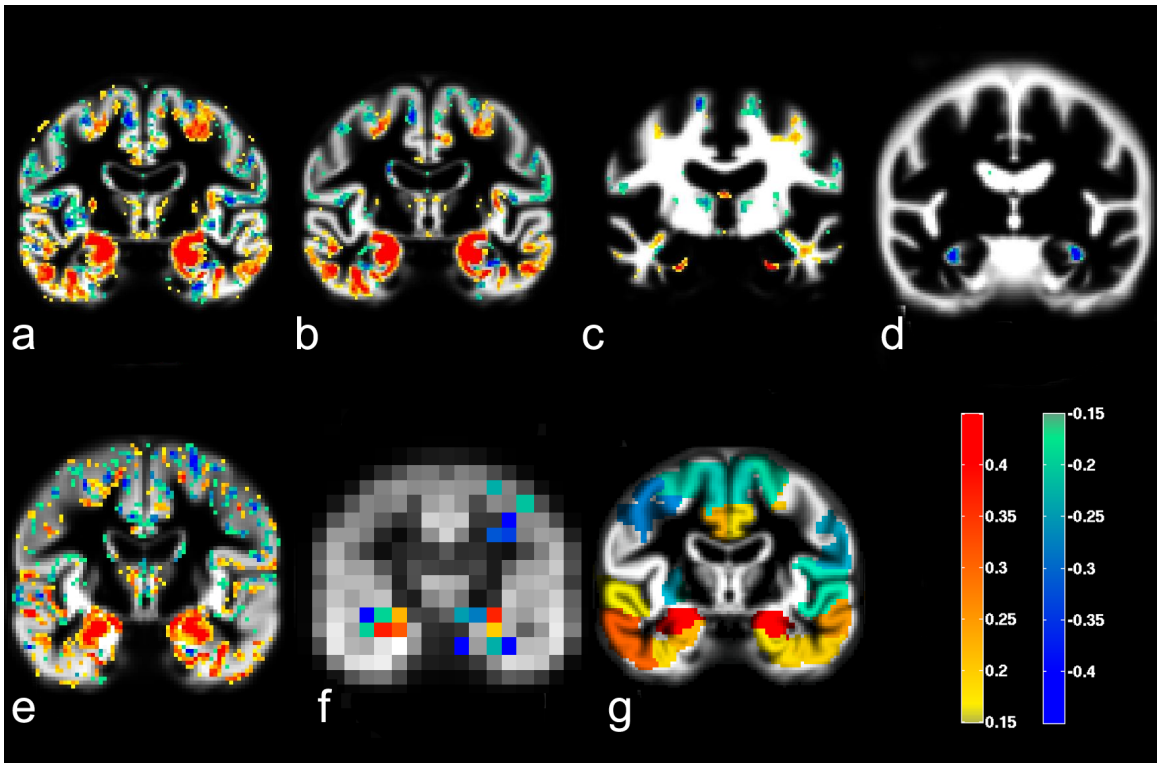


FIGURE 2.13 : Coefficients de l'hyperplan séparateur optimal pour la comparaison *CN vs AD* avec : *Voxel-Direct-D-gm* (a), *Voxel-Direct-D-all* (b-d), *Voxel-Direct-S-gm* (e), *Voxel-STAND-D-gm* (f) et *Voxel-Atlas-D-gm* (g). Les images représentent les coefficients normalisés de l'OHM superposé avec les cartes de probabilité des tissus. La coupe coronale correspond à $y = 9$ mm dans l'espace MNI. Pour des raisons de visualisation, seuls les coefficients supérieurs à 0.15 en valeur absolue sont représentés. Dans les régions en couleurs chaudes, une atrophie augmente la probabilité d'être classé comme AD ou MCI_c. En ce qui concerne les régions en couleur froide, c'est l'inverse.

influence dans la fonction de classification. Inversement, si w_i est grand, la i^{ieme} caractéristique va avoir un rôle important dans la classification.

Quand les données sont les intensités des voxels d'une image, chaque composante de $\mathbf{w}$ correspond également à un voxel. On peut donc représenter $\mathbf{w}$ par une image. De la même manière, lorsque les données sont les épaisseurs corticales mesurées à chaque nœud du maillage du cortex, $\mathbf{w}$ peut être représenté sur la surface corticale. Les valeurs des coefficients des OMH (*optimal margin hyperplane*) sont représentées par les figures 2.13 à 2.16. Par abus de langage, nous appellerons OMH les coefficients de l'OMH dans la suite du manuscrit. Ces représentations permettent une analyse qualitative des caractéristiques utilisées par le classifieur.

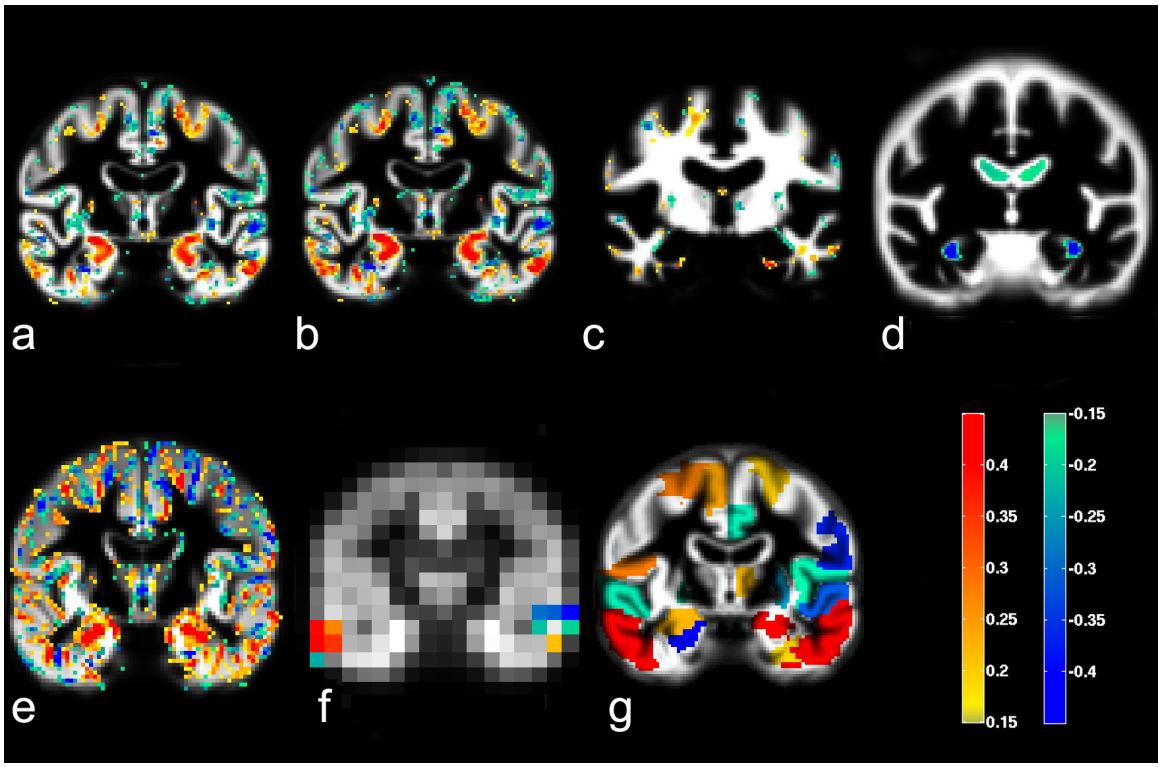


FIGURE 2.14 : Coefficients de l'hyperplan séparateur optimal pour la comparaison *CN vs* MCI_c avec : *Voxel-Direct-D-gm* (a), *Voxel-Direct-D-all* (b-d), *Voxel-Direct-S-gm* (e), *Voxel-STAND-D-gm* (f) et *Voxel-Atlas-D-gm* (g) (Le lecteur est prié de se référer à la figure 2.13 pour une description complète de la figure).

Les figures 2.13 et 2.14 montrent les OHM respectivement pour les comparaisons *CN vs AD* et *CN vs* MCI_c pour les méthodes *Voxel*. De manière générale, les distributions des régions discriminantes obtenues avec *CN vs AD* et *CN vs* MCI_c sont similaires. Pour la méthode *Voxel-Direct-D-gm*, ce sont principalement le lobe temporal médian (hippocampe, amygdale et gyrus parahippocampique), les gyri temporaux inférieurs et moyens du lobe temporal, le gyrus cingulaire postérieur et le gyrus frontal moyen postérieur. Dans une moindre mesure, nous trouvons le lobule pariétal inférieur, le gyrus supramarginal, le gyrus fusiforme, le gyrus cingulaire moyen et le thalamus. Quand les trois cartes GM, WM et CSF sont utilisées, les OHM des cartes de WM et de CSF donnent des informations qui sont déjà présentes dans l'OMH de GM (par exemple, l'élargissement des ventricules est directement lié à l'atrophie de GM). Quand la segmentation unifiée de SPM5 est utilisée à la place de DARTEL, les cartes sont beaucoup plus bruitées, les voxels utilisés pour la classification sont éparpillés et non groupés en régions anatomiques, excepté dans la région du lobe temporal médian. En ce qui concerne les OMH obtenus avec *Voxel-Atlas*, les régions discriminantes sont principalement l'hippocampe, l'amygdale, le gyrus parahippocampique, le cingulum, les gyri temporaux inférieurs et moyens

ainsi que le gyrus frontal supérieur et le gyrus frontal inférieur.

En ce qui concerne l' approche surfacique *Thickness-Atlas*, les régions sont très similaires (figure 2.15) à celles trouvée par les méthodes volumiques . Les régions obtenues avec *Thickness-Direct* (figure 2.16) sont quant à elles beaucoup moins nombreuses : le cortex entorhinal, le gyrus parahippocampique et dans une moindre mesure la partie latérale du lobe temporal, le lobule pariétal inférieur et quelques aires préfrontales.

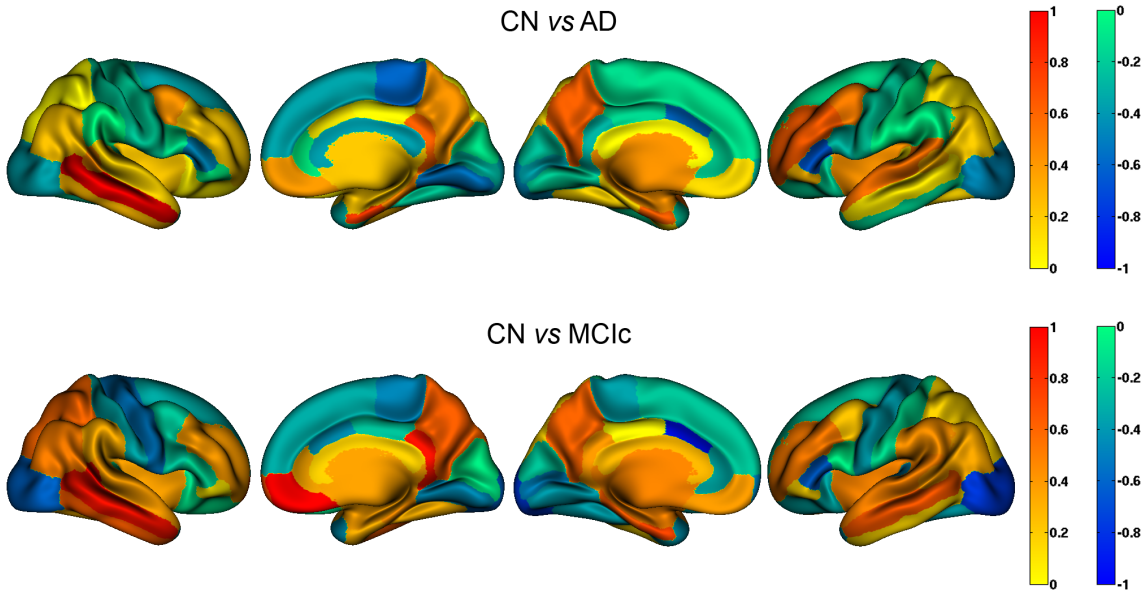


Figure 2.15 : Coefficients de l'OMH avec la méthode *Thickness-Atlas*. La ligne du haut correspond à la comparaison *CN vs AD*; la ligne du bas correspond à la comparaison *CN vs* MCI_c.

2.5.7 Hyperparamètres optimaux

Pour chaque approche, les valeurs optimales des hyperparamètres sont résumées dans les tableaux 2.13, 2.14 et 2.15.

Remarque : La méthode *Hippo-Volume* n'a pas d'hyperparamètre.

2.6 Discussion

Dans ce chapitre nous avons comparés différentes méthodes de classification de patients atteints de la maladie d'Alzheimer et de patients souffrants de troubles cognitifs légers à partir

TABLE 2.13 : Valeurs optimales des hyperparamètres (*CN vs AD*).

	Méthode	*CN vs AD* Hyperparamètres optimaux
1.1.1 a	*Voxel-Direct-D-gm*	$\log_{10}(C) = -3.5$
1.1.1 b	*Voxel-Direct-D-all*	$\log_{10}(C) = -4.5$
1.1.2 a	*Voxel-Direct-S-gm*	$\log_{10}(C) = -3.5$
1.1.2 b	*Voxel-Direct-S-all*	$\log_{10}(C) = -4.0$
1.2.1 a	*Voxel-Direct_VOI-D-gm*	$\log_{10}(C) = -1.5$
1.2.1 b	*Voxel-Direct_VOI-D-all*	$\log_{10}(C) = -2.0$
1.2.2 a	*Voxel-Direct_VOI-S-gm*	$\log_{10}(C) = -1.5$
1.2.2 b	*Voxel-Direct_VOI-S-all*	$\log_{10}(C) = -1.0$
1.3.1 a	*Voxel-STAND-D-gm*	$\log_{10}(C) = 0.5; t = 0.88$
1.3.1 b	*Voxel-STAND-D-all*	$\log_{10}(C) = -2.0; t = 0.90$
1.3.2 a	*Voxel-STAND-S-gm*	$\log_{10}(C) = -1.5; t = 0.84$
1.3.2 b	*Voxel-STAND-S-all*	$\log_{10}(C) = -0.5; t = 0.90$
1.3.3 a	*Voxel-STAND-Sc-gm*	$\log_{10}(C) = -3.0; t = 0.54$
1.3.3 b	*Voxel-STAND-Sc-all*	$\log_{10}(C) = -4.0; t = 0.06$
1.4.1 a	*Voxel-Atlas-D-gm*	$\log_{10}(C) = 0.5$
1.4.1 b	*Voxel-Atlas-D-all*	$\log_{10}(C) = 0.5$
1.4.2 a	*Voxel-Atlas-S-gm*	$\log_{10}(C) = 1.0$
1.4.2 b	*Voxel-Atlas-S-all*	$\log_{10}(C) = 0.5$
1.5.1 a	*Voxel-COMPARE-D-gm*	$\log_{10}(C) = 2.5; \sigma = 700; n = 20$
1.5.1 b	*Voxel-COMPARE-D-all*	$\log_{10}(C) = 2.0; \sigma = 300; n = 16$
1.5.2 a	*Voxel-COMPARE-S-gm*	$\log_{10}(C) = 1.5; \sigma = 500; n = 6$
1.5.2 b	*Voxel-COMPARE-S-all*	$\log_{10}(C) = 2.0; \sigma = 600; n = 98$
2.1	*Thickness-Direct*	$\log_{10}(C) = -5.0$
2.2	*Thickness-Atlas*	$\log_{10}(C) = 0.0$
2.3	*Thickness-ROI*	$-$
3.1.1	*Hippo-Volume-F*	$-$
3.1.2	*Hippo-Volume-S*	$-$
3.2	*Hippo-Shape*	$\log_{10}(C) = -1.5$

Table 2.14 : Valeurs optimales des hyperparamètres (*CN vs MCI$_c$*).

	Méthode	*CN vs MCI_c* Hyperparamètres optimaux
1.1.1 a	*Voxel-Direct-D-gm*	$\log_{10}(C) = -3.5$
1.1.1 b	*Voxel-Direct-D-all*	$\log_{10}(C) = -4.0$
1.1.2 a	*Voxel-Direct-S-gm*	$\log_{10}(C) = -3.5$
1.1.2 b	*Voxel-Direct-S-all*	$\log_{10}(C) = -4.0$
1.2.1 a	*Voxel-Direct_VOI-D-gm*	$\log_{10}(C) = -1.0$
1.2.1 b	*Voxel-Direct_VOI-D-all*	$\log_{10}(C) = -1.5$
1.2.2 a	*Voxel-Direct_VOI-S-gm*	$\log_{10}(C) = -1.0$
1.2.2 b	*Voxel-Direct_VOI-S-all*	$\log_{10}(C) = -2.0$
1.3.1 a	*Voxel-STAND-D-gm*	$\log_{10}(C) = -1.5; t = 0.80$
1.3.1 b	*Voxel-STAND-D-all*	$\log_{10}(C) = -2.5; t = 0.64$
1.3.2 a	*Voxel-STAND-S-gm*	$\log_{10}(C) = -2.0; t = 0.44$
1.3.2 b	*Voxel-STAND-S-all*	$\log_{10}(C) = -3.5; t = 0.22$
1.3.3 a	*Voxel-STAND-Sc-gm*	$\log_{10}(C) = -1.5; t = 0.54$
1.3.3 b	*Voxel-STAND-Sc-all*	$\log_{10}(C) = -3.5; t = 0.18$
1.4.1 a	*Voxel-Atlas-D-gm*	$\log_{10}(C) = 1.5$
1.4.1 b	*Voxel-Atlas-D-all*	$\log_{10}(C) = 0.5$
1.4.2 a	*Voxel-Atlas-S-gm*	$\log_{10}(C) = 1.0$
1.4.2 b	*Voxel-Atlas-S-all*	$\log_{10}(C) = 0.5$
1.5.1 a	*Voxel-COMPARE-D-gm*	$\log_{10}(C) = 1.5; \sigma = 500; n = 5$
1.5.1 b	*Voxel-COMPARE-D-all*	$\log_{10}(C) = 2.0; \sigma = 300; n = 20$
1.5.2 a	*Voxel-COMPARE-S-gm*	$\log_{10}(C) = 1.0; \sigma = 200; n = 117$
1.5.2 b	*Voxel-COMPARE-S-all*	$\log_{10}(C) = 2.0; \sigma = 1000; n = 5$
2.1	*Thickness-Direct*	$\log_{10}(C) = -5.0$
2.2	*Thickness-Atlas*	$\log_{10}(C) = -0.5$
2.3	*Thickness-ROI*	–
3.1.1	*Hippo-Volume-F*	–
3.1.2	*Hippo-Volume-S*	–
3.2	*Hippo-Shape*	$\log_{10}(C) = -1.0$

TABLE 2.15 : Valeurs optimales des hyperparamètres (MCI_{nc} *vs* MCI_c).

	Méthode	*MCI_nc vs MCI_c* **Hyperparamètres optimaux**
1.1.1 a	*Voxel-Direct-D-gm*	$\log_{10}(C) = -5.0$
1.1.1 b	*Voxel-Direct-D-all*	$\log_{10}(C) = -5.0$
1.1.2 a	*Voxel-Direct-S-gm*	$\log_{10}(C) = -5.0$
1.1.2 b	*Voxel-Direct-S-all*	$\log_{10}(C) = -5.0$
1.2.1 a	*Voxel-Direct_VOI-D-gm*	$\log_{10}(C) = 0.0$
1.2.1 b	*Voxel-Direct_VOI-D-all*	$\log_{10}(C) = -5.0$
1.2.2 a	*Voxel-Direct_VOI-S-gm*	$\log_{10}(C) = -5.0$
1.2.2 b	*Voxel-Direct_VOI-S-all*	$\log_{10}(C) = -5.0$
1.3.1 a	*Voxel-STAND-D-gm*	$\log_{10}(C) = -2.0; t = 0.72$
1.3.1 b	*Voxel-STAND-D-all*	$\log_{10}(C) = -5.0; t = 0.10$
1.3.2 a	*Voxel-STAND-S-gm*	$\log_{10}(C) = -1.0; t = 0.90$
1.3.2 b	*Voxel-STAND-S-all*	$\log_{10}(C) = -3.5; t = 0.34$
1.3.3 a	*Voxel-STAND-Sc-gm*	$\log_{10}(C) = -1.5; t = 0.86$
1.3.3 b	*Voxel-STAND-Sc-all*	$\log_{10}(C) = -2.0; t = 0.64$
1.4.1 a	*Voxel-Atlas-D-gm*	$\log_{10}(C) = -5.0$
1.4.1 b	*Voxel-Atlas-D-all*	$\log_{10}(C) = -5.0$
1.4.2 a	*Voxel-Atlas-S-gm*	$\log_{10}(C) = -5.0$
1.4.2 b	*Voxel-Atlas-S-all*	$\log_{10}(C) = -5.0$
1.5.1 a	*Voxel-COMPARE-D-gm*	$\log_{10}(C) = 0.0; \sigma = 400; n = 4$
1.5.1 b	*Voxel-COMPARE-D-all*	$\log_{10}(C) = 2.5; \sigma = 1000; n = 89$
1.5.2 a	*Voxel-COMPARE-S-gm*	$\log_{10}(C) = 1.5; \sigma = 100; n = 128$
1.5.2 b	*Voxel-COMPARE-S-all*	$\log_{10}(C) = 1.0; \sigma = 100; n = 104$
2.1	*Thickness-Direct*	$\log_{10}(C) = -5.0$
2.2	*Thickness-Atlas*	$\log_{10}(C) = -0.5$
2.3	*Thickness-ROI*	–
3.1.1	*Hippo-Volume-F*	–
3.1.2	*Hippo-Volume-S*	–
3.2	*Hippo-Shape*	$\log_{10}(C) = -5.0$

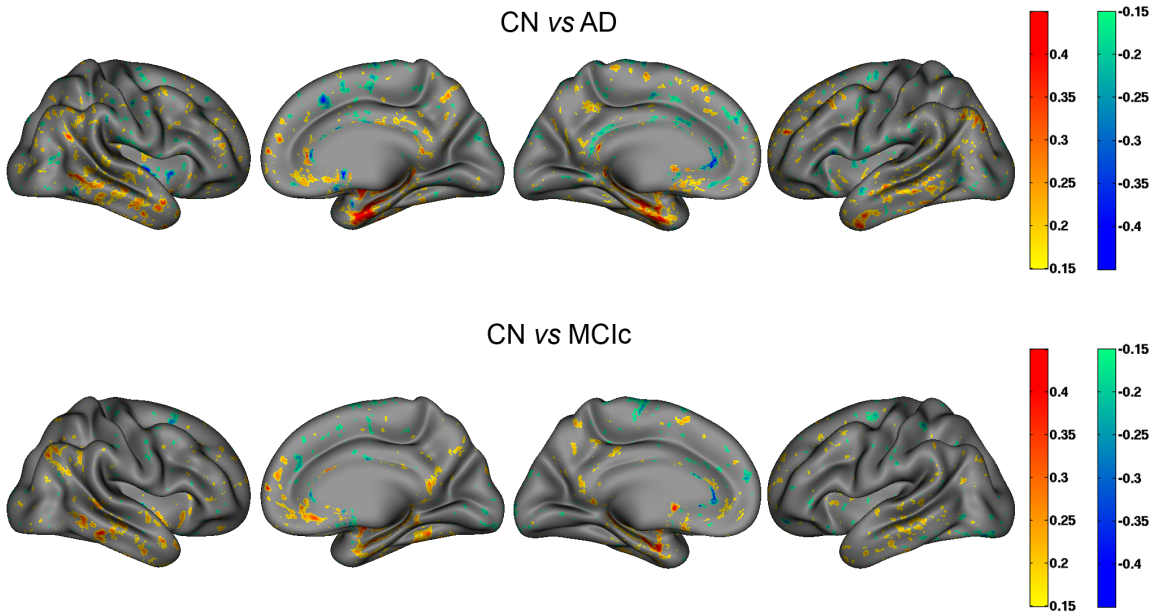


FIGURE 2.16 : Coefficients de l'OMH avec la méthode *Thickness-Direct*. La ligne du haut correspond à la comparaison *CN vs AD*; la ligne du bas correspond à la comparaison *CN vs* MCI_c.

d'IRM anatomiques pondérées en T_1. Afin d'évaluer et de comparer les performances de chaque méthode, trois expériences de classifications ont été réalisées : *CN vs AD*, *CN vs* MCI_c et MCI_{nc} *vs* MCI_c.

La population d'étude a été aléatoirement divisée en deux ensembles de même taille : un ensemble d'apprentissage et un ensemble de test. Pour chaque méthode, les valeurs optimales des hyperparamètres ont été estimées à l'aide d'une recherche en grille et d'une évaluation de type LOOCV sur l'ensemble d'apprentissage. Ces valeurs optimales ont ensuite servi à entraîner le classifieur sur le groupe d'apprentissage et à le tester sur l'ensemble test. De cette manière, nous obtenons des estimations non biaisées des performances de classification de chaque méthode.

2.6.1 Classification *AD vs CN*

Toutes les méthodes testées dans ce chapitre ont obtenu des performances de classification significativement meilleures que le hasard pour la classification de patients AD et de témoins. Toutes ces méthodes, à l'exception de *Voxel-COMPARE* et *Hippo* ont également obtenu des résultats meilleurs que le hasard dans la détection d'AD prodromaux (MCI_c). Pour la détection de patients AD, les méthodes sont très sensibles et spécifiques; en ce qui concerne la détection d'AD prodromaux la sensibilité est beaucoup plus faible.

Les résultats obtenus pour la comparaison *AD vs CN* avec les méthodes *Voxel-Atlas* et *Voxel-COMPARE* sont plus faibles que ceux reportés dans les papiers originaux ; Fan et al. [2008b] rapportent un taux de classification de 94% pour *COMPARE* et Magnin et al. [2009] rapportent une sensibilité de 92% et une spécificité de 97% pour leur méthode utilisant l'atlas AAL. Ces différences peuvent avoir différentes causes. Tout d'abord, dans les papiers originaux, les hyperparamètres ont été estimés sur l'ensemble de test. Cela peut entraîner du sur-apprentissage sur l'ensemble de test entraînant ainsi une surestimation de la sensibilité et de la spécificité. Dans notre étude, nous avons pris soin de séparer l'ensemble d'apprentissage et l'ensemble de test afin d'éviter tout biais de ce type. Ces différences de performances rapportées pourraient également provenir de la différence entre les populations d'étude utilisées (la taille de la population, l'état d'avancement de la maladie). En particulier, la base de données ADNI inclut un grand nombre de sujets avec des lésions vasculaires contrairement à la population d'étude utilisée par Magnin et al. [2009]. Enfin, ces différences peuvent également provenir de l'étape de prétraitement des images. Nous avons vu dans cette étude que l'étape de segmentation et de recalage pouvait avoir un impact important sur les résultats. Davatzikos et al. [2008c] et Fan et al. [2008a] utilisent les cartes RAVENS [Goldszal et al., 1998]. Autrement dit, leurs étapes de segmentation et de recalage sont différentes des nôtres, ce qui pourrait conduire à des résultats différents. Cependant, l'objectif de ce chapitre est de comparer différentes stratégies de classification. C'est pour cette raison que nous avons utilisé les mêmes étapes de prétraitement pour toutes les approches. Comme la plupart de ces méthodes utilisent SPM, nous avons choisi d'utiliser ce logiciel. Il est possible que d'autres algorithmes de recalage tels que HAMMER [Shen & Davatzikos, 2002] améliorent les performances de classification mais ce n'est pas l'objet de notre étude.

Les résultats obtenus avec les méthodes *Voxel-STAND* et *Voxel-Direct* sont similaires à ceux rapportés dans les papiers originaux [Vemuri et al., 2008] et [Klöppel et al., 2008b]. Ceci provient probablement du fait que Vemuri et al. [2008] ont également utilisé deux ensembles indépendants pour l'apprentissage et le test. Quant à Klöppel et al. [2008b], ils ne mentionnent pas d'optimisation d'hyperparamètres. En ce qui concerne la méthode *Thickness-ROI*, les résultats obtenus (69% en sensibilité et 94% de spécificité) sont plus faibles que ceux obtenus par Desikan et al. [2009] (100% en sensibilité et spécificité). Cela pourrait être dû au fait que dans [Desikan et al., 2009] le classifieur est entraîné sur une population différente (des patients avec un CDR=0.5) provenant d'une base de données également différente (la base de données OASIS).

Les résultats obtenus avec *Hippo-Volume* sont similaires avec ceux rapportés pour la base ADNI par une de nos précédentes études [Chupin et al., 2009a]. Les sensibilités et spécificités obtenues sont néanmoins plus faibles qu'une étude précédente avec une autre population d'étude [Colliot et al., 2008a] (84% en sensibilité et spécificité pour la classification *CN vs AD*). Cette différence peut être due à différents facteurs [Chupin et al., 2009a] :

- ADNI est une étude multicentrique tandis que les données utilisée dans [Colliot et al., 2008a] proviennent d'une même machine.

- La population utilisée dans ADNI inclut un grand nombre de patients avec des lésions vasculaires.

La petite différence que l'on observe avec les résultats rapportés dans [Chupin et al., 2009b] provient probablement de la différence dans la procédure d'estimation des performances de classification : nous utilisons deux groupes séparés à la place d'une validation LOOCV. En ce qui concerne l'analyse de forme de l'hippocampe, *Hippo-Shape*, les résultats obtenus dans cette étude sont plus faibles que dans le papier original [Gerardin et al., 2009] (86% pour la comparaison *CN vs AD*). Cela est probablement dû au petit nombre de sujets utilisés dans l'étude précédente. De plus, l'estimation des performances de classification dans le papier original utilise un LOOCV. Enfin, cela peut aussi provenir du fait que dans la comparaison de méthodes, tous les sujets ont été utilisés sans tenir aucun compte du contrôle qualité des segmentations de l'hippocampe [Chupin et al., 2009a].

À notre connaissance, le problème de classification *CN vs* MCI_c n'a été abordé que par Desikan et al. [2009], Davatzikos et al. [2008a] ; Fan et al. [2008b,a] ont effectué la classification entre les témoins et les MCI sans faire de distinction entre les converteurs et les non-converteurs. Or, les MCI ne sont pas tous des AD prodromaux ; il n'est donc pas possible de comparer leurs résultats à ceux obtenus lors de la comparaison *CN vs* MCI_c. Desikan et al. [2009] ont classé les témoins et les MCI qui ont converti dans les deux ans qui suivirent leur inclusion. Ils obtiennent un taux de classification de 91%. C'est nettement plus élevé que ce que l'on obtient avec la même méthode pour la comparaison *CN vs* MCI_c (*Thickness-ROI* : sensibilité de 65% et spécificité de 94%).

2.6.2 Prédiction de la conversion des patients MCI

Pour la prédiction de la conversion des patients MCI, aucune méthode testée dans cette étude n'obtient des performances significativement supérieures au hasard. Les meilleures performances sont obtenues avec : *Voxel-STAND* (sensibilité de 57% et spécificité de 78%), *Voxel-COMPARE* (sensibilité de 62% et spécificité de 67%) et *Hippo-Volume* (sensibilité de 62% et spécificité de 69%). Ces trois méthodes ont restreint leur analyse à une partie du cerveau. Les méthodes *Voxel-STAND* et *Voxel-COMPARE* restreignent leur analyse à l'aide d'étapes de sélection de *features*. Les régions sélectionnées par ces deux méthodes sont principalement celles du lobe temporal médian. En ce qui concerne la méthode *Hippo-Volume*, la restriction est intrinsèque à la méthode puisque par définition elle n'utilise que les hippocampes pour l'analyse.

Même avec ces trois méthodes, les performances en termes de classification restent très faibles. Cela vient très probablement du fait que le groupe des MCI non converteurs (MCI_{nc}) est un groupe très hétérogène. Certains patients de ce groupe convertissent vers la maladie d'Alzheimer peu de temps après la fin du suivi et sont en fait également des AD prodromaux tandis que d'autres patients MCI_{nc} vont longtemps rester cliniquement stables. Les méthodes de classification devraient principalement se focaliser en premier lieu sur la classification d'AD prodromaux. Ils constituent un groupe beaucoup mieux défini.

À notre connaissance, la classification MCI_{nc} *vs* MCI_c a uniquement été abordée par Misra et al. [2009] et Querbes et al. [2009]. Misra et al. [2009] ont considéré la conversion dans les 12 premiers mois qui suivirent l'inclusion tandis que Querbes et al. [2009] ont considéré la conversion dans les 24 premiers mois. Ils ont obtenu de meilleurs résultats. Ces différences ont probablement les mêmes origines que celles expliquées dans le paragraphe précédent :

- utilisation de deux groupes séparés pour l'apprentissage et le test
- utilisation d'étapes de prétraitements différentes

Querbes et al. [2009] ont également utilisé une étape de sélection de variables ; cela pourrait aussi expliquer les différences remarquées.

2.6.3 Hippocampe ou cerveau entier ?

Pour la comparaison *CN vs AD*, les méthodes utilisant le cerveau dans son ensemble ou du moins le cortex dans son ensemble sont nettement plus spécifiques (plus de 90% de spécificité) que celles utilisant uniquement l'hippocampe (spécificité comprise entre 63% et 84%). En ce qui concerne la détection d'AD prodromaux (*CN vs* MCI_c), les approches basées sur la segmentation de l'hippocampe sont compétitives avec celles utilisant le cerveau entier. Ces dernières sont vraisemblablement plus intéressantes pour l'analyse de stades avancés de la maladie. L'atrophie est alors en effet beaucoup plus étendue à ces stades. De plus, un grand nombre de sujets de la base ADNI ont des lésions vasculaires. Ces lésions sont probablement en partie prises en compte par les méthodes utilisant le cerveau dans son ensemble.

Pour les stades moins avancés, une alternative à ces approches serait de considérer un ensemble de régions sélectionnées *a priori* à la place d'une analyse du cerveau entier ou de l'hippocampe seul. Par exemple, l'approche *Thickness-ROI* obtient des résultats au moins aussi bons que l'approche « cerveau entier » pour la détection de sujets AD prodromaux.

De plus, même si les approches utilisant l'hippocampe seul obtiennent de moins bons résultats que les autres, elles restent d'un grand intérêt pour les cliniciens. En effet, elles donnent aux cliniciens des mesures directes et facilement interprétables (exemple le volume

de l'hippocampe), ce qui n'est pas le cas des approches « cerveau entier » qui basent leur classification sur des combinaisons complexes de différentes régions du cerveau.

Enfin, les approches basées sur l'hippocampe et les approches « cerveau entier » sont non seulement les plus différentes dans leur principe de base mais également en terme de résultats (indice de Jaccard). Cependant, leur combinaison à l'aide d'un MKL ne donne pas de meilleurs résultats.

2.6.4 Le recalage : un modèle complètement déformable est-il avantageux ?

L'utilisation de DARTEL pour l'étape de recalage améliore les résultats de classification dans six cas et les détériore dans seulement deux cas. Ceci est cohérent avec de précédentes études sur le recalage. Klein et al. [2009] et Yassa & Stark [2009] ont en effet rapporté que DARTEL obtenait de meilleurs recouvrements et était plus sensible pour les études VBM (*voxel-based morphometry*). En particulier, l'utilisation d'un modèle complètement déformable donne de meilleurs recalages du lobe temporal médian [Yassa & Stark, 2009 ; Bergouignan et al., 2009]. L'hippocampe étant principalement atteint dans la maladie d'Alzheimer, nous nous attendions à ce qu'un meilleur recalage de l'hippocampe conduise à de meilleurs taux de classification.

2.6.5 Utilité des cartes de substance blanche et de CSF dans la classification

Dans leur version originale, certaines des méthodes testées dans cette étude, comme celles de Vemuri et al. [2008], Fan et al. [2007] ou de Magnin et al. [2009], utilisaient les trois types de tissus (GM, WM et CSF) pour la classification tandis que d'autres, comme celle de Klöppel et al. [2008b], n'utilisaient que les cartes de substance grise. Nous avons pour cette raison testé chacune de ces méthodes avec d'une part les cartes de substance grise uniquement et d'autre part avec les cartes des trois tissus. Notre objectif n'est pas de déterminer si la substance blanche contient de l'information utile au diagnostic. Elle en contient très probablement. Notre seul objectif était de savoir si l'ajout des cartes de WM et de CSF conduisait à de meilleurs résultats pour ces méthodes spécifiques.

De manière générale, l'ajout des cartes de substance blanche et de liquide cérébrospinal n'améliore pas les résultats de classification. Ajouter ces cartes augmente la dimension de l'espace des caractéristiques ; cet ajout peut donc avoir pour conséquence une instabilité du classifieur et un risque plus important de surapprentissage. Ce problème est bien connu dans le milieu de l'apprentissage statistique, c'est la *curse of dimensionality*. De plus, les personnes

âgées sont pour la plupart atteintes de lésions de la substance blanche (leucoaraïose ou autre). De telles lésions peuvent altérer la segmentation des tissus [Levy-Cooperman et al., 2008]. Elles peuvent également fausser le diagnostic clinique (démences mixes). L'ajout de la substance blanche dans la classification peut donc augmenter le bruit. Les cartes de substance grise sont des caractéristiques plus robustes, et ce, même si les lésions de la substance blanche altèrent l'étape de segmentation des tissus [Levy-Cooperman et al., 2008].

L'ajout de la substance blanche et du liquide cérébrospinal pourrait améliorer les performances de classification dans deux cas :

1. les méthodes utilisant des étapes de sélection de caractéristiques sont plus à même d'éliminer le bruit et de ne garder que les caractéristiques les plus discriminantes ;

2. dans le cas des méthodes utilisant un atlas, l'ajout de WM et du CSF pourrait compenser les erreurs de segmentation des régions.

Dans tous les cas, l'amélioration des performances n'est que très faible.

2.6.6 Faut-il faire de la sélection de variables ?

L'objectif principal de la sélection de variables est de conserver pour la classification uniquement les caractéristiques discriminantes et de diminuer la dimension de l'espace des caractéristiques. Dans notre comparaison, deux méthodes utilisent des étapes de sélection de caractéristiques : *Voxel-STAND* et *Voxel-COMPARE*. Au final, ces méthodes n'obtiennent pas de meilleurs résultats que les autres plus naïves. Leurs résultats sont plus sensibles à l'ensemble d'apprentissage choisi. En effet la sélection de caractéristiques peut être considérée comme une étape d'apprentissage. Cette étape augmente la taille de l'ensemble des fonctions de classification possibles et donc le risque de sur-apprentissage. Une manière plus robuste de diminuer le nombre de caractéristiques serait de faire une sélection *a priori*.

De plus, la sélection de caractéristiques peut s'avérer coûteuse en temps de calcul ; l'ajout d'hyperparamètres liés à ces étapes rend la recherche en grille plus difficile. Par rapport aux méthodes *Voxel-Direct* et *Voxel-Atlas* qui ne font pas de sélection de *features*, les méthodes *Voxel-STAND* et *Voxel-COMPARE* sont très coûteuses en temps de calcul. Elles nécessitent des semaines de calculs alors que les deux premières ne requièrent que quelques minutes. Ceci est principalement dû au nombre d'hyperparamètres.

Néanmoins, la sélection de caractéristiques peut s'avérer utile dans deux cas particuliers :

1. lors de l'ajout des cartes de WM et de CSF : pour diminuer la dimension de l'espace des données et éliminer le bruit lié aux lésions de la substance blanche ;

2. lors de l'étude de la conversion (MCI_{nc} *vs* MCI_c) : seulement une petite partie du cerveau est discriminante.

2.6.7 Influence de l'âge sur le taux de classification

De manière générale, nous avons trouvé que les témoins les plus âgés et que les patients les plus jeunes étaient plus souvent mal-classés. Cela peut avoir différentes origines :

1. lors du vieillissement cérébral normal, la substance grise et la substance blanche s'atrophient et la quantité de CSF augmente [Good et al., 2001 ; Salat et al., 2004] ;

2. des altérations des tissus ainsi qu'une diminution de leur contraste en IRM, généralement liées au vieillissement normal, augmentent artificiellement le taux d'atrophie mesurée dans l'image [Salat et al., 2009] ;

3. les sujets âgés sont plus propices aux lésions de la substance blanche ; ces lésions peuvent altérer l'étape de segmentation des tissus [Levy-Cooperman et al., 2008] ;

4. les sujets âgés sont plus facilement atteints de démences mixes [Zekry et al., 2002].

2.6.8 Les hyperplans séparateurs

Les hyperplans séparateurs optimaux (OMH) obtenus avec des SVM linéaires peuvent être facilement représentés comme nous l'avons mentionné dans le paragraphe 2.5.6. Ils permettent d'avoir des informations sur les régions du cerveau utilisées pour la classification. Il faut néanmoins être prudent : cette analyse n'est que qualitative et en aucun cas quantitative.

Avec les méthodes *Voxel-Direct-D*, *Voxel-Atlas* et *Thickness-Atlas*, les régions pour lesquelles une atrophie de tissus augmentait la probabilité d'être classé AD ou MCI_c sont cohérents avec les distributions d'atrophies rapportées par les études morphométriques. Ces régions incluent le lobe temporal médian, le gyrus temporal moyen, le gyrus temporal inférieur [Chételat & Baron, 2003 ; Good et al., 2002 ; Busatto et al., 2003 ; Rusinek et al., 2004 ; Tapiola et al., 2008], le cingulaire postérieur [Karas et al., 2004 ; Chételat et al., 2005 ; Laakso et al., 1998], ainsi que le gyrus frontal moyen [Whitwell et al., 2007], le gyrus fusiforme et le thalamus [Karas et al., 2003 ; Chételat et al., 2005]. En ce qui concerne les méthodes corticales, les régions concernées sont principalement : le lobe temporal médian, les gyri temporaux inférieurs et moyens, le cingulaire postérieur et dans une moindre mesure certaines régions pariétales et frontales et les régions latérales du cortex occipital. Ceci est cohérent avec les précédentes études sur l'épaisseur corticale [Thompson et al., 2004 ; Lerch et al., 2005 ; McDonald et al., 2009].

2.7 Conclusion

En conclusion, nous avons comparé différentes méthodes de classification automatiques d'images anatomiques pondérées en T_1 pour l'aide au diagnostic de la maladie d'Alzheimer. Cette étude de comparaison de méthodes est effectuée sur une grande population d'étude issue de la base de données ADNI. La grande majorité des méthodes ont réussi à classer les patients atteints de la maladie d'Alzheimer et les sujets témoins avec de bonnes performances. Cependant, les résultats obtenus pour la détection des AD prodromaux sont nettement moins bons.

Nous avons testé différents *features* pour la classification. De façon générale, l'utilisation de *features* surfaciques n'a pas conduit à une amélioration notable des performances par rapport au cas volumique, mais a augmenté le temps de calcul de façon importante. Par ailleurs, les approches restreintes à l'hippocampe ont généralement été moins performantes pour la classification *AD vs CN*. Différentes stratégies de sélection de *features* ont été comparées. Elles n'ont pas amélioré les résultats de façon substantielle par rapport aux méthodes directes. Enfin, l'analyse visuelle des hyperplans séparateurs a montré qu'ils manquent de régularité spatiale et de cohérence avec l'anatomie sous-jacente.

Chapitre 3

Régularisation spatiale et anatomique des machines à vecteurs supports

La classification d'images cérébrales est un problème de classification en grande dimension : la dimension d des données (*i.e.* le nombre de voxels) est beaucoup plus grande que le nombre N de sujets utilisés pour l'apprentissage ($N \ll d$).

Pour se rendre compte de ce que cela signifie, faisons l'hypothèse suivante. Nous supposons que la distribution des données admet une fonction de densité dans un sous-espace de dimension au moins égale au nombre de sujets. Dans ce cas, pour toute partition de l'ensemble des sujets en deux sous-groupes, il existe presque sûrement[1] un hyperplan qui sépare les deux groupes. En particulier, il existe presque sûrement un hyperplan qui sépare les sujets sains des sujets malades. Cela vient du fait que l'espace des matrices inversibles est dense dans l'espace des matrices. La matrice des données (de taille $d \times N$) est donc presque sûrement une matrice de rang N.

Il faut donc être prudent dans le choix de l'hyperplan pour ne pas faire du surapprentissage. On parle de surapprentissage lorsque la fonction de classification apprise est trop spécifique à l'échantillon d'apprentissage et n'est pas généralisable. Pour éviter un tel comportement, il est nécessaire d'ajouter de l'information *a priori* dans le classifieur. Cette information peut être sur la structure des données et/ou sur la pathologie. Les méthodes vues dans les chapitres précédents le font en ajoutant une étape avant la classification elle même, par exemple en découpant l'image en régions à l'aide d'un atlas, en lissant et éventuellement sous-échantillonnant les images, ou en utilisant des techniques de *clustering*.

[1] Avec probabilité un.

Dans ce chapitre nous allons nous attacher à introduire de l'information *a priori* directement dans le classifieur SVM à l'aide d'opérateurs de régularisation. Nous verrons qu'en remplaçant la régularisation du SVM linéaire par une régularisation spectrale, nous pouvons intégrer dans le SVM diverses contraintes spatiales et/ou anatomiques. Nous montrons le lien entre cette régularisation et les noyaux de diffusion.

Ce chapitre est organisé de la façon suivante. Dans la section suivante, (section 3.1), nous faisons un rappel sur les SVM linéaires avant d'introduire les opérateurs de régularisation. Nous verrons alors à la section 3.2 que le cadre des opérateurs de régularisation permet d'intégrer différents types d'informations via la définition de la notion de proximité. La section 3.3 présente le premier modèle de proximité, la proximité spatiale. Autrement dit, deux *features* sont proches si et seulement s'ils sont spatialement proches. Nous proposons alors, à la section 3.4, un modèle un peu plus complexe de proximité que nous appellerons proximité anatomique. Deux *features* seront considérés comme proches s'ils appartiennent à la même structure cérébrale ou s'ils sont connectés anatomiquement ou fonctionnellement (ex : via des réseaux de fibres, synchronies cérébrales ou corrélation du signal IRMf). Dans la section 3.5, nous verrons comment combiner ces deux types de proximités (spatiales et anatomiques). Pour finir, une discussion des méthodes est présentée à la section 3.6.

3.1 Introduction d'*a priori* dans les SVM

3.1.1 Rappels sur le SVM linéaire

Dans cette étude, les caractéristiques analysées sont calculées soit en chaque voxel de l'image, soit en chaque nœud du maillage du cortex. Par exemple ce peut être des cartes de probabilité de substance grise dans le premier cas ou des cartes d'épaisseur corticale dans le cas surfacique. Les méthodes présentées dans ce chapitre peuvent également s'appliquer à d'autres modalités telles que l'IRM fonctionnelle ou l'IRM de diffusion.

Hypothèses de travail. Toutes les images et surfaces sont recalées dans un espace stéréotaxique (par exemple avec [Shen & Davatzikos, 2002; Ashburner, 2007; Auzias et al., 2009]) commun à tous les sujets, comme c'est le cas dans la grande majorité des études interindividuelles [Klöppel et al., 2008b; Vemuri et al., 2008; Cuingnet et al., 2010; Fan et al., 2007; Lao et al., 2004; Querbes et al., 2009].

