Mathematical methods of image analysis for cross-sectional and longitudinal population studies

Jean-Baptiste Fiot



Université Paris-Dauphine
École Doctorale de Dauphine

Méthodes mathématiques d'analyse d'image pour les études de population transversales et longitudinales.

Mathematical methods of image analysis for cross-sectional and longitudinal population studies.

THÈSE

Pour l'obtention du titre de

Docteur en Sciences - Spécialité Mathématiques Appliquées
(Arrêté du 7 Août 2006)

Présentée par

Jean-Baptiste FIOT

Soutenue publiquement le 17 septembre 2013 devant le jury composé de

Directeur de thèse :	**Laurent D. Cohen** Directeur de recherche CNRS, Université Paris-Dauphine France
Rapporteurs :	**Olivier Colliot** Chargé de recherche CNRS, ICM France **Xavier Pennec** Directeur de recherche, INRIA Sophia Antipolis France
Examinateurs :	**Matthew B. Blaschko** Assistant Professor, École Centrale Paris, France **Jurgen Fripp** Research Scientist, CSIRO, Brisbane, Australia **Joan Alexis Glaunès** Maître de conférence, Université Paris Descartes, France **Gabriel Peyré** Chargé de recherche CNRS, Université Paris-Dauphine, France

L'université n'entend donner aucune approbation ni improbation aux opinions émises dans les thèses : ces opinions doivent être considérées comme propres à leurs auteurs.

Acknowledgments

Fist, I would like to thank the members of my jury:

- Laurent Cohen, my PhD advisor, for offering invaluable advice and generous help during all my PhD,
- Olivier Colliot and Xavier Pennec, for accepting to review my manuscript and taking the time to give me great feedback that truly helped me to improve this manuscript,
- Jurgen Fripp, for spending a huge amount of time to collaborate and giving me relevant technical advice,
- Gabriel Peyré, for organizing workshops, sharing his knowledge and giving me advice,
- Matthew B. Blaschko and Joan Alexis Glaunès for doing me the honor to be in my jury.

Then I would like to thank my other collaborators:

- Parnesh Raniga for his collaboration on the work of lesion segmentation, his enthusiasm and for sharing his knowledge of medical applications,
- François-Xavier Vialard, for helping me to discover the field of computational anatomy, and sharing valuable insights about the inner workings of the scientific world,
- Laurent Risser, for his great technical insights and enthusiasm,
- Hugo Raguet, for sharing his knowledge, his hard work and dedication.

Now, I would like to thank several people who contributed and helped me at various levels of my thesis:

- Nicolas Schmidt for sharing his knowledge on a broad variety of mathematical topics and being always ready to help,

- Samuel Vaiter for sharing expertise from his own research and IT knowledge,

- Raphaël Prévost for our discussions and proofreading parts of my manuscript,

- the post-docs from our image group Sira Ferradans, Giacomo Nardi and Vincent Duval for their discussions,

- Oliver Salvado, group leader at CSIRO, for believing in my research projects, finding funding, and helping me to set up an international collaboration,

- Pierrick Bourgeat, Nicholas Dowson, Oscar Acosta, David Raffelt, Jason Dowling and Vincent Doré for their discussions,

- my friends Antoine Labatie and Olivier Comas for giving me advice from their own past PhD experience,

- all the people I shared an office with in Paris (Hugo, Nicolas, Samuel, Loïc, Vito, Simona, Joana, Mauricio), and in Brisbane (Zoé, Cyril, Pierre, Rémy, Olivier, Hugo, Alex, Ales, Kaikai, Elsa, Florence, David, Chris, Vincent, William, Eric, Mario, Zim, ...).

Finally, I would like to warmly thank my family, in particular my parents, and my friends for supporting me during my projects.

Méthodes mathématiques d'analyse d'image pour les études de population transversales et longitudinales.

Résumé En médecine, les analyses de population à grande échelle ont pour but d'obtenir des informations statistiques pour mieux comprendre des maladies, identifier leurs facteurs de risque, développer des traitements préventifs et curatifs et améliorer la qualité de vie des patients.

Dans cette thèse, nous présentons d'abord le contexte médical de la maladie d'Alzheimer, rappelons certains concepts d'apprentissage statistique et difficultés rencontrées lors de l'application en imagerie médicale. Dans la deuxième partie, nous nous intéressons aux analyses *transversales*, c-a-d ayant un seul point temporel. Nous présentons une méthode efficace basée sur les séparateurs à vaste marge (SVM) permettant de classifier des lésions dans la matière blanche. Ensuite, nous étudions les techniques d'apprentissage de variétés pour l'analyse de formes et d'images, et présentons deux extensions des Laplacian eigenmaps améliorant la représentation de patients en faible dimension grâce à la combinaison de données d'imagerie et cliniques. Dans la troisième partie, nous nous intéressons aux analyses *longitudinales*, c-a-d entre plusieurs points temporels. Nous quantifions les déformations des hippocampes de patients via le modèle des larges déformations par difféomorphismes pour classifier les évolutions de la maladie. Nous introduisons de nouvelles stratégies et des régularisations spatiales pour la classification et l'identification de marqueurs biologiques.

Mots clés : Imagerie médicale, Analyse de population, Maladie d'Alzheimer, Traitement d'image, Apprentissage de variétés, Modèle prédictif, Régularisation, Marqueur biologique

Mathematical methods of image analysis for cross-sectional and longitudinal population studies.

Abstract In medicine, large scale population analysis aim to obtain statistical information in order to understand better diseases, identify their risk factors, develop preventive and curative treatments and improve the quality of life of the patients.

In this thesis, we first introduce the medical context of Alzheimer's disease, recall some concepts of statistical learning and the challenges that typically occur when applied in medical imaging. The second part focus on *cross-sectional* studies, i.e. at a *single time point*. We present an efficient method to classify white matter lesions based on support vector machines. Then we discuss the use of manifold learning techniques for image and shape analysis. Finally, we present extensions of Laplacian eigenmaps to improve the low-dimension representations of patients using the combination of imaging and clinical data. The third part focus on *longitudinal* studies, i.e. *between several time points*. We quantify the hippocampus deformations of patients via the large deformation diffeomorphic metric mapping framework to build disease progression classifiers. We introduce novel strategies and spatial regularizations for the classification and identification of biomarkers.

Keywords: Medical imaging, Population analysis, Alzheimer's disease, Image processing, Manifold learning, Predictive model, Regularization, Biomarker

Contents

List of Figures

List of Tables

List of Notations

Dimensionality reduction notations

Δ Distance matrix . . . 126

General notations

K Mercer Kernel . . . 92

$\| \quad \|_K$ Norm on $\mathcal{H}_K$ . . . 92

n Number of inputs / observations . . . 42

λ Regularization parameter . . . 92

$\mathcal{H}_K$ Reproducing Kernel Hilbert Space . . . 92

Machine learning notations

DSC Dice score . . . 97

FN False negative . . . 97

FP False positive . . . 97

ℓ_{hinge} Hinge loss . . . 92

$\mathbf{x}$ Input / observation / sample (element of $\mathcal{X}$) . . . 42

$\mathcal{X}$ Input space . . . 42

y Output (element of $\mathcal{Y}$) . . . 42

$\mathcal{Y}$ Output space . . . 42

f Prediction function . . . 42

$Sens$ Sensitivity . . . 97

$\mathcal{F}(\mathcal{X}, \mathcal{Y})$ Set of all prediction functions from $\mathcal{X}$ to $\mathcal{Y}$ . . . 42

ξ Slack variable . . . 92

$Spec$ Specificity . . . 97

TN True negative . . . 97

TP True positive . . . 97

Image and registration notations

$\star$ Convolution operator . . . 161

ϕ Deformation from $t = 0$ to $t = 1$ . . . 161

ϕ_{t_1,t_2} Deformation from $t = t_1$ to $t = t_2$ . . . 161

div Divergence operator . . . 161

Ω Image domain (subset of $\mathbb{R}^2$ or $\mathbb{R}^3$) . . . 181

F Follow-up image . . . 162

grad Gradient operator . . . 161

Jac Jacobian operator . . . 161

P_t Momentum at time t . . . 161

S Screening image . . . 162

I Source image . . . 161

J Target image . . . 161

ω Voxel (element of Ω) . . . 181

v_t Velocity field at time t . . . 161

List of Acronyms

Alzheimer's disease

AD	Alzheimer's disease
NC	normal control
MCI	mild cognitive impairment
s-MCI	stable mild cognitive impairment
p-MCI	progressive mild cognitive impairment
MMSE	mini-mental state exam
ApoE	apolipoprotein E
AIBL	Australian imaging biomarker and lifestyle
ADNI	Alzheimer's disease neuroimaging initiative
ptau	phosphorylated tau

Brain imaging

MR	magnetic resonance
T1-w	T1-weighted
T2-w	T2-weighted
FLAIR	fluid attenuated inversion recovery
PD	proton density
PiB	Pittsburgh compound B marker
FDG	fluorodeoxyglucose
F-18	fluorine-18
PET	positron emission tomography
WM	white matter
GM	grey matter

CSF	cerebro-spinal fluid
WMH	white matter hyper-intensities
ROI	region of interest

Computational anatomy

RR	rigid registration
NRR	non-rigid registration
VBM	voxel-based morphometry
DBM	deformation-based morphometry
TBM	tensor-based morphometry
SVF	stationary velocity field
FFD	free-form deformation
LDDMM	large deformation diffeomorphic metric mapping

Manifold learning

DR	dimensionality reduction
NLDR	non linear dimensionality reduction
MDS	multi dimension scaling
PCA	principal component analysis
kPCA	kernel principal component analysis
ISOMAP	isometric mapping
LLE	local linear embeddings
LEM	Laplacian eigenmaps
HEM	Hessian eigenmaps
DM	diffusion maps
LTSA	local tangent space alignment

SVD	singular value decomposition
GRAM	geodesic registration on anatomical manifold
LEAP	learning embeddings for atlas propagation
NMI	normalized mutual information

Data analysis, statistical and machine learning

ERM	empirical risk minimization
RKHS	reproducing kernel Hilbert space
SVM	support vector machines
kNN	k nearest neighbours
TP	true positive
TN	true negative
FP	false positive
FN	false negative
DSC	Dice score
EM	expectation-maximisation
RLS	regularized least squares
LM-BFGS	limited memory Broyden-Fletcher-Goldfarb-Shanno
HANSO	hybrid algorithm for non-smooth optimization
GFB	generalized forward-backward
M-FISTA	monotonous fast iterative shrinkage thresholding algorithm
EDA	exploratory data analysis
DOF	degrees of freedom

Introduction

En médecine, les analyses de population ont pour but d'obtenir des informations d'ordre statistique permettant une meilleure interprétation de l'état de santé d'un patient particulier. Les informations obtenues dans le cadre d'une étude de population vont de la fréquence d'apparition d'une maladie, la recherche de symptômes, la recherche de causes de certains états biologiques, l'identification de facteurs de risques, etc. Grâce à ces informations statistiques, les équipes médicales peuvent détecter au plus tôt voire prédire l'apparition de maladies chez un nouveau patient, développer des traitements curatifs ou préventifs plus efficaces, et améliorer les conditions de vie des malades.