Soit $\mathbf{x}_s$ les données d'un sujet $s \in \mathcal{S}$. Dans le cas des images 3D, il y a deux manières de considérer $\mathbf{x}_s$:

(i) soit comme un élément de $\mathbb{R}^d$ où d est le nombre de voxels de l'image (point de vue discret),

(ii) soit comme une fonction (de carré sommable) à valeurs réelles définie sur un compact de $\mathbb{R}^3$ (point de vue continu).

Les deux points de vue discret et continu seront abordés dans ce chapitre. De la même manière, dans le cas surfacique, $\mathbf{x}_s$ peut être considéré :

(i) soit comme un élément de $\mathbb{R}^d$ où d est le nombre de nœuds du maillage cortical,

(ii) soit comme une fonction (de carré sommable) à valeurs réelles définie une variété riemannienne de dimension 2 (la surface corticale).

Soit $\mathcal{V}$ le domaine des images ou des surfaces. On notera v un élément de $\mathcal{V}$. Autrement dit, v correspond à un voxel ou à un nœud du maillage cortical dans le cas discret et v correspond à une position dans le cas continu. Ainsi $\mathcal{X} = L^2(\mathcal{V})$ muni du produit scalaire canonique est l'*input space*.

Nous considérons un groupe de N sujets avec leurs données correspondantes $(\mathbf{x}_s)_{s\in[1,N]} \in \mathcal{X}^N$ ainsi que leur label $(y_s)_{s\in[1,N]} \in \{-1,1\}^N$ (typiquement leur diagnostic, en d'autres termes malade ou sain).

Le SVM linéaire résout le problème d'optimisation suivant [Vapnik, 1995 ; Schölkopf & Smola, 2001 ; Shawe-Taylor & Cristianini, 2004] :

$$\left(\mathbf{w}^{\text{opt}}, b^{\text{opt}}\right) = \underset{\mathbf{w}\in\mathcal{X}, b\in\mathbb{R}}{\arg\min} \frac{1}{N}\sum_{s=1}^{N} \ell_{\text{hinge}}\left(y_s\left[\langle \mathbf{w}, \mathbf{x}_s\rangle + b\right]\right) + \lambda\|\mathbf{w}\|^2 \tag{3.1}$$

où $\lambda \in \mathbb{R}^+$ est le paramètre de régularisation et ℓ_{hinge} est la fonction de perte appelée *hinge loss function* et définie par :

$$\ell_{\text{hinge}} : \; u \in \mathbb{R} \mapsto \max(0, 1-u)$$

Lorsque l'on utilise un SVM linéaire, l'espace des caractéristiques (*feature space*) est le même que l'espace des données (*input space*). Par conséquent, lorsque les données sont les voxels d'une image 3D, les composantes de $\mathbf{w}^{\text{opt}}$ correspondent également à des voxels. De la même manière, dans le cas des surfaces, les éléments de $\mathbf{w}^{\text{opt}}$ peuvent être représentés comme les valeurs des nœuds d'un maillage cortical. Pour être cohérent avec l'anatomie, si deux voxels $v^{(1)} \in \mathcal{V}$ et $v^{(2)} \in \mathcal{V}$ sont proches selon la topologie de $\mathcal{V}$, leur poids dans le classifieur SVM, respectivement $w^{\text{opt}}_{v^{(1)}}$ et $w^{\text{opt}}_{v^{(2)}}$, devrait être similaire. En d'autres termes, si $v^{(1)} \in \mathcal{V}$ et $v^{(2)} \in \mathcal{V}$ correspondent à des régions voisines, ils devraient jouer un rôle similaire dans la fonction de classification. Cependant, ce n'est pas garanti avec un SVM linéaire standard (comme par exemple dans [Klöppel et al., 2008b]) puisque la régularisation n'est pas une régularisation spatiale. L'objectif principal de ce chapitre est de présenter des méthodes capables de contraindre $\mathbf{w}^{\text{opt}}$ pour que l'hyperplan correspondant soit régularisé spatialement.

3.1.2 *A priori* et SVM

Régulariser spatialement un SVM revient à ajouter de l'information *a priori* sur la distribution spatiale des caractéristiques (*i.e.* voxels et points du maillage). Il y a, à notre connaissance, trois grandes manières d'ajouter de l'information *a priori* dans un SVM.

3.1.2.1 Construction directe du noyau

Dans un SVM, toute l'information utilisée pour la classification est encodée dans le noyau du SVM. Ainsi, la première façon d'introduire de l'information *a priori* dans le SVM est de directement construire le noyau du SVM en fonction de l'*a priori* [Schölkopf & Smola, 2001]. Mais cela implique de connaître au moins une métrique ou une fonction d'affinité sur l'ensemble $\mathcal{X}$ qui soit consistante avec l'*a priori* voulu.

3.1.2.2 Invariances locales

Une autre façon d'aborder ce problème est de forcer la fonction de classification à être localement invariante à certaines transformations. Cela peut être fait de trois manières [Decoste & Schölkopf, 2002] : (i) en construisant un noyau localement invariant par certaines transformations, (ii) en utilisant des vecteurs supports virtuels (*virtual support vectors*) ou en utilisant une combinaison des deux méthodes précédentes.

Noyau localement invariant. Étant donnée une famille de transformations dépendant d'un seul paramètre $t : \{\mathcal{L}_t\}$, Schölkopf et al. [1998] proposent d'utiliser le noyau suivant :

$$K(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1 \left[\gamma I + (1-\gamma) \frac{1}{|S_{\text{train}}|} \sum_{s \in \mathcal{S}_{\text{train}}} d\mathbf{x}_s d\mathbf{x}_s^T \right]^{-1} \mathbf{x}_2 \tag{3.2}$$

avec γ un paramètre pour ajuster l'importance de l'invariance, $\mathcal{S}_{\text{train}}$ l'ensemble des sujets utilisés pour l'apprentissage et avec pour un sujet s,

$$d\mathbf{x}_s = \left. \frac{\partial}{\partial t} \right|_{t=0} \mathcal{L}_t \mathbf{x}_s = \lim_{t \to 0} \frac{1}{t} \left(\mathcal{L}_t \mathbf{x}_s - \mathbf{x}_s \right)$$

Ce noyau force l'hyperplan obtenu par le classifieur à être localement invariant à la famille de transformations choisies. Cette approche peut être vue comme un cas particulier de l'approche par les opérateurs de régularisation que nous verrons dans la section suivante (section 3.1.3). Le SVM avec le noyau proposé par Schölkopf et al. [1998] peut être vu comme un SVM linéaire standard (équation (3.1)) auquel on a rajouté un terme de régularisation :

$$\frac{1}{|S_{\text{train}}|} \sum_{s \in \mathcal{S}_{\text{train}}} \left(\left. \frac{\partial}{\partial t} \left(\langle \mathbf{w}, \mathcal{L}_t \mathbf{x}_s \rangle + b \right) \right|_{t=0} \right)^2$$

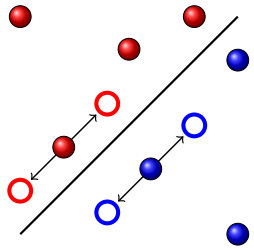


FIGURE 3.1 : Illustration de la méthode des vecteurs supports virtuels. Afin de rendre la fonction de classification localement invariante à certaines transformations (ici la translation), de nouveau exemples (cercles non rempli) sont générés à partir des vecteurs supports de l'ensemble d'apprentissage. Ce sont les vecteurs supports virtuels.

Chapelle & Schölkopf [2002] ont étendu cette approche aux noyaux non-linéaires.

Virtual SV. L'invariance peut aussi être forcée en générant artificiellement des données de l'ensemble d'apprentissage (figure 3.1) à partir des données existantes en leur appliquant des transformations. Ces nouvelles données sont appelées vecteurs supports virtuels (*virtual support vectors*) [Schölkopf et al., 1996].

Kernel jittering. La dernière approche est une combinaison des deux précédentes proposée par Decoste & Schölkopf [2002]. Elle est appelée *kernel jittering*. L'idée est de modifier le noyau K en un noyau K^J pour avoir une fonction de classification invariante selon certaines transformations. Cette modification est faite en utilisant des vecteurs virtuels suivant la procédure suivante (étant donné un noyau K) :

1. pour deux sujets s_1 et s_2, on considère parmi le vecteur $\mathbf{x}_{s_1}$ et tous les vecteurs générés à partir de $\mathbf{x}_{s_1}$ celui qui est le plus proche de $\mathbf{x}_{s_2}$ pour la métrique induite par K. Soit $\tilde{\mathbf{x}}_{s_1}$ ce vecteur

2. le nouveau noyau est alors : $K^J(\tilde{\mathbf{x}}_{s_1}, \mathbf{x}_{s_2}) = K(\tilde{\mathbf{x}}_{s_1}, \mathbf{x}_{s_2})$

La principale difficulté de ces approches est de définir l'ensemble des transformations.

3.1.2.3 Régularisation

La dernière manière d'ajouter de l'information *a priori* dans un SVM est d'utiliser les opérateurs de régularisation [Schölkopf & Smola, 2001 ; Smola & Schölkopf, 1998]. L'idée est de rendre la fonction de classification lisse selon certains critères définis *a priori*. L'approche que nous proposons se situe dans ce cadre.

3.1.3 Les opérateurs de régularisation

L'objectif principal est de régulariser spatialement la fonction de classification du SVM. Cette fonction peut s'écrire :

$$\operatorname{sgn}\left(f(\mathbf{x}_s) + b\right) \tag{3.3}$$

avec $f \in \mathbb{R}^{\mathcal{X}}$. Ceci se fait en utilisant un opérateur de régularisation [Smola & Schölkopf, 1998; Schölkopf & Smola, 2001] :

Définition 3.1.1 (Opérateur de régularisation) *On appelle opérateur de régularisation une application linéaire P d'un espace $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ dans un espace muni d'un produit scalaire $(\mathcal{D}, \langle \cdot, \cdot \rangle_{\mathcal{D}})$.*

Pour pouvoir continuer, nous avons besoin de la notion de fonction de Green d'un opérateur de régularisation :

Définition 3.1.2 (fonction de Green) *Soit P un opérateur de régularisation. $G : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ est une fonction de Green[2] de l'opérateur P ssi :*

$$\forall f \in \mathcal{F},\ \forall \mathbf{x} \in \mathcal{X},\ f(\mathbf{x}) = \langle P(G(\mathbf{x}, \cdot)), P(f) \rangle_{\mathcal{D}} \tag{3.4}$$

Ceci mène à la proposition suivante :

Proposition 3.1.3 *Si P admet au moins une fonction de Green G, alors :*

(i) G est un noyau semi-défini positif;

(ii) l'équation :

$$\left(f^{\text{opt}}, b^{\text{opt}}\right) = \underset{f \in \mathcal{F}, b \in \mathbb{R}}{\arg\min} \frac{1}{N} \sum_{s=1}^{N} \ell_{\text{hinge}}\left(y_s \left[f(\mathbf{x}_s) + b\right]\right) + \lambda \|P(f)\|_{\mathcal{D}}^2 \tag{3.5}$$

est équivalente au problème de minimisation d'un SVM avec G comme noyau.

De plus, comme, dans nos hypothèses de travail, le SVM utilisé est un SVM linéaire, la fonction f est, à une constante près, le dual d'un élément de l'espace des données (*input space*). Le problème d'optimisation (3.5) permet donc d'introduire de manière naturelle de la régularisation spatiale sur f dans le SVM via la définition de P.

Remarque : La plupart du temps, l'espace $\mathcal{F}$ est un espace à noyau reproduisant (RKHS) de noyau K et $\mathcal{D} = \mathcal{F}$. On a donc : si P est borné et si $P^\dagger P$ est inversible (par exemple si P est injectif et compact), alors P admet une fonction de Green :

$$G = (P^\dagger P)^{-1} K \tag{3.6}$$

où $P^\dagger$ dénote l'opérateur adjoint de P.

[2] En réalité on fait ici un petit abus de langage : G ainsi défini est l'opérateur de Green de $P^\dagger P$.

Démonstration P est borné. P admet alors un unique adjoint $P^\dagger$. Soit $(f, \mathbf{x}) \in \mathcal{F} \times \mathcal{X}$

$$\begin{aligned}\left\langle P\left(\left(P^\dagger P\right)^{-1} K(\mathbf{x},\cdot)\right), P(f)\right\rangle_{\mathcal{D}} &= \left\langle P^\dagger P\left(\left(P^\dagger P\right)^{-1} K(\mathbf{x},\cdot)\right), f\right\rangle_{\mathcal{D}} \\ &= \langle K(\mathbf{x},\cdot), f\rangle_{\mathcal{D}} \\ &= f(\mathbf{x})\end{aligned}$$

■

Il reste maintenant à définir un opérateur de régularisation P adapté à notre problème. C'est ce que nous allons voir dans la section suivante.

3.2 Cadre de la régularisation laplacienne

La régularisation spatiale nécessite la définition de la proximité entre les éléments de $\mathcal{V}$. Ceci peut se faire via la définition d'un graphe dans le cas discret ou la définition d'une métrique dans le cas continu.

3.2.1 Graphes

3.2.1.1 Cadre

Lorsque l'ensemble $\mathcal{V}$ est fini, les graphes pondérés représentent un cadre naturel souvent utilisé [Geman & Geman, 1984 ; Shi & Malik, 2000] pour prendre en compte de l'information spatiale. En effet, les voxels d'une image de cerveau peuvent être considérés comme les nœuds d'un graphe qui modéliserait leur proximité (figure 3.2). Ce graphe peut être par exemple la connectivité des voxels (6, 18 or 26) ou un graphe plus sophistiqué permettant de prendre en compte d'autres informations.

Nous avons opté pour l'opérateur de régularisation suivant :

$$\begin{array}{rccc} P: & \mathcal{F} = \mathcal{L}(\mathbb{R}^{\mathcal{V}}, \mathbb{R}) & \rightarrow & \mathcal{F} \\ & \mathbf{w}^* & \mapsto & \left(e^{\frac{1}{2}\beta L}\mathbf{w}\right)^* \end{array} \tag{3.7}$$

avec L le laplacien du graphe [Chung, 1992] et $\mathbf{w}^*$ le dual[3] de $\mathbf{w}$. Le paramètre β contrôle la taille de la régularisation. Le problème d'optimisation s'écrit alors :

$$\left(\mathbf{w}^{\text{opt}}, b^{\text{opt}}\right) = \underset{\mathbf{w}\in\mathcal{X}, b\in\mathbb{R}}{\arg\min} \frac{1}{N}\sum_{s=1}^{N} \ell_{\text{hinge}}\left(y_s\left[\langle \mathbf{w}, \mathbf{x}_s\rangle + b\right]\right) + \lambda \|e^{\frac{1}{2}\beta L}\mathbf{w}\|^2 \tag{3.8}$$

[3] Le dual $\mathbf{w}^* \in \mathcal{L}(\mathbb{R}^{\mathcal{V}}, \mathbb{R})$ d'un vecteur $\mathbf{w} \in \mathcal{X}$ est par définition l'application linéaire définie par : $\forall \mathbf{x} \in \mathcal{X}, \mathbf{w}^*(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x}\rangle$.

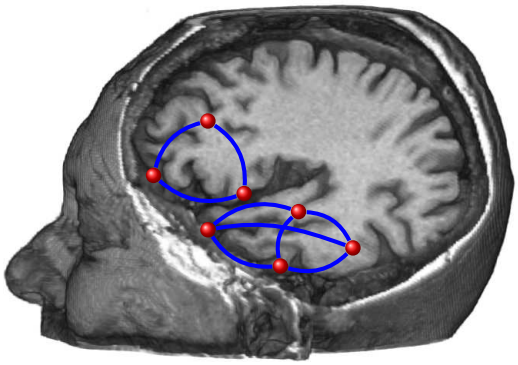


FIGURE 3.2 : La proximité entre les voxels ou régions du cerveau peut être modélisée par un graphe pondéré. Les voxels sont les nœuds du graphe (en rouge). Les liens entre les nœuds codent la proximité.

Une telle régularisation pénalise exponentiellement les composantes à hautes fréquences. Ainsi, la régularisation force le classifieur à considérer comme similaires les voxels fortement connectés selon la matrice d'adjacence du graphe. D'après la section précédente (3.1.3), ce nouveau problème de minimisation (3.8) est équivalent à un problème de minimisation d'un SVM. Le nouveau noyau est donné par K_β :

$$K_\beta(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1^T e^{-\beta L} \mathbf{x}_2 \tag{3.9}$$

C'est un noyau de la chaleur, également appelé noyau de diffusion sur un graphe.

3.2.1.2 Calcul de la matrice de Gram

Une fois la notion de proximité définie, pour pouvoir résoudre le problème d'optimisation 3.8, on a besoin de calculer la matrice de Gram définie par $\left(K_\beta\left(\mathbf{x}_i, \mathbf{x}_j\right)\right)_{i,j}$. Le calcul de cette matrice nécessite de calculer $e^{-\beta L}\mathbf{x}_s$ pour tous les sujets s de l'ensemble d'apprentissage.

Il existe de nombreuses méthodes pour le calcul d'**exponentielles de matrices**. Moler & Van Loan [2003] regroupent ces méthodes en quatre grandes catégories : (i) les méthodes basées sur des séries tronquées, (ii) celles basées sur les résolutions d'équations différentielles, (iii) les méthodes polynomiales et (iv) les méthodes basées sur des décompositions de matrices.

Parmi les méthodes de type **séries**, il y a essentiellement la méthode d'approximation par série de Taylor qui consiste à approximer l'exponentielle d'une matrice par sa série de Taylor tronquée à l'ordre p et l'approximation de Padé. L'approximation de Padé nécessite l'inversion de la matrice « dénominateur » ; elle n'est donc pas adaptée aux cas de grandes matrices. Les méthodes basées sur la **résolution d'une équation différentielle** calculent $e^{-\beta L\mathbf{x}_s}$ en résolvant

l'équation d'inconnue $\mathbf{y}$:

$$\begin{cases} \dfrac{d}{dt}\mathbf{y} = -\beta L\mathbf{y} \\ \mathbf{y}(0) = \mathbf{x}_s \end{cases}$$

Ces méthodes sont très proches des méthodes de type séries et ne présentent pas à notre connaissance d'avantage particulier par rapport aux précédentes. Quant aux **méthodes polynomiales** (ex : Caley-Hamilton, interpolation de Lagrange, interpolation de Newton), la grande majorité nécessite la connaissance du polynôme caractéristique. Or la plupart du temps, nous ne connaissons pas ce polynôme. Nous verrons par la suite que, dans notre cas, quand le polynôme caractéristique est connu, nous connaissons également les vecteurs propres. On utilise alors l'approche par diagonalisation. Il existe d'autres approches utilisant des **décompositions de matrices** telles que la décomposition de Jordan mais elles présentent souvent des problèmes numériques.

La spécificité de notre problème est que L est une matrice carrée de taille d avec $d \sim 10^6$. Il est donc impossible de calculer directement $e^{-\beta L}$ numériquement. Il est même en général impossible de stocker cette matrice, et ce, même si L est très creuse sauf pour des très petites valeurs de β. Pour cette raison, nous utiliserons, dans le cas général (où nous ne connaissons pas le polynôme caractéristique), l'approximation par série de Taylor. Une manière de calculer la matrice de Gram est donc d'utiliser la décomposition en série de Taylor jusqu'à l'ordre p :

$$e^{-\beta L}\mathbf{x}_s \approx \sum_{k=0}^{p} \frac{1}{k!} \left(-\beta L\right)^k \mathbf{x}_s \tag{3.10}$$

Il reste maintenant à choisir p. L'erreur résiduelle pour la norme spectrale obtenue avec l'approximation (3.10) est bornée par (figure 3.3) (ex : [Smola & Kondor, 2003]) :

$$\frac{(\|L\|_2\beta)^{p+1}}{(p+1)!} \tag{3.11}$$

Remarque : Il faut se méfier de l'approximation par série de Taylor. Moler & Van Loan [2003] donnent un exemple frappant où une telle approximation peut conduire à des résultats catastrophiques. Considérons la matrice suivante

$$M = \begin{pmatrix} -49 & 24 \\ -64 & 31 \end{pmatrix} \tag{3.12}$$

Une approximation correcte à 6 décimales de l'exponentielle de cette matrice est :

$$e^M \approx \begin{pmatrix} -0.735759 & 0.551819 \\ -1.471518 & 1.103638 \end{pmatrix} \tag{3.13}$$

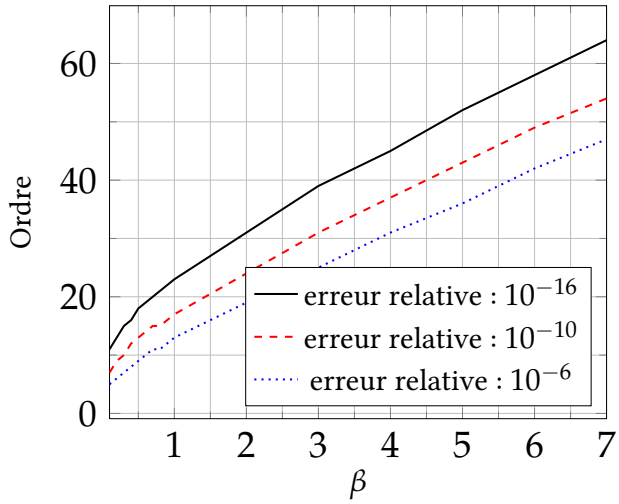


FIGURE 3.3 : Ordre de la décomposition en série de Taylor en fonction de β selon l'erreur relative souhaitée (pour un laplacien normalisé).

Cependant, si l'on utilise l'approximation par série de Taylor, lorsque la précision arithmétique est de 32bits (*single precision*), le résultat tend vers :

$$\begin{pmatrix} 0.234995 & -0.182978 \\ 1.04608 & -1.906229 \end{pmatrix} \tag{3.14}$$

Si l' on cherche à calculer directement :

$$e^{M} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \tag{3.15}$$

on obtient une erreur relative de plus de 250% en norme 1. Pour éviter ce type d'erreurs, une approche consiste à choisir un entier n tel que $\left\| \frac{1}{n} \beta L \right\|_1 \leq 1$ et à utiliser la formule suivante :

$$e^{-\beta L} = \left(e^{-\frac{\beta}{n} L} \right)^n \tag{3.16}$$

Cette approche est connue sous le nom de *scaling and squaring* [Moler & Van Loan, 2003 ; Higham, 2005]. L'erreur résiduelle pour la norme spectrale est alors bornée par :

$$\left(1 + \frac{1}{(p+1)!} \right)^n - 1 \tag{3.17}$$

Démonstration Soit la matrice $M = \sum_{k=0}^{p} \frac{1}{k!} \left(-\frac{\beta}{n} L \right)^k \mathbf{x}_s$. D'après l'équation (3.11), il existe une matrice ϵ telle que : $M = e^{-\frac{\beta}{n} L} + \epsilon$ et $\|\epsilon\|_2 \leq \frac{(\|L\|_2 \beta / n)^{p+1}}{(p+1)!}$. Notons que la matrice L est symétrique. Sa norme infinie et sa

norme d'indice 1 sont donc égales. Par conséquent, la norme spectrale de L est inférieure à sa norme d'indice 1. Or comme $\left\|\frac{1}{n}\beta L\right\|_1 \leq 1$, on a : $\|\epsilon\|_2 \leq \frac{1}{(p+1)!}$.

On a également :

$$\begin{aligned} M^n &= \left(e^{-\frac{\beta}{n}L} + \epsilon\right)^n \\ M^n - e^{-\beta L} &= \left(e^{-\frac{\beta}{n}L} + \epsilon\right)^n - e^{-\beta L} \end{aligned}$$

La norme spectrale étant matricielle, en utilisant l'inégalité triangulaire, on a :

$$\left\|M^n - e^{-\beta L}\right\|_2 \leq \left(\left\|e^{-\frac{\beta}{n}L}\right\|_2 + \|\epsilon\|_2\right)^n - \left\|e^{-\frac{\beta}{n}L}\right\|_2^n$$

Or la matrice L est semi-définie positive et β est positif donc : $\left\|e^{-\frac{\beta}{n}L}\right\|_2 \leq 1$. On obtient alors l'inégalité (3.17).

■

3.2.2 Variétés riemanniennes

3.2.2.1 Cadre

Dans ce chapitre, lorsque l'ensemble $\mathcal{V}$ est continu, il peut être considéré comme une variété riemannienne compacte de dimension 2 (exemple : surface) ou de dimension 3 (exemple : cas euclidien). Le tenseur de métrique modélise alors la notion de proximité entre les *features*. Un espace compact est complet. Or la notion de noyau de la chaleur existe pour les variétés riemanniennes complètes [Jost, 2008 ; Lafferty & Lebanon, 2005]. Ainsi, la régularisation laplacienne présentée dans le paragraphe précédent (3.2.1) peut être étendue aux variétés riemanniennes compactes [Lafferty & Lebanon, 2005]. De la même manière que pour les graphes, nous optons pour l'opérateur de régularisation suivant :

$$P : \ \mathbf{w}^* \in \mathcal{F} = \mathcal{L}(U, \mathbb{R}) \mapsto \left(e^{\frac{1}{2}\beta\Delta}\mathbf{w}\right)^* \in \mathcal{F} \tag{3.18}$$

où U est un sous-espace des fonctions sur $\mathcal{V}$ de carrés sommables et Δ représente l'opérateur de Laplace-Beltrami et $t \mapsto e^{-t\Delta}$ le noyau de la chaleur avec conditions de Dirichlet homogènes aux bords. Le nouveau problème d'optimisation est aussi équivalent à un problème d'optimisation d'un SVM avec comme noyau :

$$K_\beta(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1^T e^{-\beta\Delta}\mathbf{x}_2 \tag{3.19}$$

3.2.2.2 Calcul de la matrice de Gram

Une fois la notion de proximité définie, pour pouvoir résoudre le problème d'optimisation (3.5), il faut calculer la matrice de Gram. Le calcul de cette matrice nécessite le calcule de $e^{-\beta\Delta}\mathbf{x}_s$

pour tous les sujets s de l'ensemble d'apprentissage. Le nombre de voxels de l'image ($d \sim 10^6$) rend la décomposition de l'opérateur de Laplace-Beltrami impossible numériquement. Une autre manière de calculer $e^{-\beta\Delta}\mathbf{x}_s$ est de considérer ce vecteur comme étant la solution au temps $t = \beta$ de l'équation de la chaleur d'inconnue $\mathbf{u} \in \mathbb{R}^{\mathcal{V}\times\mathbb{R}}$, avec comme conditions aux bords, les conditions homogènes de Dirichlet :

$$\begin{cases} \dfrac{\partial \mathbf{u}}{\partial t} - \Delta\mathbf{u} = 0 \\ \mathbf{u}(t=0) = \mathbf{x}_s \end{cases} \tag{3.20}$$

L'opérateur de Laplace-Beltrami est défini par Jost [2008] :

$$\Delta\mathbf{u} = \frac{1}{\sqrt{\det g}} \sum_{j=1}^{d_\mathcal{M}} \frac{\partial}{\partial v_j} \left(\sum_{i=1}^{d_\mathcal{M}} h_{ij} \sqrt{\det g} \frac{\partial \mathbf{u}}{\partial v_i} \right) \tag{3.21}$$

avec $d_\mathcal{M}$, la dimension de la variété, g le tenseur de métrique et h le tenseur inverse de g. La résolution de l'équation différentielle de la chaleur peut se faire avec une approche variationnelle.

Nous venons de voir dans cette section les opérateurs de régularisation que nous utiliserons et la manière de calculer la matrice de Gram dans le cadre général. Dans les sections suivantes, sections 3.3 et 3.4, nous présentons différents modèles de proximités qui correspondent à différents types de graphes et de distances : une régularisation spatiale et une régularisation anatomique. Nous proposons ensuite dans la section 3.5 de combiner ces deux régularisations. Afin d'avoir des temps de calculs raisonnables, pour chaque modèle de proximité, nous proposons une méthode de calcul de la matrice de Gram adaptée au modèle.

3.2.3 Lien avec les approches à noyau de diffusion

L'opérateur utilisé pénalise exponentiellement les hautes fréquences. Une telle régularisation conduit à l'utilisation d'un SVM à noyau de diffusion. L'utilisation de tels noyaux en classification par SVM a été proposée par Kondor & Lafferty [2002]. Lafferty & Lebanon [2005] ont proposé une approche similaire dans un cadre continu. Ces noyaux sont un cas particulier de régularisation spectrale [Smola & Kondor, 2003]. La régularisation spectrale a également été utilisée dans d'autres domaines tels que l'imagerie satellitaire (exemple : [Gomez et al., 2008]) ou encore la bioinformatique [Vert & Kanehisa, 2003 ; Lanckriet et al., 2004 ; Tsuda & Noble, 2004].

L'approche présentée dans ce chapitre diffère des précédentes. Dans notre approche, les nœuds du graphe sont les *features*, par exemple les voxels, tandis que dans les approches citées, les nœuds du graphe sont les objets à classer. À ma connaissance, cette approche n'a pas été appliquée aux images cérébrales auparavant, mais seulement dans le cadre de la classification de *microarray* par Rapaport et al. [2007].

3.3 Régularisation spatiale

Dans cette section, nous considérons le cas de la régularisation basée sur la proximité spatiale uniquement. En d'autres termes, deux voxels (ou nœuds d'un maillage) sont proches si et seulement s'ils sont proches spatialement.

3.3.1 Cas volumique

Quand $\mathcal{V}$ représente l'ensemble des voxels de l'image (cas discret), la manière la plus simple est d'utiliser le graphe de connectivité des voxels de l'image comme graphe pour la régularisation spatiale. De la même manière, dans le cas continu, si $\mathcal{V}$ est un sous-ensemble compact de $\mathbb{R}^3$, la proximité est définie à l'aide de la distance euclidienne. Dans les deux cas, régulariser est presque équivalent à prétraiter les données avec un lissage gaussien d'écart type $\sigma = \sqrt{\beta}$ voxels [Kondor & Lafferty, 2002] (voir remarque pour le « presque »). La version discrète correspond à une résolution par différences finies de l'équation de la chaleur. La complexité de calcul de la matrice de Gram est donc, en termes de nombre d'opérations, en : $O\left(Nd\log(d)\right)$.

Remarque : La différence entre la régularisation utilisée et le lissage en amont des données par un filtre gaussien est due aux conditions aux bords. Dans le cadre de notre approche nous considérons un sous ensemble compact de $\mathbb{R}^3$ avec les conditions aux bords de Dirichlet homogènes. Le lissage gaussien, quant à lui, est l'opérateur de Green de l'équation de la chaleur non pas sur $\mathcal{V}$ mais sur $\mathbb{R}^3$ tout entier (avec quelques hypothèses à l'infini). Quitte à agrandir l'image en ajoutant des zéros aux bords de l'image, cela influe très peu dans notre cas. Nous donnons quelques détails sur cette approximation dans l'annexe B.

3.3.2 Cas surfacique

Le graphe de connectivité n'est pas applicable directement au cas surfacique. En effet, un tel graphe serait trop dépendant de la triangulation de la surface corticale. Ce problème peut être résolu en repondérant le graphe de connectivité avec des poids conformaux (figure 3.4). Nous avons préféré adopter, dans ce chapitre, le point de vue continu pour le cas des surfaces : la surface corticale est une variété riemannienne compacte de dimension 2. Nous utilisons alors l'opérateur de régularisation défini à l'équation (3.18). En effet, le laplacien est un opérateur intrinsèque et ne dépend pas de la paramétrisation de la surface corticale.

Le noyau de la chaleur a déjà été utilisé pour le lissage de la surface corticale, notamment dans [Andrade et al., 2001 ; Cachia et al., 2003 ; Chung et al., 2003 ; Chung, 2004 ; Chung et al., 2005a,b]. Nous ne détaillerons donc pas les calculs et invitons le lecteur à se référer à ces

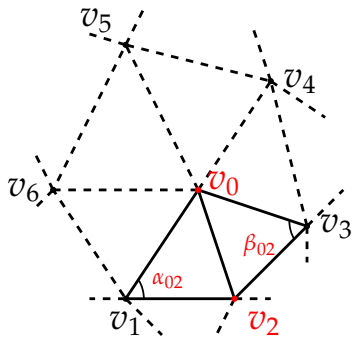


FIGURE 3.4 : Poids conformaux : $A_{i,j} = \cot(\alpha_{ij}) + \cot(\beta_{ij})$.

papiers pour plus de détails sur l'implémentation. Dans les grandes lignes, [Andrade et al., 2001 ; Cachia et al., 2003 ; Chung et al., 2003] utilisent des méthodes de différences finies ou d'éléments finis. Nous utiliserons l'implémentation décrite dans [Chung et al., 2005a,b]. Elle est basée sur l'approximation paramétrique [Rosenberg, 1997] au premier ordre du noyau de la chaleur. Cette implémentation est disponible en ligne[4].

De la même manière que dans le cas gaussien, le paramètre de diffusion β règle la taille σ^2 du noyau de lissage : $\sigma = \sqrt{\beta}$. Le calcul de la matrice de Gram nécessite $O(N\beta d)$ opérations.

3.4 Régularisation anatomique

Dans cette section, nous considérons un autre type de proximité que nous appelons proximité anatomique. Deux voxels sont considérés comme proches s'ils appartiennent au même réseau. Par exemple, deux voxels peuvent être proches s'ils appartiennent à la même région anatomique ou fonctionnelle (définie par un atlas probabiliste). Cela peut être vu comme une connectivité « courte-distance ». Une autre possibilité est celle de la proximité « longue-distance » où deux voxels spatialement éloignés sont proche anatomiquement (connectés par un faisceau de fibres) ou fonctionnellement (en utilisant les réseaux fonctionnels obtenus en IRMf).

Nous considérons dans cette section uniquement le cas discret. Pour plus de clarté, la régularisation est présentée uniquement dans le cas volumique. Elle est directement applicable au cas surfacique.

[4] `http://www.stat.wisc.edu/~mchung/softwares/hk/hk.html`

3.4.1 Graphe de régularisation

Soient $(\mathcal{A}_1, \cdots, \mathcal{A}_R)$ les R régions d'intérêt (ROI) d'un atlas probabiliste. Soit $p(v \in A_r)$ la probabilité que le voxel v appartienne à la région $\mathcal{A}_r$. Avec ces notations, la probabilité que les voxels $v^{(1)}$ et $v^{(2)}$ appartiennent à la même région est donnée par :

$$\sum_{r=1}^{R} p\left(\left(v^{(1)}, v^{(2)}\right) \in \mathcal{A}_r^2\right) \tag{3.22}$$

Nous supposons que, pour deux voxels $v^{(1)}$ et $v^{(2)}$ vérifiant $v^{(1)} \neq v^{(2)}$, nous ayons :

$$p\left(\left(v^{(1)}, v^{(2)}\right) \in \mathcal{A}_r^2\right) = p\left(v^{(1)} \in \mathcal{A}_r\right) p\left(v^{(2)} \in \mathcal{A}_r\right) \tag{3.23}$$

Soit $E \in \mathbb{R}^{d \times R}$ la matrice stochastique à droite définie par :

$$E_{i,r} \;=\; p\left(v^{(i)} \in \mathcal{A}_r\right) \tag{3.24}$$

Alors, pour tout $v^{(i)} \neq v^{(j)}$, la (i,j)-ème entrée de la matrice d'adjacence $A = EE^T$ correspond à la probabilité que les deux voxels $v^{(i)}$ et $v^{(j)}$ appartiennent à la même région.

Pour pouvoir modéliser les connexions « longues-distances » (structurelles ou fonctionnelles), on peut considérer une matrice semi-définie positive C de taille $R \times R$ telle que la (r_1, r_2)-ème entrée corresponde à la probabilité que les régions $\mathcal{A}_{r_1}$ et $\mathcal{A}_{r_2}$ de l'atlas soient connectées. La matrice d'adjacence devient alors :

$$A = ECE^T \tag{3.25}$$

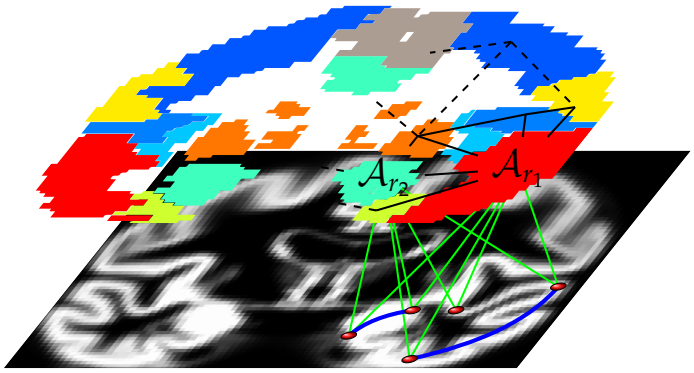


FIGURE 3.5 : La proximité anatomique encodée par un graphe. Les poids des connexions entre les nœuds (points rouges) sont représentés par des arcs bleus. Ce sont les éléments de la matrice d'adjacence A. Ils sont fonction de la probabilité d'appartenance (en vert) aux régions $\mathcal{A}_r$ d'un atlas (matrice E) et du lien (en noir) entre les régions (matrice C).

Le laplacien est donc de la forme [Chung, 1992] :

$$L = D - A \tag{3.26}$$

avec D une matrice diagonale. Pour calculer la matrice de Gram, nous avons besoin de calculer l'exponentielle de la matrice laplacienne $-\beta L$ (cf. section 3.2.1). Le calcul de e^A et de e^D est aisé mais A et D ne commutant pas, cela ne peut servir directement pour le calcul de $e^{-\beta L}$. Une possibilité est d'utiliser la formule du produit de Lie-Trotter[5]. Nous préférons utiliser le laplacien normalisé $\tilde{L}$ [Chung, 1992] :

$$\tilde{L} = I_d - D^{-\frac{1}{2}} E C E^T D^{-\frac{1}{2}} \tag{3.27}$$

On peut réécrire cette équation :

$$\tilde{L} = I_d - \tilde{E}\tilde{E}^T \tag{3.28}$$

avec : $\tilde{E} = D^{-\frac{1}{2}} E C^{\frac{1}{2}}$.

L'utilisation du laplacien normalisé a un intérêt évident sur le plan numérique : elle permet de s'assurer que les deux membres de l'addition commutent. Elle a également un intérêt sur le plan de la modélisation lorsque L ne représente pas une version discrétisée de l'opérateur de Laplace-Beltrami. En effet, comme nous l'avons vu dans la section 3.2.1, une telle regularisation pénalise exponentiellement les hautes fréquences. Autrement dit, elle tend à pénaliser fortement les variations de $\mathbf{w}^{\text{opt}}$ sur des clusters homogènes du graphe. Ces clusters peuvent être vus comme une segmentation *soft min-cut* du graphe. L'utilisation du Laplacien normalisé conduit à des clusters qui peuvent être considérés comme une segmentation *soft normalized-cut* Shi & Malik [2000], ce qui est préférable.

3.4.2 Calcul de la matrice de Gram

3.4.2.1 Formulation : cas général

Comme nous l'avons vu à la section 3.2.1.2, il y a différentes manières de calculer des exponentielles de matrices [Moler & Van Loan, 2003]. L'une d'entre elles, fréquemment utilisée, est la diagonalisation. On diagonalise la matrice $\tilde{L}$ et dans cette nouvelle base, l'exponentielle de matrice est obtenue directement en prenant les exponentielles des termes diagonaux. Malheureusement, dans le cas général, la complexité de la diagonalisation est cubique. Il est donc impossible de diagonaliser numériquement directement la matrice L. Heureusement, dans notre cas, il suffit de trouver une base de $\left(\ker \tilde{E}\tilde{E}^T\right)^{\perp}$ constituée de vecteurs propres de $\tilde{L}$. C'est ce que nous faisons dans ce qui suit.

La matrice $\tilde{E}^T\tilde{E}$ est une matrice réelle symétrique. Soient X une matrice orthogonale de taille $R \times R$ et Λ une matrice diagonale de même taille telles que :

$$X^T \tilde{E}^T \tilde{E} X = \Lambda \tag{3.29}$$

[5]Soit M_1 et M_2 deux matrices réelles carrées de taille $d \times d$, alors : $e^{M_1+M_2} = \lim_{n\to\infty} \left(e^{\frac{1}{n}M_1} e^{\frac{1}{n}M_2}\right)^n$.

Soit k le rang de la matrice $\tilde{E}^T\tilde{E}$. Sans perte de généralité, supposons que les seuls éléments non nuls de Λ soient les k-premiers éléments diagonaux. Nous les noterons : $\Lambda_{1,1}, \cdots, \Lambda_{k,k}$. Soit $\tilde{X}$ la matrice $d \times k$ définie de la façon suivante. Le r-ème vecteur colonne de $\tilde{X}$, noté $\tilde{X}_r$, est donné par :

$$\tilde{X}_r = \Lambda_{r,r}^{-\frac{1}{2}} \tilde{E} X_r \tag{3.30}$$

avec X_r le r-ème vecteur colonne de X. Définissons alors $\tilde{\Lambda}$ comme étant la matrice diagonale de taille $k \times k$ vérifiant :

$$\tilde{\Lambda}_{r,r} = 1 - \Lambda_{r,r} \tag{3.31}$$

$\left(\tilde{X}_r\right)_{r=1,\cdots,k}$ est alors une base orthonormée de vecteurs propres de $\left(\ker \tilde{E}\tilde{E}^T\right)^{\perp}$ et les valeurs propres correspondantes sont : $(\Lambda_{r,r})_{r=1,\cdots,k}$. L'exponentielle de matrice est alors obtenue à l'aide de la relation suivante :

$$e^{-\beta\tilde{L}} = \tilde{X} e^{-\beta\tilde{\Lambda}} \tilde{X}^T + e^{-\beta}\left[I_d - \tilde{X}\tilde{X}^T\right] \tag{3.32}$$

3.4.2.2 Formulation : cas d'un atlas binaire

Quand l'atlas utilisé est un atlas binaire, en d'autres termes lorsque $p(v \in \mathcal{A}_r) \in \{0, 1\}$, la formulation de l'exponentielle de matrice est plus explicite et son calcul beaucoup plus simple.

Soit $d^{(r)}$ le nombre de voxels de la région $\mathcal{A}_r$. Quitte à renuméroter les voxels, nous supposons que les voxels sont ordonnés par régions. Plus précisément, nous supposons que les voxels $v^{(1)}, \cdots, v^{\left(d^{(1)}\right)}$ appartiennent à la région $\mathcal{A}_1$, que les voxels $v^{\left(d^{(1)}+1\right)}, \cdots, v^{\left(d^{(1)}+d^{(2)}\right)}$ appartiennent à $\mathcal{A}_2$ et ainsi de suite. La matrice d'adjacence A est alors une matrice diagonale par blocs et vérifie :

$$A = \left(\mathbf{1}_{d^{(1)}}\mathbf{1}_{d^{(1)}}^T\right) \oplus \left(\mathbf{1}_{d^{(2)}}\mathbf{1}_{d^{(2)}}^T\right) \oplus \cdots \oplus \left(\mathbf{1}_{d^{(R)}}\mathbf{1}_{d^{(R)}}^T\right) \tag{3.33}$$

avec $\mathbf{1}_{d^{(r)}}$ le vecteur colonne de taille $d^{(r)}$ constitué uniquement de uns. On obtient alors l'exponentielle de la matrice $\tilde{L}$ avec la relation suivante :

$$e^{-\beta\tilde{L}} = e^{-\beta\tilde{L}^{(1)}} \oplus e^{-\beta\tilde{L}^{(2)}} \oplus \cdots \oplus e^{-\beta\tilde{L}^{(R)}} \tag{3.34}$$

avec, pour tout $r \in [1, R]$:

$$e^{-\beta\tilde{L}^{(r)}} = e^{-\beta} I_{d^{(r)}} + \left(1 - e^{-\beta}\right) \underbrace{\left[\frac{1}{d^{(r)}}(\mathbf{1}_{d^{(r)}}\mathbf{1}_{d^{(r)}}^T)\right]}_{\text{opérateur : moyenne sur une region}} \tag{3.35}$$

Comportement asymptotique. Dans le cas $\beta = 0$, c'est équivalent à un SVM linéaire standard sans aucune régularisation anatomique. Dans le cas limite $\beta = +\infty$, la régularisation revient à remplacer chaque voxel par la valeur moyenne des voxels de la région à laquelle il appartient. On se retrouve alors dans une approche similaire à celles de Lao et al. [2004] et Magnin et al. [2009]. Les cas $\beta \in \mathbb{R}^{+*}$ sont des cas intermédiaires.

3.4.2.3 Complexité

Le calcul de la matrice de Gram requiert donc principalement (i) le calcul de $D^{-\frac{1}{2}}$, ce qui est peu coûteux puisque D est diagonale et (ii) la diagonalisation d'une matrice carrée de taille R. Cette diagonalisation est elle aussi peu coûteuse puisqu'en pratique $R \sim 10^2$. Le coût de calcul de la matrice de Gram en terme d'opérations est alors en :

$$O\left(NRd + R^3\right)$$

Dans le cas particulier de l'atlas binaire, le nombre d'opérations nécessaires est beaucoup plus faible. Si $R < d$, la complexité devient : $O\left(Nd\right)$.

Comparativement à la méthode de calcul de la matrice de Gram par développement en série de Taylor (section 3.2.1), la méthode de calcul présentée dans cette section a, en ce qui concerne son implémentation, une complexité plus faible. Nous voulions nous assurer que le temps de calcul était également plus faible. La difficulté dans l'évaluation du temps de calcul est que la méthode basée sur le développement en série de Taylor est itérative. De plus le nombre d'itérations dépend de l'erreur relative souhaitée tandis que l'erreur obtenue avec la méthode présentée dans cette section n'est pas contrôlée.

Nous avons donc commencé par évaluer l'erreur d'estimation de la matrice de Gram. La principale source d'erreur vient de la diagonalisation de la matrice de taille $R \times R$. Nous nous sommes donc focalisés sur cette erreur. D'après la documentation de LAPACK (la librairie utilisée), l'erreur sur l'estimation des valeurs propres est ($\epsilon \approx 10^{-16}$) : $\epsilon \left\| E^T E \right\|_2$ avec ϵ la précision numérique. Quant à l'erreur d'angle sur les vecteurs propres, elle est donnée par : $\epsilon \left\| E^T E \right\|_2 \kappa_2$ avec κ_2 le conditionnement de $E^T E$. En pratique, avec les matrices utilisées dans nos expériences (incluant l'estimation du temps de calcul), nous obtenons des erreurs estimées inférieures à 10^{-13}. Pour être sûr de ne pas favoriser la méthode présentée dans ce paragraphe, nous avons donc choisi de fixer l'erreur à 10^{-10} pour la méthode utilisant le développement en série de Taylor.

Pour les différentes méthodes, nous avons estimé les temps de calcul de la matrice de Gram en fonction du nombre de voxels. Nous avons pour cela pris $R = 116$ régions (le nombre de régions de l'atlas AAL). Pour chaque nombre de voxels fixé, nous avons effectué 100 tirages aléatoires de matrices stochastiques à droite pour estimer le temps de calcul moyen. L'écart type est trop faible ($< 1\%$) pour pouvoir être représenté sur le graphique. Les calculs sont effectués avec un processeur de 3.6 GHz avec 2 Gb de RAM. Les temps de calculs des différentes méthodes sont présentés dans la figure 3.6. En pratique, on gagne plus d'un facteur 10 en temps de calcul (près d'un facteur 20 dans nos expériences du chapitre 4).

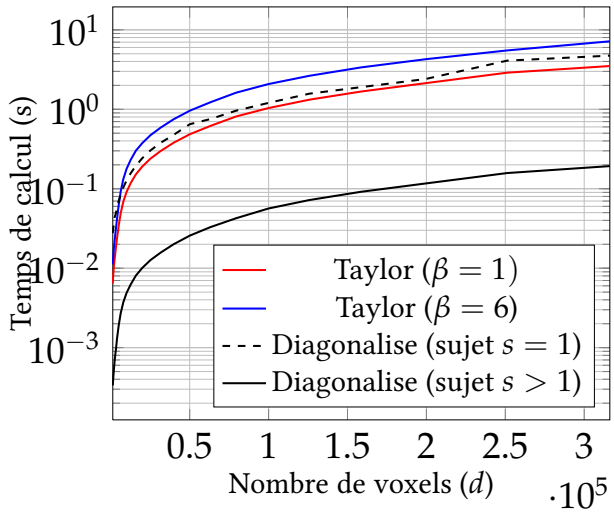


FIGURE 3.6 : Temps de calcul de $e^{-\beta L}\mathbf{x}_s$ avec la régularisation anatomique en fonction du nombre de voxels pour la méthode de développement en série de Taylor et pour la méthode par diagonalisation. Dans la première, le temps de calcul dépend de β. La seconde met plus de temps pour le premier sujet. Les calculs sont effectués avec un processeur de 3.6 GHz avec 2 Gb de RAM.

3.4.3 Choix du paramètre de diffusion

La régularisation utilisée pénalise exponentiellement les composantes de hautes fréquences du graphe de régularisation. Plus précisément, chaque composante est pondérée par $e^{-\beta\mu}$ où μ est la valeur propre correspondante. Dans l'approche décrite précédemment, les valeurs propres et vecteurs propres du graphe de régularisation sont connus. Il est donc possible de choisir le paramètre β en fonction.

La méthode décrite dans cette section s'applique directement au cas surfacique. Malheureusement le faible coût en termes de calculs a été obtenu au détriment de la proximité spatiale. Dans la section suivante (3.5), nous verrons comment combiner ces deux approches.

3.5 Combiner les régularisations spatiales et anatomiques

Dans les deux sections précédentes, nous avons vu comment définir une régularisation spatiale (section 3.3) et une régularisation anatomique (section 3.4). Dans cette section nous proposons deux manières de combiner ces deux proximités.

Il y a principalement deux approches différentes pour combiner les proximités spatiales et anatomiques. La première approche consiste à considérer les deux types de proximités comme deux concepts totalement distincts. La combinaison des deux peut alors se faire en sommant les termes de régularisation. C'est ce que nous verrons dans la section 3.5.1. Une autre façon de voir la combinaison est de la considérer comme une modification locale de la topologie induite par la proximité spatiale prenant en compte l'information anatomique. Nous en proposons une première approche discrète dans la section 3.5.2. Nous en donnerons ensuite une version continue à la section 3.5.3.

3.5.1 Somme des termes de régularisation

3.5.1.1 Le problème d'optimisation

Dans la suite du manuscrit, lorsqu'il y aura confusion possible entre la régularisation spatiale et la régularisation anatomique, nous ajouterons des indices pour différencier la régularisation spatiale ($_\mathrm{s}$) de la régularisation anatomique ($_\mathrm{a}$). Par exemple, L_s est le laplacien du graphe de régularisation spatiale et L_a celui de la régularisation anatomique. L'approche la plus directe pour combiner ces deux régularisations est de sommer les opérateurs de régularisation. On obtient alors le problème d'optimisation suivant :

$$\left(\mathbf{w}^{\mathrm{opt}}, b^{\mathrm{opt}}\right) = \underset{\mathbf{w}\in\mathcal{X}, b\in\mathbb{R}}{\arg\min} \frac{1}{N}\sum_{s=1}^{N} \ell_{\mathrm{hinge}}\left(y_s\left[\langle \mathbf{w}, \mathbf{x}_\mathrm{s}\rangle + b\right]\right) + \lambda_\mathrm{s}\|e^{\frac{\beta_\mathrm{s}}{2}L_\mathrm{s}}\mathbf{w}\|^2 + \lambda_\mathrm{a}\|e^{\frac{\beta_\mathrm{a}}{2}L_\mathrm{a}}\mathbf{w}\|^2 \tag{3.36}$$

avec λ_s et λ_a des paramètres de régularisation. Dans la suite, nous n'avons considéré que le cas $\lambda_\mathrm{s} = \lambda_\mathrm{a} = \lambda$.

La somme de deux matrices définies positives est une matrice définie positive (puisqu'elles forment un cône). Le problème d'optimisation (3.36) est donc, d'après la section 3.2.1, celui d'un SVM de noyau :

$$K_{\beta_\mathrm{a},\beta_\mathrm{s}}(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1^T\left(e^{\beta_\mathrm{a}L_\mathrm{a}} + e^{\beta_\mathrm{s}L_\mathrm{s}}\right)^{-1}\mathbf{x}_2 \tag{3.37}$$

3.5.1.2 Calcul de la matrice de Gram : cas général

Comme nous l'avons vu dans les sections précédentes (sections 3.3 et 3.4), les exponentielles de matrices $e^{\beta_a L_a}\mathbf{x}_s$ et $e^{\beta_s L_s}\mathbf{x}_s$ se calculent facilement. Il est donc possible de calculer $\left(e^{\beta_a L_a} + e^{\beta_s L_s}\right)^{-1}\mathbf{x}_s$ à l'aide d'un gradient conjugué [Golub & Van Loan, 1996].

Proposition 3.5.1 *Avec les notations précédentes, si de plus nous supposons que les deux laplaciens sont normalisés, le nombre d'itérations du gradient conjugué nécessaire pour obtenir* $\left(e^{\beta_a L_a} + e^{\beta_s L_s}\right)^{-1}\mathbf{x}_s$ *avec une erreur inférieure à* η *est au plus de :*

$$\left\lceil \log\left(\frac{\eta}{2}\right)\left(\log\left(\frac{\sqrt{e^{\beta_a}+e^{2\beta_s}}-\sqrt{2}}{\sqrt{e^{\beta_a}+e^{2\beta_s}}+\sqrt{2}}\right)\right)^{-1}\right\rceil \tag{3.38}$$

Démonstration Nous supposons que les deux laplaciens, L_a et L_s, sont normalisés. Leurs valeurs propres sont donc comprises dans l'intervalle $[0,2]$ [Chung, 1992]. Ce résultat s'obtient en utilisant le lien entre le quotient de Rayleigh et les valeurs propres et en utilisant l'inégalité suivante :

$$\forall (x,y) \in \mathbb{R}^2, (x-y)^2 \leq 2\left(x^2+y^2\right)$$

Le spectre de $e^{\beta_s L_s}$ est donc inclus dans l'intervalle $[1, e^{2\beta_s}]$. Puisque nous avons la relation suivante :

$$\forall x \in \mathbb{R}^d,\ \mathbf{x}^T\tilde{E}\tilde{E}^T\mathbf{x} = \|\tilde{E}^T\mathbf{x}\|_2 \geq 0$$

le spectre de $e^{\beta_a L_a}$ est inclus dans $[1, e^{\beta_a}]$.

Le conditionnement κ_2 de la matrice $\left(e^{\beta_a L_a} + e^{\beta_s L_s}\right)$ est pour la norme spectrale défini par :

$$\kappa_2 = \left\|e^{\beta_a L_a} + e^{\beta_s L_s}\right\|_2 \left\|\left(e^{\beta_a L_a} + e^{\beta_s L_s}\right)^{-1}\right\|_2$$

Le premier terme est majoré par :

$$\begin{aligned}\left\|e^{\beta_a L_a} + e^{\beta_s L_s}\right\|_2 &\leq \left\|e^{\beta_a L_a}\right\|_2 + \left\|e^{\beta_s L_s}\right\|_2 \\ &\leq e^{\beta_a} + e^{2\beta_s}\end{aligned}$$

De plus, $\left(e^{\beta_a L_a} + e^{\beta_s L_s}\right)$ étant une matrice réelle définie positive, le second terme est borné par :

$$\begin{aligned}\left\|\left(e^{\beta_a L_a} + e^{\beta_s L_s}\right)^{-1}\right\|_2^{-1} &= \min_{\mathbf{x}\in\mathbb{R}^d} \frac{\mathbf{x}^T\left(e^{\beta_a L_a} + e^{\beta_s L_s}\right)\mathbf{x}}{\mathbf{x}^T\mathbf{x}} \\ &\geq \min_{\mathbf{x}\in\mathbb{R}^d} \frac{\mathbf{x}^T e^{\beta_a L_a}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} + \min_{\mathbf{x}\in\mathbb{R}^d} \frac{\mathbf{x}^T e^{\beta_s L_s}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} \\ &\geq 2\end{aligned}$$

Par conséquent,

$$\kappa_2 \leq \frac{e^{\beta_a} + e^{2\beta_s}}{2}$$

D'après Golub & Van Loan [1996], l'erreur résiduelle, η, est majorée à la i-ème iteration par :

$$\eta \leq 2\left(\frac{\sqrt{\kappa_2}-1}{\sqrt{\kappa_2}+1}\right)^i$$

Par conséquent, en utilisant cette majoration de κ_2, on obtient :

$$\eta \leq 2\left(\frac{\sqrt{e^{\beta_a}+e^{2\beta_s}}-\sqrt{2}}{\sqrt{e^{\beta_a}+e^{2\beta_s}}+\sqrt{2}}\right)^i$$

Le nombre maximal d'itérations est donc de :

$$\left\lceil \log\left(\frac{\eta}{2}\right) \log\left(\frac{\sqrt{e^{\beta_a}+e^{2\beta_s}}-\sqrt{2}}{\sqrt{e^{\beta_a}+e^{2\beta_s}}+\sqrt{2}}\right)^{-1} \right\rceil$$

■

Proposition 3.5.2 *Avec les mêmes hypothèses, si l'on considère la factorisation suivante :*

$$e^{\beta_a L_a}+e^{\beta_s L_s} = e^{\frac{\beta_s}{2}L_s}\left(I_d + e^{\frac{-\beta_s}{2}L_s}e^{\beta_a L_a}e^{\frac{-\beta_s}{2}L_s}\right)e^{\frac{\beta_s}{2}L_s}$$

le nombre maximum d'itérations dans le gradient conjugué est alors borné par (figure 3.7) :

$$\left\lceil \log\left(\frac{\eta}{2}\right)\left(\log\left(\frac{\sqrt{1+e^{\beta_a}}-\sqrt{1+e^{-2\beta_s}}}{\sqrt{1+e^{\beta_a}}+\sqrt{1+e^{-2\beta_s}}}\right)\right)^{-1} \right\rceil \tag{3.39}$$

Démonstration On a :

$$\begin{aligned}
\left\|e^{-\frac{\beta_s}{2}L_s}e^{\beta_a L_a}e^{-\frac{\beta_s}{2}L_s}\right\|_2 &= \left\|e^{-\frac{\beta_s}{2}L_s}e^{\beta_a L_a}e^{-\frac{\beta_s}{2}L_s}\right\|_2 \\
&\leq \left\|e^{-\frac{\beta_s}{2}L_s}\right\|_2^2 \left\|e^{\beta_a L_a}\right\|_2 \\
&\leq e^{\beta_a}
\end{aligned}$$

et

$$\begin{aligned}
\left\|\left(e^{-\frac{\beta_s}{2}L_s}e^{\beta_a L_a}e^{-\frac{\beta_s}{2}L_s}\right)^{-1}\right\|_2^{-1} &= \min_{\mathbf{x}\in\mathbb{R}^d}\frac{\mathbf{x}^T e^{\frac{-\beta_s}{2}L_s}e^{\beta_a L_a}e^{\frac{-\beta_s}{2}L_s}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} \\
&= \min_{\mathbf{x}\in\mathbb{R}^d}\frac{\mathbf{x}^T e^{\beta_a L_a}\mathbf{x}}{\mathbf{x}^T e^{\beta_s L_s}\mathbf{x}} \\
&= \min_{\mathbf{x}\in\mathbb{R}^d}\frac{\mathbf{x}^T e^{\beta_a L_a}\mathbf{x}}{\mathbf{x}^T\mathbf{x}}\frac{\mathbf{x}^T\mathbf{x}}{\mathbf{x}^T e^{\beta_s L_s}\mathbf{x}} \\
\left\|\left(e^{-\frac{\beta_s}{2}L_s}e^{\beta_a L_a}e^{-\frac{\beta_s}{2}L_s}\right)^{-1}\right\|_2^{-1} &\leq e^{-2\beta_s}
\end{aligned}$$

Le conditionnement $\tilde{\kappa}_2$ de $\left(I_d + e^{-\frac{\beta_s}{2}L_s}e^{\beta_a L_a}e^{-\frac{\beta_s}{2}L_s}\right)$ est alors borné par :

$$\tilde{\kappa}_2 \leq \frac{1+e^{\beta_a}}{1+e^{-2\beta_s}} \tag{3.40}$$

Par conséquent, la borne sur le nombre d'itérations devient :

$$\left\lceil \log\left(\frac{\eta}{2}\right)\left(\log\left(\frac{\sqrt{1+e^{\beta_a}}-\sqrt{1+e^{-2\beta_s}}}{\sqrt{1+e^{\beta_a}}+\sqrt{1+e^{-2\beta_s}}}\right)\right)^{-1}\right\rceil$$

Quand $\beta_s \geq \frac{1}{2}\beta_a$, cette borne est plus petite que la précédente.

■

Si la complexité du calcul de $e^{-\beta L}$ est proportionnelle à β (comme c'est le cas dans l'approche surfacique) et si de plus $\beta_s \geq \frac{1}{2}\beta_a$, en utilisant la factorisation suivante :

$$e^{\beta_a L_a} + e^{\beta_s L_s} = e^{\frac{\beta_s}{2}L_s}\left(I_d + e^{\frac{-\beta_s}{2}L_s} e^{\beta_a L_a} e^{\frac{-\beta_s}{2}L_s}\right) e^{\frac{\beta_s}{2}L_s}$$

on obtient donc une meilleure borne sur le nombre maximal d'opérations (figure 3.7).

En particulier, si l'on considère le cas de la régularisation avec un atlas binaire, le spectre de $\tilde{L}_a$ est $\mathrm{Sp}\left(\tilde{L}_a\right) = \{1\}$. Pour β_a dans l'intervalle $[0,6]$ et une erreur résiduelle η inférieure à 10^{-4}, le nombre d'itérations ne dépassera pas 100, et ce, quel que soit β_s. Cette vérification nous montre que le calcul de la matrice de Gram est tout à fait réalisable. En pratique, dans nos expériences du chapitre 4, le nombre d'itérations ne dépasse pas 43.

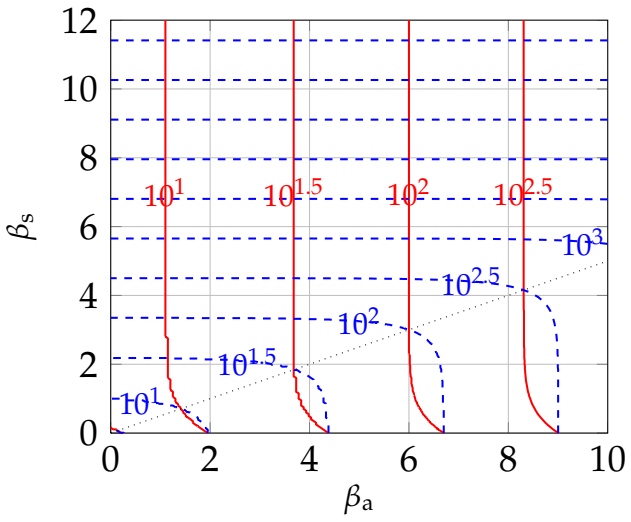


FIGURE 3.7 : Borne sur le nombre d'itérations pour le gradient conjugué d'après l'inégalité (3.38) (en pointillés bleus) et l'inégalité (3.39) (ligne pleine rouge) pour une erreur résiduelle fixée à $\eta = 10^{-4}$.

3.5.1.3 Calcul de la matrice de Gram : cas gaussien

Dans le cas volumique, si la proximité spatiale est encodée par le graphe de connectivité des voxels de l'image (connectivité 6), alors la matrice L_s est diagonalisable dans une base orthonormée avec comme matrice de passage Q la partie imaginaire d'une matrice extraite de la matrice de la transformée de Fourier discrète (TFD).

En effet, en considérant le graphe de connectivité comme le produit cartésien de trois graphes de connectivité en dimension 1, on obtient le résultat ci-dessous. Soit d_i la taille de l'image (en voxels) dans la dimension i. Les valeurs propres de L sont données par :

$$\mu_\alpha = \sum_{i=1}^{3} 4\sin^2\left(\frac{\alpha_i \pi}{2\,(d_i+1)}\right) \tag{3.41}$$

pour tout $\alpha \in [1, d_1] \times [1, d_2] \times [1, d_3]$. Les coordonnées des vecteurs propres correspondants $\mathbf{u}^\alpha = \left(u^\alpha_{\mathbf{j}}\right)_{\mathbf{j}}$, vérifient :

$$u^\alpha_{\mathbf{j}} = \prod_{i=1}^{3} \sqrt{\frac{2}{d_i+1}} \sin\left(\frac{j_i \alpha_i \pi}{d_i+1}\right) \tag{3.42}$$

et ce, pour tout $\mathbf{j} = (j_1, j_2, j_3) \in [1, d_1] \times [1, d_2] \times [1, d_3]$. La matrice Q est la matrice dont les colonnes sont ces vecteurs propres. Par conséquent, multiplier un vecteur colonne par Q ne requiert que $O(d \log(d))$ opérations en utilisant la transformée de sinus discrète (TSD).

Soit S la matrice diagonale telle que : $L_s = QSQ$. On a donc, en utilisant l'équation (3.32) :

$$e^{\beta_a L_a} + e^{\beta_s L_s} = e^{\beta_s QSQ} + \tilde{X}\left[e^{\beta_a \tilde{\Lambda}} - e^{\beta_a} I_R\right]\tilde{X}^T + e^{\beta_a} I_d \tag{3.43}$$

Proposition 3.5.3 (Égalité matricielle de Woodbury [Golub & Van Loan, 1996]) *Soient $n \geq k \geq 1$. Soient $A \in \mathbb{R}^{n\times n}$, $C \in \mathbb{R}^{k\times k}$, $U \in \mathbb{R}^{n\times k}$ et $V \in \mathbb{R}^{k\times n}$. Si les matrices A, C et $\left(C^{-1} + VA^{-1}U\right)$ sont inversibles, alors on a l'égalité suivante :*

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U\left(C^{-1} + VA^{-1}U\right)^{-1} VA^{-1} \tag{3.44}$$

En utilisant l'égalité matricielle de Woodbury (équation (3.44)), on a :

$$\left(e^{\beta_a L_a} + e^{\beta_s L_s}\right)^{-1} = \mathbf{D} - \mathbf{D}\tilde{X}\left[\left(e^{\beta_a \tilde{\Lambda}} - e^{\beta_a} I_R\right)^{-1} + \tilde{X}^T \mathbf{D} \tilde{X}\right]^{-1} \tilde{X}^T \mathbf{D} \tag{3.45}$$

avec :

$$\mathbf{D} = Q\left(e^{\beta_s S} + e^{\beta_a} I_d\right)^{-1} Q$$

En termes de complexité, les étapes les plus coûteuses en nombre d'opérations sont le calcul de $\mathbf{D}\tilde{X}$ et la multiplication par $\mathbf{D}$. Par conséquent, en utilisant l'équation (3.4.2.3), la complexité est en :

$$O\left((N+R)d\log_2(d) + R^3\right)$$

3.5.1.4 Le choix des paramètres β_s et β_a

Les paramètres β_s et β_a sont choisis en utilisant les remarques des deux sections précédentes.

3.5.2 Modification du graphe de régularisation

3.5.2.1 Graphe de régularisation

La deuxième approche consiste à modifier localement le graphe de régularisation pour qu'il prenne en compte l'information de type anatomique.

Nous supposons que nous avons un atlas anatomique de R régions : $\{\mathcal{A}_r\}_{r=1,\cdots,R}$. En chaque élément $v \in \mathcal{V}$, nous avons une distribution de probabilité $p_{\text{atlas}}(\cdot|v) \in \mathbb{R}^{\mathcal{A}}$. Cette distribution contient l'information provenant de l'atlas au niveau du voxel v.

Nous proposons de prendre comme graphe de régularisation le graphe défini de la façon suivante. Deux voxels u et v sont voisins si et seulement s'ils sont voisins dans l'image (connectivité 6). Le poids de l'arc entre u et v est alors :

$$A_{u,v} = \exp\left(\frac{-\chi^2(p(\cdot|u), p(\cdot|v))^2}{2\sigma^2}\right) \tag{3.46}$$

avec $\sigma \in \mathbb{R}^{+*}$ un paramètre et χ^2 la distance du χ^2 définie par :

$$\chi^2(p(\cdot|u), p(\cdot|v))^2 = \frac{1}{2}\sum_{r=1}^{R}\frac{(p(\mathcal{A}_r|u) - p(\mathcal{A}_r|v))^2}{p(\mathcal{A}_r|u) + p(\mathcal{A}_r|v)} \tag{3.47}$$

Nous n'avons pas considéré cette approche dans le cas surfacique. Une possibilité serait de pondérer les poids de la matrice d'adjacence (équation 3.46) par les poids conformaux.

3.5.2.2 Calcul de la matrice de Gram

Dans cette approche, nous utilisons le développement en série de Taylor présenté à la section 3.2.1 pour le calcul de la matrice de Gram. Le calcul reste rapide puisque dans cette approche, la matrice d'adjacence du graphe de régularisation est très creuse.

3.5.2.3 Choix de σ et du paramètre de diffusion

En ce qui concerne σ, afin d'éviter l'optimisation a posteriori d'un hyperparamètre supplémentaire, nous l'avons choisi comme étant égal à l'écart type (estimé) de $\chi^2(p(\cdot|u), p(\cdot|v))$.

En ce qui concerne le paramètre de diffusion β c'est un peu plus délicat. Nous n'avons, dans cette approche, aucune information sur les valeurs propres du laplacien du graphe de

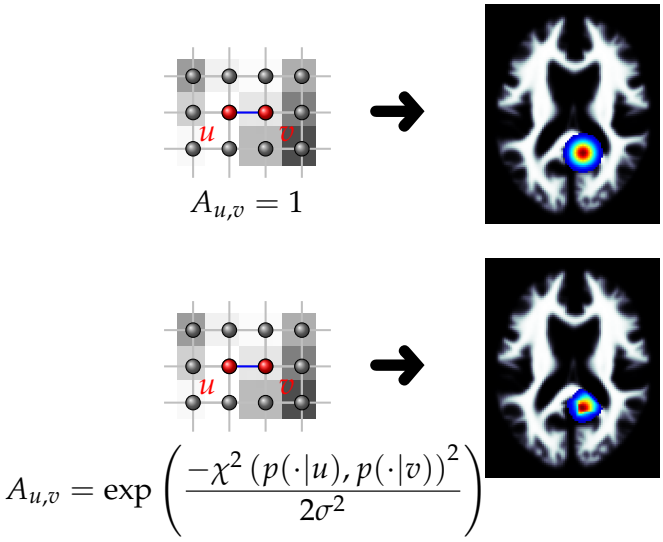


Figure 3.8 : Lorsque le graphe de régularisation est le graphe de connectivité de l'image (ligne du haut), la régularisation revient à lisser les données avec un noyau gaussien. La ligne du bas montre le graphe de régularisation proposé. Il prend en compte les différents types de tissus.

régularisation. Nous essayons alors de nous rapprocher au maximum du cas gaussien et de fixer β de la même manière que dans le cas gaussien. Pour cela, nous normalisons les poids de la matrice d'adjacence pour qu'ils valent un en moyenne.

3.5.3 Variétés riemanniennes

Dans cette section, l'objectif est de prendre en compte différentes informations *a priori* telles que des informations sur les tissus (GM, WM, CSF), des informations provenant d'un atlas anatomique ou fonctionnel ou encore des informations sur la localisation spatiale des voxels. Nous verrons que c'est possible en utilisant une variété statistique avec la métrique de Fisher. Nous donnerons ensuite des détails sur l'implémentation de la matrice de Gram.

3.5.3.1 Métrique de Fisher

Considérons une position $v \in \mathbb{R}^3$. Les images avec lesquelles nous travaillons sont recalées. On connaît donc, lorsque l'on considère la position v, la véritable localisation à une erreur de recalage près. Cette information sur la localisation peut s'exprimer sous la forme d'une densité de probabilité $x \in \mathbb{R}^3 \mapsto p_{\text{loc}}(x|v)$. Cette dernière encode l'information sur la localisation spatiale de v. Un exemple simple est $p_{\text{loc}}(\cdot|v) \sim \mathcal{N}(v, \sigma^2_{\text{loc}})$. On peut voir cette information comme un indice de confiance sur la localisation v.

Nous supposons également que nous avons un atlas anatomique ou fonctionnel de R régions : $\{\mathcal{A}_r\}_{r=1,\cdots,R}$. En chaque élément v, nous avons une distribution de probabilités $\mathcal{A}_r \in \mathcal{A} \mapsto p_{\text{atlas}}(\mathcal{A}|v)$. Cette distribution contient l'information provenant de l'atlas au niveau du voxel v.

Ainsi, en chaque point position $v \in \mathbb{R}^3$, nous avons une information sur la position spatiale et une information de type anatomique sur la région dans laquelle on se situe qui peut être considérée comme une densité de probabilité $p(\cdot|v) \in \mathbb{R}^{\mathcal{A}\times\mathbb{R}^3}$. Ainsi nous considérons la famille paramétrée (de paramètre $v \in \mathbb{R}^3$) de distribution de probabilités suivante : $\mathcal{M} = \left\{p(\cdot|v) \in \mathbb{R}^{\mathcal{A}\times\mathbb{R}^3}\right\}_{v\in\mathbb{R}^3}$. En d'autres termes, dans cette section, les voxels ne sont pas considérés comme tels mais chaque voxel est décrit par des distributions de probabilités informant sur les régions de l'atlas probable en ce voxel et sur la localisation spatiale. Pour simplifier les calculs, nous supposons dans la suite[6] que les probabilités sont indépendantes, autrement dit que : $p = p_{\text{loc}}p_{\text{atlas}}$.

Dans la suite du manuscrit, nous supposerons que p est suffisamment lisse selon v et ne s'annule pas sur $\mathbb{R}^3$. Nous supposerons également que la matrice d'information de Fisher est définie en chaque point $v \in \mathbb{R}^3$. Avec ces hypothèses, la famille $\mathcal{M}$ peut être considérée comme une variété différentielle où le paramètre $v \in \mathbb{R}^3$ joue localement un rôle de système de coordonnées [Amari et al., 1987]. La notion de proximité se définit alors assez naturellement en utilisant la métrique de Fisher comme l'ont fait Amari et al. [1987]. Cette métrique a été utilisée dans le cadre de l'apprentissage par SVM par Lafferty & Lebanon [2005]. Munie de cette métrique, $\mathcal{M}$ est une variété riemannienne [Amari et al., 1987] appelée **variété statistique** ou *statistical manifold.*

Pour des raisons principalement de clarté, nous présentons cette régularisation uniquement dans le cadre volumique. Elle pourrait également être utilisée dans le cadre surfacique avec quelques petites modifications. Le tenseur g de la métrique de Fisher est alors défini par :

$$g_{ij}(v) = \mathbf{E}_v\left[\frac{\partial \log p(\cdot|v)}{\partial v_i}\frac{\partial \log p(\cdot|v)}{\partial v_j}\right],\ 1 \leq i,j \leq 3 \tag{3.48}$$

$\mathcal{M}$, munie de la métrique de Fisher, est une variété riemannienne [Amari et al., 1987]. L'ensemble $\mathcal{V}$ étant compact et $v \mapsto p(\cdot|v)$ continue, le sous-ensemble $\left\{p(\cdot|v) \in \mathbb{R}^{\mathcal{A}\times\mathbb{R}^3}\right\}_{v\in\mathcal{V}} \subset \mathcal{M}$ est compact. On se retrouve alors dans les conditions de la section 3.2.2.

Proposition 3.5.4 *Si l'on suppose que $p_{\text{loc}}(\cdot|v)$ est isotrope, nous avons :*

$$g_{ij}(v) = g_{ij}^{\text{atlas}}(v) + \delta_{ij}\int_{u\in\mathbb{R}^3} p_{\text{loc}}(u|v)\left(\frac{\partial \log p_{\text{loc}}(u|v)}{\partial v_i}\right)^2 du \tag{3.49}$$

où δ_{ij} est de symbole de Kronecker et g^{atlas} est le tenseur de métrique obtenu lorsque : $p(\cdot|v) = p_{\text{atlas}}(\cdot|v)$.

[6] même si cette hypothèse est forcément fausse

Démonstration D'après l'équation (3.48), si l'on suppose que :

$$p(\cdot|v) = p_{\text{atlas}}(\cdot|v)p_{\text{loc}}(\cdot|v)$$

alors nous pouvons écrire :
$1 \leq i,j \leq 3,\ \forall v \in \mathcal{V}$:

$$g_{ij}(v) = \sum_{r=1}^{R} \int_{u\in\mathbb{R}^3} p_{\text{loc}}(u|v)p_{\text{atlas}}(\mathcal{A}_r|v) \cdot \frac{\partial \log p_{\text{loc}}(u|v)p_{\text{atlas}}(\mathcal{A}_r|v)}{\partial v_i} \cdot \frac{\partial \log p_{\text{loc}}(u|v)p_{\text{atlas}}(\mathcal{A}_r|v)}{\partial v_j} du \tag{3.50}$$

Ce qui donne en développant :
$1 \leq i,j \leq 3,\ \forall v \in \mathcal{V}$:

$$\begin{aligned} g_{ij}(v) \quad = \quad & g_{ij}^{\text{atlas}}(v) \\ & + \int_{u\in\mathbb{R}^3} p_{\text{loc}}(u|v) \frac{\partial \log p_{\text{loc}}(u|v)}{\partial v_i} \frac{\partial \log p_{\text{loc}}(u|v)}{\partial v_j} du \\ & + \left(\sum_{r=1}^{R} p_{\text{atlas}}(\mathcal{A}_r|v) \frac{\partial \log p_{\text{atlas}}(\mathcal{A}_r|v)}{\partial v_j} \right) \int_{u\in\mathbb{R}^3} p_{\text{loc}}(u|v) \frac{\partial \log p_{\text{loc}}(u|v)}{\partial v_i} du \\ & + \left(\sum_{r=1}^{R} p_{\text{atlas}}(\mathcal{A}_r|v) \frac{\partial \log p_{\text{atlas}}(\mathcal{A}_r|v)}{\partial v_i} \right) \int_{u\in\mathbb{R}^3} p_{\text{loc}}(u|v) \frac{\partial \log p_{\text{loc}}(u|v)}{\partial v_j} du \end{aligned} \tag{3.51}$$

avec, $1 \leq i,j \leq 3,\ \forall v \in \mathcal{V}$:

$$g_{ij}^{\text{atlas}}(v) = \sum_{r=1}^{R} p_{\text{atlas}}(\mathcal{A}_r|v) \frac{\partial \log p_{\text{atlas}}(\mathcal{A}_r|v)}{\partial v_i} \frac{\partial \log p_{\text{atlas}}(\mathcal{A}_r|v)}{\partial v_j}$$

Si on suppose de plus que $p_{\text{loc}}(\cdot|v)$ est istotrope, les termes suivants sont nuls :

$$\int_{u\in\mathbb{R}^3} p_{\text{loc}}(u|v) \frac{\partial \log p_{\text{loc}}(u|v)}{\partial v_i} du = 0$$

et pour $i \neq j$:

$$\int_{u\in\mathbb{R}^3} p_{\text{loc}}(u|v) \frac{\partial \log p_{\text{loc}}(u|v)}{\partial v_i} \frac{\partial \log p_{\text{loc}}(u|v)}{\partial v_j} du = 0$$

■

Proposition 3.5.5 *Si de plus* $p_{\text{loc}}(\cdot|v) \sim \mathcal{N}(v, \sigma_{\text{loc}}^2 I_3)$, *on obtient :*

$$g_{ij}(v) = g_{ij}^{\text{atlas}}(v) + \frac{\delta_{ij}}{\sigma_{\text{loc}}^2} \tag{3.52}$$

Démonstration On utilise l'équation (3.49) et on effectue une intégration par parties.

■

Remarque : Le second terme $\frac{\delta_{ij}}{\sigma_{\text{loc}}^2}$ assure que $g_{ij}(v)$, la matrice d'information de Fisher en v, est bien définie.

3.5.3.2 Calcul de la matrice de Gram : résolution de l'équation de la chaleur

Dans ce paragraphe, l'indice du sujet s est fixé. Pour résoudre l'équation de la chaleur (3.20), une possibilité est d'utiliser une approche variationnelle [Druet et al., 2004]. Nous utilisons les éléments finis d'ordre 1 sur un maillage rectangulaire $\left\{\psi^{(i)}\right\}$ pour la discrétisation de l'espace et les différences finies avec un schéma explicite pour la discrétisation en temps. Soient Δ_x et Δ_t respectivement le pas en espace et le pas en temps. Soient $U(t)$ les coordonnées de $u(t)$, U^n celles de $u(t = n\Delta_t)$ et U^0 celles de $\mathbf{x}_s$. L'approche variationnelle donne :

$$\begin{cases} \mathbf{M}\dfrac{dU}{dt}(t) + \mathbf{K}U(t) = 0 \\ U(t=0) = U^0 \end{cases} \tag{3.53}$$

où $\mathbf{K}$ est la matrice de rigidité et $\mathbf{M}$ la matrice de masse. La matrice de rigidité, $\mathbf{K}$, est définie par :

$$\mathbf{K}_{i,j} = \int_{v\in\mathcal{V}} \left\langle \nabla_{\mathcal{M}}\psi^{(i)}(v), \nabla_{\mathcal{M}}\psi^{(j)}(v) \right\rangle_{\mathcal{M}} d\mu_{\mathcal{M}} \tag{3.54}$$

Plus précisément, comme nous utilisons les éléments finis $\left\{\psi^{(i)}\right\}_i$ obtenus en translatant la fonction $\psi^{(0)}$ centrée en zéro et définie par :

$$\psi^{(0)}(x,y,z) = \left(1 - \frac{|x|}{\Delta_x}\right)\left(1 - \frac{|y|}{\Delta_x}\right)\left(1 - \frac{|z|}{\Delta_x}\right)$$

pour tout $(x,y,z) \in [-\Delta_x, +\Delta_x]^3$ et zéro partout ailleurs.

Comme la matrice de rigidité $\mathbf{K}$ vérifie :

$$\begin{aligned} \mathbf{K}_{i,j} &= \int_{v\in\mathcal{V}} \left\langle \nabla_{\mathcal{M}}\psi^{(i)}(v), \nabla_{\mathcal{M}}\psi^{(j)}(v) \right\rangle_{\mathcal{M}} d\mu_{\mathcal{M}} \\ &= \int_{v\in\mathcal{V}} h(v)\left(\nabla\psi^{(1)}(v), \nabla\psi^{(2)}(v)\right)\sqrt{\det(g(v))}dx \end{aligned}$$

On pose : $\tilde{h} = \sqrt{\det(g)}\cdot h$. En utilisant la formule des trapèzes pour approcher ces intégrales, on a :

$$\begin{aligned} \mathbf{K}_{i,i} \approx \Delta_x \big(&\left[\tilde{h}_{11}(0,0,0) + \tilde{h}_{22}(0,0,0) + \tilde{h}_{33}(0,0,0)\right] \\ &+\tfrac{1}{2}\left[\tilde{h}_{11}(+\Delta_x,0,0) + \tilde{h}_{22}(0,+\Delta_x,0) + \tilde{h}_{33}(0,0,+\Delta_x)\right] \\ &+\tfrac{1}{2}\left[\tilde{h}_{11}(-\Delta_x,0,0) + \tilde{h}_{22}(0,-\Delta_x,0) + \tilde{h}_{33}(0,0,-\Delta_x)\right]\big) \end{aligned}$$

Dans la suite, sans perte de généralité, nous supposons que $\psi^{(i)} = \psi^{(0)}$.

Pour $\psi^{(j)}$ centrée en $(\epsilon_1\Delta_x, \epsilon_2\Delta_x, 0)$, $(\epsilon_1,\epsilon_2) \in \{\pm 1\}^2$:

$$\mathbf{K}_{i,j} \approx -\epsilon_1\epsilon_2\frac{\Delta_x}{4}\left[\tilde{h}_{12}(\epsilon_1\Delta_x,0,0) + \tilde{h}_{21}(0,\epsilon_2\Delta_x,0)\right]$$

Pour $\psi^{(j)}$ centrée en $(\epsilon_1\Delta_x, 0, \epsilon_3\Delta_x)$, $(\epsilon_1, \epsilon_3) \in \{\pm 1\}^2$:

$$\mathbf{K}_{i,j} \approx -\epsilon_1\epsilon_3\frac{\Delta_x}{4}\left[\tilde{h}_{13}(\epsilon_1\Delta_x, 0, 0) + \tilde{h}_{31}(0, 0, \epsilon_3\Delta_x)\right]$$

Pour $\psi^{(j)}$ centrée en $(0, \epsilon_2\Delta_x, \epsilon_3\Delta_x)$, $(\epsilon_2, \epsilon_3) \in \{\pm 1\}^2$:

$$\mathbf{K}_{i,j} \approx -\epsilon_2\epsilon_3\frac{\Delta_x}{4}\left[\tilde{h}_{23}(0, \epsilon_2\Delta_x, 0) + \tilde{h}_{32}(0, 0, \epsilon_3\Delta_x)\right]$$

Et pour $\psi^{(j)}$ centrée en $(\epsilon\Delta_x, 0, 0)$, $\epsilon \in \{\pm 1\}$:

$$\mathbf{K}_{i,j} \approx -\frac{\Delta_x}{2}\left[\tilde{h}_{11}(0, 0, 0) + \tilde{h}_{11}(\epsilon\Delta_x, 0, 0)\right]$$

Les autres éléments non nuls de $\mathbf{K}$ sont obtenus en permutant les dimensions. Tout le reste est nul.

Quant à la matrice de masse $\mathbf{M}$, elle vérifie :

$$\mathbf{M}_{i,j} = \int_{v\in\mathcal{V}} \psi^{(i)}(v)\psi^{(j)}(v)d\mu_{\mathcal{M}} \tag{3.55}$$

L'approximation de l'intégrale en utilisant la règle des trapèzes donne : $\mathbf{M}_{i,j} \approx \delta_{ij} \det g\left(v^{(i)}\right)$.

La discrétisation en temps utilise les différences finies avec un schéma explicite. Ainsi, pour $n \in \mathbb{N}$, U^{n+1} vérifie :

$$\mathbf{M}U^{n+1} = (\mathbf{M} - \Delta_t\mathbf{K})\,U^n \tag{3.56}$$

Δ_x est fixé par la résolution spatiale de l'IRM. Δ_t doit alors être choisi de sorte à ce que la condition de Courant-Friedrichs-Lewy (CFL) soit respectée. Dans notre cas, cette condition s'écrit :

$$\Delta_t \leq 2(\max\lambda_i)^{-1}$$

avec λ_i les valeurs propres du problème général aux valeurs propres suivant : $\mathbf{K}U = \lambda\mathbf{M}U$. La complexité en terme de nombre d'opérations est donc en :

$$O\left(N\beta(\max_i \lambda_i)d\right)$$

Afin de calculer le pas de temps le moins coûteux en termes de temps de calcul, nous calculons la valeur propre la plus grande à l'aide la méthode des puissances [Golub & Van Loan, 1996]. Avec nos données, pour $\sigma_{\text{loc}} = 5$, $\lambda_{\max} \approx 15.4$ et pour $\sigma_{\text{loc}} = 10$, $\lambda_{\max} \approx 46.5$.

3.5.3.3 Choix du paramètre de diffusion

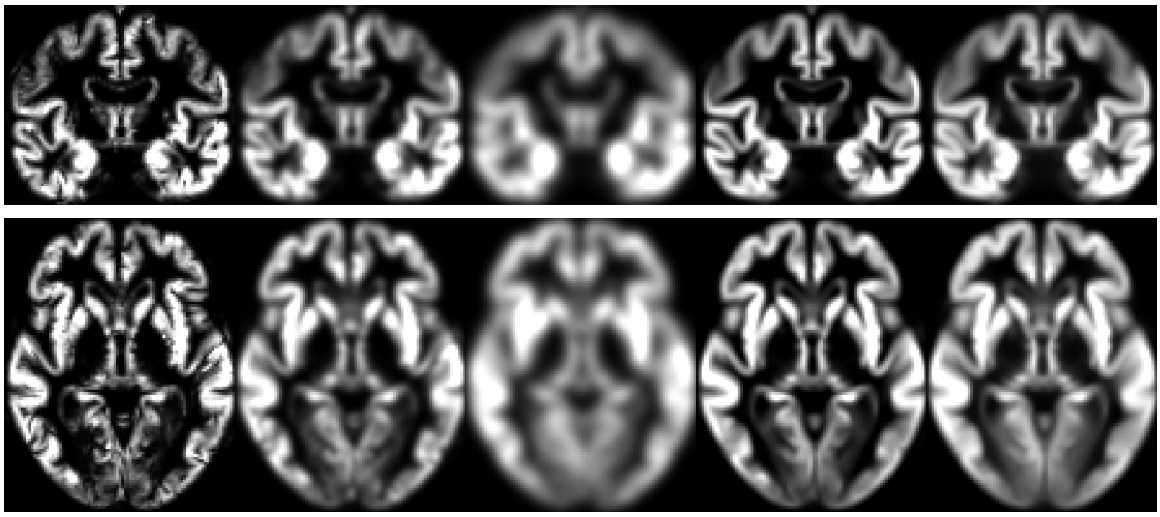
Dans ce paragraphe nous nous focalisons sur le choix du paramètre de diffusion β. Nous supposons σ_{loc} fixé *a priori*. L'évaluation du spectre de l'opérateur de Laplace-Beltrami est très délicate compte tenu de la taille des images. Nous adoptons donc la procédure suivante pour choisir le paramètre de diffusion.

Pour que le tenseur de métrique soit comparable avec celui dans le cas gaussien (spatial uniquement), nous normalisons g par :

$$\left(\frac{1}{|\mathcal{V}|}\int_{u\in\mathcal{V}}\frac{1}{3}\text{tr}\left(g^{\frac{1}{2}}(u)\right)du\right)^2$$

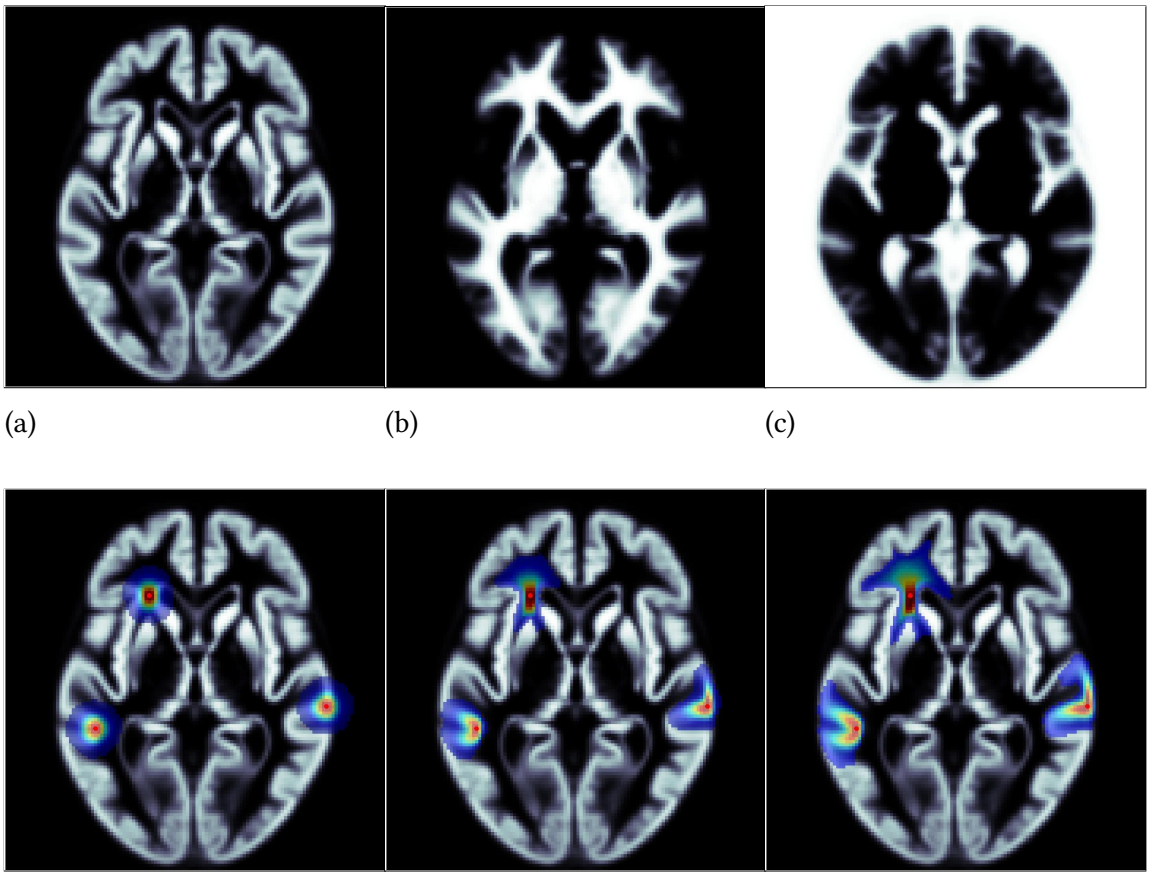
Les valeurs propres de g sont calculées en chaque voxel à l'aide de la méthode de résolution des équations du 3ème degré de Cardano. Ainsi cette étape de normalisation est peu coûteuse en temps de calcul.

Le paramètre β est ensuite choisi pour être équivalent au paramètre de diffusion d'un lissage gaussien, c'est à dire $\beta = \sigma^2$, où σ désigne l'écart type du noyau de lissage gaussien.