L'origine et le type de données observées lors des analyses de population peut être extrêmement variable : concentrations chimiques après une prise de sang, scores cognitifs suite à un questionnaire, échantillons cellulaires provenant d'une biopsie, etc. Les techniques d'imageries médicales (imagerie par résonance magnétique, scanner, rayons X, ultrasons, etc) ont connu un essor considérable lors des dernières années. En effet, celles-ci permettent de collecter des informations *in-vivo* de manière *non-invasive*.

Pour analyser des populations par le biais d'images médicales, l'utilisation de méthodes automatiques ou semi-automatiques est incontournable. En effet, lorsqu'un seul patient est étudié, la quantité d'information à analyser et la difficulté du traitement sont telles que des techniques de mathématiques appliquées sont déjà utilisées sur divers problèmes, comme par exemple la segmentation d'organes. Le développement de méthodes numériques pour analyser des populations d'images médicales à grande échelle est donc a fortiori inévitable, et de nombreuses méthodes statistiques et d'apprentissage (*machine learning*) ont été introduites dans la communauté scientifique.

⋆
⋆ ⋆

Motivations cliniques dans la littérature

Les travaux présentés dans cette thèse ont été réalisés en partenariat entre différents instituts : le laboratoire CEREMADE de l'Université Paris Dauphine à Paris en France, le CSIRO à Brisbane en Australie, et l'Institut des mathématiques de Toulouse en France. Le contexte clinique de ces travaux est l'étude de la maladie d'Alzheimer. Nous présentons dans cette section différentes motivations cliniques dans la littérature, afin de mieux situer nos travaux présentés par la suite.

La maladie d'Alzheimer fut découverte au début du XX^{e} siècle par Aloïs Alzheimer. Les symptômes de la première patiente diagnostiquée incluaient des déficits de mémoire, de compréhension, une aphasie (perte du langage), et une perte du sens de l'orientation. La maladie d'Alzheimer est aujourd'hui une maladie répandue à grande échelle et affectant une large part de la population mondiale. De nombreuses études scientifiques cherchent à analyser différents aspects de la maladie.

Création d'outils de diagnostic

Une première classe d'études vise à construire un modèle d'évolution de la maladie. La création d'un tel modèle peut avoir différents objectifs

- identifier les étapes de l'évolution de la maladie,
- trouver des tendances dans les populations.

La connaissance des étapes typiques de l'évolution de la maladie permet de situer un patient d'un point de vue clinique et de l'informer de son évolution la plus probable. Néanmoins, il existe une variabilité dans les évolutions. Par exemple, certains patients atteints de troubles cognitifs légers vont développer la maladie alors que d'autres resteront stables. Il est donc important d'essayer d'identifier les tendances dans les populations, et tenter de comprendre pourquoi certains patients ont des évolutions différentes. Pour évaluer les états ou évolutions des patients, plusieurs étapes sont nécessaires, notamment

- la définition de différents états cliniques,
- la création d'outils de diagnostic (ou d'aide au diagnostic).

Les outils de diagnostic ainsi créés ont pour but d'associer un état clinique à un patient à partir de données observées. Dans le cas de la maladie d'Alzheimer, différents types de données peuvent être utilisés pour la construction de ces outils : des mesures provenant de tests sanguins, des résultats de tests cognitifs, des indicateurs de style de vie, des données d'imagerie, etc. Il est intéressant d'évaluer la quantité d'information contenue dans chaque type de données, ainsi que de définir des techniques permettant de les combiner. En combinant plusieurs types de données, on peut en effet espérer une amélioration des performances des outils de diagnostic, puisque ceux-ci disposent de plus d'information.

Pour créer ces outils de diagnostic, des techniques de mathématiques appliquées et notamment d'apprentissage statistique sont souvent utilisées. De plus, on distingue deux types d'études

1. transversales, et
2. longitudinales.

Les études transversales analysent des populations à un instant temporel, alors que les études longitudinales analysent les évolutions entre différents instants.

Identification et quantification de biomarqueurs

Une deuxième classe d'études vise à identifier et quantifier des *biomarqueurs*. Les biomarqueurs, ou marqueurs biologiques, sont des substances, mesures ou indicateurs d'un état biologique. Les biomarqueurs peuvent exister avant l'apparition de symptômes cliniques. En 2010, Jack et al. ont introduit un modèle hypothétique de la maladie d'Alzheimer, où différents biomarqueurs permettent d'évaluer l'évolution de la maladie chez un patient [Jack 2010]. Parmi les biomarqueurs que l'on cherche à caractériser à l'aide d'images de patients, on peut citer

- les biomarqueurs structuraux,
- les biomarqueurs fonctionnels.

La recherche de biomarqueurs structuraux est basée sur l'hypothèse que la maladie affecte les structures anatomiques cérébrales. L'hippocampe est l'une des structure particulièrement étudiées. La recherche de biomarqueurs fonctionnels est basée sur l'hypothèse que la maladie affecte le comportement fonctionnel et neuronal du cerveau. De nombreuses études ont par la suite été publiées, dans le but de vérifier la validité des hypothèses de ce modèle. Ce modèle a récemment été mis à jour [Jack 2013].

Identification et quantification de facteurs de risque

Une troisième classe d'études vise à identifier et quantifier des *facteurs de risque*. Un facteur de risque est une variable corrélée avec une maladie. Par définition, un facteur de risque n'est pas nécessairement une cause de maladie. Néanmoins, la compréhension des facteurs de risque peut aider le développement de traitements préventifs.

Dans le cas de la maladie d'Alzheimer, plusieurs facteurs de risques potentiels sont étudiés dans la littérature. On peut citer l'âge, le sexe, les facteurs génétiques (en particulier le gène codant pour l'ApoE), les hyper-intensités de la substance blanche, etc. Dans le but d'identifier si ces facteurs potentiels ont un réel impact sur la progression de la maladie, il est intéressant de pouvoir les quantifier de manière précise, rapide et automatique.

Analyse exploratoire

Une quatrième classe d'études est dite *exploratoire*. L'analyse exploratoire est un processus particulier d'utilisation et d'analyse des données. En analyse classique, le processus est

$$\text{Problème} \rightarrow \text{Données} \rightarrow \text{Modèle} \rightarrow \text{Analyse} \rightarrow \text{Conclusions.}$$

En analyse exploratoire, le raisonnement diffère

$$\text{Problème} \rightarrow \text{Données} \rightarrow \text{Analyse} \rightarrow \text{Modèle} \rightarrow \text{Conclusions.}$$

L'analyse exploratoire fut notamment initiée par Tukey dans [Tukey 1962]. Elle a pour objectifs de

- découvrir les structures sous-jacentes,
- identifier les variables importantes,
- détecter les données aberrantes et les anomalies,
- tester des suppositions issues des données,
- développer des modèles minimaux,
- etc.

En imagerie médicale, et en particulier pour dans le cadre de l'analyse de populations, ces techniques sont intéressantes pour identifier des tendances et faire des hypothèses sur les régions d'intérêt potentiellement liées à l'évolution de maladies. Il faut noter que les outils de visualisation sont importants pour l'analyse exploratoire.

Aspects méthodologiques

Dans cette thèse, nous utilisons différentes méthodes et stratégies d'analyse de données.

Construction de modèles statistiques

En mathématiques appliquées, de nombreux modèles ont été proposés pour faire de l'apprentissage automatique à partir de base de données. L'idée de base est très simple : elle consiste à prédire certaines informations sur un nouveau patient à partir d'observations passées d'autres patients. Un problème classique est celui de la classification, où l'on peut par exemple chercher à prédire le diagnostic à partir de l'observation d'une image du patient. Cependant, certains challenges se posent lorsque l'on cherche à utiliser ces techniques sur des bases de données d'images médicales (malédiction de la dimension, malédiction de la grande quantité de donnée (big data), classes déséquilibrées, etc). Dans cette thèse, nous étudions différentes stratégies pour faire face à ces challenges.

Définition des descripteurs et des distances

L'utilisation de modèles d'apprentissage nécessite

1. la définition de descripteurs,
2. la définition d'une distance adéquate.

Dans cette thèse, nous étudions des descripteurs encodant de l'information à différents niveaux : au niveau du voxel (i.e. du voisinage anatomique), du patient, et de l'évolution du patient.

Dans le cas de descripteurs au niveau du patient, que ce soit en statique pour dans les études transversales, ou en dynamique dans les études longitudinales, il est possible de mesurer des écarts entre patients en utilisant une distance euclidienne, ou en construisant des distances plus évoluées, basées sur les déformations. Nous verrons qu'une "simple" distance euclidienne permet déjà d'obtenir certaines informations. Dans la littérature, de nombreux modèles d'*anatomie numérique* ont été développés. Nous étudions en particulier des algorithmes de création d'atlas, permettant de représenter les patients dans le même espace, et la création de descripteurs à partir du modèle des larges déformations par difféomorphismes (LDDMM).

Nombre de paramètres et réduction de dimension

La réduction de dimension est une technique très populaire permettant de diminuer la dimension des observations d'un système. Dans le cadre du traitement d'images, elle peut avoir pour but

- de faciliter la visualisation et permettre une analyse exploratoire,

- de régulariser des algorithmes d'apprentissage.

Dans cette thèse, nous utilisons la réduction de dimension pour la visualisation et l'analyse exploratoire. Nous avons également réalisé certaines expériences utilisant les Laplacian SVM [Belkin 2004], où la régularisation est utilisée pour régulariser l'apprentissage. Les résultats obtenus avec cette méthode sont intéressants sur des données synthétiques, cependant sur des données réelles nos premiers résultats n'étant pas meilleurs que ceux obtenus avec des SVM, ils ne sont pas présentés ici.

L'importance des régularisations

De nombreux problèmes d'imagerie médicale (recalage, segmentation, apprentissage, etc) peuvent être formulés sous une forme variationnelle. La solution recherchée est exprimée comme minimisant une fonctionnelle, et appartenant à un certain espace. Cependant, pour contrôler la régularité de la solution, des régularisations sont introduites.

De plus, la majorité de ces problèmes sont en fait mal posés (par opposition à la notion de problème bien posé au sens d'Hadamard), et l'on cherche à obtenir une solution au problème régularisé. Un exemple typique de problème mal posé consiste à vouloir optimiser un nombre de paramètres supérieurs au nombre d'observations d'un système. L'idée d'introduire des régularisations pour la résolution de problèmes mal posés est largement étudiée dans la littérature, on peut par exemple citer les développements de Tikhonov [Tikhonov 1943, Tikhonov 1963, Tikhonov 1977]. Dans la littérature, de nombreuses régularisations ont été introduites pour résoudre divers problèmes en traitement d'image et imagerie médicales.

Dans cette thèse, nous introduisons des régularisations spatiales pour la classification et l'identification de biomarqueurs structuraux.

Plan et contributions

Le manuscrit de cette thèse s'organise en trois parties.