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

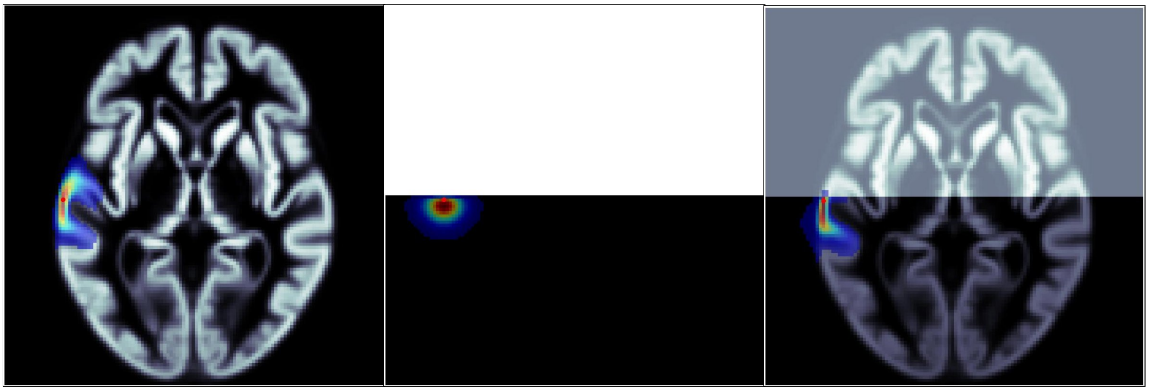
FIGURE 3.9 : Carte de probabilité de substance grise d'un sujet témoin : (a) carte originale, (b) carte lissée avec un filtre gaussien de FWHM de 4 mm, (c) carte lissée avec un filtre gaussien de FWHM de 8 mm, (d)-(e) carte lissée avec $e^{-\frac{\beta}{2}\Delta}$ où Δ représente l'opérateur de Laplace-Beltrami de la variété statistique et β correspond à un filtre gaussien de FWHM respectivement de 4 mm et 8 mm FWHM. La figure 3.10 présente une coupe de l'atlas utilisé pour la métrique de Fisher.



(a) (b) (c)

(d) (e) (f)

Figure 3.10 : Illustration sur une coupe 2D de la régularisation dans le cadre de la métrique de Fisher. La rangée du haut (a-c) représente l'atlas utilisé comme information *a priori* pour la régularisation. Il s'agit d'une coupe d'un atlas de tissus – substance grise (a), substance blanche (b) et autre (c). La ligne du bas représente le résultat de l'équation de la chaleur de la somme de trois Dirac (points rouges) superposés à la carte (a). Le paramètre β est choisi de sorte à être équivalent à un lissage gaussien de 8 mm. On représente la solution de l'équation de la chaleur pour différentes valeurs de σ_{loc} : (d) $\sigma_{\text{loc}} = 1$ mm, (e) $\sigma_{\text{loc}} = 5$ mm, (f) $\sigma_{\text{loc}} = 10$ mm. Le paramètre σ_{loc} quantifie la confiance dans le recalage. Par conséquent, plus σ_{loc} est grand, plus la régularisation se base sur les tissus. Le cas limite $\sigma_{\text{loc}} \to 0$ est l'équation de la chaleur dans l'espace euclidien (grâce à la normalisation du tenseur).



(a) (b) (c)

FIGURE 3.11 : Illustration sur une coupe 2D de la régularisation dans le cadre de la métrique de Fisher pour différents atlas pour β équivalent à un lissage gaussien de 8 mm et σ_{loc} (cf. figure 3.10). La condition initiale est un Dirac (représenté par un point rouge). (a) l'atlas utilisé présenté à la figure 3.10 ; (b) l'atlas utilisé est composé de deux régions : moitié haute de l'image (région blanche) et moitié basse de l'image (région noire) ; (c) la métrique prend en compte les deux atlas précédents.

3.6 Discussion

3.6.1 Autres régularisations spectrales

3.6.1.1 Noyau de diffusion : cas particulier de régularisation spectrale

Dans ce chapitre, nous forçons le classifieur à considérer comme similaires des voxels fortement connectés ou très proches selon la métrique choisie. Ceci est effectué en utilisant un terme de régularisation qui pénalise exponentiellement les composantes à hautes fréquences. Nous avons déjà mentionné à la section 3.2.1 qu'une telle régularisation est un cas particulier de régularisation spectrale [Smola & Kondor, 2003] sans donner plus de détails. Revenons maintenant sur ce point.

Dans le cas discret, la matrice laplacienne est une matrice réelle symétrique. Elle est donc diagonalisable dans une base orthonormale (on sait également que ses valeurs propres sont positives ou nulles). Dans le cas continu, l'espace dans lequel on travaille étant compact, l'opérateur de Laplace-Beltrami est un opérateur linéaire compact. D'après le théorème spectral, il admet donc une base orthonormale complète de vecteurs propres. Que ce soit dans le cadre discret ou dans le cadre continu, soit $\{\psi_i\}_i$ une telle base de vecteurs propres du laplacien et $\{\lambda_i\}_i$ les valeurs propres correspondantes.

La pénalisation utilisée pour régulariser le SVM (équations (3.7) et (3.18)) peut se réécrire :

$$\left\| e^{\frac{1}{2}\beta\Delta}\mathbf{w} \right\|^2 = \sum_i \langle \mathbf{w}, \psi_i \rangle r\left(\lambda_i\right) \langle \mathbf{w}, \psi_i \rangle \tag{3.57}$$

avec $r : x \mapsto e^{\beta x}$.

On a donc bien une régularisation spectrale [Smola & Kondor, 2003] : l'image est décomposée suivant une base de vecteurs propres du laplacien ; les hautes fréquences sont ensuite pénalisées à l'aide d'une fonction r. La pénalisation utilisée dans ce chapitre est une pénalisation exponentielle, ce qui conduit au noyau de diffusion. D'autres pénalisations auraient pu être utilisées comme le laplacien régularisé ($r : x \mapsto 1 + \epsilon x, \epsilon$ paramètre), la marche aléatoire à p-step ($r : x \mapsto (a - \lambda)^{-p}$, a paramètre), ou encore l'inverse du cosinus ($x :\mapsto \left(\cos \lambda \frac{\pi}{4}\right)^{-1}$) (figure 3.12).

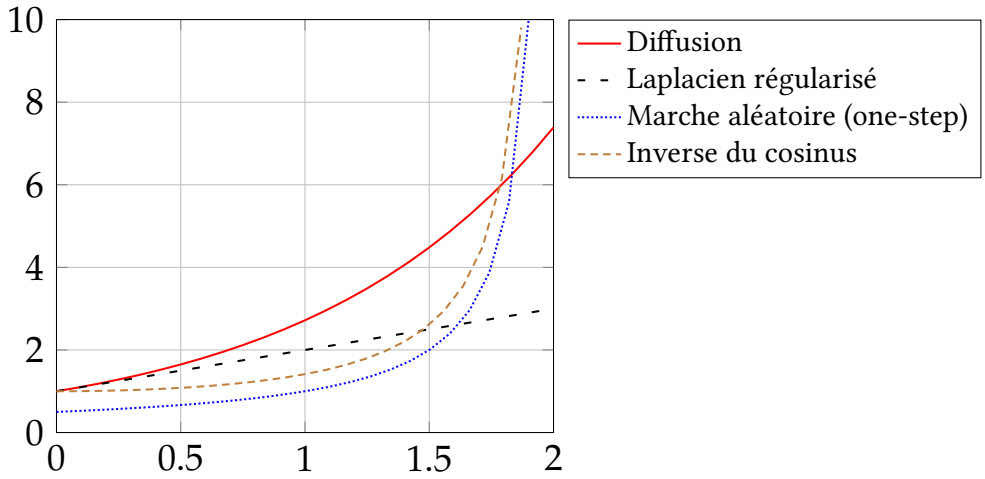


Figure 3.12 : Fonction de régularisation $r(x)$ pour la diffusion ($\beta = 1$), le laplacien régularisé ($\epsilon = 1$), la marche aléatoire ($a = 2, p = 1$) et l'inverse du cosinus. Nous regardons uniquement sur l'intervalle $[0, 2]$ puisque le spectre d'un laplacien régularisé est compris dans cet intervalle.

3.6.1.2 Choix de la diffusion

Comme nous venons de le voir dans la section précédente (3.6.1.1), la régularisation de type diffusion est un cas particulier de régularisation spectrale. Ce n'est pas pour des raisons computationnelles que nous avons choisi la diffusion. Notons en particulier que, lorsque le calcul de la matrice de Gram s'effectue par diagonalisation du laplacien, l'utilisation d'une autre régularisation spectrale n'entraîne pas de difficultés supplémentaires en termes de calculs. La raison principale d'un tel choix de régularisation est simplement que le noyau de diffusion généralise le lissage gaussien, prétraitement fréquemment utilisé [Fan et al., 2007 ; Vemuri et al., 2008].

3.6.1.3 Le laplacien régularisé

La pénalisation. Parmi les différentes régularisations possibles, nous avons également testé celle du laplacien régularisé. Cette régularisation a l'avantage d'être très naturelle. En effet, on souhaite que $\mathbf{w}^{\text{opt}}$ soit spatialement lisse et comme nous le verrons dans ce qui suit, utiliser un laplacien régularisé revient à pénaliser par $\|\nabla \mathbf{w}\|^2$. Nous allons en dire quelques mots dans cette section.

Afin d'avoir un paramètre de régularisation qui varie dans un intervalle borné, nous avons considéré la fonction de régularisation r :

$$r : x \mapsto \epsilon + (1-\epsilon)x \tag{3.58}$$

avec le paramètre $\epsilon \in]0,1]$. Cela correspond à l'utilisation dans le cas discret de l'opérateur de régularisation P défini par :

$$P : \ \mathbf{w}^* \in \mathcal{F} = \mathcal{L}(\mathbb{R}^{\mathcal{V}}, \mathbb{R}) \mapsto \left([\epsilon I_d + (1-\epsilon)L]^{\frac{1}{2}} \mathbf{w} \right)^* \in \mathcal{F} \tag{3.59}$$

En d'autres termes, dans le cas discret, on regarde le problème d'optimisation suivant :

$$\underbrace{\min_{\mathbf{w}\in\mathcal{X}, b\in\mathbb{R}} \frac{1}{N} \sum_{s=1}^{N} \ell_{\text{hinge}} \left(y_s \left[\langle \mathbf{w}, \mathbf{x}_s \rangle + b \right] \right) + \lambda\epsilon \|\mathbf{w}\|^2}_{\text{SVM linéaire standard}} + \underbrace{\lambda(1-\epsilon)\mathbf{w}^T L \mathbf{w}}_{\text{Régularisation spatiale}} \tag{3.60}$$

De manière similaire, dans le cadre continu, on considère le problème d'optimisation suivant :

$$\underbrace{\min_{\mathbf{w}\in\mathcal{X}, b\in\mathbb{R}} \frac{1}{N} \sum_{s=1}^{N} \ell_{\text{hinge}} \left(y_s \left[\langle \mathbf{w}, \mathbf{x}_s \rangle + b \right] \right) + \lambda\epsilon \|\mathbf{w}\|^2}_{\text{SVM linéaire standard}} + \underbrace{\lambda(1-\epsilon)\|\nabla \mathbf{w}\|^2}_{\text{Régularisation spatiale}} \tag{3.61}$$

Les équations 3.60 et 3.61 sont équivalentes à celle du SVM linéaire standard (équation 3.1) avec comme contrainte supplémentaire, le terme $\mathbf{w}^T L \mathbf{w}$ dans le cas discret ou son équivalent continu $\|\nabla \mathbf{w}\|^2$.

Remarque : Dans le cas limite $\epsilon = 1$, nous retrouvons l'équation du SVM linéaire standard.

Notons que, dans le cas discret, si A correspond à la matrice d'adjacence du graphe de régularisation, la contrainte supplémentaire $\mathbf{w}^T L \mathbf{w}$ peut s'écrire :

$$\mathbf{w}^T L \mathbf{w} = \frac{1}{2} \sum_{i,j} A_{i,j} (w_i - w_j)^2$$

Cette équation permet de se rendre compte que le terme de pénalisation ajouté force les composantes fortement connectées d'après la matrice d'adjacence à avoir le même poids. Dans le cas continu, on reconnaît l'**énergie de Dirichlet**, $\|\nabla \mathbf{w}\|^2$. Elle force les composantes proches à avoir le même poids.

Calcul de la matrice de Gram. D'après la section 3.1.3, c'est équivalent à un SVM de noyau :

$$K_\epsilon(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1^T(\epsilon I_d + (1-\epsilon)L)^{-1}\mathbf{x}_2 \tag{3.62}$$

C'est un **noyau de diffusion de Von Neumann** [Shawe-Taylor & Cristianini, 2004].

Dans le cas continu, il suffit de remplacer L par l'opérateur de Laplace-Beltrami, Δ. De manière générale, dans le cas discret, on calcule la matrice de Gram à l'aide du gradient conjugué. Dans les cas particuliers vus précédemment, on peut utiliser les mêmes techniques, à savoir : la diagonalisation à l'aide de la FFT dans le cas de la grille régulière, la diagonalisation ou plus simplement le lemme d'inversion matricielle (égalité matricielle de Woobury – équation 3.44) dans le cas de l'atlas.

Dans le cas continu, le calcul de $(\epsilon \cdot id + (1-\epsilon)\Delta)^{-1}\mathbf{x}_s$ pour un sujet s s'effectue en résolvant le **problème de Helmholtz** aux limites de Dirichlet homogènes, d'inconnue $\mathbf{u}$:

$$-\Delta\mathbf{u} + k^2\mathbf{u} = \mathbf{x}_s \tag{3.63}$$

avec $k = \sqrt{\frac{\epsilon}{1-\epsilon}}$. Pour plus de détails sur les hypothèses de ce problème, le lecteur peut se référer au cours d'Allaire [2005]. Il pourra alors vérifier en utilisant une approche variationnelle et à l'aide du théorème de Lax-Milgram que cette équation est bien posée au sens de Hadamard. Autrement dit, il existe une unique solution de la formulation variationnelle de l'équation 3.63 et cette solution dépend continûment de $\mathbf{x}_s$. La résolution de ce problème s'effectue en utilisant les éléments finis du premier ordre. L'inversion de la matrice de rigidité est faite à l'aide d'un gradient conjugué dans le cas général ou d'une FFT dans le cas euclidien.

Choix du paramètre ϵ. La principale difficulté est la même qu'avec la régularisation exponentielle : comment choisir le paramètre de régularisation ? Dans le cas discret, lorsque l'approche utilise la diagonalisation du laplacien, le paramètre ϵ, tout comme le paramètre β dans le cas de la diffusion, peut être choisi en fonction des vecteurs propres et des valeurs propres correspondantes. Si ce n'est pas le cas, nous nous efforçons de nous rapprocher au mieux du cas continu euclidien.

Nous venons de voir, dans le cas des conditions limites de Dirichlet homogènes, que le problème de Helmholtz est bien posé au sens de Hadamard. Pour cela nous avons utilisé une approche variationnelle. L'avantage de cette approche est qu'elle donne un cadre de résolution numérique de l'équation. L'inconvénient principal est que l'on a peu d'informations directement exploitables sur la solution. Or, mis à part les conditions aux bords, la pénalisation par énergie de Dirichlet est invariante par translation. On devrait donc, comme nous l'avons vu précédemment avoir une interprétation de la solution comme le résultat d'un lissage de $\mathbf{x}_s$. C'est ce que nous allons regarder ici. Cela revient à chercher un **opérateur de Green** de

l'équation de Helmholtz. Rappelons qu'un opérateur de Green est une solution de l'équation d'inconnue $T \in \mathcal{D}'(\mathring{\mathcal{V}})$ (espace des distributions sur $\mathring{\mathcal{V}}$) :

$$-\Delta T + k^2 T = \delta \tag{3.64}$$

où δ est la distribution de Dirac. On suppose que l'on a toujours les conditions de Dirichlet homogènes aux bords. Notons que dans ce cas, si G est un opérateur de Green du problème de Helmholtz, on a alors $G * \mathbf{x}_s$ solution du problème de Helmholtz (3.63). Affranchissons nous de ces conditions limites un moment et considérons des conditions limites plus naturelles : les conditions de rayonnement de Sommerfeld (en dimension 3).

$$\lim_{r\to\infty} r\left(\frac{\partial}{\partial r} - ik\right) T(r) = 0 \tag{3.65}$$

Autrement dit, on veut que la solution tende suffisamment vite vers 0 à l'infini. Dans ce cas, en utilisant la transformée de Fourier et en passant en coordonnées sphériques après avoir remarqué l'invariance par rotation du problème, on a :

$$\forall \mathbf{r} \in \mathbb{R}^3,\ G(\mathbf{r}) = \frac{e^{-k\|\mathbf{r}\|}}{4\pi \parallel \mathbf{r} \parallel} \tag{3.66}$$

est une solution fondamentale. Autrement dit, une solution de (3.63) s'écrit sous la forme

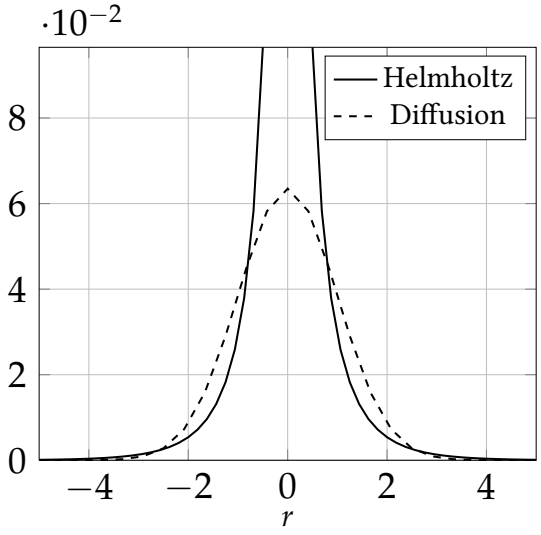


FIGURE 3.13 : Formes fondamentales de l'équation de Helmholtz et de l'équation de diffusion en dimension 3 pour des distances caractéristiques égales à 1. Ces formes fondamentales étant isotropes, on donne leurs valeurs en fonction du rayon r.

$G * \mathbf{x}_s + \gamma$ avec γ une fonction harmonique de l'équation de Helmholtz rectifiant la valeur de $G * \mathbf{x}_s$ aux bords $\partial\mathcal{V}$. La solution est donc une version lissée avec un lissage de **longueur**

caractéristique $\sigma_c = \frac{1}{k}$ corrigée aux bords. Pour plus de détails, le lecteur pourra se référer par exemple au livre de Nédélec [2001] ou à celui de Myint-U & Debnath [2007]. La longueur caractéristique du noyau de lissage σ_c est alors (cf. figure 3.14) :

$$\sigma_c = \frac{1}{k} = \sqrt{\frac{1-\epsilon}{\epsilon}} \tag{3.67}$$

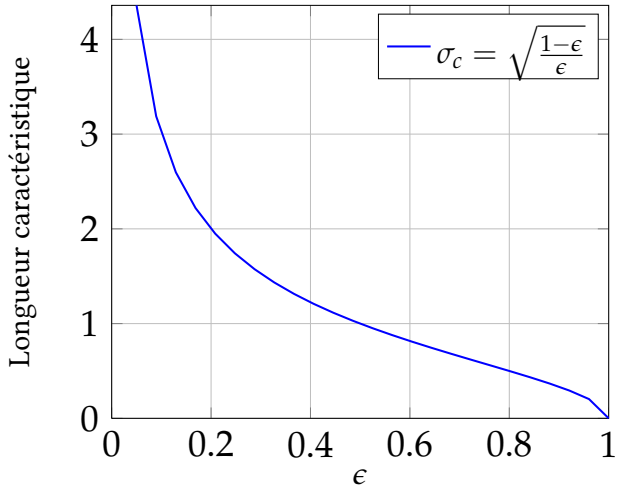


Figure 3.14 : Longueur caractéristique du noyau de lissage en fonction du paramètre de régularisation ϵ.

Remarque : Quand $\epsilon \to 0$ on obtient l'**équation de Poisson** et le noyau de Green est en $\frac{1}{\|\mathbf{r}\|}$. Il est alors invariant par changement d'échelle. On est dans le cas où la seule régularisation est l'énergie de Dirichlet.

Remarque : Notons que la même démarche pour un problème à une dimension aboutit au noyau laplacien [Schölkopf & Smola, 2001 ; Shawe-Taylor & Cristianini, 2004].

En conclusion la régularisation à l'aide d'un laplacien régularisé à l'avantage d'avoir un terme de pénalisation explicite et facile à comprendre. La principale raison pour laquelle nous avons préféré le noyau de diffusion est qu'il étend la notion de lissage gaussien. Il y a d'autres raisons comme la singularité en 0 du noyau de lissage dans le cas de la pénalisation par l'énergie de Dirichlet. Le noyau de lissage est très fortement piqué en 0 (figure 3.13), ce qui donne des résultats de lissage moins habituels (figure 3.15). En ce qui concerne les performances de classification, nous avons testé le laplacien régularisé uniquement dans le cas de la proximité spatiale et dans le cas de la proximité anatomique sur la même base de données que dans le chapitre précédent (tableau 2.3). Les résultats en termes de classification sont légèrement moins bons qu'avec une régularisation de type diffusion mais ces différences ne sont pas statistiquement significatives.

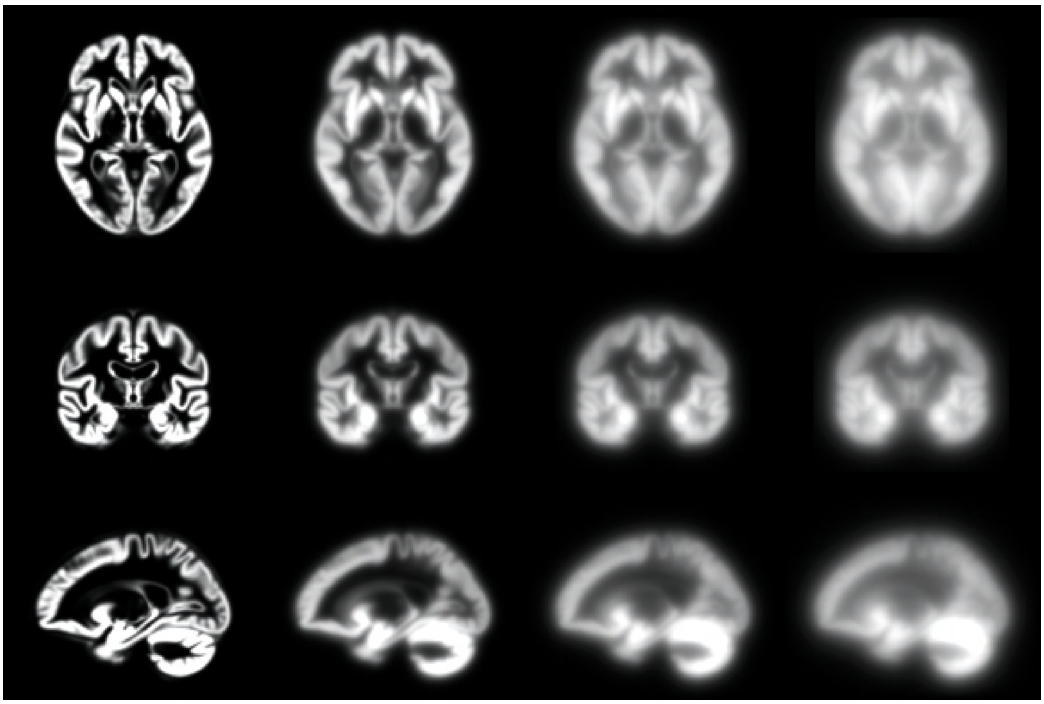


FIGURE 3.15 : De gauche à droite, solutions de l'équation de Helmholtz (laplacien régularisé) pour des valeurs d'ϵ choisies telles que les distances caractéristiques soient respectivement de 0, 2, 4 et 6 voxels (de gauche à droite – 1 voxel = $1.5mm$). Les coupes axiales, coronales et sagittales correspondent à $z = -18mm$, $y = 9mm$ et $x = 24mm$ dans l'espace MNI.

3.6.2 Gestion de la grande dimension

3.6.2.1 Régularisation quadratique

Le problème de classification d'images cérébrales est un problème de classification à grandes dimensions. Autrement dit, la dimension de l'espace des données est beaucoup plus grande que le nombre de sujets ($N \ll d$). Pour traiter ce problème nous avons restreint l'espace de recherche à un espace de dimension N (où N désigne le nombre de sujets) grâce aux machines à vecteurs supports. Nous avons également ajouté des connaissances *a priori* afin de guider l'algorithme d'apprentissage. Ces connaissances exploitant la structure de l'image et certaines informations anatomiques sont incorporées dans le SVM à l'aide des opérateurs de régularisation. La régularisation force le SVM à considérer comme similaires des voxels proches *a priori*.

L'idée sous-jacente de la régularisation est la suivante. Il est impossible d'inférer de l'information sur l'ensemble des voxels de l'image directement à partir des données (en utilisant par exemple des outils telles que la régression linéaire ou l'analyse linéaire discriminante) sans avoir de l'information sur ces voxels. En revanche si l'on sait que les voxels sont fortement

corrélés, alors le nombre de « degrés de liberté » de la séparatrice que l'on cherche est faible par rapport au nombre de sujets. En exploitant cette information, il est possible de faire de la classification en régularisant fortement à l'aide de l'information *a priori* sur notre problème. D'une certaine manière, lorsque la régularisation est quadratique comme celle utilisée dans ce chapitre, on force les voxels que l'on considère comme fortement corrélés à être moyennés.

3.6.2.2 Parcimonie

Une autre approche pour les problèmes de grande dimension fréquemment utilisée depuis la méthode du LASSO de Tibshirani [1996] est d'utiliser la parcimonie : on cherche parmi l'ensemble des voxels un petit nombre de voxels explicatifs. Pour cela, si l'on réutilise la notation, l'idée est d'utiliser comme pénalisation $\|\mathbf{w}\|_0$, c'est à dire le nombre de coefficients non nuls de $\mathbf{w}$. Malheureusement une telle pénalisation n'est pas convexe. Cette contrainte est relaxée en utilisant comme pénalisation : $\|\mathbf{w}\|_1$. Malgré la relaxation, une telle pénalisation force le nombre de composantes non nulles de $\mathbf{w}$ à être faible.

Greenshtein & Ritov [2004] ont montré dans le cadre du modèle linéaire général que la consistance de la régression linéaire avec une pénalisation ℓ_1, autrement dit le LASSO [Tibshirani, 1996], requiert que la norme ℓ_1 de la solution réelle soit en $O\left(\sqrt{\frac{N}{\log d}}\right)$. Autrement dit, une telle pénalisation n'est adaptée qu'au cas où la solution réelle est parcimonieuse. Or, dans le cadre de notre travail, les données sont des images. Il semble peu probable que les différences entre deux populations ne mettent en jeux qu'un petit nombre de voxels.

Nous avons tout de même testé la régression logistique avec une pénalisation ℓ_1 pour la classification de patients atteints de la maladie d'Alzheimer (cf. chapitre 2). Pour des raisons de temps de calcul nous l'avons uniquement testé avec les données d'épaisseur corticale de la méthode *Thickness-Atlas*. Il aurait été plus intéressant de tester sur les données brutes et non pas sur les données regroupées en régions, mais les temps de calcul étaient trop longs. Les résultats obtenus étaient moins bons pour les trois comparaisons *CN vs AD*, *CN vs* MCI_c et MCI_{nc} *vs* MCI_c.

En revanche une hypothèse qui semble réaliste est de supposer que l'hyperplan séparateur est constant par morceaux. En d'autres termes, cela revient à supposer que $\nabla\mathbf{w}$ est parcimonieux. Il serait donc possible d'utiliser comme pénalisation la variation totale, $\|\mathbf{w}\|_{TV}$, définie par :

$$\|\mathbf{w}\|_{TV} = \int_{v\in\mathcal{V}} \|\nabla\mathbf{w}\|_2 d\mu(v)$$

La variation totale est principalement utilisée en débruitage d'images [Rudin et al., 1992]. Une telle pénalisation force l'hyperplan séparateur à être constant par morceaux. Une autre manière de voir la pénalisation par la variation totale en 2D est de la voir comme une pénalisation de la

longueur des courbes de niveau (ex. [Mallat, 2001]). On peut voir également cette régularisation comme la variante ℓ_1 de l'énergie de Dirichlet.

Malheureusement la principale difficulté avec toutes ces approches est le problème d'optimisation. Un avantage de la régularisation quadratique est qu'elle satisfait le théorème du représentant. Ce n'est plus le cas lorsque la pénalisation est de type ℓ_1. Dans certain cas particuliers comme le LASSO [Tibshirani, 1996], des approches par ensembles actifs permettent de gérer facilement les grandes dimensions. Ce genre d'approches ne fonctionne plus pour des pénalisations de type « taux de variation ». La combinaison des régularisations ℓ_1 et quadratique a été proposée avec par exemple l'*elastic net* de Zou & Hastie [2005]. Nous n'avons pas abordé ce point.

3.6.2.3 Apprentissage semi-supervisé

Une autre approche possible pour gérer les problèmes de grandes dimensions est de ne pas faire uniquement de l'inférence inductive mais de l'inférence transductive. Il s'agit en d'autres termes de faire de l'apprentissage semi-supervisé. Il s'agit donc de faire à la fois de la classification sur les données d'apprentissage et du *clustering* sur les données tests.

Pour cela, on peut considérer le ***transductive SVM*** (TSVM) [Vapnik, 1995 ; Joachims, 1999]. Le problème de classification associé est le suivant :

$$\begin{aligned} \min_{y_1^*,\cdots,y_{N_{\text{test}}}^*,\mathbf{w},b,\xi_1,\cdots,\xi_{N_{\text{train}}},\xi_1^*,\cdots,\xi_{N_{\text{test}}}^*} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{s=1}^{N_{\text{train}}}\xi_s + C^*\sum_{s=1}^{N_{\text{test}}}\xi_s^* \\ s.t.: \quad & y_s[\langle \mathbf{w}, \mathbf{x}_s\rangle + b] \geq 1 - \xi_s \\ & y_s^*[\langle \mathbf{w}, \mathbf{x}_s^*\rangle + b] \geq 1 - \xi_s^* \\ & \xi_s, \xi_s^* \geq 0 \end{aligned}$$

en indiquant par des astérisques les sujets de l'ensemble de test. Notons qu'à la différence du SVM, le TSVM est un problème d'optimisation combinatoire ! Joachims [1999] propose une méthode pour le résoudre. Nous utiliserons son implémentation sans rentrer dans les détails. Nous l'avons testé sur les mêmes problèmes de comparaison qu'au chapitre 2 sur les cartes de concentration de substance grise recalées avec DARTEL. Les taux de classification obtenus sont de 90% (sensibilité 90%, spécificité 90%) pour la comparaison *CN vs AD*, 81% (sensibilité 70%, spécificité 85%) pour la comparaison *CN vs* MCI_c et 72% (sensibilité 62%, spécificité 78%) pour la comparaison MCI_{nc} *vs* MCI_c. Ils sont similaires à ceux du SVM.

Notons que la régularisation proposée dans ce chapitre peut s'appliquer directement au TSVM. Le TSVM n'apportant pas d'amélioration importante, nous nous sommes restreints au SVM par souci de clarté.

3.6.3 Différents modèles de proximité

3.6.3.1 Une proximité anatomique et une proximité spatiale

Dans ce chapitre, nous avons commencé par considérer la proximité la plus simple : la proximité spatiale. Deux voxels sont proches s'ils sont proches dans l'espace selon la distance euclidienne dans le cas volumique ou selon la distance géodésique de la surface du cortex dans le cas de l'épaisseur corticale. Cette proximité permet d'avoir des hyperplans spatialement lisses. Une telle proximité prend en compte la structure de l'image mais pas spécifiquement l'anatomie du cerveau.

Pour prendre en compte les informations d'un plus haut niveau, nous avons fait l'hypothèse que ces informations peuvent être représentées sous la forme d'un atlas probabiliste. Ces informations peuvent être par exemple des informations sur les tissus considérés (substance grise, substance blanche, CSF) ou sur des régions anatomo-fonctionnelles. Il est possible de considérer cette proximité comme celle qui considère que deux voxels sont proches s'ils appartiennent à un même réseau. Par exemple, deux voxels sont proches s'ils appartiennent à la même région anatomique ; ce peut être vu comme une connectivité à courte distance. Un autre exemple est celui des connectivités à longue distance qui modélisent le fait que deux voxels peuvent être proches s'ils sont connectés anatomiquement par des faisceaux de fibres ou s'ils sont connectés sur le plan fonctionnel.

Cette proximité est directement applicable au cas surfacique et s'implémente très rapidement. Malheureusement, cette souplesse et cette efficacité sont au prix de la perte de l'information de proximité spatiale. Nous nous sommes donc efforcés de combiner ces deux proximités

3.6.3.2 Combiner les proximités spatiales et anatomiques

Il y a deux points de vue totalement différents pour combiner les proximités spatiales et anatomiques.

Le premier point de vue consiste à considérer ces différentes proximités comme émanant de deux concepts complètement distincts. La combinaison des deux proximités est alors faite en sommant simplement les termes de régularisation. Cette approche est principalement appropriée pour la modélisation de connectivité longue distance.

Une autre approche pour combiner les deux proximités spatiale et anatomique et de considérer qu'une telle combinaison peut se voir comme une modification locale de la topologie induite par l'information spatiale afin de respecter l'information anatomique. Puisque les images sont des éléments discrets, de telles modifications locales peuvent être modélisées par des distances entre histogrammes telles que la distance du χ^2 ou la divergence de Kullback-Leibler. Notons que la divergence de Kullback-Leibler est reliée à la métrique de Fisher (ex.

[Lafferty & Lebanon, 2005]). Nous avons proposé cette approche à la section 3.5.2. Cependant, le cerveau est intrinsèquement un objet continu (à la résolution de l'image) ; il semble donc plus intéressant de modéliser des modifications locales à l'aide d'une approche continue.

Nous avons donc proposé un cadre de régularisation permettant de prendre en compte diverses informations telles que des informations sur les tissus, sur les régions anatomiques ou encore des informations sur la localisation spatiale. Dans cette approche, l'idée clé est de ne pas considérer les voxels comme tels mais de décrire chaque voxel par une distribution de probabilités donnant de l'information sur la région anatomique, sur les types de tissus ainsi que sur la localisation. La distance entre deux voxels est alors donnée par la métrique de Fisher. Le choix de cette métrique a l'avantage de nous permettre de travailler dans un cadre connu (les variétés riemanniennes).

En ce qui concerne l'information spatiale encodée dans cette approche, on peut la voir comme un indice de confiance sur la localisation. Cette information spatiale pourrait être adaptée plus spécifiquement à l'algorithme de recalage par exemple en faisant varier σ_{loc} suivant la position dans l'image ou en prenant des modèles plus compliqués.

Il y a deux améliorations possibles à l'approche utilisant la métrique de Fisher. Premièrement, cette approche ne permet pas, telle qu'elle est énoncée, de prendre en compte des connectivités à longues distances : on s'est en effet restreint à des informations locales. Deuxièmement, dans sa version actuelle, l'approche utilisant la métrique de Fisher n'est pas bien adaptée pour les atlas binaires pour deux raisons : les hypothèses de régularité et les problèmes de discrétisation. Les problèmes de discrétisation sont illustrés par la figure 3.16. Le tenseur de métrique est évalué en chaque voxel. Par conséquent, la norme du tenseur de métrique est très grande à la frontière entre deux régions. Le processus de diffusion est donc interrompu ou fortement modifié sur une bande d'une largeur de deux voxels. Or, dans nos applications (chapitre 4), deux voxels correspondent à trois millimètres au moins ; ce n'est pas négligeable par rapport à l'épaisseur du cortex et à la taille de certaines structures. Sous-échantillonner les images n'est pas forcément la meilleure option compte tenu de leur taille.

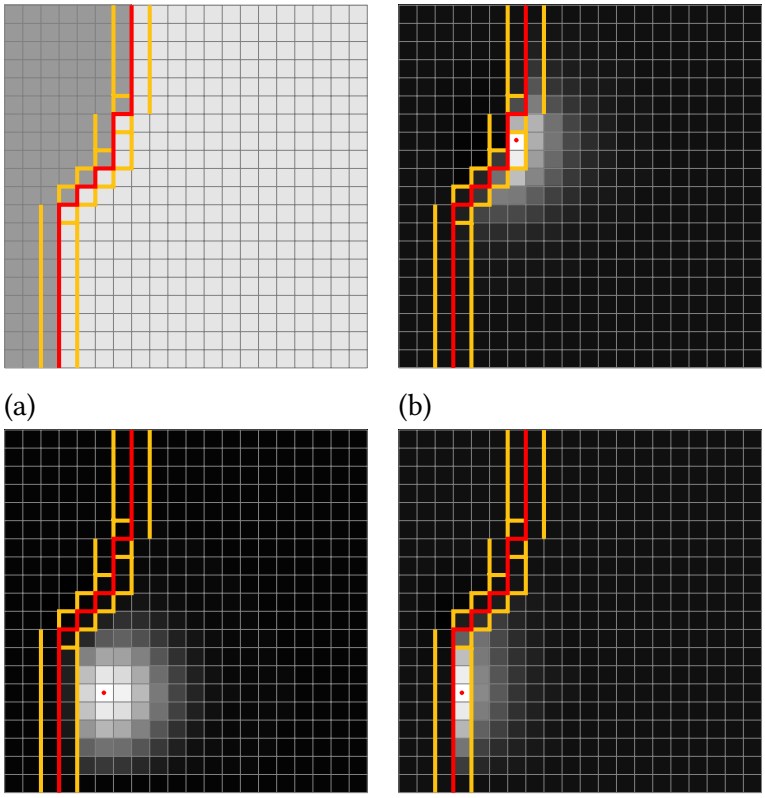


Figure 3.16 : Limite de la méthode de Fisher. Illustration sur un exemple en 2D. On considère deux régions (gris clair et gris foncé) séparées par une frontière tracée en rouge (a). À cause de la discrétisation, il n'y a pas ou peu de diffusion entre deux voxels séparés par une ligne jaune ou rouge. De (b) à (d), quelque exemples de diffusion de Dirac (FWHM$\sim$ 3mm, $\sigma_{\text{loc}} = 5$ mm); les distributions de Dirac sont représentés par des points rouges.

3.6.4 Perspectives

3.6.4.1 Gestion des variables confondantes

Dans ce chapitre nous ne nous sommes pas intéressés aux variables confondantes. Nous avons fait, du moins implicitement, l'hypothèse que la valeur de chaque voxels d'une image d'un sujet est une fonction du label de ce sujet (malade ou sain) et d'un bruit qui comprend la variabilité interindividuelle ainsi que les erreurs de mesures. Or ce n'est en général pas le cas. Il y a d'autres facteurs tels que l'âge du sujet par exemple qui peuvent influer. Par conséquent, il se peut que la séparatrice que l'on obtient entre les deux groupes de sujets ne prenne pas seulement en compte les différences de labels des groupes mais d'autres caractéristiques que l'on appelle **variables confondantes** .

Pour la gestion de l'âge en particulier, Vemuri et al. [2008] et Querbes et al. [2009] proposent

d'ajouter l'âge comme une dimension supplémentaire avant la classification. Cette approche revient à chercher une fonction de classification f de la forme :

$$f(\mathbf{x}_s, a_s) = \operatorname{sgn}\left(\langle \mathbf{w}, \mathbf{x}_s \rangle + \mu a_s + b\right) \tag{3.68}$$

avec a_s l'âge du sujet s, $\mathbf{w} \in \mathcal{X}$ et μ et b deux réels. La concaténation de l'âge avec les données de l'image nous semble pas approprié en grande dimension (ce qui est le cas de [Vemuri et al., 2008]). La difficulté réside dans le poids à donner à la variable « âge ». Nous proposons deux d'approches différentes pour traiter ce problème. Ces approches sont au stade d'ébauche.

Projection sur l'espace orthogonal. L'idée est d'essayer de rendre le score du SVM indépendant de l'âge. Nous supposons les données centrées. Une première approche consisterait donc à annuler $\mathbf{E}_s\left[a_s \mathbf{x}_s\right]$. Cela revient à projeter les données sur l'espace nul des variables confondantes. Si on note Y la matrice de taille $N \times n_c$ avec n_c le nombre de variables confondantes, le projecteur P_c sur l'espace nul des variables confondantes est défini par :

$$P_c = I_N - Y(Y^T Y)^{-1} Y^T$$

en particulier, pour l'âge uniquement, $Y = \mathbf{a}$, on a :

$$P_c = I_N - \mathbf{a}(\mathbf{a}^T \mathbf{a})^{-1} \mathbf{a}^T$$

Approche par régularisation. Une autre approche utilisant le cadre de la régularisation consisterait à pénaliser la fonction de classification par sa covariance avec l'âge. Si l'on note $\mathbf{a} \in \mathbb{R}^N$ le vecteur colonne des âges des sujets et $X \in \mathbb{R}^{N \times d}$ dont les lignes sont les vecteurs $\left(\mathbf{x}_s^T\right)_{s \in \mathcal{S}_{\text{train}}}$, une estimation de la covariance est, pour des données centrées :

$$\hat{\mathbf{E}}_s\left[a_s \langle \mathbf{w}^{\text{opt}}, \mathbf{x}_s \rangle\right] = \frac{1}{N} \mathbf{a}^T X \mathbf{w}$$

L'équation à optimiser devient alors :

$$\begin{aligned}\left(\mathbf{w}^{\text{opt}}, b^{\text{opt}}\right) = \underset{\mathbf{w} \in \mathcal{X}, b \in \mathbb{R}}{\arg\min} \quad & \tfrac{1}{N} \textstyle\sum_{s=1}^{N} \ell_{\text{hinge}}\left(y_s\left[\langle \mathbf{w}, \mathbf{x}_s \rangle + b\right]\right) \\ & + \lambda \left[(1-\epsilon)\|e^{\frac{1}{2}\beta L}\mathbf{w}\|^2 + \tfrac{\epsilon}{N^2}\|\mathbf{a}^T X \mathbf{w}\|^2\right]\end{aligned}$$

avec ϵ un paramètre qui permet de régler l'importance de la régularisation. D'après la section 3.2.1, c'est un SVM de noyau :

$$K_{\beta,\epsilon}(\mathbf{x}_{s_1}, \mathbf{x}_{s_2}) = {}^t\mathbf{x}_{s_1} \left[(1-\epsilon)e^{\beta L}\mathbf{w} + \frac{\epsilon}{N^2} X^T \mathbf{a}\mathbf{a}^T X\right]^{-1} \mathbf{x}_{s_2}$$

Or, en utilisant l'égalité matricielle de Woodbury (équation (3.44)), on a :

$$\left[(1-\epsilon)e^{\beta L}\mathbf{w} + \epsilon X^T \mathbf{a}\mathbf{a}^T X\right]^{-1} =$$
$$\frac{1}{(1-\epsilon)}e^{-\beta L} - \frac{1}{(1-\epsilon)^2}e^{-\beta L}X^T\mathbf{a}\left[\frac{N^2}{\epsilon} + \frac{1}{(1-\epsilon)}\mathbf{a}^T X e^{-\beta L}X^T\mathbf{a}\right]^{-1}\mathbf{a}^T X e^{-\beta L}$$

Notons que la complexité du calcul de la matrice de Gram est équivalente à celle du calcul de la matrice de Gram du noyau de diffusion.

3.6.4.2 Prise en compte du stade de la maladie

Un autre point qui n'est pas pris en compte dans les approches que nous proposons est le degré d'avancement de la maladie. Lors de la phase d'apprentissage, on souhaiterait pénaliser un patient mal classé d'autant plus que son degré d'atteinte est important.

Deux approches sont possibles. La première consisterait à faire de la régression en utilisant un score correspondant au degré d'avancement de la maladie (par exemple le MMS pour la maladie d'Alzheimer ou l'ADASCog) et à utiliser le régresseur obtenu comme un classifieur à l'aide d'un seuil. Nous avons testé cette approche pour la comparaison *CN vs AD*. En utilisant le MMS, nous avons obtenu des scores de classification similaires aux autres classifieurs ($\approx 84\%$) vus dans le chapitre 2. L'inconvénient est que la régression est un problème statistiquement plus difficile. En outre le score d'atteinte varie peu chez les sujets sains par rapport aux sujets malades.

Une autre approche consisterait à prendre des valeurs du paramètre C du SVM différentes pour chaque sujet afin de pénaliser plus les sujets fortement atteints et mal classés. L'utilisation du SVM avec des valeurs de C dépendantes des observations a été proposée par Schmidt [1997]. Nous n'avons pas encore testé cette approche.

3.7 Conclusion

Dans ce chapitre, nous avons proposé de régulariser spatialement les machines à vecteurs supports pour la classification en neuroimagerie. L'idée de base était de contraindre le SVM à prendre en compte la structure des images ainsi que des informations sur l'anatomie. Nous avons pour cela fait l'hypothèse que toutes les images était recalées dans un espace commun à tous les sujets. La régularisation utilisée passe par une pénalisation quadratique qui revient à lisser les données à l'aide d'un noyau de diffusion défini sur un graphe ou sur une variété riemannienne.

Dans le chapitre suivant (chapitre 4), nous appliquerons les méthodes de régularisation décrites dans ce chapitre dans le cadre de la maladie d'Alzheimer et des accidents vasculaires cérébraux.

Chapitre 4

Applications à la maladie d'Alzheimer et aux accidents vasculaires cérébraux

La grande majorité des méthodes de classification utilisées en neuroimagerie (chapitres 1 et 2) ne prend pas en compte la distribution spatiale des *features*. Par conséquent, les classifieurs obtenus avec de telles méthodes ne sont pas nécessairement cohérents avec l'anatomie. Nous avons donc proposé dans le chapitre précédent (chapitre 3) des méthodes capables de contraindre le SVM pour que l'hyperplan séparateur obtenu soit régularisé spatialement et anatomiquement.

Dans ce chapitre nous présentons deux applications de ces méthodes : la maladie d'Alzheimer (section 4.1) et les accidents vasculaires cérébraux (section 4.2).

L'application aux accidents vasculaires cérébraux a été effectuée en collaboration avec le Dr Charlotte Rosso et le Dr Yves Samson du service d'urgences cérébro-vasculaires de l'Hôpital de la Pitié Salpêtrière.

4.1 Application à la maladie d'Alzheimer

La maladie d'Alzheimer est la principale cause de démences neurodégénératives [Blennow et al., 2006]. De nombreuses études de groupes basées sur la volumétrie de certaines régions d'intérêt (ex : [Good et al., 2001 ; Chételat & Baron, 2003]), sur des analyses de type *voxel-based morphometry* (VBM) (ex : [Good et al., 2001 ; Whitwell et al., 2008]) ou encore sur des mesures d'épaisseur corticale (ex : [Thompson et al., 2004]) ont montré que l'atrophie cérébrale chez les sujets atteints de la maladie d'Alzheimer (AD) et chez les AD prodromaux n'était pas focale mais spatialement distribuée. L'atrophie touche de nombreuses régions du cerveau dont principalement : le cortex entorhinal, les hippocampes, les structures temporales latérales et inférieures ainsi que le cingulaire antérieur et postérieur. L'utilisation d'outils de classification

multivariée apparaît donc adapté à l'étude de cette pathologie. Comme nous l'avons vu dans les deux chapitres précédents (chapitres 1 et 2), différentes méthodes ont été proposées pour ce problème. Cependant, la majorité de ces méthodes ne prend pas en compte la distribution spatiale des *features*, ce qui conduit à des hyperplans séparateurs souvent bruités et qui ne respectent pas complètement l'anatomie.

Dans ce chapitre nous proposons donc d'utiliser les SVM régularisés spatialement pour la classification d'IRM anatomiques de sujets atteints de la maladie d'Alzheimer et de sujets sains. Plus particulièrement, nous appliquons les SVM régularisés spatialement aux problèmes de classification utilisés pour la comparaison de méthodes (chapitre 2) : la classification entre les témoins (CN) et les patients AD probables (*CN vs AD*), la classification entre les témoins et les sujets MCI qui convertissent vers la maladie d'Alzheimer durant les 18 premiers mois suivant l'inclusion (*CN vs* MCI_c) et à la prédiction de la conversion chez les patients MCI (MCI_{nc} *vs* MCI_c).

4.1.1 Données

4.1.1.1 Participants

Nous avons utilisé pour cette étude la même base de données que celle utilisée pour la comparaison de méthodes de classification (chapitre 2). Au total, 509 sujets de la base ADNI ont ainsi été sélectionnés. Parmi ces sujets, on dénombre 162 témoins (CN - *cognitively normal elderly controls*), 137 patients atteints de la maladie d'Alzheimer (AD - *Alzheimer's Disease*), 76 patients souffrants d'un déclin cognitif léger qui ont converti vers la maladie d'Alzheimer dans les 18 mois qui suivirent l'IRM initiale (MCI_c) ainsi que 134 autres MCI qui n'ont pas converti vers la maladie d'Alzheimer durant ces 18 premiers mois (MCI_{nc}). Les caractéristiques cliniques et démographiques de la population d'étude sont rappelées dans le tableau 4.1.

4.1.1.2 Acquisition des IRM

Nous avons utilisé les mêmes images que dans le chapitre 2. Par conséquent, les images sont des IRM anatomiques acquises à 1.5T lors de la visite de *baseline* quand celles-ci sont disponibles. Lorsque ce n'est pas le cas, nous prenons celles de la visite de *screening*. Le protocole d'acquisition des IRM est détaillé dans [Jack et al., 2008].

Dans le protocole d'acquisition des données d'ADNI, les sujets sont scannés deux fois à chaque visite. Pour chaque sujet, nous considérons donc uniquement celle des deux IRM qui a la meilleure qualité selon les membres d'ADNI. Nous utilisons les images ayant subi les prétraitements suivants : *grad-warp* et correction des inhomogénéités de champ B_1.

TABLE 4.1 : Caractéristiques cliniques et démographiques de la population d'étude. Les valeurs sont indiquées comme : moyenne± écart-type [intervalle].

Ensemble	Diag.	Nb.	Age	Genre	MMS
appr.	CN	81	76.1 ± 5.6 [60 – 89]	38 M / 43 F	29.2 ± 1.0 [25 – 30]
	AD	69	75.8 ± 7.5 [55 – 89]	34 M / 35 F	23.3 ± 1.9 [18 – 26]
	MCI_c	39	74.7 ± 7.8 [55 – 88]	22 M / 17 F	26.0 ± 1.8 [23 – 30]
	MCI_{nc}	67	74.3 ± 7.3 [58 – 87]	42 M / 25 F	27.1 ± 1.8 [24 – 30]
test	CN	81	76.5 ± 5.2 [63 – 90]	38 M / 43 F	29.2 ± 0.9 [26 – 30]
	AD	68	76.2 ± 7.2 [57 – 91]	33 M / 35 F	23.2 ± 2.1 [20 – 27]
	MCI_c	37	74.9 ± 7.0 [57 – 87]	21 M / 16 F	26.9 ± 1.8 [24 – 30]
	MCI_{nc}	67	74.7 ± 7.3 [58 – 88]	42 M / 25 F	27.3 ± 1.7 [24 – 30]
total	CN	162	76.3 ± 5.4 [60 – 90]	76 M / 86 F	29.2 ± 1.0 [25 – 30]
	AD	137	76.0 ± 7.3 [55 – 91]	67 M / 70 F	23.2 ± 2.0 [18 – 27]
	MCI_c	76	74.8 ± 7.4 [55 – 88]	43 M / 33 F	26.5 ± 1.9 [23 – 30]
	MCI_{nc}	134	74.5 ± 7.2 [58 – 88]	84 M / 50 F	27.2 ± 1.7 [24 – 30]

Les 509 images proviennent de 41 centres différents. Il n'y a aucun autre critère d'exclusion.

4.1.1.3 Extractions des caractéristiques

Dans le chapitre précédent, nous avons proposé un cadre de régularisation spatiale pour le cas volumique et pour le cas surfacique. Afin d'illustrer ces deux cas, nous utilisons pour la classification des images cérébrales, soit des cartes de concentration de substance grise, soit des cartes d'épaisseurs corticales.

Cartes de concentration de substance grise. Pour l'analyse à partir des cartes de concentration de substance grise, toutes les images sont préalablement segmentées en substance grise (GM), substance blanche (WM) et liquide cérébrospinal (CSF) à l'aide de la segmentation unifiée de SPM5 (Statistical Parametric Mapping, London, UK) [Ashburner & Friston, 2005]. Ces cartes sont ensuite recalées à l'aide de l'algorithme DARTEL [Ashburner, 2007]. Toutes les cartes recalées, que ce soit par la segmentation unifiée de SPM5 ou par DARTEL, ont été modulées par le jacobien de leur transformation. Cela permet de préserver les quantités de tissus. Aucun lissage spatial n'a été réalisé au préalable.

Épaisseur corticale. Les mesures d'épaisseurs corticales sont réalisées avec le logiciel d'analyse FreeSurfer (Massachusetts General Hospital, Boston, MA). Les détails techniques de ce logiciel sont décrits principalement dans les papiers suivants : [Sled et al., 1998 ; Dale et al., 1999 ; Fischl et al., 1999a,b ; Fischl & Dale, 2000]. Toutes les cartes d'épaisseurs corticales sont recalées sur le *template* par défaut de FreeSurfer.

4.1.2 Régularisations utilisées

Dans le chapitre 3, nous avons effectué les classifications pour chacun des types de régularisation présentés. Par conséquent, nous avons testé les régularisations suivantes.

4.1.2.1 Régularisation spatiale

Nous avons utilisé la régularisation spatiale présentée à la section 3.3. Dans la suite du manuscrit, nous y ferons référence par *Voxel-Regul-Spatial* dans le cas volumique et par *Thickness-Regul-Spatial* dans le cas surfacique.

4.1.2.2 Régularisation anatomique

Pour la régularisation anatomique (section 3.4), nous avons utilisé l'atlas binaire AAL (*Automatic Anatomical Labeling*) de Tzourio-Mazoyer et al. [2002] pour le cas volumique. Cet atlas est composé de 116 régions d'intérêt. Cette approche sera appelée *Voxel-Regul-Atlas*.

En ce qui concerne le cas surfacique, nous avons utilisé l'atlas binaire cortical de Desikan et al. [2006]. Cet atlas est composé de 68 régions d'intérêt. L'approche correspondante sera appelée *Thickness-Regul-Atlas*.

4.1.2.3 Combinaison des régularisations spatiales et anatomiques

Nous avons testé les deux manières proposées pour combiner la régularisation spatiale et la régularisation anatomique (section 3.5). L'approche qui consiste à sommer les deux termes de régularisation utilise également l'atlas AAL dans le cas volumique (*Voxel-Regul-CombineSum*) et l'atlas de Desikan et al. [2006] dans le cas surfacique (*Thickness-Regul-CombineSum*).

Quant à l'approche qui consiste à combiner les deux régularisations en modifiant la métrique euclidienne à l'aide des connaissances anatomiques, nous l'avons appliquée avec comme seule information de type atlas les *templates* de concentration de substance grise, de substance blanche et de liquide cérébrospinal. Nous l'appellerons *Voxel-RegulCombineFisher*.

4.1.2.4 Sans régularisation

Afin d'évaluer l'impact de la régularisation sur les performances de classification, nous avons également effectué les classifications sans régularisation spatiale *Voxel-Direct* et *Thickness-Direct*.

À savoir, *Voxel-Direct* correspond à l'approche qui consiste à utiliser l'intensité des voxels des cartes de concentration de substance grise directement comme caractéristiques pour la classification. La classification est effectuée à l'aide d'un SVM linéaire (standard). De la même manière, *Thickness-Direct* correspond à l'approche qui consiste à utiliser les mesures d'épaisseur corticale en chaque point du maillage directement comme caractéristiques de classification pour un SVM linéaire.

4.1.3 Hyperplans séparateurs optimaux

La fonction de classification obtenue avec un SVM linéaire est, à une constante près, le signe du produit scalaire du vecteur des caractéristiques avec le vecteur $\mathbf{w}^{\text{opt}}$, un vecteur orthogonal à l'hyperplan séparateur optimal. Par conséquent, si $\left|w_i^{\text{opt}}\right|$, la valeur absolue de la i-ème composante de $\mathbf{w}^{\text{opt}}$, est faible comparativement aux autres composantes $\left(w_j^{\text{opt}}\right)_{j\neq i}$, la i-ème caractéristique aura alors une faible influence dans la fonction de classification. Inversement, si $\left|w_i^{\text{opt}}\right|$ est relativement grand, la i-ème caractéristique jouera un rôle important dans la fonction de classification. Ainsi les poids $\mathbf{w}^{\text{opt}}$ permettent d'évaluer qualitativement la cohérence de la fonction de classification avec l'anatomie et avec la pathologie.

Dans un but de clarté, nous avons choisi de ne pas montrer de manière exhaustive tous les hyperplans obtenus pour toutes les comparaisons et toutes les régularisations mais seulement quelques uns dans le cadre de la comparaison *AD vs CN*. Dans toutes ces expériences, le paramètre C du SVM est fixé à un. Les cartes associées aux hyperplans séparateurs optimaux obtenus sont présentées par les figures 4.1 et 4.2. Les régions en couleurs chaudes correspondent aux régions pour lesquelles une atrophie augmente la probabilité d'être classé comme AD. Pour les régions en bleu, c'est l'inverse.

La figure 4.1.a montre les coefficients de l'hyperplan obtenus avec *Voxel-Direct*. Quand aucune régularisation spatiale n'est utilisée, les cartes sont bruitées et manquent de cohérence spatiale. Les figures 4.1.b et 4.1.c montrent les hyperplans séparateurs pour la régularisation spatiale dans le cas volumique. Les cartes sont plus lisses et par conséquent plus cohérentes spatialement. Cependant, elles ne respectent pas toujours la topologie du cortex. Une telle régularisation peut, par exemple, mélanger des voxels du lobe temporal avec des voxels du frontal et des voxels du lobe pariétal.

Les régions pour lesquelles une atrophie augmente la probabilité d'être classé comme AD sont principalement : le lobe temporal médian (hippocampe, amygdale, gyrus parahippocampique), les gyri temporaux inférieurs et moyens, le gyrus cingulaire postérieur ainsi que le gyrus frontal moyen postérieur.

Les résultats avec une régularisation à la fois spatiale et anatomique, *Voxel-Regul-Combine-Fisher*, sont présentés figures 4.1.d et 4.1.e. Les cartes sont plus cohérentes avec l'anatomie du cerveau. En particulier, le lissage respecte plus la topologie du cortex. Dans nos expériences, le paramètre de régularisation FWHM $= 4$ mm a semblé le meilleur compromis entre lissage et préservation de la forme des structures (figure 4.1.d).

La figure 4.2 représente les coefficients de l'hyperplan séparateur optimal obtenu avec *Thickness-Regul-Atlas*. Notons que lorsqu'aucune régularisation spatiale n'est ajoutée (cas $\beta = 0$), *Thickness-Regul-Atlas* est identique à l'approche *Thickness-Direct*. Les cartes sont alors, de la même manière que dans le cas volumique, bruitées et manquent de cohérence spatiale (figure 4.2.a). Plus on ajoute de la régularisation spatiale, plus les voxels d'une même région ont tendance à être considérés comme similaires par le classifieur (figure 4.2.b-d).

Les régions pour lesquelles une atrophie augmente la probabilité d'être classé comme AD sont similaires à celles trouvées dans le cas de la régularisation anatomique et de la régularisation spatiale.

4.1.4 Performances de classification

4.1.4.1 Expériences

Dans ce chapitre nous utilisons les SVM régularisés spatialement pour la classification d'IRM anatomiques de sujets atteints de la maladie d'Alzheimer et de sujets sains. Plus précisément, nous appliquons les SVM régularisés spatialement aux problèmes de classification utilisés pour la comparaison de méthodes (chapitre 2).

Le premier est la classification entre les témoins (CN) et les patients AD probables (*CN vs AD*). Le deuxième problème est la classification entre les témoins et les sujets MCI qui convertissent vers la maladie d'Alzheimer durant les 18 premiers mois suivants l'inclusion (*CN vs* MCI_c). Il correspond au problème de détection de patients AD prodromaux. La troisième expérience est la classification des sujets MCI_{nc} versus les sujets MCI_c (MCI_{nc} *vs* MCI_c). Cela correspond à la prédiction de la conversion chez les patients MCI.

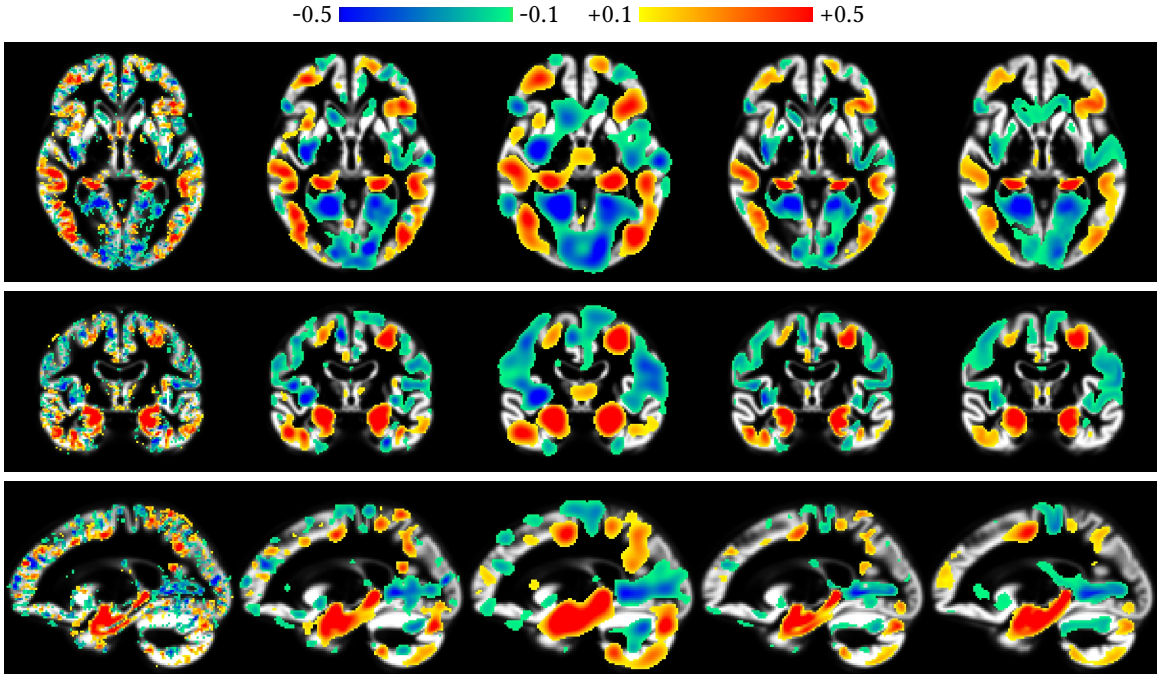


FIGURE 4.1 : Vecteur de poids du SVM $\mathbf{w}^{\text{opt}}$ normalisé par la norme infinie pour les méthodes : (a) *Voxel-Direct*, (b) *Voxel-Regul-Spatial* with FWHM $= 4$ mm, (c) *Voxel-Regul-Spatial* avec FWHM ~ 8 mm, (d) *Voxel-Regul-CombineFisher* avec FWHM ~ 4 mm, ($\sigma_{\text{loc}} = 10$), (e) *Voxel-Regul-CombineFisher* avec FWHM ~ 8 mm ($\sigma_{\text{loc}} = 10$).

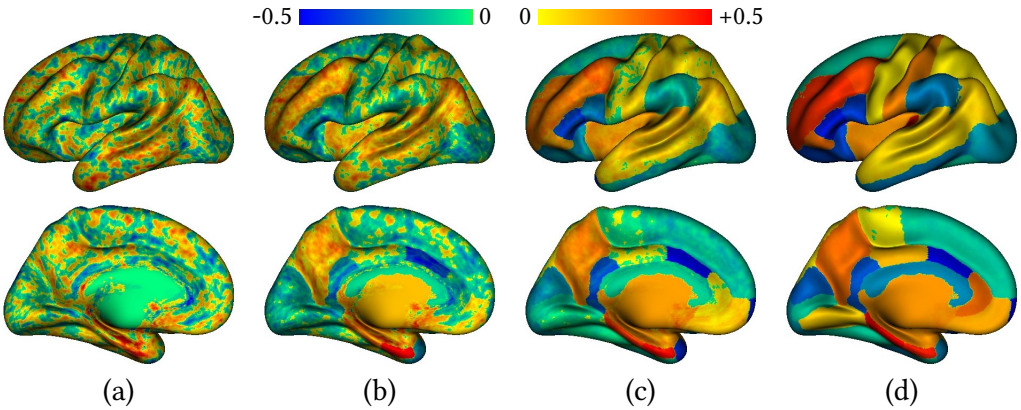


FIGURE 4.2 : Vecteur de poids du SVM $\mathbf{w}^{\text{opt}}$ normalisé par la norme infinie pour la méthode *Thickness-Regul-Atlas*. De (a) à (d), $\beta = 0, 2, 4, 6$.

4.1.4.2 Évaluation

Nous avons estimé les performances de classification suivant le même protocole que dans le chapitre 2 et sur les mêmes images. Ainsi, afin d'obtenir une évaluation non biaisée de ces performances, le groupe de participants est divisé en deux sous-groupes de tailles identiques : un groupe d'apprentissage et un groupe de test. Le groupe d'apprentissage est utilisé pour déterminer les valeurs optimales des hyperparamètres de chaque méthode et pour entraîner le classifieur. Le groupe de test est uniquement utilisé pour l'évaluation des performances de classification. L'ensemble d'apprentissage est utilisé pour estimer par validation croisée (CV - *cross validation*) les valeurs optimales des hyperparamètres. On effectue une recherche en grille pour trouver les paramètres optimaux. La recherche en grille est effectuée sur les intervalles/ensembles suivants : $C = 10^{-5.0}, 10^{-4.5}, \cdots, 10^{3.0}$, $\beta \in \{\alpha/\mu | \alpha \in \{0, 0.25, \cdots, 6\}, \mu \in \mathrm{Sp}(L)\}$, FWHM $= 0, 2, \cdots, 8$ mm et $\sigma_{\mathrm{loc}} = 5, 10$ mm.

Pour chaque approche, l'ensemble optimal d'hyperparamètres est utilisé pour entraîner le classifieur sur l'ensemble d'apprentissage ; les classifieurs obtenus sont alors évalués à l'aide du groupe de test.

4.1.4.3 Résultats

Les résultats des différentes expériences de classifications sont résumés dans les tableaux 4.2, 4.3 et 4.4 respectivement pour les comparaisons *CN vs AD*, *CN vs MCI*$_{\mathrm{c}}$ et *MCI*$_{\mathrm{nc}}$ *vs MCI*$_{\mathrm{c}}$.

CN vs AD. Les performances obtenues pour la comparaison *CN vs AD* sont résumées dans le tableau 4.2. Les taux de bonnes classifications sont compris entre 87% (*Voxel-Regul-CombineSum*) et 91% (*Voxel-Regul-CombineFisher*) dans le cas volumique. Ils ne sont pas significativement différents. Sans régularisation, on obtient 89%.

Quant au cas surfacique, *Thickness-Direct* atteint 83% de bonnes classifications ; la régularisation spatiale obtient 84% de bonnes classifications et la régularisation anatomique 85%. La combinaison des deux régularisations, *Thickness-Regul-CombineSum*, donne 87% de bonnes classifications.

CN vs MCI$_c$. Les performances obtenues pour la comparaison *CN vs MCI$_c$* sont rapportées dans le tableau 4.3. Les taux de classifications obtenus sont compris entre 78% et 84%.

MCI$_{nc}$ vs MCI$_c$. Les performances obtenues pour la comparaison *MCI$_{nc}$ vs MCI$_c$* sont rapportées dans le tableau 4.4. Aucune approche n'obtient des résultats significativement meilleurs que le hasard.

TABLE 4.2 : Performances de classification en termes de taux de bonnes classifications, de sensibilité et de spécificité pour la comparaison *CN vs AD*.

Méthode	**Taux de bonnes classifications**	**Sensibilité**	**Spécificité**
Voxel-Direct	89%	81%	95%
Voxel-Regul-Spatial	89%	85%	93%
Voxel-Regul-Atlas	90%	82%	96%
Voxel-Regul-CombineFisher	91%	88%	93%
Voxel-Regul-CombineSum	87%	82%	90%
Thickness-Direct	83%	74%	90%
Thickness-Regul-Spatial	84%	79%	88%
Thickness-Regul-Atlas	85%	82%	86%
Thickness-Regul-CombineSum	87%	83%	90%

TABLE 4.3 : Performances de classification en termes de taux de bonnes classifications, de sensibilité et de spécificité pour la comparaison *CN vs* MCI_c.

Méthode	**Taux de bonnes classifications**	**Sensibilité**	**Spécificité**
Voxel-Direct	84%	57%	96%
Voxel-Regul-Spatial	81%	62%	90%
Voxel-Regul-Atlas	84%	59%	95%
Voxel-Regul-CombineFisher	81%	65%	89%
Voxel-Regul-CombineSum	82%	62%	91%
Thickness-Direct	83%	54%	96%
Thickness-Regul-Spatial	82%	57%	94%
Thickness-Regul-Atlas	78%	51%	90%
Thickness-Regul-CombineSum	82%	57%	94%

Table 4.4 : Performances de classification en termes de taux de bonnes classifications, de sensibilité et de spécificité pour la comparaison MCI_{nc} *vs* MCI_c.

Méthode	Taux de bonnes classifications	Sensibilité	Spécificité
Voxel-Direct	64%	100%	0%
Voxel-Regul-Spatial	67%	49%	78%
Voxel-Regul-Atlas	64%	100%	0%
Voxel-Regul-CombineFisher	67%	49%	78%
Voxel-Regul-CombineSum	69%	54%	78%
Thickness-Direct	70%	32%	91%
Thickness-Regul-Spatial	67%	30%	88%
Thickness-Regul-Atlas	66%	27%	88%
Thickness-Regul-CombineSum	68%	22%	94%

4.1.4.4 Influence du paramètre β

Afin de mieux comprendre l'influence du paramètre de diffusion β sur les performances de classification, nous évaluons le taux de bonne classification en fonction de β pour la comparaison *CN vs AD*. Pour cela, le paramètre C du SVM est fixé à un. Les différents taux de classification en fonction de β sont résumés par la figure 4.3.

De manière générale, l'ajout de régularisation spatiale ou anatomique semble améliorer les taux de classification. Cependant, une régularisation trop forte conduit à une détérioration des performances.

4.1.5 Discussion

Dans cette section, nous avons utilisé les différentes approches de régularisation proposées au chapitre 3 pour le problème de classification d'IRM pondérés en T_1 de patients atteints de la maladie d'Alzheimer et de sujets sains témoins. Plus précisément, nous avons repris les expériences de classification utilisées dans la comparaison de méthodes présentée au chapitre 2 : la classification entre les témoins et les AD probables, la détection d'AD prodromaux et la prédiction de la conversion.

L'ajout d'un terme de régularisation spatiale dans le SVM conduit à des fonctions de classification moins bruitées et plus lisses spatialement. Lorsque ce terme de régularisation prend également en compte des informations de type anatomiques, les hyperplans séparateurs

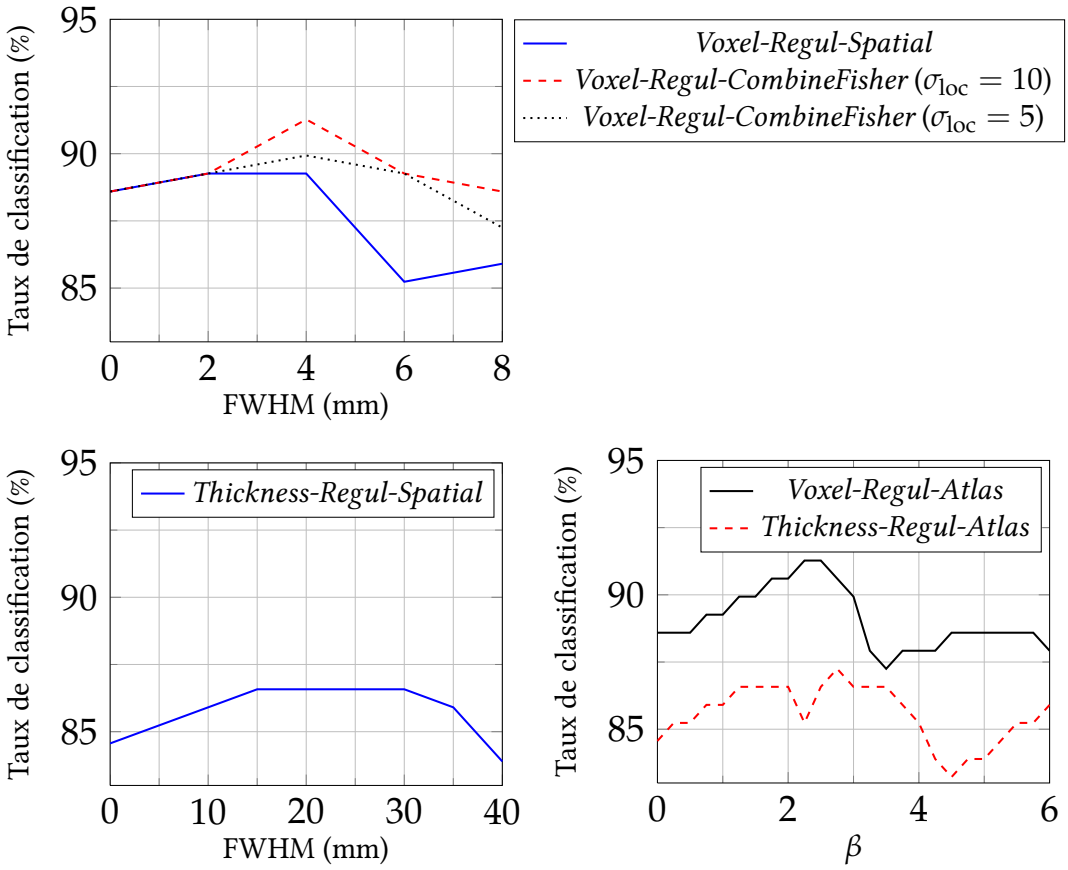


FIGURE 4.3 : Taux de bonne classification en fonction du paramètre de diffusion.

obtenus respectent mieux l'anatomie et/ou la pathologie. Par exemple dans l'approche utilisant la métrique de Fisher, si l'on utilise comme information de type atlas les informations sur les types de tissus, la régularisation lisse la fonction de classification tout en évitant de mélanger les voxels correspondant à des tissus distincts. Ce lissage respecte donc mieux les circonvolutions cérébrales. Lorsque l'information anatomique est de plus haut niveau (par exemple à partir d'un atlas probabiliste anatomo-fonctionnel), la régularisation proposée rend la fonction de classification plus cohérente avec l'information anatomique *a priori* en contraignant les voxels d'une même région à être considérés de la même manière par le classifieur.

Les performances de classifications sont comparables et même parfois légèrement meilleures que celles rapportées dans le chapitre de comparaison de méthodes. Nous avons utilisé les mêmes sujets, les mêmes images et le même protocole de validation qu'au chapitre 2. Avec les mêmes caractéristiques de classification, les méthodes *Voxel-Direct*, *Voxel-Atlas*, *Voxel-STAND* et *Voxel-COMPARE* obtiennent des taux de classification pour la comparaison *AD vs CN* respectivement de 89%, 86%, 81% et de 86% alors que les taux de classification obtenus

avec les approches régularisées spatialement sont compris entre 87% et 91%. Il en est de même pour les approches surfaciques où la régularisation ne détériore pas les performances de classification et les améliore même légèrement les résultats.

En ce qui concerne les autres comparaisons, et principalement la détection d'Alzheimer prodromaux, l'utilisation de la régularisation donne des résultats similaires. Les méthodes *Voxel-Direct*, *Voxel-Atlas*, *Voxel-STAND* et *Voxel-COMPARE* obtiennent pour cette comparaison des taux de classification respectivement de 84%, 75%, 81% et 71% tandis que les approches régularisées spatialement ont des taux de classification compris entre 81% et 84%.

Contrairement au cas de la classification des AD probables, la régularisation spatiale ne semble pas améliorer les résultats en termes de performances de classification pour le problème de classification de patients AD prodromaux. Ce résultat peut être dû au faible nombre de sujets mais il est également possible que la régularisation spatiale quadratique ne soit pas suffisante pour l'étude de stades plus précoces de la maladie. L'atrophie touchant moins de régions [Braak & Braak, 1991], il se peut qu'il faille ajouter à cette régularisation spatiale de la parcimonie afin de sélectionner uniquement les régions atteintes.

En conclusion, l'utilisation de la régularisation proposée dans le chapitre précédent (chapitre 3) permet en remplaçant la régularisation du SVM linéaire standard par une régularisation basée sur des *a priori* spatiaux et anatomiques d'obtenir des hyperplans séparateurs plus cohérents avec l'anatomie, et ce, sans perte de performances.

4.2 Application aux accidents vasculaires cérébraux

Dans cette section, nous proposons une utilisation du SVM régularisé pour détecter les altérations de diffusion associées aux bons ou mauvais pronostics de patients atteints d'accidents vasculaires cérébraux (AVC) ischémiques sylviens.

Après une introduction sur les accidents vasculaires cérébraux ischémiques (section 4.2.1), nous proposons une étude statistique des coefficients de l'hyperplan obtenu avec un SVM régularisé pour détecter des différences entre deux populations (section 4.2.2). Nous avons testé cette approche sur un exemple synthétique (section 4.2.2). Nous l'avons ensuite appliquée pour analyser une population de 72 patients atteints d'AVC afin de détecter les altérations de diffusion associées au pronostic (section 4.2.4). Une discussion des méthodes et résultats est présentée dans la section 4.2.5.

4.2.1 Les AVC ischémiques

Les accidents vasculaires cérébraux (AVC) sont la deuxième principale cause de décès selon le rapport 2004 de l'organisation mondiale de la santé (OMS). La définition de l'accident vasculaire cérébral donnée par l'OMS est la suivante. Un accident vasculaire cérébral est un déficit brutal d'une fonction cérébrale focale sans autre cause apparente qu'une cause vasculaire. L'atteinte de la fonction cérébrale peut être globale (coma, hémorragie méningée). Les symptômes doivent durer plus de 24 heures. L'évolution peut se faire vers la mort ou vers la régression totale, partielle ou nulle des déficits fonctionnels. Les AVC sont d'origine hémorragique ou ischémique. La grande majorité (80%) des AVC sont d'origine ischémique [Donnan et al., 2008].

4.2.1.1 Terminologie de l'AVC ischémique

L'accident vasculaire ischémique résulte d'une baisse du débit sanguin cérébral suffisamment sévère et prolongée pour que se constitue en son sein une zone de nécrose irréversible appelée *infarctus cérébral*. À la périphérie de cet infarctus, l'ischémie est plus modérée et la baisse du débit sanguin entraîne une modification du métabolisme qui est réversible si l'ischémie est levée suffisamment rapidement. Dans le cas contraire, cette zone va se nécroser. Cette zone est appelée *pénombre ischémique*. En périphérie de cette zone se trouve une zone appelée *oligémie*. L'oligémie est une réduction du débit sanguin mais qui ne modifie pas le métabolisme.

4.2.1.2 Mécanisme vasculaires à l'origine d'un infarctus cérébral

La diminution du débit sanguin cérébral à l'origine de lésions ischémiques peut résulter de causes et mécanismes divers. Les trois mécanismes principaux sont : le mécanisme embolique artérioartériel ou d'origine cardiaque, le mécanisme hémodynamique et l'atteinte des artères perforantes.

Mécanisme embolique. Le mécanisme embolique, surtout évoqué par l'apparition brutale du déficit neurologique, semble être le plus souvent impliqué dans la pathogénie des infarctus cérébraux. Il peut s'agir d'une embolie fibrinoplaquettaire à partir d'un thrombus[1] blanc résultant de l'adhésion des plaquettes sur la plaque d'athérosclérose[2], d'une embolie fibrinocruorique provenant de la fragmentation d'un thrombus mural à partir d'une plaque d'athérosclérose ulcérée, d'un thrombus formé dans une cavité cardiaque ou encore, ce qui est plus rare, de la migration à travers un foramen ovale perméable d'un thrombus veineux

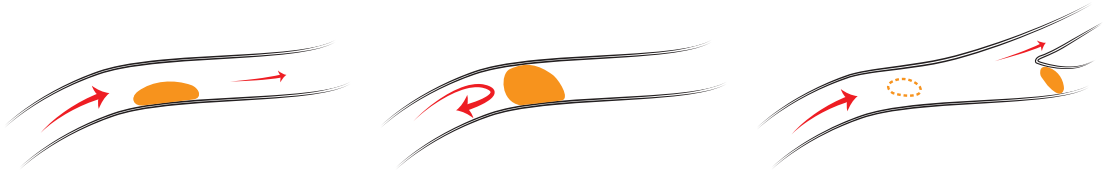
[1] Caillot formé par un vaisseau sanguin.

[2] L'athérome est un dépôt de plaques riches en cholestérol sur la paroi interne des artères, finissant par provoquer l'athérosclérose.

profond. On parle dans ce dernier cas d'embolie paradoxale. D'autres mécanismes emboliques existent mais ils sont beaucoup plus rares.

Mécanisme hémodynamique. L'accident hémodynamique est lui surtout évoqué lorsque la symptomatologie neurologique déficitaire est fluctuante, en particulier lorsque cette fluctuation clinique est corrélée aux changements de positions ou si elle est associée à une diminution de la pression artérielle, et ce, qu'elle qu'en soit la cause. Ce type de mécanisme s'observe parfois en cas de rétrécissement sévère d'une grosse artère à destinée cérébrale, que ce rétrécissement soit d'origine athéromateuse ou non, comme c'est le cas dans certaines dissections artérielles. Il est par ailleurs possible que ce type de mécanisme soit mis en évidence lors de l'infarctus en rapport avec un hémodétournement tel par exemple un vol sous-clavier ou encore à l'occasion d'un choc cardiogénique. En dehors d'infarctus siégeant dans le territoire des gros vaisseaux, ce type de mécanisme serait plutôt responsable du développement d'infarctus jonctionnels, à savoir des infarctus touchant de manière préférentielle la jonction de deux territoires artériels. Les infarctus consécutifs à un choc cardiogénique sont quant à eux plus volontiers des infarctus bilatéraux, parfois de type jonctionnel ou encore touchant préférentiellement les noyaux gris centraux.

Atteinte des artères perforantes. L'atteinte des artères perforantes est le plus souvent consécutive à une pathologie de la paroi artérielle sous la forme d'une lipohyalinose dans le contexte d'une hypertension artérielle ou d'un diabète. La pathologie de ces petites artères se traduit cliniquement par le développement de lésions ischémiques dites lacunaires (infarctus cérébraux de petite taille) ou par la survenue d'hémorragies profondes. Il semblerait que l'infarctus soit déterminé par l'obturation de l'une des branches perforantes profondes mais le mécanisme précis de la constitution de tels infarctus demeure encore discuté. Chez certains patients, ces lésions lacunaires peuvent être multiples. Dans ce contexte, la survenue d'une nouvelle lésion est parfois de diagnostic difficile, les séquences d'IRM de diffusion permettraient d'en faciliter le dépistage.



(a) Sténose (b) Thrombose (c) Embolisme

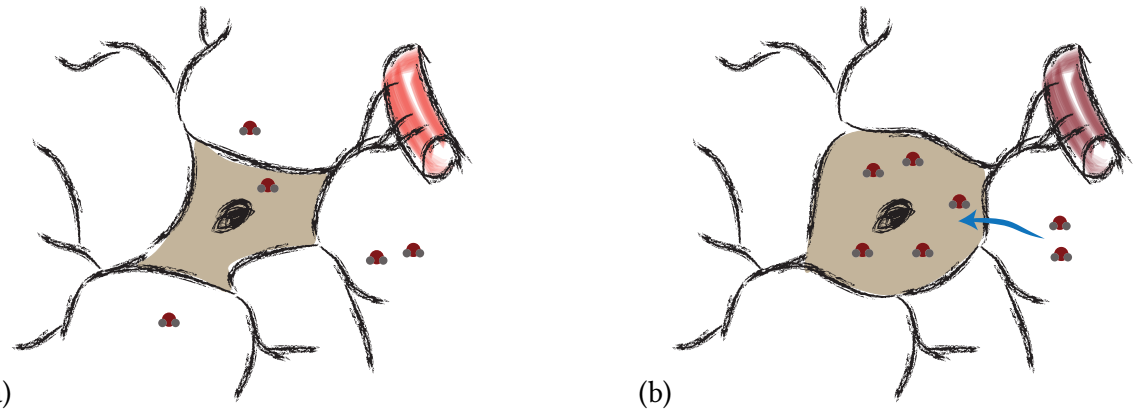
Figure 4.4 : Les causes d'accident vasculaire ischémique.

4.2.1.3 Mécanisme cellulaire de l'infarctus cérébral

Le mécanisme cellulaire de l'infarctus cérébral est très complexe et est encore relativement mal connu. Regardons deux modélisations assez simples des mécanismes qui sont à la base de notre étude.

Augmentation du taux d'acide lactique. La diminution du débit sanguin entraîne une diminution de dioxygène O_2 et de glucose $C_6H_{12}O_6$. La synthèse d'ATP[3] diminue donc et surtout devient anaérobique. Ceci à deux conséquences. La réaction anaérobique synthétise de l'acide lactique. Le taux d'acide lactique augment donc, ce qui concours à la mort de la cellule. La deuxième conséquence est une déplétion d'énergie. Lors d'une synthèse aérobique, une mole de glucose est convertie en 36 moles d'ADP, en revanche lors d'une synthèse anaérobique, une mole de glucose donne seulement deux moles d'ATP. En outre le taux de glucose est plus faible car le débit sanguin est plus faible.

Œdème cytotoxique. La déplétion énergétique secondaire à l'ischémie favorise la dépolarisation neuronale et gliale. La dépolarisation membranaire et la mise en jeu des récepteurs ionotropiques sensibles aux glutamates vont s'accompagner d'une entrée massive d'ions Na+ dans les cellules et de la sortie d'ions K^+. L'entrée d'ions Na+ va être associée à un afflux de molécules d'eau responsable d'un œdème cellulaire cytotoxique. La dépolarisation membranaire contribue par ailleurs à l'activation des canaux Ca^{2+} voltage-dépendants et donc à la libération de glutamate.



(a) (b)

FIGURE 4.5 : Illustration de l'œdème cytotoxique. La cellule, par exemple un astrocyte (a), en l'absence d'apport en oxygène, se remplit d'eau extracellulaire (b).

[3] adénosine triphosphate

Diminution du coefficient de diffusion de l'eau. De nombreuses études (ex : [Yang et al., 1999 ; Liu et al., 2001]) s'intéressent à l'évolution du coefficient de diffusion de l'eau dans les tissus lors d'un infarctus cérébral. La diffusivité de l'eau dans les tissus touchés par l'ischémie diminue durant les premières 48 heures après l'AVC. Les mécanismes qui sont à l'origine de cette diminution ne sont pas entièrement connus. Une des raisons de cette évolution du coefficient de diffusion de l'eau est l'œdème cytotoxique. Les molécules d'eau sont piégées à l'intérieur des cellules or la diffusivité de l'eau à l'intérieur des cellules est beaucoup plus faible à cause des membranes de la cellule mais également des macromolécules présentes à l'intérieur de la cellule telles que les protéines. Cette diminution de coefficient de diffusion est due également à d'autres facteurs tels que la diminution de perméabilité des membranes cellulaires. C'est cette diminution du coefficient de diffusion que nous exploiterons. L'IRM de diffusion permet de la quantifier.

4.2.1.4 Apport de l'imagerie de diffusion

L'imagerie pondérée en diffusion (*diffusion-weighted imaging* - DWI) est d'un intérêt considérable pour l'évaluation clinique de patients atteints d'un AVC [Chalela et al., 2007]. Elle permet l'étude de la lésion ischémique : sa localisation, sa taille et son évolution. Ainsi en corrélant ces données avec le pronostic des patients, l'IRM de diffusion permettrait d'identifier les régions associées au pronostic clinique.

L'identification de ces régions associées dès la phase aiguë est un point crucial pour trois raisons principales. Premièrement, la connaissance de ces régions pourrait permettre d'adapter la prise en charge thérapeutique et de guider les programmes de réhabilitation. De nouveaux traitements thérapeutiques pourraient cibler ces régions associées au pronostic à long terme [Heiss et al., 1999 ; Gladstone et al., 2002 ; Moustafa & Baron, 2007 ; Savitz & Fisher, 2007 ; Shuaib et al., 2007]. Deuxièmement, les interventions thérapeutiques sont plus efficaces aux phases aiguës [Moustafa & Baron, 2007 ; Konig et al., 2008]. Enfin cela pourrait également permettre de mieux informer les patients et leurs proches [Rosso et al., 2010].

De précédentes études en imagerie de tenseur (*diffusion tensor imaging* - DTI) et en IMRf ont mis en évidence le rôle clé du faisceau pyramidal [Thomalla et al., 2005 ; Cho et al., 2007 ; Kim et al., 2007 ; Yu et al., 2009] ainsi que du cortex moteur primaire [Crafton et al., 2003 ; Jaillard et al., 2005 ; Menezes et al., 2007]. Cependant, ces études étudient uniquement le pronostic moteur. L'étude d'un indice plus général du pronostic tel que l'échelle de Rankin modifiée (mRS) serait plus intéressante. De plus, ces études ont été effectuées à des stades sous-aigus ou chronique. Aucune d'entre elles n'a, à ma connaissance, détecté des différences à la phase aigüe (<48 hours) [Yu et al., 2009]. C'est ce que nous voulons étudier.

À la phase aiguë, comme nous l'avons vu précédemment, il semble logique d'étudier les cartes de coefficients apparents de diffusion. Rosso et al. [2010] suggèrent que les modifications

régionales des valeurs d'ADC représente un indice quantitatif de la sévérité d'une lésion. Nous avons donc étudié les modifications d'ADC corrélées au pronostic à trois mois.

4.2.2 Analyse statistique de l'hyperplan séparateur

4.2.2.1 L'hyperplan séparateur informe sur le pouvoir discriminant de features

La fonction de classification f^{opt} obtenue avec un SVM linéaire est donnée par :

$$\forall \mathbf{x} \in \mathcal{X}, f^{\text{opt}}(\mathbf{x}) = \langle \mathbf{w}^{\text{opt}} | \mathbf{x} \rangle + b \tag{4.1}$$

Par conséquent, si la valeur absolue de la i-ème composante du vecteur $\mathbf{w}^{\text{opt}}$, w_i^{opt}, est petite comparativement aux autres composantes $\left(\left|w_j^{\text{opt}}\right|\right)_{j \neq i}$, le i-ème *feature* va avoir une faible influence sur la fonction de classification. Inversement, si w_i est relativement grand, le i-ème *feature* va jouer un rôle important dans le classifieur.

C'est cette idée que nous avons déjà précédemment utilisée pour l'analyse qualitative des hyperplans séparateurs optimaux (section 2.5.6). Cependant, jusqu'à présent les différences observées entre deux populations à l'aide de l'hyperplan séparateur optimal déterminé par le SVM ne furent que qualitatives et non quantitatives. Nous allons l'utiliser pour localiser des différences significatives entre deux populations. En d'autres termes, on aimerait bien seuiller ces cartes afin de localiser les différences significatives entre deux groupes. Le problème est qu'il est, à ma connaissance, impossible de seuiller directement les cartes puisque la distribution des $\left(w_i^{\text{opt}}\right)_i$ est inconnue. Cette distribution étant inconnue, une alternative serait d'utiliser une approche non paramétrique tel que les tests de permutations. Mais cela pose quelques problèmes.

4.2.2.2 Les difficultés liées à l'analyse directe des coefficients de l'hyperplan

En effet l'inconvénient majeur avec l'analyse directe des coefficients de l'hyperplan séparateur optimal donné par le SVM est que l'on ne peut pas comparer directement les coefficients obtenus avec deux SVM différents lors de deux comparaisons différentes. Plus spécifiquement, soient $\mathcal{S}_1$, $\mathcal{S}'_1$, $\mathcal{S}_2$ et $\mathcal{S}'_2$ quatre groupes de sujets. Soit $\mathbf{w}^{(1)^{\text{opt}}}$ le vecteur des coefficients de l'hyperplan séparateur donné par un SVM lors de l'analyse entre les groupes $\mathcal{S}_1$ et $\mathcal{S}'_1$. De la même manière, soit $\mathbf{w}^{(2)^{\text{opt}}}$ le vecteur des coefficients de l'hyperplan obtenu par un SVM lors de la comparaison entre $\mathcal{S}_2$ et $\mathcal{S}'_2$. Si la séparation est plus grande entre les groupes $\mathcal{S}_1$ et $\mathcal{S}'_1$

qu'entre les groupes $\mathcal{S}_2$ et $\mathcal{S}'_2$, alors :

$$\left\|\mathbf{w}^{(1)^{\text{opt}}}\right\| \leq \left\|\mathbf{w}^{(2)^{\text{opt}}}\right\|$$

Pour résumer, on a deux effets qui s'opposent :

- Pour une comparaison donnée, le i-ème *feature* a un poids plus important que le j-ème si et seulement si : $w_i > w_j$
- Avec les notations précédentes, entre deux comparaisons, si l'écart entre les groupes $\mathcal{S}_1$ et $\mathcal{S}'_1$ est plus grand qu'entre les groupes $\mathcal{S}_2$ et $\mathcal{S}'_2$, alors : $\left\|\mathbf{w}^{(1)^{\text{opt}}}\right\| \leq \left\|\mathbf{w}^{(2)^{\text{opt}}}\right\|$

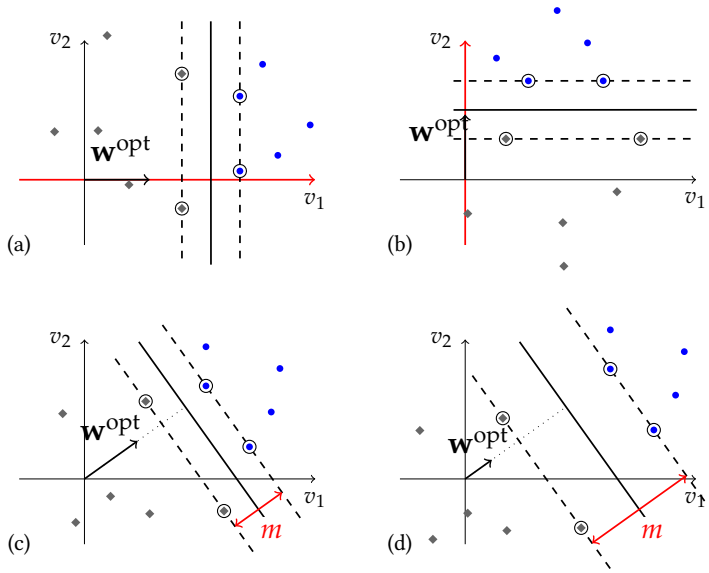


Figure 4.6 : Illustration de l'information contenue dans un SVM. L'orientation de l'hyperplan donne de l'information sur l'importance relative des *features* ((a) et (b)). Quant à la marge m, elle informe sur la séparation entre deux groupes.