1. Dans une première partie, nous présentons la problématique.
   - **Chapitre 1 :** Ce chapitre présente le contexte médical de la thèse. Il introduit le rôle de l'imagerie médicale dans le cadre de l'étude de populations. Il présente ensuite la maladie d'Alzheimer et les différentes bases de données utilisées dans cette thèse.
   - **Chapitre 2 :** Ce chapitre rappelle le contexte d'apprentissage statistique. Il introduit également les différentes difficultés rencontrées lors de l'utilisation de ces techniques en imagerie médicale.
   - **Chapitre 3 :** Ce chapitre présente les contributions de cette thèse.
2. La seconde partie traite de l'*analyse transversale* de population. Dans ce cadre, des images de plusieurs patients sont analysées, à raison d'*un point temporel par patient*, c'est-à-dire une seule image par patient.
   - **Chapitre 4 :** État de l'art. Ce chapitre présente différentes méthodes d'apprentissages de variétés (*manifold learning*), ainsi que l'application de ces méthodes en imagerie médicale pour le recalage, la segmentation, l'analyse de population et l'apprentissage statistique.
   - **Chapitre 5 :** Ce chapitre traite de la segmentation de lésions dans la matière blanche. Nos contributions dans ce chapitre sont :
     - l'introduction d'une nouvelle procédure de segmentation basée sur les séparateurs à vaste marge (SVM), incluant notamment la définition d'une nouvelle zone d'intérêt,
     - la validation sur une base de données provenant d'AIBL,
     - l'évaluation de l'apport relatif de chaque modalité,
     - l'évaluation de la performance de classification pour différents types de descripteurs,
     - la comparaison avec d'autres algorithmes de classification supervisée.
   - **Chapitre 6 :** Ce chapitre présente la réduction de dimension non-linéaire pour l'imagerie médicale. Nos contributions dans ce chapitre sont :
     - la présentation de méthodes de réduction de dimension non-linéaire,
     - l'application en imagerie médicale mono-modale et multi-modale pour la visualisation de tendances dans une population.
   - **Chapitre 7 :** Ce chapitre traite de l'utilisation de données cliniques pour améliorer la représentation de patients en faible dimension. Nos contributions dans ce chapitre sont :
     - l'introduction d'une nouvelle méthode de réduction de dimension combinant des images et des données cliniques (extension de l'algorithme des **Laplacian Eigenmaps!** (**Laplacian Eigenmaps!**)),

  - la validation de cette méthode sur une base de données provenant d'ADNI, dans le cas de données cliniques continues et discrètes.

3. La troisième partie traite de l'*analyse longitudinale* de population. Dans ce cadre, nous disposons non pas d'une mais de plusieurs images par patient, prises à différents points temporels. Le but est désormais d'analyser les évolutions des patients.

   - **Chapitre 8 :** État de l'art. Pour étudier les évolutions de formes, différentes techniques de recalage ont été proposées, permettant d'estimer des transformations réalistes biologiquement parlant. Ce chapitre présente différents modèles de transformation, ainsi que différentes manières de calculer un atlas et différentes notions de transport.
   - **Chapitre 9 :** Dans ce chapitre, nous utilisons la théorie des larges déformations par difféomorphismes (*LDDMM*), qui offre un cadre Riemannien et la possibilité de représenter l'information de déformation par le biais de vecteur tangents. Nous présentons l'utilisation des *moments initiaux* pour construire des classifieurs de progression de maladie. Nos contributions dans ce chapitre sont :
     - l'introduction d'une méthode pour la séparer les patients stables des patients progressifs, à partir de moments initiaux représentant les évolutions des hippocampes,
     - l'introduction de deux extensions pour l'algorithme de Karcher permettant (1) de calculer la forme moyenne d'une population modulo les transformations rigides et (2) d'éviter le lissage dû aux réechantillonnages successifs,
     - la validation de la méthode proposée sur une base de données de patients provenant d'ADNI,
     - une preuve de concept sur l'utilisation de sous-régions lors de la classification, permettant aux représentations locales d'avoir des meilleures performances de classification que les représentations globales.
   - **Chapitre 10 :** Dans ce chapitre, nous traitons des régularisations spatiales pour la classification de progression de la maladie d'Alzheimer et la détection de déformations hippocampales liées. Nos contributions dans ce chapitre sont :
     - l'introduction d'un modèle de classification logistique avec pénalisation spatiale,
     - la comparaison de régularisations standards (Ridge, LASSO, ElasticNet) et de régularisations spatiales (Sobolev, variation totale, et fused LASSO) pour la classification et la détection de biomarqueurs,
     - la validation sur une base de données provenant d'ADNI.

Part I

Position of the problem

Chapter 1

Clinical context

Contents

Résumé

Dans ce chapitre, nous présentons le contexte médical de cette thèse. Dans un premier temps, nous présentons le rôle de l'imagerie médicale pour l'analyse de populations à grande échelle. Nous expliquons les principaux avantages des technologies d'imagerie et d'analyse d'images dans un contexte médical. Ensuite, la maladie d'Alzheimer est présentée. Cette maladie est l'application principale utilisée dans cette thèse pour valider les méthodes développées. Finalement, nous présentons les différentes bases de données utilisées pour les expériences numériques.

Mots clés : Analyse de population, imagerie médicale, maladie d'Alzheimer

Abstract

In this chapter, we introduce the clinical context of this thesis. First, we introduce the role of medical imaging in large-scale population analysis studies. We explain some of the key advantages of imaging and image analysis technologies in a medical background. Second, Alzheimer's

disease is presented. This disease is the main application we used to validate the methods developed in this thesis. Finally, we list the different databases that were used for the numerical experiments.

Keywords: Population analysis, medical imaging, Alzheimer's disease

1.1 Role of medical imaging in population analysis studies

Over recent years, large scale medical population studies such as analysis of disease progression and cohort stratification based on imaging technologies had tremendous development. Many reasons pushed the development of imaging acquisition and analysis technologies. In particular, imaging

1. allows *in-vivo* analysis,
2. is *non-invasive*,
3. allows data acquisition in sensitive or not easily accessible areas.

The first property makes the imaging procedure particularly appealing from a medical analysis point of view. When using an in-vitro procedure, the biological medium is modified before any analysis. In contrast, in-vivo imaging procedures allow the analysis of biological media *in-place*, without any external perturbation. Invasive procedures bias the analysis of static biological medium, let alone dynamic organs. For example, fast-imaging technologies can bring insight on the understanding of the cardiac cycle by providing motion data (videos). The second property is particularly interesting from a patient point of view. In particular, when recruiting healthy patients to be healthy controls in a large-scale population study, no need to mention that non-invasive and painless procedure facilitates the recruitment process. The third property is of particular interest in the field of neuro-imaging. Indeed, the brain is an area both particularly sensitive and not easily accessible.

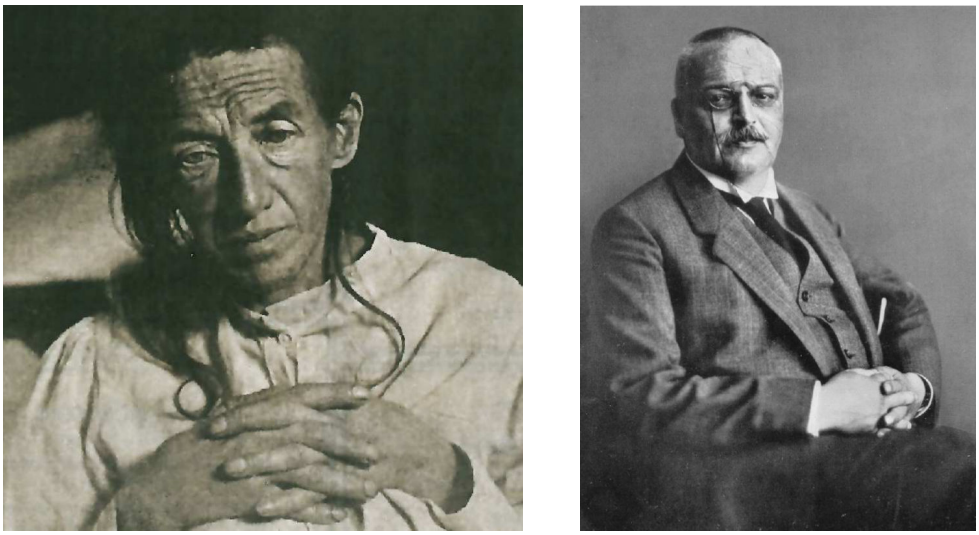
In the last decade, applied mathematics and computer science techniques have been developed and improved in order to analyze large databases of digital medical images. A widely studied example is Alzheimer's disease, which is briefly described in the next section.

1.2 Alzheimer's disease

AD is an irreversible neuro-degenerative disease that results in a loss of mental function due to the deterioration of brain tissue. It is briefly described in this section, which is partially inspired by Section 2.1.1 of [Cuingnet 2011a].

1.2.1 Symptoms and discovery

On November 26$^{\text{th}}$ 1901, Auguste D. (Fig. 1.1a) was admitted to the Frankfurt hospital. Her symptoms included a decreased understanding, decreased memory, aphasia, auditive hallucinations and a loss of the sense of directions. She received medical cared by a German neuropathologist called Alois Alzheimer (Fig. 1.1b). When Auguste D. passed away in 1903, Dr Alzheimer performed a histological study that revealed the presence of senile plaques and a particular form of neurodegeneration:



(a) (b)

Figure 1.1: Left: Auguste D., first patient diagnosed with AD. Right: Dr Alois Alzheimer.

neurofibrillary tangles (Fig. 1.2). It should be noted that up to now, these are the only entirely reliable symptoms of AD [Amieva 2007].

During the 37$^{\text{th}}$ conference of German psychiatrists in Tübingen, Dr Alois Alzheimer exposed his observations of a new type of dementia. His results were published in 1907 in an article called "Über eine eigenartige Erkankung der Hirnrinde", literally "A characteristic serious disease of the cerebral cortex". The name *Alzheimer's disease* was mentioned for the first time by Emil Kraeplin in the 8$^{\text{th}}$ edition of the "Handbook of Psychiatry" (1910).

1.2.2 Risk factors

A risk factor is a variable that is correlated with a disease. By definition, a risk factor is not necessarily a cause of the disease. Nonetheless, understanding the risk factors of AD can only be helpful in order to develop preventive treatments. Risk factors for AD can be categorized into (1) direct risk factors, (2) alterations of the cognitive reserve and (3) confounding factors.

Aging is the main risk factor of AD and other dementias (see Fig. 1.3). Then comes the genetics. In particular, the gene coding for ApoE has proven to be a risk factor [Raber 2004]. Three alleles encode this gene $\{\varepsilon_2, \varepsilon_3, \varepsilon_4\}$, leading to six possible combinations (each individual holds two alleles). The ε_4 allele increases the risk of AD. The sex of the patients is a controversial risk factor. Even though the lifetime risks can be estimated higher for women than for men (Fig. 1.4, source: `www.alz.org`), the reason of this difference is not quite understood yet.

The notion of *cognitive reserve* describes the mind's resistance to the damage of

Figure 1.2: Original drawings of neurofibrillary tangles by Dr Alois Alzheimer.

the brain. In other words, it refers to the capacity of the brain to compensate alterations related to aging or diseases. This capacity could be related to the quantity and density of neurons [Katzman 1988], or to the brain capacity to be more active and/or use cells less in areas affected by the disease [Stern 2012].

Remark. In Chapter 7, we will see that the knowledge of clinical information such as ApoE genotype can improve the low-dimension representations of the patients.

Remark (White-matter hyper-intensities (WMH)). WMH are another possible risk factor for AD and vascular dementia, with progression associated with vascular factors and cognitive decline [Lao 2008]. Methods for segmenting efficiently WMH are discussed in Chapter 5.

1.2.3 Facts and figures

AD is the most common cause of dementia among people over the age of 65, yet no prevention methods or cures have been discovered. Figure 1.3 shows the proportion of people in the U.S. with AD and other dementias, according to age and ethnicity. Figure 1.4 shows the estimated lifetime risks for AD according to age and sex. Figure 1.5 show the percentage changes for several causes of death between 2000 and 2010.

1.2.4 Alzheimer's disease model

As mentioned in Section 1.2.3, no cure for AD has been found yet to this date. As a consequence, a large number of studies aim at modeling AD to understand better its causes with the final aim of developing preventive and curing treatments. In the

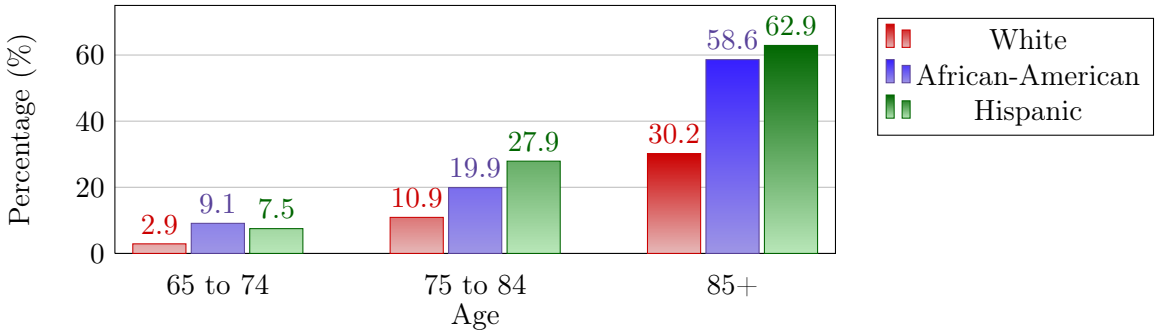


Figure 1.3: Proportion of people aged 65 and older with AD and other dementias. (Data from www.alz.org).

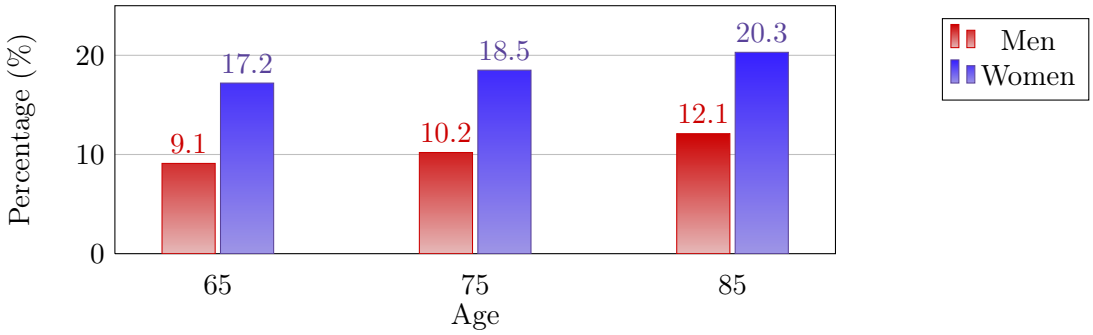


Figure 1.4: Estimated lifetime risks for AD by age and sex. (Created from data from www.alz.org).

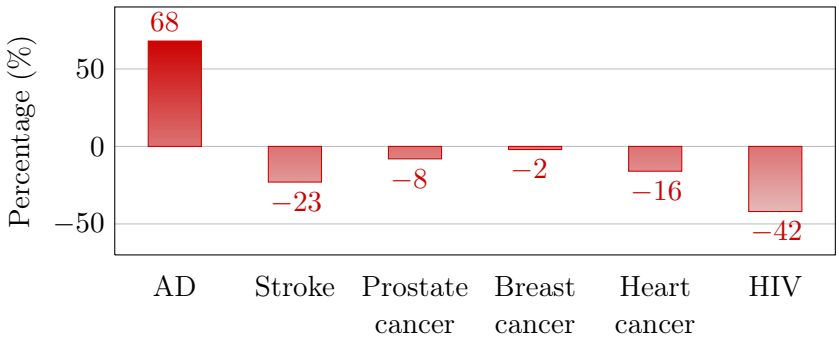


Figure 1.5: Percentage changes in selected causes of death (all ages) between 2000 and 2010. (Created from data from www.alz.org).

following section, we present some disease states that were introduced and several biomarkers that are widely studied in the scientific communities.

Definition 1.2.1 (Clinical disease stages in AD)**.** In the literature, three main clinical stages are commonly used to describe the progression of AD.

1. *Normal controls (NC)* patients show no signs of depression, mild cognitive impairment or dementia.

2. *Mild cognitive impairment (MCI)* is a brain function syndrome involving the onset and evolution of cognitive impairments beyond those expected based on the age and education of the individual, but which are not significant enough to interfere with their daily activities [Petersen 1999]. It is often found to be a transitional stage between normal aging and dementia. Although MCI can present with a variety of symptoms, when memory loss is the predominant symptom it is termed *amnestic MCI* and is frequently seen as a prodromal stage of AD [Grundman 2004].

3. *Alzheimer's disease (AD)* patients loss of mental function due to the deterioration of brain tissue. In several studies (such as ADNI, see Section 1.3.1), they have been evaluated and meet the NINCDS / ADRDA criteria for probable AD.

To determine these states or the transitions between these states, one needs to find accurate *biomarkers*.

Definition 1.2.2 (Biomarker)**.** A *biomarker*, or *biological marker*, is a substance, measurement or indicator of a biological state. Biomarkers may exist before clinical symptoms arise.

In 2010, Jack et al. introduced a hypothetical dynamic model of AD [Jack 2010]. The curves in Fig. 1.6 indicate temporal changes caused by five biomarkers of AD that were identified in this model.

1. *Amyloid beta* detected in CSF and PET amyloid imaging;

2. *Neurodegeneration* detected by rise of CSF tau species and synaptic dysfunction, measured via FDG-PET;

3. *Brain atrophy and neuron loss* measured with MR images (most notably in hippocampus, caudate nucleus, and medial temporal lobe);

4. *Memory loss* measured by cognitive assessment;

5. *General cognitive decline* measured by cognitive assessment.

Changes 1-3 are indicated by biomarkers that can be observed prior to a dementia diagnosis, whereas items 4-5 are the classic indicators of dementia diagnosis. For

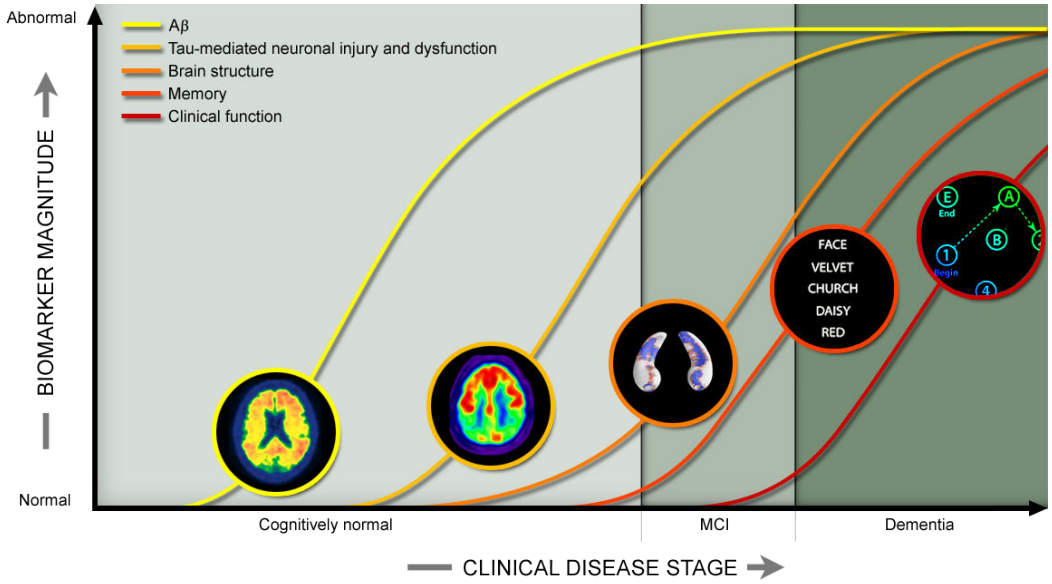


Figure 1.6: Biomarkers as indicators of dementia, as described in the hypothetical dynamic model introduced in [Jack 2010]. Illustration source: http://adni.loni.ucla.edu/study-design/background-rationale/

this reason, they are of particular interest for disease diagnosis or prediction. For example, MR imaging (e.g. T1-w) can be used to try to obtain structural information: loss of GM tissue in cortical and deep GM structures, WM degeneration, increased CSF space and enlarged ventricles (Fig. 1.8). Using PET PiB, we obtain biochemical information on increased amyloid load, neuronal loss and iron accumulation (Fig. 1.7).

The hypothetical model of [Jack 2010] received a lot of interest in the scientific communities and many studies tried to verify and evaluate it on large-scale datasets. More recently, this model was updated to take into account the latest evidence [Jack 2013].

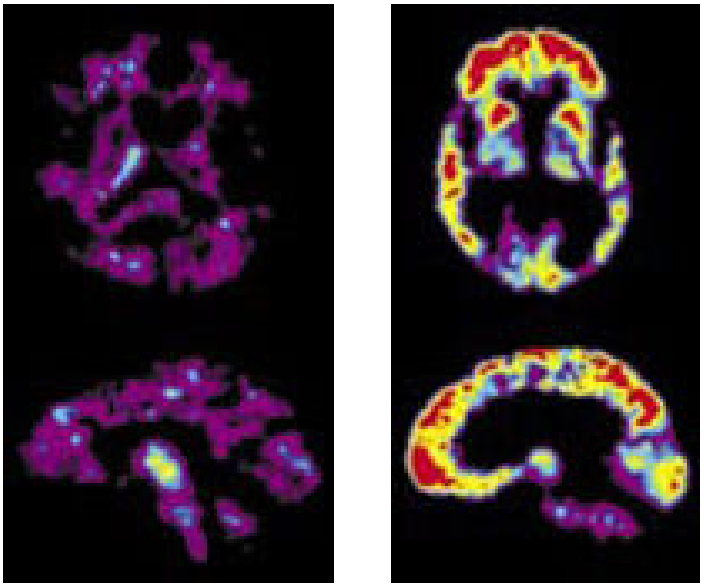
1.3 Databases used

1.3.1 Alzheimer's Disease Neuroimaging Initiative (ADNI)

The ADNI project[1] began in 2004 and was designed to find more sensitive and accurate methods to detect AD at earlier stages and mark its progress through biomarkers. The study gathered thousands of brain scans, genetic profiles, and biomarkers in blood and CSF in order to measure the progress of disease or the effects of treatment. The ADNI project is driven by several partners ans sponsors.