On ne peut donc pas faire des tests de significativité directement sur les composantes $\left\|w_i^{(1)^{\text{opt}}}\right\|$ et $\left\|w_i^{(2)^{\text{opt}}}\right\|$

4.2.2.3 Analyse des coefficients de l'hyperplan séparateur pondérés par la marge

Le SVM recherche l'hyperplan qui maximise la **marge** m entre les deux groupes. La marge m quantifie la séparation entre les deux groupes comparés. Rappelons que pour le SVM linéaire

standard (3.1), la marge m est donnée par Schölkopf & Smola [2001] :

$$m = \frac{2}{\|\mathbf{w}^{\text{opt}}\|} \tag{4.2}$$

quant à la version régularisée, (3.8),

$$m = \frac{2}{\left\|\exp\left(\frac{1}{2}\beta L\right)\mathbf{w}^{\text{opt}}\right\|} \tag{4.3}$$

Ainsi, en combinant la marge m et $\frac{|w_i^{\text{opt}}|}{\|\mathbf{w}^{\text{opt}}\|}$, on arrive à prendre en compte à la fois la séparation entre les groupes et l'importance relative des *features*.

Pour cette raison nous proposons d'analyser les coefficients du SVM pondérés par la marge, à savoir :

$$\frac{m|w_i^{\text{opt}}|}{\|\mathbf{w}^{\text{opt}}\|} \tag{4.4}$$

Pour cela, nous effectuons des tests de permutations sur $\frac{m|w_i^{\text{opt}}|}{\|\mathbf{w}^{\text{opt}}\|}$ avec comme hypothèse nulle $\mathcal{H}_0$ l'absence de relation entre le label des sujets et l'IRM. En permutant le label des sujets (en pratique 20 000 fois), et en entraînant le SVM à chaque permutation, on estime la distribution de probabilité de $\frac{m|w_i^{\text{opt}}|}{\|\mathbf{w}^{\text{opt}}\|}$ sous l'hypothèse nulle $\mathcal{H}_0$. En utilisant ces distributions de probabilité, il est alors possible de tester l'hypothèse $\mathcal{H}_0$ au niveau du voxel. Le *false discovery rate* (FDR) est utilisé pour corriger pour les comparaisons multiples en suivant la procédure de Benjamini & Hochberg [1995]. À ma connaissance, les autres analyses statistiques ne prennent pas en comte la marge dans leurs analyses (e.g. [Mourao-Miranda et al., 2005 ; Wang et al., 2007 ; Sato et al., 2009]).

4.2.3 Exemple synthétique

On teste d'abord la méthode présentée dans la section précédente sur un exemple synthétique.

4.2.3.1 Construction de l'exemple

L'exemple synthétique est construit comme ci suit (figure 4.7). Nous avons considéré deux groupes de 20 images 2D (une coupe 116×92 avec des voxels isotropes de taille 1.5 mm). Nous avons utilisé une coupe d'un *template* de substance blanche. Pour chacune des 40 images, les voxels de la substance blanche (d'après le *template*) ont une valeur aléatoire comprise entre 0 et 1. Les autres voxels ont une intensité nulle. Pour chaque image du deuxième groupe, on construit une hyperintensité $h_{\mathbf{green}}$ dans la région verte et $h_{\mathbf{red}}$ dans la région rouge (c.f. figure 4.7) de telle sorte que $h_{\mathbf{green}} + h_{\mathbf{red}} \sim \mathcal{N}(2, 0.2)$. On ajoute du bruit blanc gaussien à

toutes les images ($\mathcal{N}(0,1)$). La loi de probabilité des voxels des régions non hyperintenses de la matière blanche et celle des voxels des régions hyperintenses sont représentées par la figure 4.8.

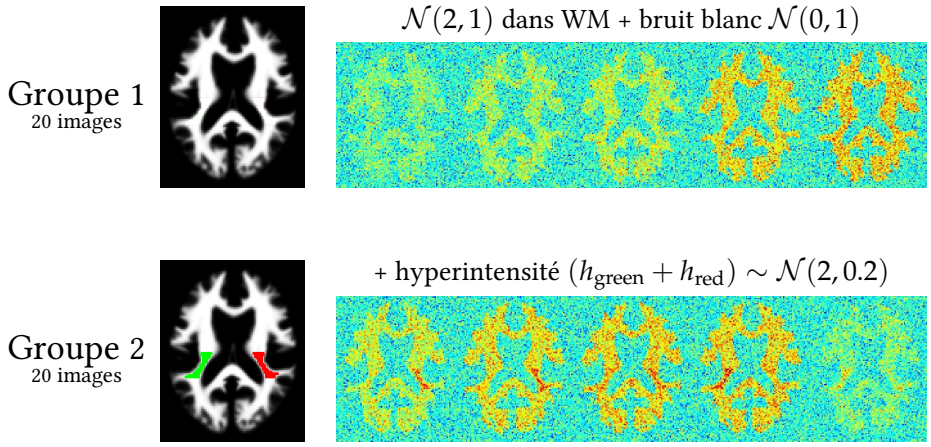


Figure 4.7 : Construction de l'exemple synthétique. La ligne du haut montre de gauche à droite la coupe du *template* de substance blanche utilisé et cinq images du premier groupe choisies aléatoirement. La ligne du bas représente la même coupe du *template* avec les régions hyperintenses à détecter (en rouge et en vert) ainsi que cinq images sélectionnées aléatoirement du deuxième groupe avec les hyperintensités.

4.2.3.2 Expériences

Analyses de groupes. Nous avons testé trois méthodes univariées et trois méthodes utilisant les SVM. Les trois analyses univariées (par tests de permutations) ont été réalisées sur les intensités des voxels : des images brutes, des images lissées avec un noyau gaussien et des images prétraitées par $e^{-\frac{\beta}{2}L}$ (avec L le laplacien du graphe utilisé pour la régularisation du SVM). Nous avons également testé trois méthodes utilisant un SVM : (i) le SVM linéaire standard sur les images brutes, (ii) le SVM linéaire standard sur les images lissées avec un noyau gaussien et (iii) le SVM régularisé spatialement sur les images brutes.

Le graphe de régularisation est celui présenté à la section 3.5.2, c'est à dire le graphe de connectivité de l'image avec comme pondération $A_{1,2}$ entre deux voxels voisins $v^{(1)}$ et $v^{(2)}$:

$$A_{1,2} = \exp\left(\frac{-\chi^2(p_{v^{(1)}}, p_{v^{(2)}})^2}{2\sigma^2}\right) \tag{4.5}$$

avec $p^{(t)}_{v^{(i)}}$ la probabilité que le i-ème voxel appartienne au tissu $t \in \mathcal{T}$ (ici « matière blanche » et « non matière blanche »). Pour éviter tout problème de l'ajustement de paramètres, σ est

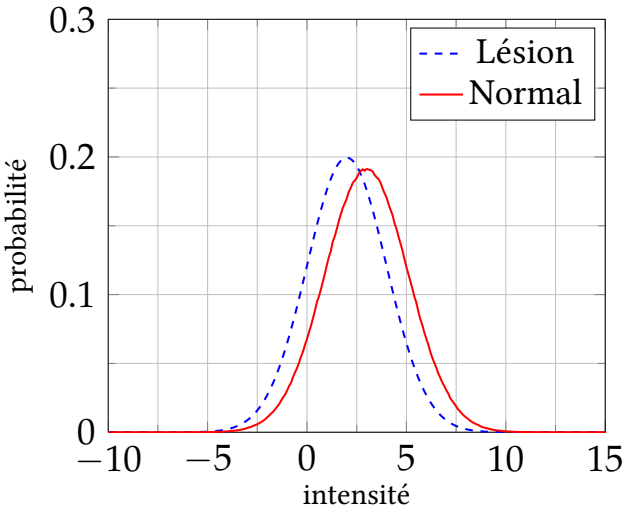


FIGURE 4.8 : Exemple synthétique. Lois de probabilité des voxels dans la substance blanche sans hyperintensité (Normal) et des voxels dans une région de la substance blanche avec hyperintensité (Lésion).

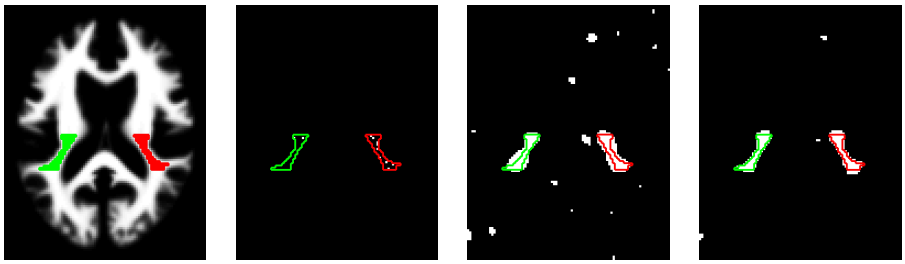
préalablement fixé comme étant égal à la variance de $\chi^2(p_{v^{(i)}}, p_{v^{(j)}})$ et le paramètre β de diffusion a été choisi pour correspondre au lissage gaussien des analyses univariées. Le nombre d'itérations des tests de permutations est 20000. Tous les tests furent corrigés avec un FDR à 5%.

4.2.3.3 Résultats

Analyses de groupes. Les résultats de ces études de groupes sont présentés dans la figure 4.9. Les trois analyses univarées n'ont détectée aucune différence significative entre les deux groupes. L'analyse à l'aide d'un SVM linéaire standard sur les images brutes n'a détecté que très peu de voxels. Le SVM avec les images lissées a détecté les deux régions mais aussi de nombreux clusters faux positifs. Le SVM régularisé spatialement a également détecté les deux régions hyperintenses mais plus précisément et aussi avec moins de clusters épars.

4.2.4 Détection des régions associées au devenir des patients

Nous avons appliqué notre analyse des coefficients de l'hyperplan (section 4.2.2) à la détection des régions associées au pronostic à trois mois des patients ayant eu un accident vasculaire. Cette étude utilise des IRM pondérées en diffusion (DWI) acquises à la phase aigüe.



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

FIGURE 4.9 : Exemple synthétique : (a) *template* WM et les régions à détecter ; détection avec un SVM linéaire (b) sur les images brutes ; (c) sur les images lissées ; (d) détection avec un SVM régularisé spatialement sur les images brutes. Toutes les autres analyses n'ont détecté aucune différence.

4.2.4.1 Sujets

Les critères d'inclusion dans cette étude sont les suivants :

- ischémie dans le territoire sylvien profond ;
- une IRM initiale 1.5 T DWI acquise dans les six premières heures après l'accident ;
- une IRM 1.5T DWI de contrôle dans les 3 jours après l'accident ;
- une évaluation clinique à trois mois à l'aide du score de Rankin modifié (mRS).

Les critères d'exclusion sont : la transformation hémorragique de l'ischémie ou le décès dans les 90 jours après l'accident. Au final, 72 patients consécutifs (âge moyen : 60 ± 14 ans $[24 - 81]$) ont été inclus dans cette étude.

Tous les patients ont été pris en charge selon la procédure de routine clinique des Urgences Cérébro-Vasculaires du groupe hospitalier Pitié-Salpêtrière. En particulier, certains sujets ont été thrombolysés. Plus particulièrement un activateur tissulaire plasminogène recombinant[4] (rt-PA) fut administré par intraveineuse aux sujets ayant un score NIHSS (National Institutes of Health Stroke Scale) supérieur à 4 sans amélioration majeure, une ischémie aigüe visible en IRM, une absence d'hémorragie ainsi que la preuve d'une occlusion intracrânienne. La fenêtre thérapeutique est de cinq heures.

Le score de Rankin modifié (mRS) est utilisé pour évaluer le pronostic à 90 jours. Le bon pronostic est défini comme l'indépendance (mRS 0, 1 ou 2 ; 39 patients) et le mauvais pronostic comme un handicap sévère (mRS de 3 à 5 ; 33 sujets). Les données démographiques de la population d'étude son résumées dans le tableau 4.5.

[4] agent utilisé pour la thrombolise

TABLE 4.5 : Caractéristiques démographiques de la population d'étude. Les valeurs sont indiquées comme moyenne $\pm$ écart type [intervalle]. mRS : score de Rankin modifié. Les p-valeurs sont obtenues avec le test T de Student.

Pronostic		*bon* mRS 0-2	*mauvais* mRS 3-5
Nombre		39	33
Age (années)	p=0.002	55.8 ± 15.0 $[24 - 81]$	64.9 ± 10.9 $[38 - 80]$
mRS (3 mois)	p<0.0001	1.2 ± 0.7 $[0 - 2]$	4.0 ± 0.8 $[3 - 5]$
NIHSS (<6h)	p<0.0001	11.4 ± 5.4 $[2 - 23]$	18.5 ± 4.9 $[6 - 30]$
NIHSS (un jour)	p<0.0001	5.5 ± 4.1 $[0 - 14]$	16.8 ± 6.0 $[6 - 35]$
Délai de la première IRM (min)	p=0.2	152 ± 66 $[66 - 360]$	164 ± 63 $[66 - 341]$

4.2.4.2 Acquisition IRM et prétraitements

Toutes les données cliniques et les données d'imagerie ont été acquises en routine clinique dans le service des Urgences Cérébro-Vasculaires du groupe hospitalier Pitié-Salpêtrière. L'étude a été approuvée par le comité d'éthique de l'Hôpital de la Pitié Salpêtrière.

Les acquisitions IRM ont été effectuées avec une unité 1.5 Tesla (General Electric Signa Horizon Echospeed). Trois séquences différentes ont été réalisées : une DWI, un FLAIR (Fluid Attenuated Inversion Recovery) et une MRA.

L'image DWI a été réalisée en axial avec une séquence d'écho de spin EPI (imagerie échoplanaire) avec réhaussement de gradient et les paramètres d'acquisition suivants : 24 coupes de 5 mm d'épaisseur avec 0.5 mm d'espace intercoupe, un temps de répétition (TR) de 2825 ms, un temps d'écho (TE) de 98.9 ms, angle d'impulsion radiofréquence de 90° (*flip angle*, un champ de vision (FOV - *field-of-view*) de 280×210 mm^2, une matrice de 96×64.

La mesure des cartes de coefficients apparents de diffusion (ADC - *apparent diffusion coefficients*) à partir des IRM de diffusion (une b-0 et une b-1000) a été réalisée à l'aide d'un logiciel commercialisé (Functool 2, General Electric, Buc, France). Les cartes d'ADC ont ensuite été normalisées dans l'espace de référence du Montreal Neurological Institute (MNI)

en utilisant le *template* T_2 de SPM5 (Statistical Parametric Mapping, Wellcome Department of Neurology, Institute of Neurology, London, UK). Pour l'analyse, toutes les lésions ont été mises artificiellement dans le même hémisphère. Nous avons utilisé les cartes d'ADC à un jour.

4.2.4.3 Expérience

Nous avons effectué des études de groupes sur les cartes d'ADC. Les analyses univariées ont été réalisées à la fois avec des tests de permutations et avec un test T de Student sur des images lissées préalablement avec un noyau gaussien (FWHM $=$ 8 mm). Nous avons également effectué des analyses de groupes à l'aide du SVM régularisé spatialement présenté à la section 3.5.2. Le graphe de régularisation est donc le graphe de connectivité de l'image avec comme pondération $A_{1,2}$ entre deux voxels voisins $v^{(1)}$ et $v^{(2)}$:

$$A_{1,2} = \exp\left(\frac{-\chi^2(p_{v^{(1)}}, p_{v^{(2)}})^2}{2\sigma^2}\right) \tag{4.6}$$

avec $p^{(t)}_{v^{(i)}}$ la probabilité que le i-ème voxel appartienne au tissus $t \in \mathcal{T}$ ($\mathcal{T} = \{GM, WM, CSF\}$). Ces probabilités sont définies dans cette étude à l'aide des *templates* de tissus de SPM. Pour éviter tout problème de l'ajustement de paramètres, σ est préalablement fixé comme étant égal à la variance de $\chi^2(p_{v^{(i)}}, p_{v^{(j)}})$ et le paramètre β de diffusion a été choisi pour correspondre au lissage gaussien des analyses univariées. Le nombre d'itérations des tests de permutations est 20000. Tous les tests ont été corrigés avec un FDR à 5%.

Nous avons également évalué le pouvoir de prédiction du pronostic à trois mois à partir des cartes d'ADC avec un SVM régularisé spatialement et avec un SVM linéaire standard. L'évaluation des performances de classification est faite à l'aide d'une procédure de validation croisée de type *leave-one-out* (LOO-CV). Le paramètre C du SVM est arbitrairement fixé à $C = 1$ pour cette étude.

4.2.4.4 Résultats

Les analyses de groupes entre les patients avec un bon pronostic et les patients avec un mauvais pronostic à partir des cartes d'ADC à un jour ont donné les résultats suivants. Les analyses univariées n'ont détecté aucune différence entre les deux groupes. Les analyses utilisant un SVM régularisé spatialement ont détecté des différences significatives d'ADC dans une région de 11cm^3 (1348 voxels). Ces modifications significatives de la valeur d'ADC sont localisées principalement (un *cluster* de 1265 voxels) dans la partie postérieure du putamen, dans la zone postérieure de la capsule interne, dans la substance blanche périventriculaire et dans la région inférieure du cortex moteur primaire. Des modifications d'ADC ont également été détectées

dans de très petits *clusters* situés dans le cortex insulaire (figure 4.10 et 4.11) (37 voxels), dans l'unciné (21 voxels) et dans le bulbe rachidien controlatéral (25 voxels).

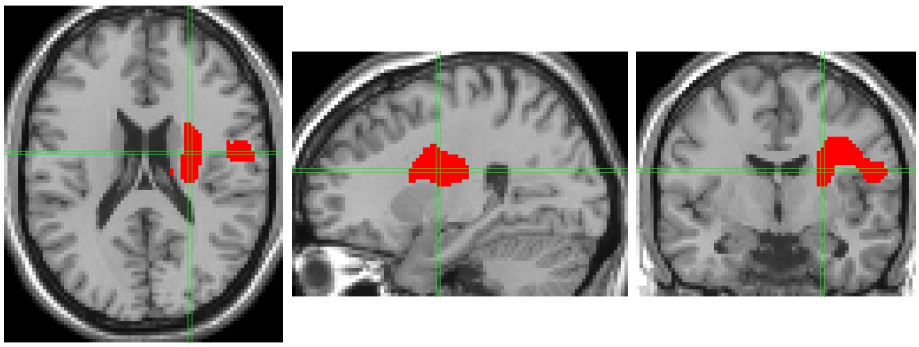


FIGURE 4.10 : Différences de groupes entre les *bons* et les *mauvais* pronostics avec un SVM régularisé spatialement à partir des cartes d'ADC à un jour (z=20 mm, x=28 mm and y=-8 mm dans l'espace standardisé du MNI).

Les taux de bonnes prédictions du pronostic à trois mois à partir des cartes d'ADC sont les mêmes avec un SVM linéaire standard et un SVM régularisé spatialement (76%).

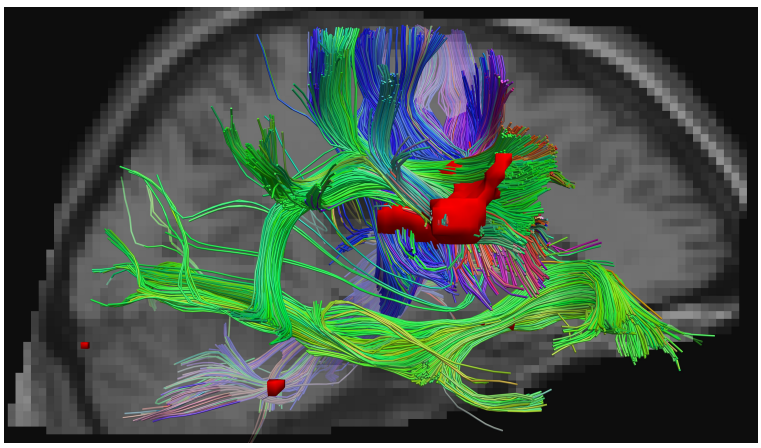


FIGURE 4.11 : En rouge : les régions détectées superposées aux faisceaux de fibres de substance blanche d'un sujet sain. Ces faisceaux ont été obtenus à partir d'imagerie de diffusion spectrale (DSI) à l'aide du logiciel Diffusion Toolkit (`http://www.trackvis.org/dtk/`). Les couleurs des fibres codent leur orientation.

4.2.5 Discussion

Dans cette section nous avons appliqué le SVM régularisé spatialement à la détection des régions du cerveau associées à la phase aiguë au pronostic à trois mois de patients ayant eu un accident vasculaire ischémique. Cette méthode a permis de détecter des changements significatifs dans le territoire sylvien profond. Plus particulièrement, ces changements concernent la substance blanche périventriculaire, le faisceau pyramidal et la partie postérieure du noyau lenticulaire. Les analyses univariées n'ont détecté aucune différence significative.

L'implication du faisceau pyramidal est cohérente avec différentes études réalisées aux phases avec des études DTI antérieures réalisées à des phases sous-aiguës [Konishi et al., 2005 ; Thomalla et al., 2005 ; Cho et al., 2007 ; Kunimatsu et al., 2007 ; Nelles et al., 2007 ; Jang et al., 2008 ; Domi et al., 2009 ; Yu et al., 2009] ou chroniques [Newton et al., 2006 ; Stinear et al., 2007 ; Schaechter et al., 2008] ainsi qu'avec des études en tomodensitométrie [Feydy et al., 2002 ; Wenzelburger et al., 2005] également réalisées à la phase chronique. Nous avons étendu ces résultats à la phase aiguë, en utilisant les valeurs d'ADC à un jour et avec une mesure de pronostic global. Ces résultats pourraient fournir des informations utiles pour guider des programmes de soin précoces. Ils pourraient également servir à donner des informations plus précises aux proches.

Les limitations de notre étude sont les suivantes. En plus du grand *cluster* localisé essentiellement dans le faisceau pyramidal, de plus petits *clusters* ont également été détectés dans les régions de l'insula, de l'unciné et dans le bulbe rachidien. Ces petits clusters peuvent être dus à un artéfact de la méthode ; notons en particulier que la méthode de correction des comparaisons multiples utilisée est le FDR. De plus la localisation de ces *clusters* n'est qu'approximative, compte tenu du fait que les images DWI acquises en phase aiguë sont sujettes aux erreurs de recalages et ont de grands voxels (coupes de 5 mm). De plus, dans sa forme actuelle, la méthode de comparaison ne prend pas en compte les covariables. Pour cette raison, l'âge et le genre n'ont pas été pris en compte en tant que covariables. Néanmoins, les différences détectées sont fortement latéralisées, ce qui laisse penser que ces modifications d'ADC sont essentiellement dues à la pathologie et non pas à l'âge.

En conclusion, notre méthode d'étude de groupe à base de SVM régularisé spatialement a permis de détecter des régions associées aux bons ou mauvais pronostics de patients ayant eu un accident cérébral sylvien profond. Ces différences détectées dès la phase aiguë impliquent principalement la substance blanche périventriculaire et le faisceau pyramidal. Cette détection a été permise en dépassant les limites de l'analyse univariée à l'aide de machines à vecteurs supports.

Conclusion

L'analyse automatique de différences anatomiques en neuroimagerie a de nombreuses applications pour la compréhension et l'aide au diagnostic de pathologies neurologiques. Récemment, il y a eu dans la communauté de neuroimagerie un intérêt croissant pour les méthodes de classification comme les SVM. De telles approches permettent de dépasser les limites de l'analyse univariée en prenant en compte des relations multivariées présentes dans les données.

* *
*

Dans cette thèse, nous nous sommes intéressés à l'apprentissage automatique par machines à vecteurs supports pour l'analyse de populations et la classification de patients en neuroimagerie.

Après une introduction sur l'apprentissage statistique, nous avons présenté un bref état de l'art des méthodes de classification d'images structurelles en neuroimagerie. Différents choix méthodologiques sont possibles à différentes étapes de la classification : définition des *features*, réduction de la dimension, choix du classifieur. Dans un deuxième temps, nous avons évalué les performances de ces différentes stratégies pour la classification automatique d'images anatomiques pondérées en T_1 pour l'aide au diagnostic de la maladie d'Alzheimer. Cette étude de comparaison a été effectuée sur une large population issue de la base de données ADNI. La grande majorité des méthodes a réussi à classer les patients atteints de la maladie d'Alzheimer des sujets témoins avec de bonnes performances. Cependant, les résultats obtenus pour la détection des AD prodromaux sont moins bons. Dans nos expériences, l'utilisation de *features* surfaciques n'a pas donné de résultats supérieurs aux approches voxelliques mais a augmenté de façon importante le temps de calcul. En revanche, l'utilisation d'une méthode de recalage complètement déformable a amélioré les résultats.

Nous avons ensuite proposé une approche permettant d'introduire des régularisations spatiales ou anatomiques dans des machines à vecteurs supports. Elle consiste à introduire de la connaissance *a priori* dans la classification sous la forme d'opérateurs de régularisation. L'idée était de contraindre le SVM à prendre en compte la structure des images ainsi que des informations sur l'anatomie. Nous avons notamment proposé des approches dans le cadre discret et dans le cadre continu. Elles sont valables dans le cas volumique ainsi que dans le cas surfacique. Les régularisations utilisées passent par des pénalisations quadratiques. Notre approche peut être reliée aux noyaux de diffusion.

Nous avons finalement appliqué cette nouvelle approche à la maladie d'Alzheimer et aux accidents vasculaires cérébraux. Les résultats montrent que cette méthode permet, en remplaçant le terme de régularisation standard des machines à vecteurs supports par un terme de régularisation spatiale, de prendre en compte la distribution spatiale et anatomique des *features*. Les classifieurs ainsi obtenus ont des hyperplans séparateurs moins bruités et plus cohérents avec l'anatomie sans perte de performances de classification. Dans le cas des accidents vasculaires cérébraux, une analyse statistique des hyperplans séparateurs a montré que des altérations de diffusion dans la région du faisceau pyramidal sont associées au mauvais pronostic des patients.

* *
*

Les perspectives de cette thèse sont diverses.

Dans le modèle de proximité anatomique que nous avons proposé, deux *features* sont considérés comme proches s'ils appartiennent à la même structure cérébrale ou s'ils sont connectés, cette connexion pouvant être de nature anatomique (ex : architecture des réseaux de fibres) ou de nature fonctionnelle (ex : synchronies cérébrales ou corrélation du signal IRMf). Cependant, nous n'avons pas testé dans cette thèse, la régularisation basée sur des **atlas de connectivité anatomique ou fonctionnelle**. Une première perspective de la thèse serait d'utiliser de tels atlas dans l'étude de pathologies.

L'approche de régularisation spatiale et/ou anatomique proposée dans la thèse n'est pas spécifique aux IRM structurelles et pourrait être appliquée à d'autres types de données comme des données fonctionnelles (IRMf, EEG, MEG, ...). Cependant, la régularisation proposée dans cette thèse est uniquement spatiale et nécessiterait une adaptation pour pouvoir prendre en compte la **dimension temporelle** des données fonctionnelles.

La régularisation que nous avons proposée dans cette thèse est quadratique. Un des grands avantages d'une telle régularisation est l'utilisation de l'astuce du noyau : il est ainsi possible de travailler dans des espaces de faibles dimensions. Mais il se peut que l'on ait besoin de parcimonie. Si l'on considère, par exemple, la maladie d'Alzheimer, il est possible que la régularisation spatiale quadratique ne soit pas suffisante pour l'étude des stades précoces de la maladie. L'atrophie touchant moins de régions [Braak & Braak, 1991], il se peut qu'il faille ajouter à cette régularisation spatiale de la parcimonie afin de sélectionner uniquement les régions atteintes. Comme nous l'avons mentionné dans la discussion du chapitre 2, l'utilisation d'une pénalisation de type LASSO [Tibshirani, 1996] ne nous semble pas appropriée : il semble peu probable que les différences entre deux populations ne mettent en jeu qu'un petit nombre de voxels, et ce, d'autant plus que les données sont sujettes aux erreurs de recalage. Utiliser la combinaison des régularisations ℓ_1 et quadratique (***elastic net*** [Zou & Hastie, 2005]) est une première possibilité. Une autre approche qui semble réaliste est de **pénaliser par la variation totale** de l'hyperplan séparateur. Cela reviendrait à contraindre l'hyperplan à être constant par morceaux. En d'autres termes, cela revient à contraindre $\nabla \mathbf{w}^{\text{opt}}$ à être parcimonieux.

Dans certaines pathologies, l'utilisation d'une seule modalité d'imagerie n'est pas suffisante pour le diagnostic. Il est parfois nécessaire d'utiliser de l'information provenant de plusieurs modalités d'imagerie ; on parle alors de **multimodalité**. On peut notamment citer certaines formes d'épilepsie comme l'épilepsie de la face médiale du lobe temporal avec sclérose de l'hippocampe. La sclérose de l'hippocampe est définie sur le plan anatomopathologique. Sur le plan neuroradiologique, elle apparaît comme une atrophie de l'hippocampe visible en IRM pondérée en T_1 et/ou un hypersignal en IRM pondérée en T_2 (présence de glie) [Duncan, 1997]. La multimodalité est également cruciale pour des problèmes de segmentation de lésions comme les lésions vasculaires[5] [Anbeek et al., 2004].

Dans le cadre de la classification d'images, son utilisation est problématique. Une difficulté réside dans l'augmentation de la dimension de l'espace des données. Plusieurs approches ont été proposées par exemple en concaténant les données [Fan et al., 2008b] ou en combinant les noyaux obtenus avec les différentes modalités [Zhang et al., 2011]. Ces approches uniquement guidées par les données (*data-driven*) nous semblent incertaines. L'augmentation des dimensions entraîne inéluctablement une augmentation de la variabilité du modèle. Par conséquent, il est nécessaire d'accroître le nombre de sujets de la base d'apprentissage ou d'ajouter des *a priori* (spécifiques à la pathologie étudiée).

Les méthodes présentées dans cette thèse reposent sur le principe de l'analyse voxel-à-voxel. Elles sont, par conséquent, dépendantes de l'étape de recalage. Malgré des avancées

[5]Notons que dans le cas de la segmentation, (i) la gestion de la multimodalité est spécifique à une pathologie précise et (ii) il n'y a, en général, pas de problème de grandes dimensions.

importantes, le recalage intersujets reste un problème difficile. Nous nous sommes efforcés de prendre en compte ce problème intrinsèque aux approches voxel-à-voxel et de pallier les erreurs de recalage en introduisant, dans la phase d'apprentissage, un terme de régularisation spatiale. Nous avons vu, dans l'approche utilisant la métrique de Fisher, que ce terme peut être considéré comme un indice local de confiance en la méthode de recalage. Il semblerait pertinent **d'adapter** cet indice et plus généralement **le SVM à l'algorithme de recalage utilisé**.

Une autre approche envisageable pour pallier les erreurs de recalage est celle suivie entre autres par Mangin et al. [2004b]. Elle consiste à ne pas travailler directement avec les voxels de l'image mais à utiliser des invariants ou **primitives de plus hauts niveaux**. Une telle analyse pourrait nécessiter une modélisation spécifique de ces primitives. Cette modélisation peut être « non-structurée », par exemple des ensembles de points, ou « structurée » comme les graphes. Les analyses pourrait alors se faire à l'aide d'autres noyaux comme des noyaux pour ensembles [Wolf & Shashua, 2003 ; Kondor & Jebara, 2003 ; Cuturi & Vert, 2005] ou des noyaux pour graphes [Tsuda et al., 2002 ; Gartner et al., 2003 ; Mahé et al., 2004 ; Neuhaus & Bunke, 2006 ; Bach, 2008].

Liste des publications

Publications dans des revues à comité de lecture

- R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias ; S. Lehéricy, M.-O. Habert, M. Chupin, H. Benali, O. Colliot and the Alzheimer's Disease Neuroimaging Initiative, Automatic classification of patients with Alzheimer's disease from structural MRI : A comparison of ten methods using the ADNI database. Neuroimage, 2010, In Press

- M. Chupin, E. Gerardin, R. Cuingnet, C. Boutet, L. Lemieux, S. Lehéricy, H. Benali, L. Garnero, O. Colliot and the Alzheimer's Disease Neuroimaging Initiative, Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data from ADNI. Hippocampus, 2009, 6, 19 : 579-587

- E. Gerardin ; G. Chételat, M. Chupin, R. Cuingnet, B. Desgranges, H.S. Kim, M. Niethammer, B. Dubois, S Lehéricy, L. Garnero, F. Eustache, O. Colliot, and the Alzheimer's Disease Neuroimaging Initiative, Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging. Neuroimage, 2009, 4, 47 : 1476-1486

Articles de conférences à comité de lecture

- R. Cuingnet, M. Chupin, H. Benali and O. Colliot, Spatial and anatomical regularization of SVM for brain image analysis. Proceedings of the Neural Information Processing Systems conference (NIPS 2010), 2010, to appear [poster - Travel Award]

- R. Cuingnet, C. Rosso, S. Lehéricy, D. Dormont, H. Benali, Y. Samson and O. Colliot, Spatially regularized SVM for the detection of brain areas associated with stroke outcome. Medical Image Computing Computer Assisted Intervention (MICCAI). Lecture Notes in Computer Science. 2010 ;13 (Pt 1) :316-23 [oral - Young Scientist Award]

- R. Cuingnet, M. Chupin, H. Benali and O. Colliot, Spatial prior in SVM-based classification of brain images. Proceedings of the SPIE Symposium on Medical Imaging 2010 San Diego, 7624, 76241L (2010) [oral]

- M. Chupin, R. Cuingnet, L. Lemieux, S. Lehéricy, H. Benali, L. Garnero and O. Colliot, Fully Automatic Hippocampus Segmentation Discriminates between Alzheimer's Disease and Normal Aging ? Data from the ADNI database. In Proceedings of the International Workshop on the Computational Anatomy and Physiology of the Hippocampus (CAPH'08, MICCAI'08), 2008 35-45. [oral, speaker M. Chupin]

Résumés présentés lors de conférences

- R. Cuingnet, E. Gérardin, J. Tessieras, G. Auzias, S. Lehéricy, M.O. Habert, M. Chupin, H. Benali, O. Colliot and the Alzheimer's Disease Neuroimaging Initiative, Comparison of ten methods to automatically detect Alzheimer's disease from structural MRI. In Proceedings of the 16th International Conference on Functional Mapping of the Human Brain, 2010, Barcelona, Spain. [poster]

- C. Rosso, G. Auzias, R. Cuingnet, So. Crozier, E. Bardinet, S. Lehéricy, S. Baillet and Y. Samson, Acute and follow-up MCA infarct probability maps in stroke patients with MCA occlusion. Proceedings of the ISMRM 17th Annual Scientific Meeting and Exhibition, Honolulu, Hawaii, 2009 CD-ROM. [poster]

- R. Cuingnet, O. Colliot, B. Magnin, M. Chupin, B. Dubois, S. Lehéricy, H. Benali and the Alzheimer's Disease Neuroimaging Initiative, Detection of prodromal Alzheimer's disease using whole-brain atlas-based classification. Proceedings of the 15th International Conference on Functional Mapping of the Human Brain, 2009, San Francisco, Californie, USA. [poster]

- E. Gerardin, M. Chupin, R. Cuingnet, B. Dubois, S. Lehéricy, L. Garnero, O. Colliot and the Alzheimer's Disease Neuroimaging Initiative, SVM classification of patients with Alzheimer's disease and mild cognitive impairment using hippocampal shape features. Proceedings of the 15th International Conference on Functional Mapping of the Human Brain, 2009, San Francisco, Californie, USA. [poster]

Financements

Ce travail de thèse a pu être effectué à grâce aux financements suivants :

- allocation moniteur polytechnicien (AMX) ;
- allocation nationale de recherche HM-TC (ANR # ANR-09-EMER-006) ;
- « Programme Hospitalier de Recherche Clinique EVAL-USINV' » (n° AOM 03 008).

Annexe A

Une borne d'apprentissage simple

A.1 Énoncé du problème

La proposition 1.2.4 et plus particulièrement l'inégalité (1.5) qui en découle est centrale dans notre approche. Elle permet de comprendre l'approche en apprentissage statistique consistant à minimiser l'erreur empirique régularisée (problème d'optimisation (1.6)). Même si le schéma de démonstration est classique (il suit [Bartlett et al., 2002]), il n'existe pas à ma connaissance de référence directe permettant d'éviter la démonstration.

Récapitulons les hypothèses que nous avons. Les observations $\mathbf{x}$ appartiennent à un ensemble noté $\mathcal{X}$. On se donne un espace de Hilbert $\mathcal{H}$ et une fonction $\phi \in \mathcal{H}^{\mathcal{X}}$. L'ensemble de recherche est $\mathcal{F} = \mathcal{E} \circ \phi$ avec $\mathcal{E}$ un sous ensemble de $\mathcal{H}^*$ l'espace dual de $\mathcal{H}$. Quant à la fonction de perte ℓ, c'est une fonction décroissante de la marge, k-lipschitzienne et on suppose que la perte est bornée de borne c.

On rappelle que l'on note $R_\ell[f]$ le risque d'une fonction de classification et $R_\ell^N[f]$ son risque empirique. L'estimateur ERM est noté $\hat{f}_N$. L'erreur de Bayes est notée R_ℓ^*.

L'objectif est de montrer que, pour tout $\delta > 0$, on a avec probabilité au moins $1 - \delta$:

$$R_\ell\left[\hat{f}_N\right] - R^* \leq \left(\inf_{f \in \mathcal{F}} R_\ell[f] - R_\ell^*\right) + 2k\mathcal{R}_{N,\phi(X)}(\mathcal{E}) + 2c\sqrt{\frac{\ln(1/\delta)}{2N}} \tag{A.1}$$

avec $\mathcal{R}$ la complexité de Rademacher [Bartlett et al., 2002].

A.2 Quelques outils

A.2.1 La complexité de Rademacher

Afin de mesurer la complexité de l'ensemble de recherche, nous utilisons la complexité de Rademacher introduite dans le contexte de l'apprentissage statistique par Bartlett et al. [2002]. L'avantage de cette mesure est qu'elle est relative à la distribution des observations X.

Définition A.2.1 (Complexité de Rademacher) *La complexité de Rademacher $\mathcal{R}_{N,Z}(\mathcal{F})$ de l'ensemble $\mathcal{F}$ par rapport à la distribution de la variable aléatoire Z pour un échantillon de taille N i.i.d., $Z_1, \cdots, Z_N$, est l'espérance de la variable aléatoire $\hat{\mathcal{R}}_{N,Z}(\mathcal{F})$:*

$$\mathcal{R}_{N,Z}(\mathcal{F}) = \mathbf{E}_Z \left[\hat{\mathcal{R}}_{N,Z}(\mathcal{F}) \right] \tag{A.2}$$

avec $\hat{\mathcal{R}}_{N,Z}(\mathcal{F})$ défini comme l'espérance conditionnelle :

$$\hat{\mathcal{R}}_{N,Z}(\mathcal{F}) = \mathbf{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left| \frac{2}{N} \sum_{s=1}^{N} \sigma_s f(Z_s) \right| \middle| Z_1, \cdots, Z_N \right] \tag{A.3}$$

où $\sigma_1, \cdots, \sigma_N$ sont N variables aléatoires indépendantes uniformes à valeurs dans $\{\pm 1\}$.

Par abus de langage, nous noterons parfois $\mathcal{R}_{N,Z}(f)$ à la place de $\mathcal{R}_{N,Z}(\mathcal{F})$.

A.2.2 Inégalités de concentration

Nous avons également besoin d'inégalités de concentration. Une variable aléatoire est dite concentrée si elle s'écarte peu de son espérance. L'inégalité principale est celle de McDiarmid [1989].

Théorème A.2.2 (McDiarmid) *Soient $Z_1, \cdots, Z_N$ des variables aléatoires indépendantes à valeurs dans un ensemble $\mathcal{Z}$. Soit f une fonction de $\mathcal{Z}^N \to \mathbb{R}$ telle que :*

$$\exists (c_i)_i \in \mathbb{R}^N, \forall i, \quad \sup_{\mathbf{z}_1, \cdots, \mathbf{z}_N, \mathbf{z}_i'} |f(\mathbf{z}_1, \cdots, \mathbf{z}_N) - f(\mathbf{z}_1, \cdots, \mathbf{z}_{i-1}, \mathbf{z}_i', \mathbf{z}_{i+1}, \cdots, \mathbf{z}_N)| \leq c_i$$

alors pour tout $\epsilon > 0$:

$$\mathbf{P} \left\{ f(Z_1, \cdots, Z_N) - \mathbf{E} \left[f(Z_1, \cdots, Z_N) \right] \geq \epsilon \right\} \leq e^{\frac{-2\epsilon^2}{\sum_{i=1}^N c_i^2}} \tag{A.4}$$

Corrolaire A.2.3 *Avec les même hypothèses et notations, pour tout $\delta > 0$, on a avec probabilité au moins $1 - \delta$:*

$$f(Z_1, \cdots, Z_N) - \mathbf{E} \left[f(Z1, \cdots, Z_N) \right] \leq \sqrt{\frac{\sum_{i=1}^N c_i^2}{2} \ln \frac{1}{\delta}} \tag{A.5}$$

A.3 Démonstration

Le schéma de démonstration peut se trouver par exemple à la page 96 de [Shawe-Taylor & Cristianini, 2004].

A.3.1 Différence entre le risque et le risque empirique

A.3.1.1 Application de l'inégalité de McDiarmid

On regarde la fonction g définie par :

$$\begin{array}{rccc} g: & (\mathcal{X}\times\mathcal{Y})^N & \to & \mathbb{R} \\ & \big((\mathbf{x}_1,y_1),\cdots,(\mathbf{x}_N,y_N)\big) & \mapsto & \sup\limits_{f\in\mathcal{F}}\left|R[f]-\dfrac{1}{N}\sum\limits_{s=1}^{N}\ell\big(y_sf(\mathbf{x}_s)\big)\right| \end{array}$$

Soient $\big((\mathbf{x}_1,y_1),\cdots,(\mathbf{x}_N,y_N),(\mathbf{x}'_N,y'_N)\big)\in\mathcal{X}^{N+1}$. La fonction ℓ étant à valeurs dans $[0,c]$:

$$\left|R[f]-\frac{1}{N}\sum_{s=1}^{N}\ell\big(y_sf(\mathbf{x}_s)\big)\right|\leq\left|R[f]-\frac{1}{N}\sum_{s=1}^{N-1}\ell\big(y_sf(\mathbf{x}_s)\big)-\frac{\ell\big(y'_Nf(\mathbf{x}'_N)\big)}{N}\right|+\frac{c}{N}$$

On a donc en passant aux suprema :

$$g\big((\mathbf{x}_1,y_1),\cdots,(\mathbf{x}_N,y_N)\big)\leq g\big((\mathbf{x}_1,y_1),\cdots,(\mathbf{x}'_N,y'_N)\big)+\frac{c}{N}$$

L'ordre des arguments de g ne comptant pas, en notant $\mathbf{z}_i=(\mathbf{x}_i,y_i)$, on obtient :

$$\forall i,\ \sup_{\mathbf{z}_1,\cdots,\mathbf{z}_N,\mathbf{z}'_i}|g(\mathbf{z}_1,\cdots,\mathbf{z}_N)-g(\mathbf{z}_1,\cdots,\mathbf{z}_{i-1},\mathbf{z}'_i,\mathbf{z}_{i+1},\cdots,\mathbf{z}_N)|\leq\frac{c}{N}$$

Par conséquent, d'après l'inégalité de McDiarmid (théorème A.2.2), pour tout $\delta>0$, on a avec probabilité au moins $1-\delta$:

$$\sup_{f\in\mathcal{F}}\Big|R[f]-R^N_\ell[f]\Big|\leq\mathbf{E}_X\left[\sup_{f\in\mathcal{F}}\Big(\big|R[f]-R^N_\ell[f]\Big|\right]+\sqrt{\frac{c^2}{2N}\ln\frac{1}{\delta}}\tag{A.6}$$

A.3.1.2 Symétrisation

L'objetif est de borner $\mathbf{E}_X\left[\sup_{f\in\mathcal{F}}\big(R[f]-R^N_\ell[f]\big)\right]$. En utilisant la même démonstration que celle à aux pages 96-97 de [Shawe-Taylor & Cristianini, 2004] (symétrisation à l'aide des variables aléatoires de Rademacher), la valeur absolue ne changeant rien, on obtient l'inégalité suivante :

$$\mathbf{E}_X\left[\sup_{f\in\mathcal{F}}\Big|R[f]-R^N_\ell[f]\Big|\right]\leq\mathcal{R}_{N,(X,Y)}(\ell)$$

La fonction ℓ est k-lipschiztienne donc, d'après le théorème de Ledoux-Talagrand :

$$\mathcal{R}_{N,(X,Y)}(\ell) \leq 2k\mathbf{E}_{(X,Y)}\left[\mathbf{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{N}\sum_{s=1}^{N}\sigma_s Y_s f(X_s)\right| \middle| (X_1,Y_1),\cdots,(X_N,Y_N)\right]\right]$$

Or pour $(Y_s)_s \in \{\pm 1\}^N$ fixé, $(\sigma_s)_s$ étant indépendants et identiquement distribués de distribution **uniforme** (donc symétrique) sur $\{\pm 1\}$, $\left(\sigma'_s\right)_s$ avec pour tout s, $\sigma'_s = Y_s\sigma_s$, sont également indépendants et identiquement distribuées de distribution uniforme sur $\{\pm 1\}$. Par conséquent, on a :

$$\begin{aligned}
\mathcal{R}_{N,(X,Y)}(\ell) &\leq 2k\mathbf{E}_{(X,Y)}\left[\mathbf{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{N}\sum_{s=1}^{N}\sigma_s f(X_s)\right| \middle| (X_1,Y_1),\cdots,(X_N,Y_N)\right]\right] \\
&\leq 2k\mathbf{E}_X\left[\mathbf{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{N}\sum_{s=1}^{N}\sigma_s f(X_s)\right| \middle| X_1,\cdots,X_N\right]\right] \\
&\leq 2k\mathcal{R}_{N,X}(\mathcal{F}) \\
\mathcal{R}_{N,(X,Y)}(\ell) &\leq 2k\mathcal{R}_{N,\phi(X)}(\mathcal{E}) \qquad \text{(A.7)}
\end{aligned}$$

En injectant l'inégalité (A.7) dans (A.6), on a pour tout $\delta > 0$, avec probabilité au moins $1 - \frac{\delta}{2}$:

$$\sup_{f\in\mathcal{F}}\left|R[f] - R_\ell^N[f]\right| \leq 2k\mathcal{R}_{N,\phi(X)}(\mathcal{E}) + \sqrt{\frac{c^2}{2N}\ln\frac{1}{\delta}} \qquad \text{(A.8)}$$

A.3.2 Difference entre le risque et l'erreur de Bayes

D'après l'inégalité (A.8), on a pour tout $\delta > 0$ avec probabilité au moins $1 - \delta$

$$R_\ell\left[\hat{f}_N\right] \leq R_\ell^N\left[\hat{f}_N\right] + 2k\mathcal{R}_{N,\phi(X)}(\mathcal{E}) + \sqrt{\frac{c^2}{2N}\ln\frac{1}{\delta}}$$

Or par définition de $\hat{f}_N$, on a pour tout $f \in \mathcal{F}$: $R_\ell^N\left[\hat{f}_N\right] \leq R_\ell^N[f]$ d'où :

$$R_\ell\left[\hat{f}_N\right] \leq R_\ell^N[f] + 2k\mathcal{R}_{N,\phi(X)}(\mathcal{E}) + \sqrt{\frac{c^2}{2N}\ln\frac{1}{\delta}}$$

On a également d'après l'inégalité (A.8),

$$R_\ell^N[f] \leq R_\ell[f] + 2k\mathcal{R}_{N,\phi(X)}(\mathcal{E}) + \sqrt{\frac{c^2}{2N}\ln\frac{1}{\delta}}$$

On déduit des deux inégalités précédentes que pout tout $\delta > 0$, on a avec probabilité au moins $1 - \delta$:

$$R_\ell\left[\hat{f}_N\right] \leq R_\ell[f] + 2k\mathcal{R}_{N,\phi(X)}(\mathcal{E}) + c\sqrt{\frac{2}{N}\ln\frac{1}{\delta}}$$

En passant à la limite et en soustrayant de part et d'autre de l'inégalité le risque de Bayes R^*, on a ce que l'on cherchait. ■

Annexe B

Retour sur certaines approximations

B.1 Approximation par un lissage gaussien

Dans la section 3.3, nous avons approximé la solution de l'équation de la chaleur sur un parallélépipède rectangle de $\mathbb{R}^3$ par celle sur l'espace tout entier. Regardons si cette approximation est justifiée. Pour les solutions de l'équation de la chaleur avec différentes conditions aux bords, le lecteur pourra se référer à [Myint-U & Debnath, 2007].