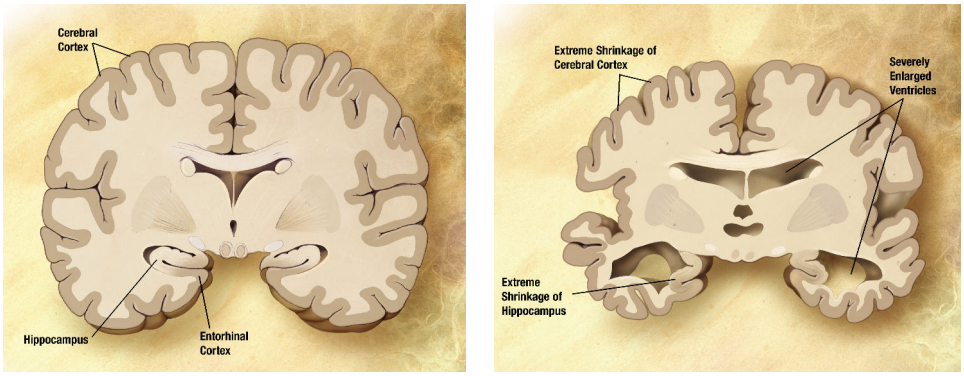
ADNI uses brain-imaging techniques, such as PET, including fluorodeoxyglucose (FDG)-PET (which measures glucose metabolism in the brain); F-18-PET using a

[1] http://www.adni-info.org/



(a) NC (b) AD

Figure 1.7: Detection of amyloid beta $A\beta$ using PET imaging



(a) Preclinical AD (b) Severe AD

Figure 1.8: Morphological brain changes related to AD

radioactive compound (Florbetapir ^{18}F which contains the radionuclide fluorine-18 (F-18)) that measures brain amyloid accumulation; and structural MR. One goal of the ADNI study is to track the progression of the disease using biomarkers to assess the brain's structure and function over the course of the disease states.

ADNI also aim to define biomarkers for use in clinical trials and to determine the best way to measure the treatment effects of AD therapeutics.

After the success of the first step (ADNI 1), ADNI 2 began in 2011. The goal of ADNI 2 is to identify who may be at risk for AD. ADNI now also aims to track the disease progression and define tests to measure the effectiveness of potential interventions.

Remark. The ADNI database is used in Chapters 7, 9 and 10.

1.3.2 Australian Imaging, Biomarker and Lifestyle Flagship Study of Ageing (AIBL)

The AIBL[2] study [Ellis 2009] aims to improve understanding of the causes and diagnosis of Alzheimer's disease, and helps develop preventative strategies. Launched on November 14^{th} 2006, the AIBL study is a prospective longitudinal study of aging comprised of patients with AD, MCI and NC. The study will help researchers develop and confirm a set of diagnostic markers biomarkers and psychometrics that can be used to objectively monitor disease progression and to develop hypotheses about diet and lifestyle factors that might delay the onset of this disease. Successful completion of this work will enable the design and conduct of extensive cohort studies that may lead to clinically proven preventative strategies for AD. The AIBL has four research streams

1. *Biomarkers* Blood samples have been taken from each participant for testing ranging from clinical pathology screening to novel biomarker examinations, to differentiate between those with and without AD.

2. *Clinical & Cognitive* A comprehensive neuropsychological assessment has been carried out and includes cognitive and mood tests, assessment of vital signs, collection of medical history (personal and family) and medication information, and questionnaires about lifestyle factors.

3. *Lifestyle* The lifestyle research stream is examining diet and exercise through questionnaires, monitoring and DEXA scans in a subset of participants.

4. *Neuroimaging* Participants undergo scans using the structural neuroimaging with MR Imaging and beta amyloid imaging with PiB PET methods.

Compared to ADNI, the AIBL database also provides FLAIR images, which are particularly interesting for the segmentation of WMH (see Chapter 5).

Remark. The AIBL database is used in Chapters 5 and 6.

[2] http://www.aibl.csiro.au

Chapter 2

Challenges in large-scale population studies

Contents

Résumé

Dans le Chapitre 1, nous avons introduit le contexte médical de cette thèse. En particulier, nous avons mentionné le rôle des technologies d'imagerie pour l'analyse de populations. Dans ce chapitre, nous décrivons brièvement les concepts d'*apprentissage automatique* utilisés pour construire des modèles prédictifs à partir d'un ensemble d'observations. Ensuite, nous expliquons certaines difficultés pouvant apparaître lorsque ces techniques sont appliquées en imagerie. Finalement, nous mentionnons plusieurs stratégies numériques couramment utilisées dans la littérature et dans les chapitres suivants.

Mots clés : Statistiques, apprentissage automatique, modèle prédictif, fonction de perte, minimisation du risque, malédiction de la dimension, classes déséquilibrées, sélection de modèle, sélection de paramètre, validation croisée, régularisation

Abstract

In Chapter 1, we introduced the medical background of this thesis. In particular, we mentioned the role of imaging technologies for population studies. In this chapter, we briefly describe the *machine learning* concepts used to build predictive models from a set of observations. Then we explain several challenges that can occur when these techniques are applied to imaging datasets. Finally, we mention several numerical strategies that are commonly used in the literature and in the following chapters.

Keywords: Statistics, Machine Learning, Predictive model, Loss function, Risk minimization, Curse of Dimensionality, Skewed Classes, Model Selection, Parameter Selection, Cross-validation, Regularization

2.1 Learning predictive models [1]

2.1.1 Position of the problem

We are given n pairs $\{(\mathbf{x}_i, y_i)\}_{1\leq i\leq n} \in \mathcal{X}\times\mathcal{Y}$, assumed to be independent realizations of a couple of random variables $(\mathbf{X}, \mathbf{Y})$ following the law $p_{\mathbf{XY}}$. The inputs $\mathbf{x}_i$ belong to the *input space* $\mathcal{X}$ which is typically $\mathbb{R}^d$. The outputs y_i belong to the *output space* $\mathcal{Y}$ which is typically a finite set or $\mathbb{R}$.

Example 2.1.1. In medical imaging, the $\mathbf{x}_i$ can be images (MR, PET, ultrasound, X-ray, ...), cognitive scores, genotype information, etc. The y_i can be diagnosis, disease progression, presence or absence of lesions/tumors, etc.

The goal of statistical learning is to predict the output y associated to a new input $\mathbf{x}$, assuming the pair $(\mathbf{x}, y)$ is a new realization of $(\mathbf{X}, \mathbf{Y})$ following $p_{\mathbf{XY}}$ and independent of the previous observations $\{(\mathbf{x}_i, y_i)\}_{1\leq i\leq n}$. To do so, one aims to build a *prediction function*.

Definition 2.1.1 (Prediction function or decision function)**.** A *prediction function* is a mesurable function $f\colon \mathcal{X} \to \mathcal{Y}$. We note $\mathcal{F}(\mathcal{X}, \mathcal{Y})$ the set of all prediction functions from $\mathcal{X}$ to $\mathcal{Y}$.

The main challenge in building an accurate prediction function arises from the fact that the law $p_{\mathbf{XY}}$ is generally unknown. The concept of *supervised learning* is to use the set of observations $\{(\mathbf{x}_i, y_i)\}_{1\leq i\leq n}$ (called *training set*) to build the prediction function. A *learning algorithm* is therefore a function that given a training set outputs a prediction function. Usually, a learning algorithm builds the prediction function via the evaluation of a loss function on a training set.

Definition 2.1.2 (Loss function)**.** A *loss function* is a function $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ which is used to assess the quality of a predicted output compared to the real output.

Definition 2.1.3 (Ground truth)**.** The real output used to evaluate the performance of a learning algorithm is called the *ground truth*.

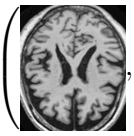
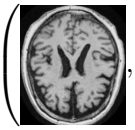
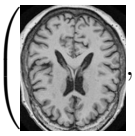
Definition 2.1.4 (Classification)**.** When $\operatorname{card}\mathcal{Y} < \infty$, the elements of $\mathcal{Y}$ are called *classes* and the learning problem is referred as a *classification* problem.

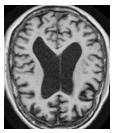
Example 2.1.2 (Loss function for classification)**.** When solving a classification problem, one loss function that can be used is

$$\ell^{0/1}(y, y') \overset{\text{def.}}{=} \begin{cases} 1 & \text{if } y \neq y', \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

[1] This section is inspired by the class "Introduction to statistical learning" by Jean-Yves Audibert from the MSc "Math, Vision, Learning" of ENS Cachan, France.

Example 2.1.3 (Disease classification problem in medical imaging)**.** In medical imaging, given a set of couples image-diagnosis $\{(\mathbf{x}_i, y_i)\}_{1 \leq i \leq n} =$

$\left\{ \left(\right.\right.$  $, AD \left.\right) , \left(\right.$  $, AD \left.\right) , \ldots , \left(\right.$  $, NC \left.\right) , \left(\right.$  $, NC \left.\right) \right\} ,$

building a disease classifier consists in building a function predicting the diagnosis of a new image, such as  .

Definition 2.1.5 (Regression)**.** When $\operatorname{card} \mathcal{Y} = \infty$ (typically $\mathcal{Y} = \mathbb{R}^n$), the prediction problem is called a *regression* problem.

Example 2.1.4 (L_p regression)**.** When $\mathcal{Y} = \mathbb{R}$ and the loss function

$$\ell^p(y, y') \stackrel{\text{def.}}{=} |y - y'|^p \,, \tag{2.2}$$

where $p \geq 1$, the problem is called L_p regression. In the case $p = 2$, it is also called *least-square regression.*

Definition 2.1.6 (Risk or generalization error)**.** The *risk* of a prediction function $f \colon \mathcal{X} \to \mathcal{Y}$ is defined by

$$R(f) \stackrel{\text{def.}}{=} \mathbb{E}\left[\ell\left(\mathbf{Y}, f(\mathbf{X})\right)\right] \tag{2.3}$$

$$= \int_{\mathcal{X} \times \mathcal{Y}} \ell(f(\mathbf{x}), y) dp_{\mathbf{XY}}(\mathbf{x}, y) \,. \tag{2.4}$$

where $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is a loss function.

Definition 2.1.7 (Target function)**.** A target function is a prediction function minimizing the risk.

Remark. A target function does not necessarily exist. The Theorem 2.1.1 gives sufficient conditions for this target function to exist, its proof is given in Appendix A.1.1.

Theorem 2.1.1. *Let us assume that for all* $\mathbf{x} \in \mathcal{X}$ *the infimum* $\inf_{y \in \mathcal{Y}} \mathbb{E}_{p_{\mathbf{XY}}(dY|\mathbf{x})} \ell(Y, y)$ *is reached. Then a function* $f^* \colon \mathcal{X} \to \mathcal{Y}$ *such that, for all* $\mathbf{x} \in \mathcal{X}$ $f^*(\mathbf{x})$ *is a minimizer of* $y \mapsto \mathbb{E}_{p_{\mathbf{XY}}(dY|\mathbf{x})} \ell(Y, y)$*, is a target function.*

$$\forall \mathbf{x} \in \mathcal{X}, \quad f^*(\mathbf{x}) \in \operatorname*{argmin}_{y \in \mathcal{Y}} \mathbb{E}_{p_{\mathbf{XY}}(dY|\mathbf{x})} \ell(Y, y) \quad \Rightarrow \quad f^* \in \operatorname*{argmin}_{\mathcal{F}(\mathcal{X}, \mathcal{Y})} R \,. \tag{2.5}$$

The previous theorem gave some sufficient conditions for a target function to exist in the general case. The next theorem gives a target function in the case of the least-square regression, as well as the excess of risk for another prediction function. Its proof is given in Appendix A.1.2.

Theorem 2.1.2 (Target function in least-square regression)**.** *In least square regression, the function*

$$\eta^* \colon \mathbf{x} \in \mathcal{X} \mapsto \mathbb{E}[\mathbf{Y}|\mathbf{X} = \mathbf{x}] = \int_{\mathcal{Y}} y \; dp_{\mathbf{XY}}(y|\mathbf{x}) \tag{2.6}$$

is a target function. Besides

$$\forall \eta \colon \mathcal{X} \to \mathbb{R}, \quad R(\eta) = R(\eta^*) + \mathbb{E}(\eta - \eta^*)^2 \,, \tag{2.7}$$

where $\mathbb{E}(\eta - \eta^*)^2 = \int_{\mathcal{X}} (\eta(\mathbf{x}) - \eta^*(\mathbf{x}))^2 dp_{\mathbf{X}}(\mathbf{x})$.

Now let us see a similar theorem that gives the target functions in classification with the $\ell^{0/1}$ loss. First we define the sign function which is convenient for binary classification, and then give the theorem, which proof is given in Appendix A.1.3.

Definition 2.1.8 (Sign function)**.** The function $\text{sign} \colon \mathbb{R} \to \{-1, +1\}$ is defined by

$$\text{sign}(y) \overset{\text{def.}}{=} \begin{cases} +1 & \text{if } y \geq 0, \\ -1 & \text{otherwise.} \end{cases} \tag{2.8}$$

Theorem 2.1.3 (Target function in classification)**.** *In classification with the* $\ell^{0/1}$ *loss, the target functions are the functions* f^* *such that*

$$\forall \mathbf{x} \in \mathcal{X}, \quad f^* \in \underset{y \in \mathcal{Y}}{\text{argmax}} \; p_{\mathbf{XY}}(\mathbf{Y} = y|\mathbf{X} = \mathbf{x}) \,. \tag{2.9}$$

Moreover in binary classification ($\mathcal{Y} = \{-1, +1\}$)*, the function*

$$f^* \colon \mathbf{x} \mapsto \text{sign}(\eta^*(\mathbf{x})) \tag{2.10}$$

is a target function, where η^* *is a target function for the least-square regression.*

So far, we have seen that to build a predictive model, it is convenient to define a loss function to evaluate how accurate is a prediction for a given $\mathbf{x}$. To evaluate the accuracy of a prediction function f, we have defined the risk as the average loss of f given the law $p_{\mathbf{XY}}$.

2.1.2 Empirical risk minimization

Let us recall the definition of the risk of a prediction function $f \colon \mathcal{X} \to \mathcal{Y}$

$$R(f) \overset{\text{def.}}{=} \mathbb{E}[\ell(\mathbf{Y}, f(\mathbf{X}))] = \int_{\mathcal{X} \times \mathcal{Y}} \ell(f(\mathbf{x}), y) dp_{\mathbf{XY}}(\mathbf{x}, y) \,. \tag{2.11}$$

As mentioned below, we are looking for a target function, i.e. a prediction function minimizing the risk. However, the probability $p_{\mathbf{XY}}$ is unknown, and therefore both the risk and target function are unknown. Nonetheless the risk $R(f)$ can be estimated empirically. We introduce the notion of *empirical risk* as a empirical estimation of the risk on a training set.

Definition 2.1.9 (Empirical risk)**.** The *empirical risk* of a prediction function $f\colon \mathcal{X} \to \mathcal{Y}$ is given by

$$r(f) \overset{\text{def.}}{=} \frac{1}{n}\sum_{i=1}^{n} \ell(f(\mathbf{x}_i), y_i)\,, \tag{2.12}$$

where $\{(\mathbf{x}_i, y_i)\,; \quad 1 \leq i \leq n\}$ is a training set.

Remark. The empirical risk can also be called *fitting loss* and noted $\mathcal{L}_{fit}(f)$.

Definition 2.1.10 (Empirical risk minimization ERM)**.** An *empirical risk minimization (ERM)* algorithm is an algorithm that aims to find a function f_{ERM} such that

$$f_{ERM} \in \underset{f\in\hat{\mathcal{F}}}{\operatorname{argmin}}\, r(f)\,, \tag{2.13}$$

where $\hat{\mathcal{F}} \subset \mathcal{F}(\mathcal{X}, \mathcal{Y})$ is a subset of prediction functions from $\mathcal{X}$ to $\mathcal{Y}$.

The idea to learn prediction function from ERM algorithm were studied in the work of Vapnik and Chervonenkis [Vapnik 1995]. Following this idea, it is natural to wonder if minimizing the empirical risk minimizes the real risk. In fact, choosing $\hat{\mathcal{F}} = \mathcal{F}(\mathcal{X}, \mathcal{Y})$ often leads to *overfitting*, as the resulting algorithm can have an empirical risk much lower than the real risk. In practice, one needs to choose $\hat{\mathcal{F}}$ large enough to have a good prediction but not too large to avoid overfitting. The "size" of $\hat{\mathcal{F}}$ is sometimes called *capacity* or *complexity*. This was further studied and quantified by Vapnik and Chervonenkis who introduced the notion of Vapnik-Chervonenkis dimension. Alternatively, one can add a regularization term to $r(f)$ to smooth the prediction function, as in the SVM algorithm.

Example 2.1.5 (A regularized ERM algorithm: the support vector machines (SVM))**.** The SVM are a machine learning algorithm where the loss function is the hinge loss, and the prediction functions are searched in a reproducing kernel Hilbert space (RKHS). This algorithm is described in details in Chapter 5.

Remark (Bias-variance trade-off). Now let us get back to the analysis of the real risk of an ERM prediction function. Let $\hat{f}$ be a minimizer of the risk on $\hat{\mathcal{F}}$, i.e.

$$\hat{f} \;\in\; \underset{f\in\hat{\mathcal{F}}}{\operatorname{argmin}}\, R(f). \tag{2.14}$$

Since $R(f_{ERM}) \geq R(\hat{f}) \geq R(f^*)$ (proof in Appendix A.1.5), the excess of risk of the empirical risk minimizer f_{ERM} (compared to the target prediction function) can be

decomposed into two positive terms

$$R(f_{ERM}) - R(f^*) = \underbrace{R(f_{ERM}) - R(\hat{f})}_{\text{Estimation error (bias)}} + \underbrace{R(\hat{f}) - R(f^*)}_{\text{Approximation error (variance)}} . \tag{2.15}$$

The larger $\hat{\mathcal{F}}$, the lower the approximation error, but generally the higher is the estimation error. As mentioned before, there is therefore a trade-off to find when choosing $\hat{\mathcal{F}}$, often called *bias-variance trade-off*. The term "variance" relates to the link between the estimation error and the variability of the training set, which has been assumed to be a realization of independent identically distributed random variables. A high-variance model is prone to overfitting whereas a high-bias model is prone to *underfitting*. The bias can be bounded as described in Proposition 2.1.4, proof is in Appendix A.1.4.

Proposition 2.1.4 (Bound on the estimation error)**.** *The estimation error (bias) can be bounded by*

$$R(f_{ERM}) - R(\hat{f}) \leq 2 \times \sup_{f \in \hat{\mathcal{F}}} |R(f) - r(f)| . \tag{2.16}$$

2.1.3 Local averaging

In Section 2.1.1, we have seen that the accuracy of a prediction function can be evaluated via the risk, measuring the average loss on a function assuming a probability distribution $p_{\mathbf{XY}}$. Section 2.1.2 presented ERM algorithms, based on an empirical estimation of the risk on a training set. An alternative strategy is to assume that a good prediction function is locally smooth, and therefore its value at one point is similar to the values in a neighborhood. To illustrate the idea of local averaging method, let us consider the least-square regression problem on $\mathcal{X} = \mathbb{R}^d$ with $\mathcal{Y} = [-B, B]$ with $B > 0$. The loss function is $\ell(y, y') \overset{\text{def.}}{=} (y - y')^2$. We have seen (Theorem 2.1.2) that $\eta^* \colon \mathbf{x} \mapsto \mathbb{E}(\mathbf{Y}|\mathbf{X} = \mathbf{x})$ is a target function. As the probability law $p_{\mathbf{XY}}$ is unknown, one has to estimate $\mathbb{E}(\mathbf{Y}|\mathbf{X} = \mathbf{x})$. One strategy is to average the y_i corresponding to the $\mathbf{x}_i$ close to $\mathbf{x}$. We consider learning algorithms of the form

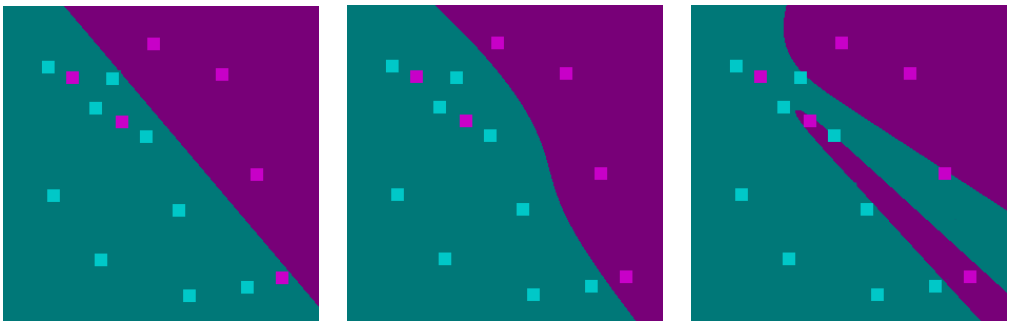
$$\hat{\eta} \colon \mathbf{x} \mapsto \sum_{i=1}^{n} w_i(\mathbf{x}) y_i, \tag{2.17}$$

where $(w_i(\mathbf{x}))_{1 \leq i \leq n} \in \mathbb{R}^n$ are weighting coefficients.

Example 2.1.6 (k nearest neighbours (kNN) algorithm)**.** For $k \in \mathbb{N}$, the kNN-algorithm considers the weights

$$w_i^{kNN}(\mathbf{x}) \overset{\text{def.}}{=} \begin{cases} \frac{1}{k} & \text{if } \mathbf{x}_i \text{ belongs to the kNN of } \mathbf{x}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.18}$$

Remark. The kNN-algorithm will be used in Chapter 5 and 7.