B.1.1 Sur un intervalle

B.1.1.1 Différence entre les deux solutions

Regardons tout d'abord le comportement sur un intervalle de $\mathbb{R}$. Soit L un réel positif. On regarde l'équation de la chaleur sur $[0, L]$ suivante d'inconnue $y \in \mathbb{R}^{[0,L]\times\mathbb{R}^+}$:

$$\begin{cases} \frac{\partial}{\partial t} y - \Delta y = 0 \\ \forall x \in [0, L],\ y(x, 0) = f(x) \\ y(0, t) = 0 \\ y(L, t) = 0 \end{cases} \tag{B.1}$$

La solution de cette équation aux dérivées partielles est :

$$\forall (x, t) \in [0, L] \times \mathbb{R}^+,\ y(x, t) = \sum_{n=1}^{\infty} A_n \sin\left(\frac{n\pi}{L}\right) e^{-\left(\frac{n\pi}{L}\right)^2 t}$$

Les coefficients de la série de Fourier, A_n, sont donnés par :

$$\forall n \in \mathbb{N}^*,\ A_n = \frac{2}{L} \int_0^L \sin\left(\frac{n\pi}{L} x\right) f(x) dx$$

Nous avons approximé cette solution par la fonction g, définie par :

$$\forall (x,t) \in [0,L] \times \mathbb{R}^+,\ g(x,t) = f * \frac{1}{\sqrt{4\pi t}} e^{-\frac{\bullet^2}{4t}}(x)$$

La fonction g est solution de l'équation de la chaleur sur $\mathbb{R}$. Dans la suite nous chercherons à quantifier la différence entre g et u.

Pour cela, considérons l'équation aux dérivées partielles d'inconnue $u \in \mathbb{R}^{[0,L]\times\mathbb{R}^+}$ suivante.

$$\begin{cases} \frac{\partial}{\partial t}u - \Delta u = 0 \\ \forall x \in [0,L],\ u(x,0) = f(x) \\ u(0,t) = g(0,t) \\ u(L,t) = g(L,t) \end{cases} \tag{B.2}$$

Il va de soi que la restriction de g à l'intervalle $[0,L]$ est solution de l'équation aux dérivées partielles (B.2). Nous allons dans ce qui suit résoudre cette équation pour avoir une relation entre g et y. Pour cela, nous introduisons une nouvelle fonction $\tilde{u}$ définie par :

$$\forall x \in [0,L],\ \tilde{u} = \frac{1}{L}\left[(L-x)g(0,t) + xg(L,t)\right] \tag{B.3}$$

Posons alors :

$$v = u - \tilde{u} \tag{B.4}$$

La nouvelle fonction v est alors solution de :

$$\begin{cases} \frac{\partial}{\partial t}v - \Delta v = -\frac{1}{L}\left[(L-x)\frac{\partial}{\partial t}g(0,t) + x\frac{\partial}{\partial t}g(L,t)\right] \\ \forall x \in [0,L],\ v(x,0) = f(x) - \frac{1}{L}\left[(L-x)f(0) + xgf(L)\right] \\ v(0,t) = 0 \\ v(L,t) = 0 \end{cases} \tag{B.5}$$

Hypothèse : dans la suite, nous faisons l'hypothèse que f est nulle aux bords. Le sytème (B.5) se réécrit donc :

$$\begin{cases} \frac{\partial}{\partial t}v - \Delta v = -\frac{1}{L}\left[(L-x)\frac{\partial}{\partial t}g(0,t) + x\frac{\partial}{\partial t}g(L,t)\right] \\ \forall x \in [0,L],\ v(x,0) = f(x) \\ v(0,t) = 0 \\ v(L,t) = 0 \end{cases} \tag{B.6}$$

La solution de cette équation est la somme du problème homogène avec condition initiale f (i.e. equation (B.1)) et du problème inhomogène avec condition initiale nulle. On a donc, pour tout $(x,t) \in [0,L] \times \mathbb{R}^+$:

$$v(x,t) = y(x,t) + \int_0^t \sum_{n=1}^{\infty} B_n(s) \sin\left(\frac{n\pi}{L}x\right) e^{-\left(\frac{n\pi}{L}\right)^2 (t-s)} ds$$

avec :

$$B_n(s) = -\frac{2}{L}\int_0^L \sin\left(\frac{n\pi}{L}x\right)\frac{1}{L}\left[(L-x)\frac{\partial}{\partial t}g(0,s) + x\frac{\partial}{\partial t}g(L,s)\right]dx$$

Ce qui nous donne la relation suivante entre y et g (rappel : g=u), d'après les équations (B.3) et (B.4), pour tout $(x,t) \in [0,L]\times\mathbb{R}^+$:

$$\begin{aligned} g(x,t) - y(x,t) &= \int_0^t \sum_{n=1}^{\infty} B_n(s)\sin\left(\frac{n\pi}{L}x\right)e^{-\left(\frac{n\pi}{L}\right)^2(t-s)}ds \\ &\quad + \frac{1}{L}\left[(L-x)g(0,t) + xg(L,t)\right] \end{aligned} \tag{B.7}$$

B.1.1.2 Borne sur la différence entre les deux solutions

Nous allons maintenant utiliser l'équation (B.7) pour borner la différence entre la solution exacte, y, et notre approximation. Posons, pour alléger les notations :

$$m_1(t) = \max\left(|g(0,t)|, |g(L,t)|\right)$$

et

$$m_2(t) = \max\left(\left|\frac{\partial}{\partial t}g(0,t)\right|, \left|\frac{\partial}{\partial t}g(0,t)\right|\right)$$

On a donc :

$$\left|\frac{1}{L}\left[(L-x)g(0,t) + xg(L,t)\right]\right| \leq m_1(t) \tag{B.8}$$

Regardons maintenant l'autre terme que nous appellerons D. L'inégalité triangulaire donne :

$$\begin{aligned} D &= \left|\int_0^t \sum_{n=1}^{\infty} B_n(s)\sin\left(\frac{n\pi}{L}x\right)e^{-\left(\frac{n\pi}{L}\right)^2(t-s)}ds\right| \\ &\leq \int_0^t \left|\sum_{n=1}^{\infty} B_n(s)\sin\left(\frac{n\pi}{L}x\right)e^{-\left(\frac{n\pi}{L}\right)^2(t-s)}ds\right| \end{aligned}$$

En utilisant l'inégalité de Cauchy-Schwartz on obtient :

$$D \leq \int_0^t \left(\sum_{n=1}^{\infty} B_n^2(s)\right)^{\frac{1}{2}}\left(\sum_{n=1}^{\infty} e^{-2\left(\frac{n\pi}{L}\right)^2(t-s)}ds\right)^{\frac{1}{2}}$$

Ce qui donne, en utilisant le fait que les termes $e^{-\left(\frac{n\pi}{L}\right)^2(t-s)}$ sont positifs :

$$D \leq \int_0^t \left(\sum_{n=1}^{\infty} B_n^2(s)\right)^{\frac{1}{2}}\sum_{n=1}^{\infty} e^{-\left(\frac{n\pi}{L}\right)^2(t-s)}ds \tag{B.9}$$

D'après l'égalité de Parseval, on a :

$$\sum_{n=1}^{\infty} B_n^2(s) \leq \frac{1}{L}\int_0^L \left|\frac{1}{L}\left[(L-x)\frac{\partial}{\partial t}g(0,s) + x\frac{\partial}{\partial t}g(L,s)\right]\right|^2 \tag{B.10}$$

$$\leq m_2(s)^2 \tag{B.11}$$

En injectant l'inégalité (B.10) dans (B.9), on obtient :

$$D \leq \max_{0\leq s\leq t}(m_2(s)) \int_0^t \sum_{n=1}^{\infty} e^{-\left(\frac{n\pi}{L}\right)^2 (t-s)} ds$$

Avec le théorème de convergence dominée de Lebesgue, l'inégalité précédente devient :

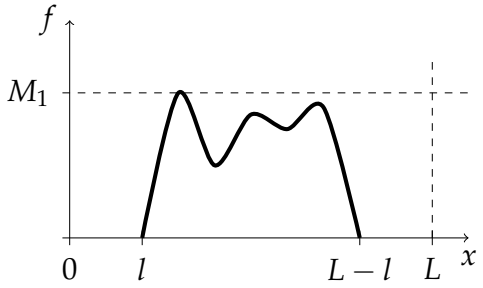
$$\begin{aligned} D &\leq \max_{0\leq s\leq t}(m_2(s)) \sum_{n=1}^{\infty} \int_0^t e^{-\left(\frac{n\pi}{L}\right)^2 (t-s)} ds \\ &\leq \max_{0\leq s\leq t}(m_2(s)) \sum_{n=1}^{\infty} \left(\frac{L}{n\pi}\right)^2 \left[1 - e^{-\left(\frac{n\pi}{L}\right)^2 t}\right] \\ &\leq \max_{0\leq s\leq t}(m_2(s)) \left(\frac{L}{\pi}\right)^2 \zeta(2) \\ &\leq \max_{0\leq s\leq t} m_2(s) \frac{L^2}{6} \end{aligned} \tag{B.12}$$

On a donc, à l'aide des inégalités (B.8) et (B.12) :

$$|g - y| \leq m_1(t) + \frac{L^2}{6} \max_{0\leq s\leq t} m_2(s) \tag{B.13}$$

Il reste maintenant à touver une borne pour m_1 et m_2.

Hypothèse : Le support de f est $[l, L-l]$.



Soient M_1 le maximum de $|f|$ et M_2 de $|\Delta f|$. Avec ces notations, on a :

$$m_1(t) \leq M_1 \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{-l}{2\sqrt{t}}\right)\right]$$

et

$$m_2(t) \leq M_2 \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{-l}{2\sqrt{t}}\right)\right]$$

En conclusion, on a, d'après (B.13) la majoration suivante de l'erreur d'approximation :

$$|g - y| \leq \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{-l}{2\sqrt{t}}\right)\right]\left[M_1 + M_2\frac{L^2}{6}\right] \tag{B.14}$$

Donc si l est suffisamment grand devant $2\sqrt{t}$ (*i.e.* taille du noyau de lissage équivalent), l'approximation est correcte.

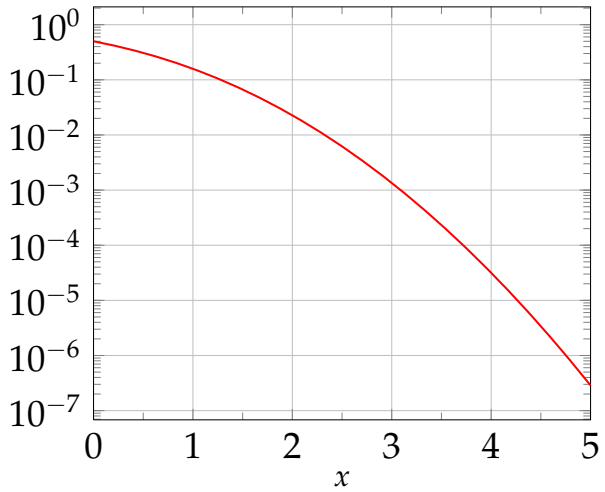


FIGURE B.1 : Graphe de la fonction : $x \mapsto \frac{1}{2}\left[1 + \mathrm{erf}(-\frac{x}{\sqrt{2}})\right]$

Remarque : Pour les plus grandes dimensions, on utilise la séparabilité de la solution suivant les dimensions de l'espace.

Annexe C

Liste des sujets de la base ADNI utilisés dans nos études

CN : TRAINING SET				CN : TRAINING SET			
CENTER ID	SUBJECT ID	SID	UID	CENTER ID	SUBJECT ID	SID	UID
2	295	13408	45109	36	1023	24338	65105
2	1261	26574	62378	41	1002	23705	65221
2	1280	26453	60057	57	643	15782	34726
3	907	19728	52782	57*	779*	18109*	80630*
3	981	20753	52777	57	934	19971	34735
5	602	15966	32674	67	19	9539	45229
6	484	18837	65567	73	312	15079	39888
6	498	15857	67058	73	386	13746	49681
6	681	18451	92306	82	363	13017	39731
7	68	10356	35791	82	640	16562	49480
7	1222	25402	60004	94	489	14121	47292
11	2	9107	35470	94	711	16553	39139
11	5	9136	32247	98	171	11460	65753
11	16	9253	32307	98	896	19439	56032
11	21	9581	32342	99	533	14938	38786
11	22	9617	32394	109	1014	24818	63480
11	23	8868	32410	114	173	11594	96322
13	502	17232	51139	114	601	15201	39851
13	575	17859	51161	126	405	14635	38828
13	1035	21984	51166	126	506	15652	38855
14	519	14488	39648	126	605	15665	38705
14	520	14473	39661	126	680	16099	38927
14	558	15789	39684	127*	260*	12419*	34385*
20	883	19459	60674	127	622	15473	34453
20	899	19329	60628	128	863	19272	98876
20	1288	26890	60665	130	232	11928	39110
22	14	9271	59376	130	969	20385	39277
22	130	11089	59525	130	1200	25090	63753
23	61	10312	31103	131	123	10960	63785
23	963	19740	89418	131	319	12322	47985
24	985	21112	62716	131	441	13681	48030
24	1063	21525	63393	133	433	13952	39947
27*	118*	11796*	34115*	133	488	14150	107935
27	403	13717	34182	136	86	13191	66414
29	824	18139	96285	136	196	13253	40261
29	843	18909	66998	941	1194	25323	63848
33	920	19288	42482	941	1195	26180	63866
33	1098	22792	42833	941	1197	25332	66463
35	48	10257	45184	941	1202	25680	63875
36	672	17131	36939	941	1203	25671	63880
36	813	18252	36980				

* : images for which the cortical thickness extraction pipeline failed. † : images for which the SPHARM computation pipeline failed.

Table C.1 : Liste des sujets CN de l'ensemble d'apprentissage de la base ADNI inclus dans nos études. Ce tableau liste les identifiants des sujets, de leur image T1 (UID) ainsi que celui de leur série.

CN : TESTING SET				CN : TESTING SET			
CENTER ID	SUBJECT ID	SID	UID	CENTER ID	SUBJECT ID	SID	UID
2	413	13893	45118	52	1250	25829	62241
2	559	14875	40675	52	1251	26231	62283
2	685	16309	40684	62	578	15035	50460
3	931	20051	53391	62	768	17527	50507
3	1021	21771	73507	62	1099	22713	50558
5	553	15527	32646	67	56	8723	35894
5	610	15727	32669	67	59	10517	35903
6	731	18321	90849	67	177	12187	34807
7	70	10950	36629	67	257	14730	34825
7	1206	25173	59982	73	89	11161	49676
11	8	9195	32265	73	311	15069	39869
13	1276	27641	62681	82	304	12557	95654
14	548	14921	39666	82	761	18119	39792
16	359	13000	96222	82	1256	26812	63157
16	538	17545	40773	94	526	14559	63469
20	97	10858	64047	94	692	17207	47296
22	66	10271	59447	94	1267	28217	80719
22	96	11006	59457	98	172	11812	65758
23	31	9785	69613	99	40	8920	34608
23	58	10335	30551	99	90	10835	35842
23	81	10813	31126	99	352	12992	34538
23	926	19390	31548	99	534	14009	34579
23	1190	24847	46418	109	876	19132	82595
23	1306	26604	46436	109	967	21137	55045
27	74	10605	34317	109	1013	22585	66150
27	120	11535	34333	114	166	11584	39811
29	845	18829	64868	114	416	13556	39837
29	866	18917	65612	127	259	12137	34363
33	516	14818	42309	127	684	16759	34459
33	734	16942	42435	130	886	19561	39173
33	741	17006	42451	131	436	14710	48021
33	923	19544	42510	131	1301	26899	63794
33	1016	21817	42773	133	493	14156	107944
33	1086	23547	42786	133	525	14991	39981
35	156	11391	39534	136	184	11974	40180
35	555	15332	39602	136	186	11774	40202
36	576	15156	36904	141	717	18413	98889
41	125	10883	35672	141	767	18337	47308
41	898	20306	34699	141	810	20274	47315
51	1123	24099	58044	141	1094	23294	47733
52	951	20352	64171				

* : images for which the cortical thickness extraction pipeline failed. † : images for which the SPHARM computation pipeline failed.

TABLE C.2 : Liste des sujets CN de l'ensemble de test de la base ADNI inclus dans nos études.

AD : TRAINING SET				AD : TRAINING SET			
CENTER ID	SUBJECT ID	SID	UID	CENTER ID	SUBJECT ID	SID	UID
2	816	18402	40732	67	812	19629	38727
2	938	19852	40981	67	828	18532	65717
2	955	20004	40755	67	1185	24635	63105
3	1059	22300	52817	67	1253	27558	55034
3	1257	27340	52791	73	565	15762	39920
5	814	18390	74592	82	1377	28495	63171
7	1248	25568	59951	94	1102	22905	63187
7	1304	26475	59911	94	1164	23871	67224
11	3	9127	32238	94	1397	31011	95663
11	10	8800	32275	94	1402	32102	66079
13	592	18419	79145	98	884	24183	56027
13	699	18366	62651	99	470	14222	34571
13	1161	24399	51486	99	492	14944	38771
14	356	12857	39630	99	1144	24218	102041
14	1095	23323	45741	109	777	18676	82577
22	129	11485	59485	114	374	13031	39818
23(†)	78(†)	10619(†)	52000(†)	114	979	21933	39860
23	84	10764	31207	126	606	17191	38910
23	139	11079	31304	126	1221	25457	48977
27	1081	25357	47169	127	844	19874	34472
29	836	18151	65014	128	1409	33787	69401
29	1184	25463	67211	128*	1430*	39199*	79858*
33	889	19296	51630	130	1290	26038	63767
35	341	12952	45217	130	1337	27584	63776
36	577	14974	36915	131	457	13976	92407
36	760	18264	38653	131	497	15315	48039
57	474	13990	34721	131	691	17266	48048
57	1371	28667	62999	133	1170	24674	89958
57	1373	28698	63009	136	299	13839	40313
57(†)	1379(†)	28761(†)	63015(†)	136	300	14136	40329
62	535	14699	50427	136	426	14581	40357
62	690	16924	50469	141	1024	22699	47749
62	730	17062	50488	141	1137	24301	48582
67	29	9904	38718	141	1152	24487	48591
67	76	10468	35912				

* : images for which the cortical thickness extraction pipeline failed. † : images for which the SPHARM computation pipeline failed.

Table C.3 : Liste des sujets AD de l'ensemble d'apprentissage de la base ADNI inclus dans nos études.

AD : TESTING SET				AD : TESTING SET			
CENTER ID	SUBJECT ID	SID	UID	CENTER ID	SUBJECT ID	SID	UID
2	1018	23128	40818	33	1281	26136	54781
5	221	11958	72129	33	1283	26144	54786
5	929	19669	74610	33	1285	26128	51589
5	1341	27673	60418	33	1308	26600	54753
6	547	16033	67316	36	759	18094	36970
6	653	16073	67325	36	1001	22691	38662
7	316	12583	36574	41	1368	27512	65249
7	1339	27414	56320	41	1391	29116	62934
11	53	10064	35487	41	1435	39186	79637
11	183	12000	32004	51	1296	26431	58024
13	996	22240	51184	53	1044	21256	64204
13	1205	25024	51543	62	793	18189	50525
14	328	12402	39621	67	110	11177	35934
16	991	21737	40795	82	1079	22650	49491
16	1263	27303	64623	94	1027	21207	49528
20	213	12386	60601	94	1090	23375	63177
22	7	9024	59367	98	149	11021	89430
22	219	12375	59535	99	372	13672	34550
22	543	14849	59544	109	1157	24711	66159
23	83	10568	31144	109	1192	25056	63504
23	93	10736	31254	114	228	11697	49736
23	916	19228	31534	126	784	19752	39013
23	1262	26314	62434	126(†)	891(†)	19386(†)	39055(†)
23	1289	26374	89939	127	431	14595	34444
24	1171	24659	63407	127	754	18515	80761
24	1307	27061	63416	127	1382	28261	66311
27	404	13866	34205	130	956	20667	39187
27	850	18554	48997	130	1201	25082	63758
27	1254	25764	47229	133	1055	22386	40029
27	1385	28133	47575	136	194	13178	40240
29	999	23248	64899	141	696	18373	82739
29	1056	22977	60742	141	790	18766	91254
33	724	17337	42401	141	852	19395	47745
33	733	16932	42426	141	853	18348	112293

* : images for which the cortical thickness extraction pipeline failed. † : images for which the SPHARM computation pipeline failed.

TABLE C.4 : Liste des sujets AD de l'ensemble de test de la base ADNI inclus dans nos études.

MCI_c : TRAINING SET				MCI_c : TRAINING SET			
CENTER ID	SUBJECT ID	SID	UID	CENTER ID	SUBJECT ID	SID	UID
2	954	19979	40745	33	906	19314	42469
2	1070	23120	40832	33	922	19341	42494
5	222	11754	54689	35	204	11661	39543
7	41	9994	35735	35	997	23184	62909
7	128	10936	36641	51	1331	29664	64153
7	344	12631	36580	52	952	20364	89953
11	856	19031	89409	52	1054	22955	62235
13	240	12308	51152	53	507	14483	80200
13	860	19237	51534	62	1299	26794	50585
22	750	17695	59553	67	243	12030	34820
22	1394	34317	68083	67	336	14023	34858
23	42	8852	31085	94	1015	21457	40764
23	388	13076	31438	94	1398	31771	63228
23	604	15182	31456	127	394	14603	34399
23	855	18561	31510	133	638	16608	67532
23	887	19087	31527	136	195	12523	40453
23	1247	25741	48858	141	982	22644	47704
27	461	15192	34232	941	1311	27408	97328
33	723	16845	42385	941	1363	28008	63898
33	725	17092	42410				

* : images for which the cortical thickness extraction pipeline failed. † : images for which the SPHARM computation pipeline failed.

Table C.5 : Liste des sujets MCI_c de l'ensemble d'apprentissage de la base ADNI inclus dans nos études.

MCI_c : TESTING SET				MCI_c : TESTING SET			
CENTER ID	SUBJECT ID	SID	UID	CENTER ID	SUBJECT ID	SID	UID
2	729	16874	40709	57	941	19985	34748
5	572	15709	32659	57	1217	25854	62985
6	1130	23457	55973	67	45	10185	35889
7	249	11911	36531	67	77	11136	68121
11	241	12088	32021	94	434	13570	39124
11	861	19476	35514	98	269	11964	65258
11	1282	26225	62637	99	54	10329	35826
13	325	13524	54666	99	111	10933	35850
14	658	17481	39702	126	1077	23496	48881
23	30	9441	31632	127	1427	37933	91127
23	625	15820	31496	128	947	19859	69074
27	179	11781	34137	130	423	15030	39134
27	256	12357	34151	133	727	18620	66358
27	1213	25492	47224	133(†)	913(†)	24646(†)	63820(†)
27	1387	28123	67202	136	695	19019	70925
33	567	14572	42371	141	915	20504	48573
41	549	15488	39507	141	1244	26845	92647
41	1412	34022	72220	941	1295	26290	63889
41	1423	36902	72233				

* : images for which the cortical thickness extraction pipeline failed. † : images for which the SPHARM computation pipeline failed.

TABLE C.6 : Liste des sujets MCI_c de l'ensemble de test de la base ADNI inclus dans nos études.

MCI_{nc} : TRAINING SET				MCI_{nc} : TRAINING SET			
CENTER ID	SUBJECT ID	SID	UID	CENTER ID	SUBJECT ID	SID	UID
3	1122	23542	52800	41	314	12492	34681
5	324	12599	32893	41	679	17077	40047
7	414	14826	36600	41	1010	23880	65231
11	326	12342	89391	41	1260	25806	65240
11	362	12678	89405	51	1072	22884	58012
11	1080	23159	35592	52	671	16062	64162
16	702	17341	40782	52	989	22476	64180
16	1028	22058	40800	52	1168	23688	65668
16	1138	24779	86046	53	621	15442	64190
22	1097	23337	59611	53	919	20422	65690
22	1351	28484	59616	57	464	14736	34708
23	126	11525	31272	57	1007	21339	47766
23	217	11731	31347	73	909	19716	75486
23	1046	22199	46397	94	1330	27038	63207
27	408	14231	37549	98	160	11224	65740
27	644	15630	34241	99	60	10478	35834
27	835	18760	35667	99	291	12065	34525
29	878	18986	64877	109	950	21165	97201
29	1038	22851	60733	114	378	12760	95689
29	1073	22828	65023	126	865	22295	39034
29	1215	25348	60747	126	1187	25143	48960
29	1218	25478	67390	127	112	11194	98858
33	511	15101	42241	127	1140	24278	63638
33	513	14673	42259	130	285	12462	39119
33	514	14663	42277	130	289	12111	39114
33	1279	26694	54758	130	449	22250	80931
33	1284	26686	54800	130	783	18023	39154
35	33	10396	45167	131	384	12790	48003
35	292	12408	39570	133	771	18575	92287
36	656	16286	36925	133	912	19884	40001
36	673	17157	36950	136	1227	26399	63839
36	748	17705	36960	141	697	18466	91236
36	869	21421	36994	141	851	19364	47871
36	976	23852	65092				

* : images for which the cortical thickness extraction pipeline failed. † : images for which the SPHARM computation pipeline failed.

Table C.7 : Liste des sujets MCI_c de l'ensemble d'apprentissage de la base ADNI inclus dans nos études.

MCI_{nc} : TESTING SET				MCI_{nc} : TESTING SET			
CENTER ID	SUBJECT ID	SID	UID	CENTER ID	SUBJECT ID	SID	UID
2	782	17835	40718	62	1182	25166	50567
2	1155	24144	40846	73	746	23286	63123
3	908	32516	62590	82	1119	23733	63148
3	1074	23536	53396	94	531	14554	49667
5	448	14032	32877	94	921	19582	49511
5	546	15566	32683	94	1293	27865	64375
5	1224	25412	60407	94	1314	27488	63197
7	101	10679	36727	98	667	15980	89496
7	293	12193	36550	99	51	10325	35820
7	698	16403	36614	99	1034	21759	47954
13	1186	25689	62657	109	1114	24702	63490
14	169	11565	40859	109	1183	24718	66168
14	557	15094	39675	114	410	13289	39825
14	563	16079	39693	114	458	14111	39846
16	769	17720	48868	114	1103	23239	49758
22	4	9234	64632	114	1106	22859	49911
22	544	14679	64673	114	1118	23803	49769
22	961	20711	59602	126	708	16897	38945
23	331	12428	31374	126	709	17326	38965
23	376	12652	31385	127	393	13200	34394
27	116	11442	34326	127	925	21560	67267
27	307	13072	34160	127	1032	22259	63633
27	485	14208	37558	128	1043	22562	69092
27	1045	22174	47211	130	102	10746	39461
33	1116	22799	42845	130	505	17292	39144
33	1309	26195	51606	133	629	15915	40955
36	945	20971	37000	133	792	18306	66374
36	1135	24501	65110	133	1031	21552	40020
36	1240	26423	65119	136	107	11707	40446
41	282	13496	39487	136	429	15534	40388
41	598	15604	40038	136	579	15952	40411
51	1131	24089	62944	136	874	22234	40420
53	389	13550	65677	141	1052	22923	47723
57	839	19188	38671				

* : images for which the cortical thickness extraction pipeline failed. † : images for which the SPHARM computation pipeline failed.

TABLE C.8 : Liste des sujets MCI_c de l'ensemble de test de la base ADNI inclus dans nos études.

Références

Allaire, G. (2005). *Analyse numérique et optimisation.* Éd. de l'École Polytechnique.

Amari, S.-I., Barndorff-Nielsen, O. E., Kass, R. E., Lauritzen, S. L., & Rao, C. R. (1987). *Differential Geometry in Statistical Inference*, volume 10. Institute of Mathematical Statistics.

Amieva, H., Andrieu, S., Berr, C., Buée, L., Checler, F., Clément, S., Dartigues, J.-F., Desgranges, B., Dubois, B., Duyckaerts, C., Joel, M.-E., Lambert, J.-C., Nourhashemi, F., Pasquier, F., Robert, P., Blanchard, F., Bloch, M.-A., Ganem-Chabenet, D., Lacomblez, L., & Saint-Jean, O. (2007). *Maladie d'Alzheimer : enjeux scientifiques, médicaux et sociétaux.* INSERM.

Anbeek, P., Vincken, K. L., van Osch, M. J. P., Bisschops, R. H. C., & van der Grond, J. (2004). Probabilistic segmentation of white matter lesions in MR imaging. *NeuroImage*, 21(3), 1037 – 1044.

Andrade, A., Kherif, F., Mangin, J., Worsley, K., Paradis, A., Simon, O., Dehaene, S., Le Bihan, D., & Poline, J. (2001). Detection of fMRI activation using cortical surface mapping. *Human Brain Mapping*, 12(2), 79–93.

Aronszajn, N. (1950). Theory of Reproducing Kernels. *Transactions of the American Mathematical Society*, 68(3), 337–404.

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *NeuroImage*, 38(1), 95–113.

Ashburner, J. & Friston, K. (2005). Unified segmentation. *NeuroImage*, 26(3), 839–851.

Ashburner, J. & Friston, K. J. (2000). Voxel-based morphometry–the methods. *NeuroImage*, 11(6), 805–21.

Auzias, G., Glaunès, J., Colliot, O., Perrot, M., Mangin, J.-F., Trouvé, A., & Baillet, S. (2009). DISCO: A coherent diffeomorphic framework for brain registration under exhaustive sulcal constraints. In G.-Z. Yang, D. Hawkes, D. Rueckert, A. Noble, & C. Taylor (Eds.), *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2009*, volume 5761 of *Lecture Notes in Computer Science* (pp. 730–738). Springer Berlin / Heidelberg.

Bach, F. (2008). Graph kernels between point clouds. In *Proceedings of the 25th international conference on Machine learning* (pp. 25–32).: ACM.

Bach, F., Lanckriet, G., & Jordan, M. I. (2004). Multiple kernel learning, conic duality, and the SMO algorithm. In *Proceedings of the 21st International Conference on Machine Learning* (pp. 41–48).

Bakkour, A., Morris, J. C., & Dickerson, B. C. (2009). The cortical signature of prodromal AD: regional thinning predicts mild AD dementia. *Neurology*, 72(12), 1048–1055.

Barber, R., Scheltens, P., Gholkar, A., Ballard, C., McKeith, I., Ince, P., Perry, R., & O'Brien, J. (1999). White matter lesions on magnetic resonance imaging in dementia with Lewy bodies, Alzheimer's disease, vascular dementia, and normal aging. *Journal of Neurology, Neurosurgery & Psychiatry*, 67(1), 66.

Bartlett, P. L., Mendelson, S., & Long, M. (2002). Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3, 2002.

Benedek, G. M. & Itai, A. (1994). Nonuniform learnability. *Journal of Computer and System Sciences*, 48(2), 311 – 323.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1), 289–300.

Benjamini, Y. & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, (pp. 1165–1188).

Bergouignan, L., Chupin, M., Czechowska, Y., Kinkingnéhun, S., Lemogne, C., Le Bastard, G., Lepage, M., Garnero, L., Colliot, O., & Fossati, P. (2009). Can voxel based morphometry, manual segmentation and automated segmentation equally detect hippocampal volume differences in acute depression? *NeuroImage*, 45(1), 29–37.

Blennow, K., de Leon, M. J., & Zetterberg, H. (2006). Alzheimer's disease. *Lancet*, 368(9533), 387–403.

Bobinski, M., De Leon, M. J., Wegiel, J., Desanti, S., Convit, A., Saint Louis, L. A., Rusinek, H., & Wisniewski, H. M. (1999). The histological validation of post mortem magnetic resonance imaging-determined hippocampal volume in alzheimer's disease. *Neuroscience*, 95(3), 721–725.

Bousquet, O., Boucheron, S., & Lugosi, G. (2004). Introduction to statistical learning theory. *Advanced Lectures on Machine Learning*, (pp. 169–207).

Braak, H. & Braak, E. (1991). Neuropathological stageing of Alzheimer-related changes. *Acta neuropathologica*, 82(4), 239–259.

Brooks, D. (2010). Imaging approaches to Parkinson disease. *Journal of Nuclear Medicine*, 51(4), 596.

Brown, M., Grundy, W., Lin, D., Cristianini, N., Sugnet, C., Furey, T., Ares, M., & Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences of the United States of America*, 97(1), 262.

Burges, C. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2), 121–167.

Busatto, G. F., Garrido, G. E., Almeida, O. P., Castro, C. C., Camargo, C. H., Cid, C. G., A, C., Buchpiguel, Furuie, S., & Bottino, C. M. (2003). A voxel-based morphometry study of temporal lobe gray matter reductions in Alzheimer's disease. *Neurobiol. Aging*, 24(2), 221–231.

Caan, M. W. A., Vermeer, K. A., van Vliet, L. J., Majoie, C. B. L. M., Peters, B., den Heeten, G., & Vos, F. (2006). Shaving diffusion tensor images in discriminant analysis: A study into schizophrenia. *Medical Image Analysis*, 10(6), 841–849.

Cachia, A., Mangin, J., Riviere, D., Kherif, F., Boddaert, N., Andrade, A., Papadopoulos-Orfanos, D., Poline, J., Bloch, I., Zilbovicius, M., et al. (2003). A primal sketch of the cortex mean curvature: a morphogenesis based approach to study the variability of the folding patterns. *IEEE Transactions on Medical Imaging*, 22(6), 754–65.

Chalela, J. A., Kidwell, C. S., Nentwich, L. M., Luby, M., Butman, J. A., Demchuk, A. M., Hill, M. D., Patronas, N., Latour, L., & Warach, S. (2007). Magnetic resonance imaging and computed tomography in emergency assessment of patients with suspected acute stroke: a prospective comparison. *The Lancet*, 369(9558), 293 – 298.

Chang, C.-C. & Lin, C.-J. (2001). *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Chapelle, O. & Schölkopf, B. (2002). Incorporating invariances in non-linear support vector machines. In *Advances in Neural Information Processing Systems*, volume 1 (pp. 609–616).

Chételat, G. & Baron, J. C. (2003). Early diagnosis of Alzheimer's disease: contribution of structural neuroimaging. *NeuroImage*, 18(2), 525–541.

Chételat, G., Landeau, B., Eustache, F., Mézenge, F., Viader, F., de la Sayette, V., Desgranges, B., & Baron, J. C. (2005). Using voxel-based morphometry to map the structural changes associated with rapid conversion in MCI: a longitudinal MRI study. *NeuroImage*, 27(4), 934–946.

Cho, S.-H., Kim, D. G., Kim, D.-S., Kim, Y.-H., Lee, C.-H., & Jang, S. H. (2007). Motor outcome according to the integrity of the corticospinal tract determined by diffusion tensor tractography in the early stage of corona radiata infarct. *Neuroscience Letters*, 426(2), 123–127.

Chung, F. R. K. (1992). *Spectral Graph Theory*. Number 92. American Mathematical Society.

Chung, M. K. (2004). *Heat kernel smoothing and its application to cortical manifolds*. Technical report, 1090. Department of Statistics, Univ of Wisconsin, Madison.

Chung, M. K., Robbins, S., & Evans, A. C. (2005a). Unified statistical approach to cortical thickness analysis. In G. E. Christensen & M. Sonka (Eds.), *Information Processing in Medical Imaging*, volume 3565 of *Lecture Notes in Computer Science* (pp. 627–38). Springer Berlin / Heidelberg.

Chung, M. K., Robbins, S. M., Dalton, K. M., Davidson, R. J., Alexander, A. L., & Evans, A. C. (2005b). Cortical thickness analysis in autism with heat kernel smoothing. *NeuroImage*, 25(4), 1256–65.

Chung, M. K., Worsley, K. J., Robbins, S., Paus, T., Taylor, J., Giedd, J. N., Rapoport, J. L., & Evans, A. C. (2003). Deformation-based surface morphometry applied to gray matter deformation. *NeuroImage*, 18(2), 198–213.

Chupin, M., Gerardin, E., Cuingnet, R., Boutet, C., Lemieux, L., Lehéricy, S., Benali, H., Garnero, L., Colliot, O., & Alzheimer's Disease Neuroimaging Initiative (2009a). Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data from ADNI. *Hippocampus*, 19(6), 579–587.

Chupin, M., Hammers, A., Liu, R. S., Colliot, O., Burdett, J., Bardinet, E., Duncan, J. S., Garnero, L., & Lemieux, L. (2009b). Automatic segmentation of the hippocampus and the amygdala driven by hybrid constraints: method and validation. *NeuroImage*, 46(3), 749–761.

Chupin, M., Mukuna-Bantumbakulu, A. R., Hasboun, D., Bardinet, E., Baillet, S., Kinkingnéhun, S., Lemieux, L., Dubois, B., & Garnero, L. (2007). Automated segmentation of the hippocampus and the amygdala driven by competition and anatomical priors: Method and validation on healthy subjects and patients with Alzheimer's disease. *NeuroImage*, 34, 996–1019.

Colliot, O., Chételat, G., Chupin, M., Desgranges, B., Magnin, B., Benali, H., Dubois, B., Garnero, L., Eustache, F., & Lehéricy, S. (2008a). Discrimination between Alzheimer disease, mild cognitive impairment, and normal aging by using automated segmentation of the hippocampus. *Radiology*, 248(1), 194–201.

Colliot, O., Chupin, M., Sarazin, M., Habert, M., Dormont, D., & Lehéricy, S. (2008b). L'apport de la neuro-imagerie dans la maladie d'Alzheimer. *PSN. Psychiatrie, sciences humaines, neurosciences*, 6(2), 68–75.

Convit, A., de Asis, J., de Leon, M. J., Tarshish, C. Y., De Santi, S., & Rusinek, H. (2000). Atrophy of the medial occipitotemporal, inferior, and middle temporal gyri in non-demented elderly predict decline to Alzheimer's disease. *Neurobiol. Aging*, 21, 19–26.

Convit, A., De Leon, M. J., Tarshish, C., De Santi, S., W. Tsui, Rusinek, H., & George, A. (1997). Specific hippocampal volume reductions in individuals at risk for Alzheimer's disease. *Neurobiol. Aging*, 18, 131–138.

Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273–297.

Coulon, O., Mangin, J. F., Poline, J. B., Zilbovicius, M., Roumenov, D., Samson, Y., Frouin, V., & Bloch, I. (2000). Structural group analysis of functional activation maps. *NeuroImage*, 11(6), 767–782.

Crafton, K. R., Mark, A. N., & Cramer, S. C. (2003). Improved understanding of cortical injury by incorporating measures of functional anatomy. *Brain*, 126(7), 1650–59.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.-O., Chupin, M., Benali, H., & Colliot, O. (2010). Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. *NeuroImage*, In Press, Corrected Proof, –.

Cuturi, M. & Vert, J. (2005). Semigroup kernels on finite sets. *Advances in Neural Information Processing Systems*, 17, 329–336.

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: I. segmentation and surface reconstruction. *NeuroImage*, 9(2), 179–194.

Davatzikos, C. (2004). Why voxel-based morphometric analysis should be used with great caution when characterizing group differences. *NeuroImage*, 23(1), 17–20.

Davatzikos, C., Fan, Y., Wu, X., Shen, D., & Resnick, S. M. (2008a). Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging. *Neurobiol. Aging*, 29(4), 514–523.

Davatzikos, C., Resnick, S., Wu, X., Parmpi, P., & Clark, C. (2008b). Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI. *NeuroImage*, 41(4), 1220–27.

Davatzikos, C., Resnick, S. M., Wu, X., Parmpi, P., & Clark, C. M. (2008c). Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI. *NeuroImage*, 41(4), 1220–1227.

de Pedro-Cuesta, J., Virues-Ortega, J., Vega, S., Seijo-Martinez, M., Saz, P., Rodriguez, F., Rodriguez-Laso, A., Rene, R., de las Heras, S., Mateos, R., Martinez-Martin, P., Manubens, J., Mahillo-Fernandez, I., Lopez-Pousa, S., Lobo, A., Regla, J., Gascon, J., Garcia, F., Fernandez-Martinez, M., Boix, R., Bermejo-Pareja, F., Bergareche, A., Benito-Leon, J., de Arce, A., & del Barrio, J. (2009). Prevalence of dementia and major dementia subtypes in spanish populations: A reanalysis of dementia prevalence surveys, 1990-2008. *BMC Neurology*, 9(1), 55.

Decoste, D. & Schölkopf, B. (2002). Training invariant support vector machines. *Machine Learning*, 46(1), 161–90.

Demiriz, A., Bennett, K., & Shawe-Taylor, J. (2002). Linear programming boosting via column generation. *Machine Learning*, 46(1), 225–254.

Desikan, R. S., Cabral, H. J., Hess, C. P., Dillon, W. P., Glastonbury, C. M., Weiner, M. W., Schmansky, N. J., Greve, D. N., Salat, D. H., Buckner, R. L., Fischl, B., & Alzheimer's Disease Neuroimaging Initiative (2009). Automated MRI measures identify individuals with mild cognitive impairment and Alzheimer's disease. *Brain*, 132(8), 2048–2057.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., & Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, 31(3), 968–980.

Devroye, L., Györfi, L., & Lugosi, G. (1996). *A probabilistic theory of pattern recognition*. Springer Verlag.

Dickerson, B. C., Bakkour, A., Salat, D. H., Feczko, E., Pacheco, J., Greve, D. N., Grodstein, F., Wright, C. I., Blacker, D., Rosas, H. D., Sperling, R. A., Atri, A., Growdon, J. H., Hyman, B. T., Morris, J. C., Fischl, B., & Buckner, R. L. (2009). The cortical signature of Alzheimer's disease: regionally specific cortical thinning relates to symptom severity in very mild to mild AD dementia and is detectable in asymptomatic amyloid-positive individuals. *Cereb. Cortex*, 19(3), 497–510.

Ding, C. & Dubchak, I. (2001). Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, 17(4), 349.

Domi, T., deVeber, G., Shroff, M., Kouzmitcheva, E., MacGregor, D. L., & Kirton, A. (2009). Corticospinal tract pre-wallerian degeneration: A novel outcome predictor for pediatric stroke on acute MRI. *Stroke*, 40(3), 780–787.

Donnan, G. A., Fisher, M., Macleod, M., & M, D. S. (2008). Stroke. *Lancet*, 371(9624), 1612 – 1623.

Druet, O., Hebey, E., & Robert, F. (2004). *Blow-up theory for elliptic PDEs in Riemannian geometry*. Princeton Univ Press.

Du, A., Schuff, N., Kramer, J., Rosen, H., Gorno-Tempini, M., Rankin, K., Miller, B., & Weiner, M. (2007). Different regional patterns of cortical thinning in Alzheimer's disease and frontotemporal dementia. *Brain*.

Dubois, B. & Albert, M. L. (2004). Amnestic MCI or prodromal Alzheimer's disease? *Lancet Neurol.*, 3(4), 246–248.

Dubois, B. et al. (2003). *Les nouveaux défis de la maladie d'Alzheimer : bilan et prospective*. Médigone.

Dubois, B., Feldman, H. H., Jacova, C., Cummings, J. L., DeKosky, S. T., Barberger-Gateau, P., Delacourte, A., Frisoni, G., Fox, N. C., Galasko, D., Gauthier, S., Hampel, H., Jicha, G. A., Meguro, K., O'Brien, J., Pasquier, F., Robert, P., Rossor, M., Salloway, S., Sarazin, M., de Souza, L. C., Stern, Y., Visser, P. J., & Scheltens, P. (2010). Revising the definition of alzheimer's disease: a new lexicon. *The Lancet Neurology*, 9(11), 1118 – 1127.

Dubois, B., Feldman, H. H., Jacova, C., DeKosky, S. T., Barberger-Gateau, P., Cummings, J., Delacourte, A., Galasko, D., Gauthier, S., Jicha, G., Meguro, K., O'Brien, J., Pasquier, F.,

Robert, P., Rossor, M., Salloway, S., Stern, Y., Visser, P. J., & Scheltens, P. (2007). Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. *Lancet Neurol.*, 6(8), 734–746.