(a) d=1 (b) d=3 (c) d=10

Figure 2.1: SVM classification in $\mathbb{R}^2$ with a polynomial kernel, depending on the degree. When d gets too high, the algorithm overfits the training data. (Figure generated from `http://www.csie.ntu.edu.tw/~cjlin/libsvm/` with $c = 100$ and $\gamma = 10$).

2.2 Challenges

2.2.1 Model selection

Machine learning problems can usually be solved with countless models of various complexity. On one hand, though a simple model might be easy to train, it might not model the full range of possibilities. On the other hand, a more complex can lead to overfitting.

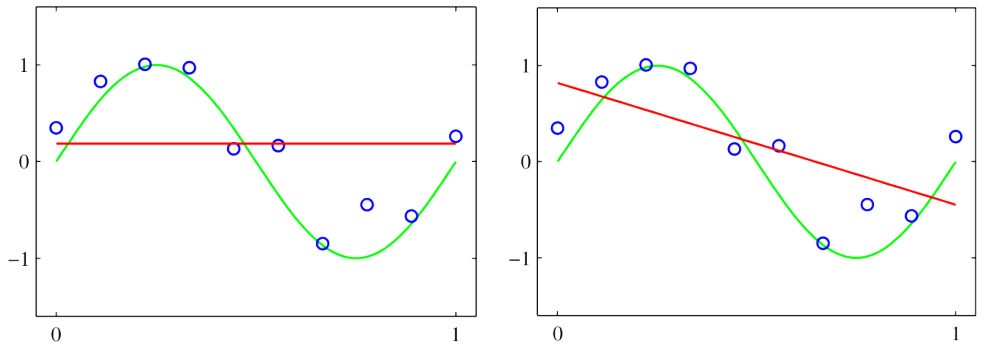
Overfitting typically occurs if the model is too complex with regard to the problem to solve (for example if it has too many parameters), if the training data is very noisy, etc. It leads to a classifier or regression function that has poor predictive performance, despite an error that can be low on the training set. In Fig. 2.1, a 2D classification problem is solved using SVM with a polynomial kernel. We notice that when the degree of the polynomial gets too high, the decision boundary suffers from strong distortions: it overfits the training data. In Fig. 2.2, a curve fitting problem is solved using least-square polynomial fitting. Similarly, when the degree gets too high, the curve is not properly approximated. To avoid or at least try to limit overfitting, several strategies are available, such as regularization, early-stopping, model priors, etc.

Remark (Robust learning). When a learning algorithm does not tend to fit noise or be sensitive to outliers, it is said to be *robust*.

2.2.2 Curse of dimensionality

When learning statistical models in spaces of high dimension, several issues can occur. These issues are said to be caused by the *curse of dimensionality*.

Designing a model for statistical learning in high dimensional can be hard, as our geometric intuitions are naturally biased. For example in Fig. 2.3, we consider



(a) d=0 (b) d=1

(c) d=3 (d) d=9

Figure 2.2: Curve fitting: in green: the true curve, in blue: the (noisy) observations, in red: the estimation using a polynome of degree d. Source: [Bishop 2007].

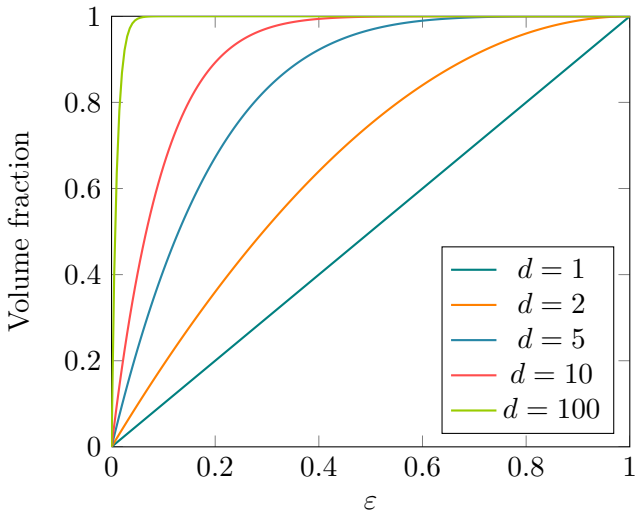


Figure 2.3: Fraction of the volume of a sphere lying in the radius range $1 - \varepsilon$ to 1, depending on the space dimension d. (Figure inspired by [Bishop 2007]).

a sphere in $\mathbb{R}^d$ and plot the fraction of the volume contained between radius 1 and radius $1 - \varepsilon$. We notice that the higher the dimension, the more the volume of a sphere is concentrated in a thin shell near the surface.

When the number of observations is much lower than the dimensionality of the space, a sampling issue occurs. Moreover, convergence proofs when the number of samples is going to infinity are no longer relevant. When the number of parameter of a model gets high, its optimization becomes complex (see Fig. 2.4).

Nonetheless, learning can still be performed in space of high dimension. The primary reason is that the intrinsic space of data is generally much lower of the dimension of the space, (this is one of key motivations for manifold learning: see Chapters 4, 6, and 7).

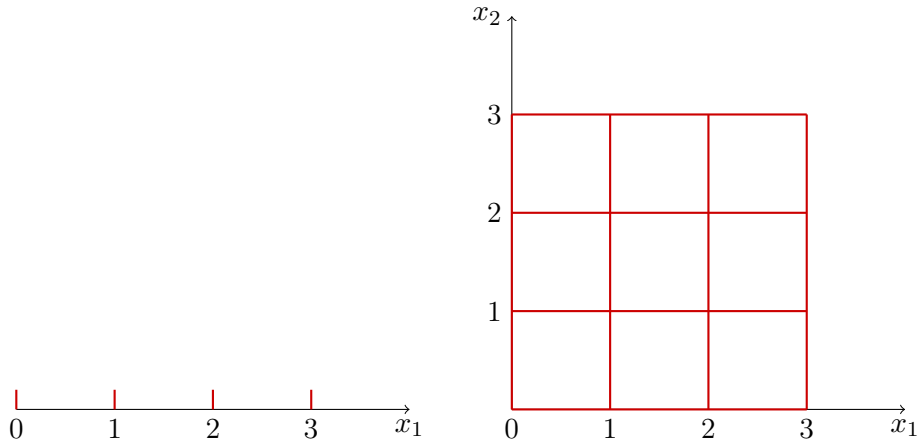
2.2.3 Classification with unbalanced training sets

In real-life applications, it sometimes happens that the training set contains much more observations belonging to one class than to the other one. In that case, the classes are said to be *skewed* and the dataset *unbalanced*. For example, in the context of AD, databases such as ADNI contain more stable MCI patients than progressive MCI ones. When a training dataset is unbalanced, a learning algorithm generally tend to be biased towards the majority class. To avoid this, several strategies are available, such as

1. forcing the training set to be balanced (even if that means not using all the data available),

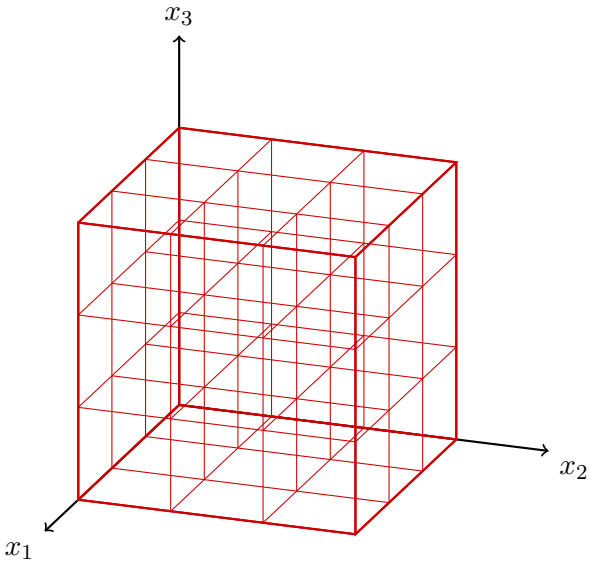
2. modifying the cost function in the optimization (e.g. by introducing weights).

Remark. These strategies will be discussed in Chapter 10.



(a) $d = 1$

(b) $d = 2$



(c) $d = 3$

Figure 2.4: Illustration of the curse of dimensionality: the number of regions of a regular grid grows exponentially with the dimension d. (Figure inspired by [Bishop 2007]).

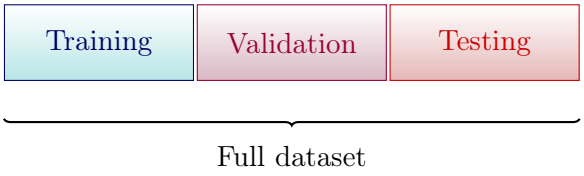


Figure 2.5: In order to properly estimate the performance of a model, a dataset is usually split into two or three parts: (1) a training set, (2) a validation set (optional), and (3) a testing set.

It is also important to notice that some methods are more robust and tend to be less biased than others. For example, the solution of an SVM classifier only depends on a subset of training vectors (the so-called *support vectors*). In that case, it is possible to create a dataset and add a large number of training examples of the same class without changing the solution of the algorithm.

2.3 Numerical strategies

2.3.1 Splitting the dataset

In machine learning, to properly estimate the quality of a model, it is important to split the dataset into several parts. As illustrated on Fig. 2.5, two or three parts are generally used

1. a training set,
2. an optional validation set, and
3. a testing set.

In the simplest case, the model is built on the *training set*, and the performance is then evaluated on the *testing set*. This splitting avoids an overestimation of the performance that usually happens when the same observations are used to build the model and to evaluate its performance.

Now when the model contains some parameters that need to be set, the performance of the model on the testing dataset should not be used to select the optimal parameters. Once again, doing so would overestimate the performance of the model. Instead, another subpart must be used to select the optimal parameters: the *validation set*.

2.3.2 Cross validation

Cross validation is a standard procedure that can be used for parameter selection. As mentioned in Section 2.3.1, the dataset needs to be split into three parts when one aims to find the optimal parameters automatically and evaluate the performance of a model. However, when the dataset does not contain a lot of observations, splitting

it into three subparts could lead to an wrong estimation of the performance of the model (e.g. underestimation if the training set is not large enough, unstable estimation if the validation or testing sets are not large enough, etc). Instead, the key concept of cross-validation is to split the dataset into training and testing sets and then split the training set into sub-training and sub-validation sets. Using these sets, the performance is evaluated for different parameters over several runs and averaged. The parameters providing the best performance are then selected. Using these parameters, the classifier is built on the whole training set. By doing so, the parameters of the model are selected automatically *without making use of the testing set*. The cross validation process is explained in details in Fig. 2.6.

Remark. The same splitting scheme can be applied on the full dataset, so that the performance is evaluated over the whole dataset (performance is averaged over several runs). This is particularly important when the number of observations on the whole dataset is low (for example as in Chapters 9 and 10).

Remark. Selecting the number of runs n_c is a trade-off between the accuracy of the parameter selection and the computational cost. A high n_c value provides more observations for the training steps, more runs, therefore a higher accuracy of the performance with the selected parameters. However, the computational cost is approximately linear in n_c.