Duchesnay, E., Cachia, A., Roche, A., Riviere, D., Cointepas, Y., Papadopoulos-Orfanos, D., Zilbovicius, M., Martinot, J., Regis, J., & Mangin, J. (2007). Classification based on cortical folding patterns. *IEEE Transactions on Medical Imaging*, 26(4), 553–565.

Duchesne, S., Bernasconi, N., Bernasconi, A., & Collins, D. (2006). MR-based neurological disease classification methodology: Application to lateralization of seizure focus in temporal lobe epilepsy. *NeuroImage*, 29(2), 557–566.

Duchesne, S., Caroli, A., Geroldi, C., Barillot, C., Frisoni, G., & Collins, D. (2008). MRI-based automated computer classification of probable AD versus normal controls. *IEEE Transactions on Medical Imaging*, 27(4), 509–20.

Duncan, J. S. (1997). Imaging and epilepsy. *Brain*, 120(2), 339–377.

Durrleman, S., Pennec, X., Trouvé, A., & Ayache, N. (2007). Measuring brain variability via sulcal lines registration: a diffeomorphic approach. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2007*, (pp. 675–682).

Durrleman, S., Pennec, X., Trouvé, A., Thompson, P., & Ayache, N. (2008). Inferring brain variability from diffeomorphic deformations of currents: an integrative approach. *Medical image analysis*, 12(5), 626–637.

Esmaeilzadeh, M., Ciarmiello, A., & Squitieri, F. (2010). Seeking Brain Biomarkers for Preventive Therapy in Huntington Disease. *CNS Neuroscience & Therapeutics*, -, –.

Fan, Y., Batmanghelich, N., Clark, C. M., Davatzikos, C., & Alzheimer's Disease Neuroimaging Initiative (2008a). Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. *NeuroImage*, 39(4), 1731–1743.

Fan, Y., Resnick, S. M., Wu, X., & Davatzikos, C. (2008b). Structural and functional biomarkers of prodromal Alzheimer's disease: a high-dimensional pattern classification study. *NeuroImage*, 41(2), 277–285.

Fan, Y., Shen, D., & Davatzikos, C. (2005). Classification of structural images via high-dimensional image warping, robust feature extraction, and SVM. In J. Duncan & G. Gerig (Eds.), *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2005*, volume 3749 of *Lecture Notes in Computer Science* (pp. 1–8). Springer Berlin / Heidelberg.

Fan, Y., Shen, D., Gur, R. C., & Davatzikos, C. (2007). COMPARE: classification of morphological patterns using adaptive regional elements. *IEEE Transactions on Medical Imaging*, 26(1), 93–105.

Ferri, C. P., Prince, M., Brayne, C., Brodaty, H., Fratiglioni, L., Ganguli, M., Hall, K., Hasegawa, K., Hendrie, H., Huang, Y., Jorm, A., Mathers, C., Menezes, P. R., Rimmer, E., & Scazufca, M. (2006). Global prevalence of dementia: a Delphi consensus study. *The Lancet*, 366(9503), 2112 – 2117.

Feydy, A., Carlier, R., Roby-Brami, A., Bussel, B., Cazalis, F., Pierot, L., Burnod, Y., & Maier, M. (2002). Longitudinal Study of Motor Recovery After Stroke: Recruitment and Focusing of Brain Activation. *Stroke*, 33(6), 1610–1617.

Fillard, P., Arsigny, V., Pennec, X., Hayashi, K., Thompson, P., & Ayache, N. (2007). Measuring brain variability by extrapolating sparse tensor fields measured on sulcal lines. *Neuroimage*, 34(2), 639–650.

Fischl, B. & Dale, A. M. (2000). Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proc. Natl Acad. Sci. USA*, 97, 11050–11055.

Fischl, B., Sereno, M. I., & Dale, A. M. (1999a). Cortical surface-based analysis: II. inflation, flattening, and a surface based coordinate system. *NeuroImage*, 9(2), 195–207.

Fischl, B., Sereno, M. I., Tootell, R. B., & Dale, A. M. (1999b). High-resolution intersubject averaging and a coordinate system for the cortical surface. *Hum. Brain Mapp.*, 8(4), 272–284.

Fox, N. C. & Schott, J. M. (2004). Imaging cerebral atrophy: normal ageing to Alzheimer's disease. *Lancet*, 363(9406), 392–394.

Fratiglioni, L., Launer, L. J., Andersen, K., Breteler, M. M. B., Copeland, J. R. M., Dartigues, J.-F., Lobo, A., Martinez-Lage, J., Soininen, H., Hofman, A., & for the Neurologic Diseases in the Elderly Research Group (2000). Incidence of dementia and major subtypes in europe: A collaborative study of population-based cohorts. *Neurology*, 54(11), Supplement 5:S10–S15.

Fratiglioni, L., Paillard-Borg, S., & Winblad, B. (2004). An active and socially integrated lifestyle in late life might protect against dementia. *The Lancet Neurology*, 3(6), 343 – 353.

Fratiglioni, L. & Wang, H.-X. (2000). Smoking and Parkinson's and Alzheimer's disease: review of the epidemiological studies. *Behavioural Brain Research*, 113(1-2), 117 – 120.

Freeborough, P. A. & Fox, N. C. (1998). MR image texture analysis applied to the diagnosis and tracking of Alzheimer's disease. *IEEE Transaction on Medical Imaging*, 27(3), 475–479.

Frisoni, G., Laakso, M., Beltramello, A., Geroldi, C., Bianchetti, A., Soininen, H., & Trabucchi, M. (1999). Hippocampal and entorhinal cortex atrophy in frontotemporal dementia and Alzheimer's disease. *Neurology*, 52(1), 91–100.

Furey, T., Cristianini, N., Duffy, N., Bednarski, D., Schummer, M., & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10), 906.

Gallagher, C., Hutchinson, P., & Pickard, J. (2007). Neuroimaging in trauma. *Current opinion in neurology*, 20(4), 403.

Gartner, T., Flach, P., & Wrobel, S. (2003). On graph kernels: Hardness results and efficient alternatives. In *Learning theory and Kernel machines: proceedings of the 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003* (pp. 129).: Springer Verlag.

Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Gerardin, E., Chételat, G., Chupin, M., Cuingnet, R., Desgranges, B., Kim, H. S., Niethammer, M., Dubois, B., Lehéricy, S., Garnero, L., Francis, E., & Colliot, O. (2009). Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging. *NeuroImage*, 47(4), 1476–1486.

Gerstner, E., Sorensen, A., Jain, R., & Batchelor, T. (2008). Advances in neuroimaging techniques for the evaluation of tumor growth, vascular permeability, and angiogenesis in gliomas. *Current opinion in neurology*, 21(6), 728.

Gladstone, D. J., Black, S. E., & Hakim, A. M. (2002). Toward Wisdom From Failure: Lessons From Neuroprotective Stroke Trials and New Therapeutic Directions. *Stroke*, 33(8), 2123–2136.

Goldszal, A. F., Davatzikos, C., Pham, D. L., Yan, M. X., Bryan, R. N., & Resnick, S. M. (1998). An image-processing system for qualitative and quantitative volumetric analysis of brain images. *J. Comput. Assist. Tomogr.*, 22(5), 827–837.

Golland, P., Grimson, W. E. L., Shenton, M. E., & Kikinis, R. (2005). Detection and analysis of statistical differences in anatomical shape. *Medical Image Analysis*, 9(1), 69–86.

Golub, G. & Van Loan, C. (1996). *Matrix computations*. Johns Hopkins University Press.

Gomez, L., Camps-Valls, G., Muñoz-Marí, J., & Calpe-Maravilla, J. (2008). Semisupervised Image Classification with Laplacian Support Vector Machines. *IEEE Geoscience and Remote Sensing Letters*, 5(3), 336–40.

Good, C. D., Johnsrude, I. S., Ashburner, J., Henson, R. N., Friston, K. J., & Frackowiak, R. S. (2001). A voxel-based morphometric study of ageing in 465 normal adult human brains. *NeuroImage*, 14(1), 21–36.

Good, C. D., Scahill, R. I., Fox, N. C., Ashburner, J., Friston, K. J., Chan, D., Crum, W. R., Rossor, M. N., & Frackowiak, R. S. (2002). Automatic differentiation of anatomical patterns in the human brain: validation with studies of degenerative dementias. *NeuroImage*, 17(1), 29–46.

Greenshtein, E. & Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of over parametrization. *Bernoulli*, 10(6), 971–988.

Guyon, I. & Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3, 1157–1182.

Guyon, I., J, W., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46, 389–422.

Hastie, T., Tibshirani, R., & Friedman, J. (2005). *The Elements of Statistical Learning: Data mining, inference and prediction*. Springer.

Heiss, W.-D., Thiel, A., Grond, M., & Graf, R. (1999). Which Targets Are Relevant for Therapy of Acute Ischemic Stroke? *Stroke*, 30(7), 1486–1489.

Helmer, C., Joly, P., Letenneur, L., Commenges, D., & Dartigues, J.-F. (2001). Mortality with dementia: Results from a french prospective community-based cohort. *Am. J. Epidemiol.*, 154(7), 642–648.

Higham, N. (2005). The scaling and squaring method for the matrix exponential revisited. *SIAM Journal on Matrix Analysis and Applications*, 26(4), 1179–96.

Hinrichs, C., Singh, V., Mukherjee, L., Xu, G., Chung, M. K., Johnson, S. C., & Alzheimer's Disease Neuroimaging Initiative (2009). Spatially augmented LP boosting for AD classification with evaluations on the ADNI dataset. *NeuroImage*, 48(1), 138–149.

Hua, S. & Sun, Z. (2001). Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, 17(8), 721.

Hua, X., Lee, S., Yanovsky, I., Leow, A. D., Chou, Y. Y., Ho, A. J., Gutman, B., Toga, A. W., Jack Jr., C. R., Bernstein, M. A., Reiman, E. M., Harvey, D. J., Kornak, J., Schuff, N., Alexander, G. E., Weiner, M. W., Thompson, P. M., & Alzheimer's Disease Neuroimaging Initiative (2009). Optimizing power to track brain degeneration in Alzheimer's disease and mild cognitive impairment with tensor-based morphometry: an ADNI study of 515 subjects. *NeuroImage*, 48(4), 668–681.

Ingalhalikar, M., Kanterakis, S., Gur, R., Roberts, T., & Verma, R. (2010). DTI Based Diagnostic Prediction of a Disease via Pattern Classification. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2010*, volume 6361 of *Lecture Notes in Computer Science* (pp. 558–565). Springer Berlin / Heidelberg.

Jaakkola, T., Diekhans, M., & Haussler, D. (2000). A discriminative framework for detecting remote protein homologies. *Journal of Computational Biology*, 7(1-2), 95–114.

Jack, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., Whitwell, J. L., Ward, C., Dale, A. M., Felmlee, J. P., Gunter, J. L., Hill, D. L., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C. S., Krueger, G., Ward, H. A., Metzger, G. J., Scott, K. T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J. P., Fleisher, A. S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., Weiner, M. W., & ADNI Study (2008). The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, 27(4).

Jack Jr, C., Slomkowski, M., Gracon, S., Hoover, T., Felmlee, J., Stewart, K., Xu, Y., Shiung, M., O'Brien, P., Cha, R., et al. (2003). MRI as a biomarker of disease progression in a therapeutic trial of milameline for AD. *Neurology*, 60(2), 253.

Jack Jr., C. R., Petersen, R. C., Xu, Y., O'Brien, P. C., Smith, G. E., Ivnik, R. J., Tangalos, E. G., & Kokmen, E. (1998). Rate of medial temporal lobe atrophy in typical aging and Alzheimer's disease. *Neurology*, 51(4), 993–999.

Jack Jr., C. R., Petersen, R. C., Xu, Y. C., Waring, S. C., O'Brien, P. C., Tangalos, E. G., Smith, G. E., Ivnik, R. J., & Kokmen, E. (1997). Medial temporal atrophy on MRI in normal aging and very mild Alzheimer's disease. *Neurology*, 49(3), 786–794.

Jagust, W. (2006). Positron emission tomography and magnetic resonance imaging in the diagnosis and prediction of dementia. *Alzheimers Dement.*, 2, 36–42.

Jaillard, A., Martin, C. D., Garambois, K., Lebas, J. F., & Hommel, M. (2005). Vicarious function within the human primary motor cortex?: A longitudinal fMRI stroke study. *Brain*, 128(5), 1122–1138.

Jang, S. H., Bai, D., Son, S. M., Lee, J., Kim, D.-S., Sakong, J., Kim, D. G., & Yang, D. S. (2008). Motor outcome prediction using diffusion tensor tractography in pontine infarct. *Annals of Neurology*, 64, 460–465.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Jorm, A. F. & Jolley, D. (1998). The incidence of dementia: A meta-analysis. *Neurology*, 51(3), 728–733.

Jost, J. (2008). *Riemannian geometry and geometric analysis*. Springer Verlag.

Juottonen, K., Laakso, M. P., Insausti, R., Lehtovirta, M., Pitkänen, A., Partanen, K., & Soininen, H. (1998). Volumes of the entorhinal and perirhinal cortices in Alzheimer's disease. *Neurobiol. Aging*, 19(1), 15–22.

Kalaria, R. N., Maestre, G. E., Arizaga, R., Friedland, R. P., Galasko, D., Hall, K., Luchsinger, J. A., Ogunniyi, A., Perry, E. K., Potocnik, F., Prince, M., Stewart, R., Wimo, A., Zhang, Z.-X., & Antuono, P. (2008). Alzheimer's disease and vascular dementia in developing countries: prevalence, management, and risk factors. *The Lancet Neurology*, 7(9), 812 – 826.

Karas, G. B., Burton, E. J., Rombouts, S. A., van Schijndel, R. A., O'Brien, J. T., Scheltens, P., McKeith, I. G., Williams, D., Ballard, C., & Barkhof, F. (2003). A comprehensive study of gray matter loss in patients with Alzheimer's disease using optimized voxel-based morphometry. *NeuroImage*, 18(4), 895–907.

Karas, G. B., Scheltens, P., Rombouts, S. A., Visser, P. J., van Schijndel, R. A., Fox, N. C., & Barkhof, F. (2004). Global and local gray matter loss in mild cognitive impairment and Alzheimer's disease. *NeuroImage*, 23(2), 708–716.

Katzman, R. (1993). Education and the prevalence of dementia and Alzheimer's disease. *Neurology*, 43(1), 13–20.

Katzman, R., Terry, R., DeTeresa, R., Brown, T., Davies, P., Fuld, P., Renbing, X., & Peck, A. (1988). Clinical, pathological, and neurochemical changes in dementia: A subgroup with preserved mental status and numerous neocortical plaques. *Annals of Neurology*, 23(2), 138–144.

Kim, D. G., Ahn, Y. H., Byun, W. M., Kim, T. G., Yang, D. S., Ahn, S. H., Cho, Y. W., & Jang, S. H. (2007). Degeneration speed of corticospinal tract in patients with cerebral infarct. *NeuroRehabilitation*, 22, 273–277.

Kimeldorf, G. & Wahba, G. (1971). Some results on Tchebycheffian spline functions* 1. *Journal of Mathematical Analysis and Applications*, 33(1), 82–95.

Klein, A., Andersson, J., Ardekani, B. A., Ashburner, J., Avants, B., Chiang, M. C., Christensen, G. E., Collins, D. L., Gee, J., Hellier, P., Song, J. H., Jenkinson, M., Lepage, C., Rueckert, D., Thompson, P., Vercauteren, T., Woods, R. P., Mann, J. J., & Parsey, R. V. (2009). Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *NeuroImage*, 46(3), 786–802.

Klöppel, S., Stonnington, C. M., Barnes, J., Chen, F., Chu, C., Good, C. D., Mader, I., Mitchell, L. A., Patel, A. C., Roberts, C. C., Fox, N. C., Jack, C. R., Ashburner, J., & Frackowiak, R. S. J. (2008a). Accuracy of dementia diagnosis–a direct comparison between radiologists and a computerized method. *Brain*, 131(11), 2969–74.

Klöppel, S., Stonnington, C. M., Chu, C., Draganski, B., Scahill, R. I., Rohrer, J. D., Fox, N. C., Jack Jr., C. R., Ashburner, J., & Frackowiak, R. S. J. (2008b). Automatic classification of MR scans in Alzheimer's disease. *Brain*, 131(3), 681–689.

Kondor, R. & Jebara, T. (2003). A kernel between sets of vectors. In *In proceedings of Twentieth International Conference on Machine Learning (ICML).*

Kondor, R. I. & Lafferty, J. D. (2002). Diffusion kernels on graphs and other discrete input spaces. In *Proc. International Conference on Machine Learning* (pp. 315–22).

Konig, I. R., Ziegler, A., Bluhmki, E., Hacke, W., Bath, P. M., Sacco, R. L., Diener, H. C., Weimar, C., & on behalf of the Virtual International Stroke Trials Archive (VISTA) Investigators (2008). Predicting long-term outcome after acute ischemic stroke: A simple index works in patients from controlled clinical trials. *Stroke*, 39(6), 1821–1826.

Konishi, J., Yamada, K., Kizu, O., Ito, H., Sugimura, K., Yoshikawa, K., Nakagawa, M., & Nishimura, T. (2005). MR tractography for the evaluation of functional recovery from lenticulostriate infarcts. *Neurology*, 64(1), 108–113.

Kunimatsu, A., Itoh, D., Nakata, Y., Kunimatsu, N., Aoki, S., Masutani, Y., Abe, O., Yoshida, M., Minami, M., & Ohtomo, K. (2007). Utilization of diffusion tensor tractography in combination with spatial normalization to assess involvement of the corticospinal tract in capsular/pericapsular stroke: Feasibility and clinical implications. *Journal of Magnetic Resonance Imaging*, 26, 1399–1404.

Laakso, M. P., Frisoni, G. B., Könönen, M., Mikkonen, M., Beltramello, A., Geroldi, C., Bianchetti, A., Trabucchi, M., Soininen, H., & Aronen, H. J. (2000). Hippocampus and entorhinal cortex

in frontotemporal dementia and Alzheimer's disease: a morphometric MRI study. *Biol. Psychiatry*, 47(12), 1056–1063.

Laakso, M. P., Soininen, H., Partanen, K., Lehtovirta, M., Hallikainen, M., Hänninen, T., Helkala, E. L., Vainio, P., & Riekkinen Sr., P. J. (1998). MRI of the hippocampus in Alzheimer's disease: sensitivity and specificity and and analysis of the incorrectly classified subjects. *Neurobiol. Aging*, 19(1), 23–31.

Lafferty, J. & Lebanon, G. (2005). Diffusion kernels on statistical manifolds. *Journal of Machine Learning Research*, 6, 129–63.

Lanckriet, G. R. G., Bie, T. D., Cristianini, N., Jordan, M. I., & Noble, W. S. (2004). A statistical framework for genomic data fusion. *Bioinformatics*, 20(16), 2626–2635.

Lao, Z., Shen, D., Xue, Z., Karacali, B., Resnick, S. M., & Davatzikos, C. (2004). Morphological classification of brains via high-dimensional shape transformations and machine learning methods. *NeuroImage*, 21(1), 46–57.

Lehéricy, S., Marjanska, M., Mesrob, L., Sarazin, M., & Kinkingnehun, S. (2007). Magnetic resonance imaging of Alzheimer's disease. *European Radiology*, 17(2), 347–362.

Lerch, J. P., Pruessner, J., Zijdenbos, A. P., Collins, D. L., Teipel, S. J., Hampel, H., & Evans, A. C. (2008). Automated cortical thickness measurements from MRI can accurately separate Alzheimer's patients from normal elderly controls. *Neurobiol. Aging*, 29(1), 23–30.

Lerch, J. P., Pruessner, J. C., Zijdenbos, A., Hampel, H., Teipel, S. J., & Evans, A. C. (2005). Focal decline of cortical thickness in Alzheimer's disease identified by computational neuroanatomy. *Cereb. Cortex*, 15(7), 995–1001.

Letenneur, L., Launer, J., Andersen, K., Dewey, M. E., Ott, A., Copeland, J. R. M., Dartigues, J.-F., Kragh-Sorensen, P., Baldereschi, M., Brayne, C., Lobo, A., Martinez-Lage, J. M., Stijnen, T., Hofman, A., & for the EURODEM Incidence Research Group (2000). Education and risk for Alzheimer's disease: Sex makes a difference EURODEM pooled analyses. *Am. J. Epidemiol.*, 151(11), 1064–1071.

Levy-Cooperman, N., Ramirez, J., Lobaugh, N. J., & Black, S. E. (2008). Misclassified tissue volumes in Alzheimer disease patients with white matter hyperintensities: importance of lesion segmentation procedures for volumetric analysis. *Stroke*, 39(4), 1134–1141.

Liu, K., Li, F., Tatlisumak, T., Garcia, J., Sotak, C., Fisher, M., & Fenstermacher, J. (2001). Regional variations in the apparent diffusion coefficient and the intracellular distribution of water in rat brain during acute focal ischemia. *Stroke*, 32(8), 1897.

Magnin, B., Mesrob, L., Kinkingnéhun, S., Pélégrini-Issac, M., Colliot, O., Sarazin, M., Dubois, B., Lehéricy, S., & Benali, H. (2009). Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI. *Neuroradiology*, 51(2), 73–83.

Mahé, P., Ueda, N., Akutsu, T., Perret, J., & Vert, J. (2004). Extensions of marginalized graph kernels. In *Proceedings of the twenty-first international conference on Machine learning* (pp.70).: ACM.

Mallat, S. (2001). *Une exploration des signaux en ondelettes*. École Polytechnique.

Mangin, J., Riviere, D., Cachia, A., Duchesnay, E., Cointepas, Y., Papadopoulos-Orfanos, D., Scifo, P., & Ochiai, T. (2004a). A framework to study the cortical folding patterns. *Neuroimage*, 23, S129–S138.

Mangin, J., Riviere, D., Coulon, O., Poupon, C., Cachia, A., Cointepas, Y., Poline, J., Bihan, D., Régis, J., & Papadopoulos-Orfanos, D. (2004b). Coordinate-based versus structural approaches to brain image analysis. *Artificial Intelligence in Medicine*, 30(2), 177–198.

Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C., & Buckner, R. L. (2007). Open Access Series of Imaging Studies OASIS: cross-sectional MRI data in young and middle aged and nondemented and and demented older adults. *J. Cogn. Neurosci.*, 19(9), 1498–1507.

Maurer, K., Volk, S., & Gerbaldo, H. (1997). Auguste D. première patiente du docteur Alzheimer. *La Recherche*, 303.

McDiarmid, C. (1989). On the method of bounded differences. *Surveys in combinatorics*, 141, 148–188.

McDonald, C. R., McEvoy, L. K., Gharapetian, L., Fennema-Notestine, C., Hagler Jr., D. J., Holland, D., Koyama, A., Brewer, J. B., Dale, A. M., & Alzheimer's Disease Neuroimaging Initiative (2009). Regional rates of neocortical atrophy from normal aging to early Alzheimer disease. *Neurology*, 73(6), 457–465.

McKeith, I., Mintzer, J., Aarsland, D., Burn, D., Chiu, H., Cohen-Mansfield, J., Dickson, D., Dubois, B., Duda, J. E., Feldman, H., Gauthier, S., Halliday, G., Lawlor, B., Lippa, C., Lopez, O. L., Machado, J. C., O'Brien, J., & Playfer, J. (2004). Dementia with lewy bodies. *The Lancet Neurology*, 3(1), 19 – 28.

Menezes, N. M., Ay, H., Zhu, M. W., Lopez, C. J., Singhal, A. B., Karonen, J. O., Aronen, H. J., Liu, Y., Nuutinen, J., Koroshetz, W. J., & Sorensen, A. G. (2007). The real estate factor: quantifying the impact of infarct location on stroke severity. *Stroke*, 38(1), 194–197.

Miller, M., Trouvé, A., & Younes, L. (2006). Geodesic shooting for computational anatomy. *Journal of Mathematical Imaging and Vision*, 24(2), 209–228.

Miller, M. & Younes, L. (2001). Group actions, homeomorphisms, and matching: A general framework. *International Journal of Computer Vision*, 41(1), 61–84.

Misra, C., Fan, Y., & Davatzikos, C. (2009). Baseline and longitudinal patterns of brain atrophy in MCI patients and and their use in prediction of short-term conversion to AD: results from ADNI. *NeuroImage*, 44(4), 1415–1422.

Moler, C. & Van Loan, C. (2003). Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM review*, 45(1), 3–49.

Mortimer, J. A. & B., G. A. (1993). Education and other socioeconomic determinants of dementia and Alzheimer's disease. *Neurology*, 43(8), S39–S44.

Mourao-Miranda, J., Bokde, A. L. W., Born, C., Hampel, H., & Stetter, M. (2005). Classifying brain states and determining the discriminating activation patterns: support vector machine on functional MRI data. *NeuroImage*, 28(4), 980–995.

Moustafa, R. & Baron, J. (2007). Pathophysiology of ischaemic stroke: insights from imaging, and implications for therapy and drug discovery. *British journal of pharmacology*, 153, S44–S54.

Mukherjee, S., Tamayo, P., Slonim, D., Verri, A., Golub, T., Mesirov, J., & Poggio, T. (1999). Support vector machine classification of microarray data. *CBCL Paper*, 182.

Muller, K., Mika, S., Ratsch, G., Tsuda, K., & Scholkopf, B. (2001). An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12(2), 181–201.

Myint-U, T. & Debnath, L. (2007). *Linear Partial Differential Equations for Scientists and Engineers*. Birkhauser Boston, 4th edition.

Nédélec, J. C. (2001). *Acoustic and electromagnetic equations: integral representations for harmonic problems*. Springer Verlag.

Nelles, M., Gieseke, J., Flacke, S., Lachenmayer, L., Schild, H., & Urbach, H. (2007). Diffusion Tensor Pyramidal Tractography in Patients With Anterior Choroidal Artery Infarcts. *AJNR Am J Neuroradiol*, 29, 488–493.

Nemetz, P. N., Leibson, C., Naessens, J. M., Beard, M., Kokmen, E., Annegers, J. F., & Kurland, L. T. (1999). Traumatic brain injury and time to onset of Alzheimer's disease: A population-based study. *Am. J. Epidemiol.*, 149(1), 32–40.

Neuhaus, M. & Bunke, H. (2006). Edit distance-based kernel functions for structural pattern classification. *Pattern Recognition*, 39(10), 1852–1863.

Newton, J. M., Ward, N. S., Parker, G. J. M., Deichmann, R., Alexander, D. C., Friston, K. J., & Frackowiak, R. S. J. (2006). Non-invasive mapping of corticofugal fibres from multiple motor areas–relevance to stroke recovery. *Brain*, 129(7), 1844–1858.

O'Meara, E. S., Kukull, W. A., Sheppard, L., Bowen, J. D., McCormick, W. C., Teri, L., Pfanschmidt, M., Thompson, J. D., Schellenberg, G. D., & Larson, E. B. (1997). Head injury and risk of Alzheimer's disease by apolipoprotein E genotype. *Am. J. Epidemiol.*, 146(5), 373–384.

Orrell, M. & Sahakian, B. (1995). Education and dementia. *BMJ*, 310(6985), 951–952.

Ott, A., Breteler, M. M. B., van Harskamp, F., Claus, J. J., van der Cammen, T. J. M., Grobbee, D. E., & Hofman, A. (1995). Prevalence of Alzheimer's disease and vascular dementia: association with education. the Rotterdam study. *BMJ*, 310(6985), 970–973.

Pearlson, G. & Marsh, L. (1999). Structural brain imaging in schizophrenia: a selective review. *Biological Psychiatry*, 46(5), 627–649.

Pennec, X., Ayache, N., & Thirion, J.-P. (2008). Landmark-based registration using features identified through differential geometry. In I. Bankman (Ed.), *Handbook of Medical Image Processing and Analysis - New edition* chapter 34, (pp. 565–578). Academic Press.

Perrot, M., Rivière, D., Tucholka, A., & Mangin, J.-F. (2009). Joint Bayesian Cortical Sulci Recognition and Spatial Normalization. In *Proc. of the 21th conference on Information Processing in Medical Imaging*, LNCS-5636 (pp. 176–187).: Springer Verlag.

Petersen, R. C., Smith, G. E., Waring, S. C., Ivnik, R. J., Tangalos, E. G., & Kokmen, E. (1999). Mild cognitive impairment: clinical characterization and outcome. *Arch. Neurol.*, 56, 303–308.

Querbes, O., Aubry, F., Pariente, J., Lotterie, J. A., Démonet, J. F., Duret, V., Puel, M., Berry, I., Fort, J. C., Celsis, P., & Alzheimer's Disease Neuroimaging Initiative (2009). Early diagnosis of Alzheimer's disease using cortical thickness: impact of cognitive reserve. *Brain*, 32(8), 2036–2047.

Raber, J., Huang, Y., & Ashford, J. W. (2004). Apoe genotype accounts for the vast majority of ad risk and ad pathology. *Neurobiology of Aging*, 25(5), 641 – 650. Challenging Views of Alzheimer's Disease - Round II.

Rakotomamonjy, A., Bach, F. R., Canu, S., & Grandvalet, Y. (2008). SimpleMKL. *J. Mach. Learn. Res.*, 9, 2491–2521.

Ramaroson, H., Helmer, C., Barberger-Gateau, P., Letenneur, L., & Dartigues, J.-F. (2003). Prévalence de la démence et de la maladie d'Alzheimer chez les personnes de 75 ans et plus : données réactualisées de la cohorte PAQUID. *Revue Neurologique*, Vol 159(4), 405–411.

Rapaport, F., Zinovyev, A., Dutreix, M., Barillot, E., & Vert, J.-P. (2007). Classification of microarray data using gene networks. *BMC Bioinformatics*, 8(1), 35.

Reitz, C., den Heijer, T., van Duijn, C., Hofman, A., & Breteler, M. (2007). Relation between smoking and risk of dementia and Alzheimer disease: The Rotterdam study. *Neurology*, 69(10), 998–1005.

Richardson, M. (2010). Current themes in neuroimaging of epilepsy: Brain networks, dynamic phenomena, and clinical relevance. *Clinical Neurophysiology*, 121(8), 1153 – 1175.

Rondeau, V., Jacqmin-Gadda, H., Commenges, D., Helmer, C., & Dartigues, J.-F. (2009). Aluminum and silica in drinking water and the risk of Alzheimer's disease or cognitive decline: Findings from 15-year follow-up of the PAQUID cohort. *Am. J. Epidemiol.*, 169(4), 489–496.

Rosen, H., Gorno-Tempini, M., Goldman, W., Perry, R., Schuff, N., Weiner, M., Feiwell, R., Kramer, J., & Miller, B. (2002). Patterns of brain atrophy in frontotemporal dementia and semantic dementia. *Neurology*, 58(2), 198.

Rosenberg, S. (1997). *The Laplacian on a Riemannian manifold: an introduction to analysis on manifolds*. Cambridge University Press.

Rosso, C., Colliot, O., Delmaire, C., Valabrègue, R., Pires, C., Crozier, S., Dormont, D., Baillet, S., Samson, Y., & Lehéricy, S. (2010). Early ADC changes in motor structurespredict outcome of acute stroke better than lesion volume. *Journal of Neuroradiology, in press*.

Rudin, L. I., Osher, S., & Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D*, 60, 259–268.

Rusinek, H., Endo, Y., Santi, S. D., Frid, D., Tsui, W. H., Segal, S., Convit, A., & de Leon, M. J. (2004). Atrophy rate in medial temporal lobe during progression of Alzheimer's disease. *Neurology*, 63(12), 2354–2359.

Salat, D. H., Buckner, R. L., Snyder, A. Z., Greve, D. N., Desikan, R. S., Busa, E., Morris, J. C., Dale, A. M., & Fischl, B. (2004). Thinning of the cerebral cortex in aging. *Cereb. Cortex*, 14(7), 721–730.

Salat, D. H., Lee, S. Y., van der Kouwe, A. J., Greve, D. N., Fischl, B., & Rosas, H. D. (2009). Age-associated alterations in cortical gray and white matter signal intensity and gray to white matter contrast. *NeuroImage*, 48(1), 21–28.

Sato, J. R., Fujita, A., Thomaz, C. E., da Graca Morais Martin, M., Mourão-Miranda, J., Brammer, M. J., & Junior, E. A. (2009). Evaluating SVM and MLDA in the extraction of discriminant regions for mental state prediction. *NeuroImage*, 46(1), 105–114.

Savitz, S. I. & Fisher, M. (2007). Future of neuroprotection for acute stroke: In the aftermath of the SAINT trials. *Annals of Neurology*, 61, 396–402.

Schaechter, J. D., Perdue, K. L., & Wang, R. (2008). Structural damage to the corticospinal tract correlates with bilateral sensorimotor cortex reorganization in stroke patients. *NeuroImage*, 39(3), 1370 – 1382.

Schmidt, M. S. (1997). Identifying speakers with support vector networks. *Computing Science and Statistics*, (pp. 305–316).

Schölkopf, B., Burges, C., & Vapnik, V. (1996). Incorporating invariances in support vector learning machines. In *Proc. Artificial Neural Networks–ICANN 1996* (pp.$\tilde{4}$7).: Springer Verlag.

Schölkopf, B., Simard, P., Smola, A., & Vapnik, V. (1998). Prior knowledge in support vector kernels. In *Proc. Conference on Advances in Neural Information Processing Systems'97* (pp. 640–46).: MIT Press.

Schölkopf, B. & Smola, A. J. (2001). *Learning with Kernels*. MIT Press.

Schölkopf, B., Tsuda, K., & Vert, J. (2004). *Kernel methods in computational biology*. The MIT press.

Shawe-Taylor, J. & Cristianini, N. (2000). *Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press.

Shawe-Taylor, J. & Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press.

Shen, D. & Davatzikos, C. (2002). HAMMER: hierarchical attribute matching mechanism for elastic registration. *IEEE Transaction on Medical Imaging*, 21(11), 1421–39.

Shi, J. & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 22(8), 888–905.

Shuaib, A., Lees, K. R., Lyden, P., Grotta, J., Davalos, A., Davis, S. M., Diener, H.-C., Ashwood, T., Wasiewski, W. W., Emeribe, U., & the SAINT II Trial Investigators (2007). NXY-059 for the Treatment of Acute Ischemic Stroke. *the New England Journal of Medicine*, 357(6), 562–571.

Sled, J. G., Zijdenbos, A. P., & Evans, A. C. (1998). A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Transactions on Medical Imaging*, 17, 87–97.

Smola, A. & Kondor, R. (2003). Kernels and regularization on graphs. In *Proc. COLT* (pp. 144).: Springer Verlag.

Smola, A. J. & Schölkopf, B. (1998). On a kernel-based method for pattern recognition, regression, approximation, and operator inversion. *Algorithmica*, 22(1/2), 211–31.

Soares, J. C. & Mann, J. J. (1997). The anatomy of mood disorders–review of structural neuroimaging studies. *Biological Psychiatry*, 41(1), 86 – 106.

Sonnenburg, S., Rätsch, G., Schäfer, C., & Schölkopf, B. (2006). Large scale multiple kernel learning. *J. Mach. Learn. Res. Arch.*, 7, 1531–1565.

Steinwart, I. (2003). Sparseness of support vector machines. *Journal of Machine Learning Research*, 4, 1071–1105.

Stern, Y. (2002). What is cognitive reserve? theory and research application of the reserve concept. *J Int Neuropsychol Soc.*, 8(3), 448–460.

Stern, Y. (2006). Cognitive reserve and alzheimer disease. *Alzheimer Dis Assoc Disord*, 20((3 Suppl 2)), S69–S74.

Stern, Y., Habeck, C., Moeller, J., Scarmeas, N., Anderson, K. E., Hilton, H. J., Flynn, J., Sackeim, H., & van Heertum, R. (2005). Brain Networks Associated with Cognitive Reserve in Healthy Young and Old Adults. *Cereb. Cortex*, 15(4), 394–402.

Stinear, C. (2010). Prediction of recovery of motor function after stroke. *The Lancet Neurology*, 9(12), 1228–1232.

Stinear, C. M., Barber, P. A., Smale, P. R., Coxon, J. P., Fleming, M. K., & Byblow, W. D. (2007). Functional potential in chronic stroke patients depends on corticospinal tract integrity. *Brain*, 130(1), 170–180.

Stone, C. (1977). Consistent nonparametric regression. *The annals of statistics*, 5(4), 595–620.

Styner, M., Oguz, I., Xu, S., Brechbuler, C., Pantazis, D., & Gerig, G. (2006). Statistical Shape Analysis of Brain Structures using SPHARM-PDM. *Insight Journal DSpace*.

Tapiola, T., Pennanen, C., Tapiola, M., Tervo, S., Kivipelto, M., Hänninen, T., Pihlajamäki, M., Laakso, M. P., Hallikainen, M., Hämäläinen, A., Vanhanen, M., Helkala, E. L., Vanninen, R., Nissinen, A., Rossi, R., Frisoni, G. B., & Soininen, H. (2008). MRI of hippocampus and entorhinal cortex in mild cognitive impairment: a follow-up study. *Neurobiol. Aging*, 29(1), 31–38.

Teipel, S. J., Born, C., Ewers, M., Bokde, A. L., Reiser, M. F., Möller, H.-J., & Hampel, H. (2007). Multivariate deformation-based analysis of brain atrophy to predict Alzheimer's disease in mild cognitive impairment. *NeuroImage*, 38(1), 13–24.

Thomalla, G., Glauche, V., Weiller, C., & Röther, J. (2005). Time course of wallerian degeneration after ischaemic stroke revealed by diffusion tensor imaging. *Journal of Neurology, Neurosurgery & Psychiatry*, 76(2), 266–268.

Thompson, P. M., Hayashi, K. M., de Zubicaray, G., Janke, A. L., Rose, S. E., Semple, J., Herman, D., Hong, M. S., Dittmer, S. S., Doddrell, D. M., & Toga, A. W. (2003). Dynamics of gray matter loss in Alzheimer's disease. *J. Neurosci.*, 23(3), 994–1005.

Thompson, P. M., Hayashi, K. M., Sowell, E. R., Gogtay, N., Giedd, J. N., Rapoport, J. L., de Zubicaray, G. I., Janke, A. L., Rose, S. E., Semple, J., Doddrell, D. M., Wang, Y., van Erp, T. G., Cannon, T. D., & Toga, A. W. (2004). Mapping cortical change in Alzheimer's disease and brain development and and schizophrenia. *NeuroImage*, 23(Suppl 1), S2–S18.

Thompson, P. M., Mega, M. S., Woods, R. P., Zoumalan, C. I., Lindshield, C. J., Blanton, R. E., Moussai, J., Holmes, C. J., Cummings, J. L., & Toga, A. W. (2001). Cortical change in Alzheimer's disease detected with a disease-specific population-based brain atlas. *Cereb. Cortex*, 11(1), 1–16.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B (Methodological)*, (pp. 267–288).

Tsuda, K., Kin, T., & Asai, K. (2002). Marginalized kernels for biological sequences. *Bioinformatics*, 18(Suppl 1), S268.

Tsuda, K. & Noble, W. S. (2004). Learning kernels from biological networks by maximizing entropy. *Bioinformatics*, 20(suppl 1), i326–333.

Turk, M. & Pentland, A. (2002). Face recognition using eigenfaces. In *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on* (pp. 586–591).

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., & Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, 15, 273–289.

Vapnik, V. (1999). An Overview of Statistical Learning Theory. *IEEE Transactions on Neural Networks*, 10(5).

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. New York Inc and Springer-Verlag.

Vellas, B., Gillette-Guyonnet, S., & Andrieu, S. (2008). Memory health clinics–a first step to prevention. *Alzheimer's and Dementia*, 4(1, Supplement 1), S144 – S149.

Vemuri, P., Gunter, J. L., Senjem, M. L., Whitwell, J. L., Kantarci, K., Knopman, D. S., Boeve, B. F., Petersen, R. C., & Jack Jr., C. R. (2008). Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies. *NeuroImage*, 39(3), 1186–1197.

Verhoeven, J., De Cock, P., Lagae, L., & Sunaert, S. (2010). Neuroimaging of autism. *Neuroradiology*, 52(1), 3–14.

Vert, J. (2002). A tree kernel to analyse phylogenetic profiles. *Bioinformatics*, 18(Suppl 1), S276.

Vert, J. & Kanehisa, M. (2003). Graph-driven feature extraction from microarray data using diffusion kernels and kernel CCA. *Advances in Neural Information Processing Systems*, (pp. 1449–1456).

Wang, H.-X., Fratiglioni, L., Frisoni, G. B., Viitanen, M., & Winblad, B. (1999). Smoking and the occurence of Alzheimer's disease: Cross-sectional and longitudinal data in a population-based study. *Am. J. Epidemiol.*, 149(7), 640–644.

Wang, Z., Childress, A. R., Wang, J., & Detre, J. A. (2007). Support vector machine learning-based fMRI data group analysis. *NeuroImage*, 36(4), 1139–1151.

Watson, R., Blamire, A., & O'Brien, J. (2009). Magnetic resonance imaging in lewy body dementias. *Dementia and geriatric cognitive disorders*, 28(6), 493–506.

Wenzelburger, R., Kopper, F., Frenzel, A., Stolze, H., Klebe, S., Brossmann, A., Kuhtz-Buschbeck, J., Golge, M., Illert, M., & Deuschl, G. (2005). Hand coordination following capsular stroke. *Brain*, 128(1), 64–74.

Whitwell, J. L., Przybelski, S. A., Weigand, S. D., Knopman, D. S., Boeve, B. F., Petersen, R. C., & Jack Jr., C. R. (2007). 3D maps from multiple MRI illustrate changing atrophy patterns as subjects progress from mild cognitive impairment to Alzheimer's disease. *Brain*, 130(7), 1777–1786.

Whitwell, J. L., Shiung, M. M., Przybelski, S. A., Weigand, S. D., Knopman, D. S., Boeve, B. F., Petersen, R. C., & Jack Jr., C. R. (2008). MRI patterns of atrophy associated with progression to AD in amnestic mild cognitive impairment. *Neurology*, 70(7), 512–520.

Wolf, L. & Shashua, A. (2003). Learning over sets using kernel principal angles. *Journal of Machine Learning Research*, 4, 913–931.

Worsley, K., Marrett, S., Neelin, P., Vandal, A., Friston, K., & Evans, A. (1996). A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping*, 4(1), 58–73.

Wright, I., McGuire, P., Poline, J., Travere, J., Murray, R., Frith, C., Frackowiak, R., & Friston, K. (1995). A voxel-based method for the statistical analysis of gray and white matter density applied to schizophrenia. *Neuroimage*, 2(4), 244–252.

Xu, Y., Jack Jr., C. R., O'Brien, P. C., Kokmen, E., Smith, G. E., Ivnik, R. J., Boeve, B. F., Tangalos, R. G., & Petersen, R. C. (2000). Usefulness of MRI measures of entorhinal cortex versus hippocampus in AD. *Neurology*, 54(9), 1760–1767.

Yang, Q., Tress, B., Barber, P., Desmond, P., Darby, D., Gerraty, R., Li, T., & Davis, S. (1999). Serial study of apparent diffusion coefficient and anisotropy in patients with acute stroke. *Stroke*, 30(11), 2382.

Yassa, M. A. & Stark, C. E. (2009). A quantitative evaluation of cross-participant registration techniques for MRI studies of the medial temporal lobe. *NeuroImage*, 44(2), 319–327.

Ye, J., Chen, K., Wu, T., Li, J., Zhao, Z., Patel, R., Bae, M., Janardan, R., Liu, H., Alexander, G., & Reiman, E. (2008). Heterogeneous data fusion for Alzheimer's disease study. In *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD '08* (pp. 1025–1033).

Yu, C., Zhu, C., Zhang, Y., Chen, H., Qin, W., Wang, M., & Li, K. (2009). A longitudinal diffusion tensor imaging study on wallerian degeneration of corticospinal tract after motor pathway stroke. *NeuroImage*, 47(2), 451 – 458.

Zekry, D., Hauw, J. J., & Gold, G. (2002). Mixed dementia: epidemiology and diagnosis and and treatment. *J. Am. Geriatr. Soc.*, 8, 1431–1438.

Zhang, D., Wang, Y., Zhou, L., Yuan, H., & Shen, D. (2011). Multimodal Classification of Alzheimer's Disease and Mild Cognitive Impairment. *NeuroImage*, In Press, Accepted Manuscript, –.

Zivadinov, R. & Cox, J. L. (2007). Neuroimaging in multiple sclerosis. In A. Minagar (Ed.), *The Neurobiology of Multiple Sclerosis*, volume 79 of *International Review of Neurobiology* (pp. 449 – 474). Academic Press.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.

Table des figures

Liste des tableaux

Index

Abstract

Brain image analyses have widely relied on univariate voxel-wise methods. In such analyses, brain images are first spatially registered to a common stereotaxic space, and then mass univariate statistical tests are performed in each voxel to detect significant group differences. However, the sensitivity of theses approaches is limited when the differences involve a combination of different brain structures. Recently, there has been a growing interest in support vector machines methods to overcome the limits of these analyses.

This thesis focuses on machine learning methods for population analysis and patient classification in neuroimaging. We first evaluated the performances of different classification strategies for the identification of patients with Alzheimer's disease based on T1-weighted MRI of 509 subjects from the ADNI database. However, these methods do not take full advantage of the spatial distribution of the features. As a consequence, the optimal margin hyperplane is often scattered and lacks spatial coherence, making its anatomical interpretation difficult. Therefore, we introduced a framework to spatially regularize support vector machines for brain image analysis based on Laplacian regularization operators. The proposed framework was then applied to the analysis of stroke and of Alzheimer's disease. The results demonstrated that the proposed classifier generates less-noisy and consequently more interpretable feature maps with no loss of classification performance.

Résumé

L'analyse automatique de différences anatomiques en neuroimagerie a de nombreuses applications pour la compréhension et l'aide au diagnostic de pathologies neurologiques. Récemment, il y a eu un intérêt croissant pour les méthodes de classification telles que les machines à vecteurs supports pour dépasser les limites des méthodes univariées traditionnelles.

Cette thèse a pour thème l'apprentissage automatique pour l'analyse de populations et la classification de patients en neuroimagerie. Nous avons tout d'abord comparé les performances de différentes stratégies de classification, dans le cadre de la maladie d'Alzheimer à partir d'images IRM anatomiques de 509 sujets de la base de données ADNI. Ces différentes stratégies prennent insuffisamment en compte la distribution spatiale des *features*. C'est pourquoi nous proposons un cadre original de régularisation spatiale et anatomique des machines à vecteurs supports pour des données de neuroimagerie volumiques ou surfaciques, dans le formalisme de la régularisation laplacienne. Cette méthode a été appliquée à deux problématiques cliniques : la maladie d'Alzheimer et les accidents vasculaires cérébraux. L'évaluation montre que la méthode permet d'obtenir des résultats cohérents anatomiquement et donc plus facilement interprétables, tout en maintenant des taux de classification élevés.