2.3.3 Regularizations

A common practice when learning model with a high number of parameters is to introduce regularizations. Different regularizations can enforce different properties in the models. For example, a widely studied class of regularizations are the ones enforcing sparsity.

Now let us see an example where regularizations can be introduced in order to be able to solve the model. The linear inverse problem consists in finding a solution $\mathbf{w} \in \mathbb{R}^p$ solution of

$$\mathbf{y} = \mathbf{X}\mathbf{w}\,, \tag{2.19}$$

where $\mathbf{y} \in \mathbb{R}^n$ represent (noisy) observations and $\mathbf{X} \in \mathbb{R}^{n \times p}$ is a linear operator. When $p = n$ and $\mathbf{X}$ of full rank, the solution can be directly computed. In that case, the problem is said to be *well-posed*, which means (as defined by Hadamard) that its solution exists, is unique, and depends continuously on its input data. Let us take one example where there is a unique solution, but the system is *ill-conditioned*.

Example 2.3.1 (Ill-conditioned problem, taken from [Ciarlet 1988])**.** Let us consider the linear system

$$\begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix} = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \mathbf{w}, \text{ of solution } \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \tag{2.20}$$

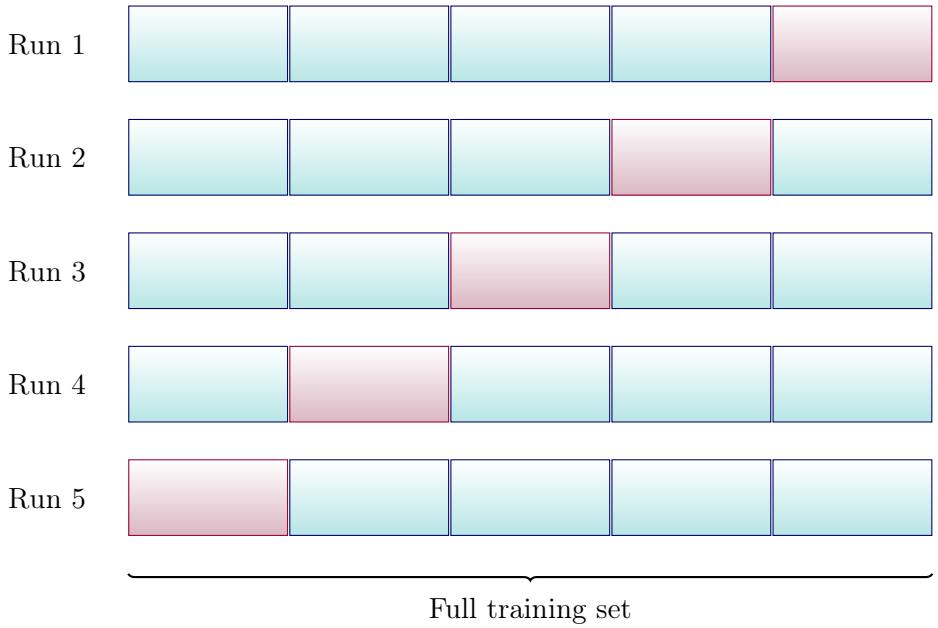


Figure 2.6: The n_c-fold cross-validation is composed of the following steps: (1) after random re-ordering, the training set is split into n_c parts, (2) a *(sub)-training set*, composed of $n_c - 1$ parts (in blue in the Figure), is used for building the model, (3) a *(sub)-validation set*, composed of the remaining part (in red in the Figure), is used for validating the model, (4) the steps (2)-(3) are repeated n_c times and performance is averaged over the different runs. The figure illustrates this concept for $n_c = 5$.

and the system with a small perturbation of in $\mathbf{y}$

$$\begin{pmatrix} 32.1 \\ 22.9 \\ 33.1 \\ 30.9 \end{pmatrix} = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \mathbf{w}, \text{ of solution } \begin{pmatrix} -9.2 \\ -12.5 \\ 4.5 \\ -1.1 \end{pmatrix}. \tag{2.21}$$

Both the original and modified system are linear, and have a unique solution ($\mathbf{X}$ is of determinant 1). However, we notice that a "small" perturbation in $\mathbf{y}$ generates a "large" perturbation in the solution.

We saw in the previous example that even with existence and uniqueness of the solution, the system can be unstable. When $p > n$, there is no uniqueness of the solution, and the problem is ill-posed. A solution of such an ill-posed problem can be computed by minimizing

$$\frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda J(\mathbf{w}), \tag{2.22}$$

where $\lambda \geq 0$ is a regularization coefficient and J a regularization function. The idea of using regularizations to solve ill-posed problems was studied (among others) by Tikhonov [Tikhonov 1943, Tikhonov 1963, Tikhonov 1977]. A classical regularization is the ridge [Hoerl 1970] penalizing the squared ℓ_2 norm of $\mathbf{w}$. The LASSO [Tibshirani 1994] uses ℓ_1 regularization to force a lot of coefficients to be (close to) zero. The *elastic net* [Zou 2005] uses a combination of both terms. More recently, non-local regularizations [Peyré 2011] were used to solve several linear inverse problems in image processing such as inpainting, super-resolution, compressed sensing, etc.

Chapter 3

Contributions of this thesis

Contents

Résumé

Dans ce chapitre, nous présentons dans un premier temps différentes motivations cliniques de notre domaine. Ceci nous permet de mieux situer les travaux effectués dans cette thèse. Ensuite, nous mentionnons certains aspects méthodologiques, liés aux challenges présentés dans le chapitre 2. Finalement, nous présentons les questions auxquelles nous avons souhaité répondre ainsi que nos contributions dans les parties II et III. Les communications écrites et orales associées à nos travaux sont également listées.

Abstract

In this chapter, we present first several clinical motivations in our domain. This enables us to understand better where stands the work we

present in this thesis. Then we mention some methodological aspects related to the challenges presented in Chapter 2. Finally, we present the problems we wanted to address as well as our contributions in Parts II and III. The oral and written communications associated with the work of this thesis are also listed.

3.1 Clinical motivations

The work presented in this thesis was realized in partnership between the CEREMADE laboratory from Paris Dauphine University in France, the Australian e-Health research centre / CSIRO in Brisbane in Australia, and the Mathematics Institute of Toulouse in France. The clinical context of this work is the study of Alzheimer's disease (AD). In this section, we present first several clinical motivations in our domain, in order to have a better understanding of the context of our work, which is presented later.

As described in Chapter 1, AD was discovered in the early twentieth century by Aloïs Alzheimer. Her first diagnosed patient had symptoms including memory and understanding deficits, aphasia (loss of the language), and the loss of the sens of directions. AD is nowadays widely spread, and affects a large proportion of the world population. Many scientific studies aim to analyze several aspects of the disease.

3.1.1 Creation of diagnostic tools

A first class of studies aim to build a disease progression model. The creation of such a model can have several goals

- identify the steps of the disease progression,
- find trends in populations.

The knowledge of the typical steps in the progression of the disease can help a clinician understand the clinical state of a patient. He might as well be able to inform the patient of his/her most probable evolution. However, there is a variability in the possible evolutions. For example, some mild cognitive impairment (MCI) patients convert to AD while others are stable. It is therefore important to analyze the trends in populations, and try to understand why some patients have different evolutions. To evaluate the possible states or evolutions of patients, several steps are required, such as

- the definition of several clinical states,
- the creation of diagnostic tools.

As we saw in Chapter 1, normal control (NC), MCI and AD are usual clinical states that are studied in the AD context. Now diagnostic tools aim at associating a clinical state to a patient based on other observed information. In the case of AD, several types of data can be used for the creation of such tools: measures from blood samples, results of cognitive tests, lifestyle indicators, imaging data, etc. It is interesting to evaluate the quantity of information contained in each type of data, as well as defining methods to be able to combine them. Indeed by combining various types of data, one can hope to increase the performance of the diagnostic tools, as

the latter would have more information available. As we saw in Chapter 2, machine learning methods can be used to build such diagnostic tools.

Population studies can be categorized into

1. cross-sectional, and
2. longitudinal studies.

Cross-sectional studies analyze populations at a single time point, whereas longitudinal studies analyze evolutions between several time points.

Remark. In this thesis, we study the combination of various types of data (which can be used for diagnosis classification, see Chapter 7), the creation of tools to classify disease progression (see Chapters 9 and 10).

3.1.2 Identification and quantification of biomarkers

A second class of studies aim to identify and quantify *biomarkers*. Biomarkers, or biological markers, are substances, measures or indicators of a clinical state. Biomarkers may exist before clinical symptoms arise. We saw in Chapter 1 that Jack et al. introduced in 2010 a hypothetical of AD, where several biomarkers can be used to evaluate the disease progression of a patient [Jack 2010]. Among the biomarkers that one can try to characterize with images, we can cite

- structural biomarkers,
- functional biomarkers.

The search for structural biomarkers is based on the assumption that the disease affects the anatomical brain structures. For example, the hippocampus is one the most studied structures. The search for functional biomarkers is based on the assumption that the disease affects the functional and neuronal activities of the brain. Many scientific studies were published in order to try to verify the validity of [Jack 2010] model. This model has been recently updated [Jack 2013].

Remark. In this thesis, we study methods to identify structural biomarkers of MCI-to-AD conversion from binary hippocampus images (see Chapters 9 and 10).

3.1.3 Identification and quantification of risk factors

A third class of studies aim to identify and quantify *risk factors*. As explained in Chapter 1, a risk factor is a variable correlated with a disease. By definition, it is not necessarily a cause of the disease. Nonetheless, understanding the risk factors can help to develop preventive treatments.

In the case of AD, several potential risk factors are studied in the literature. One can cite age, sex, genetic factors (e.g. alleles of the gene coding the apolipoprotein E (ApoE)), white matter hyper-intensities (WMH), etc. In order to identify if these potential risk factors have a real impact on the disease progression, it is important to be able to quantify them in an accurate, fast, fully-automatic and reproducible way.

Remark. In this thesis, we study the WMH segmentation, a risk factor for AD (see Chapter 5).

3.1.4 Exploratory data analysis

A fourth class of studies is called *exploratory*. Exploratory data analysis (EDA) differs in the way of using and analyzing data. In classical analysis, the process is

$$\text{Problem} \rightarrow \text{Data} \rightarrow \text{Model} \rightarrow \text{Analysis} \rightarrow \text{Conclusions}.$$

In EDA, the process is different

$$\text{Problem} \rightarrow \text{Data} \rightarrow \text{Analysis} \rightarrow \text{Model} \rightarrow \text{Conclusions}.$$

EDA was in particular initiated by [Tukey 1962]. Its aim objectives are

- discover underlying structures,
- identify important variables,
- detect aberrant and abnormal data (i.e. outliers),
- test data-driven hypotheses,
- develop minimal models,
- etc.

In medical imaging, and in particular in order to analyze populations, these techniques can be useful to identify trends and make hypotheses about regions of interest potentially related to progression of diseases. One should note that visualization techniques are important tools for EDA.

Remark. In this thesis, we study manifold learning techniques for visualization and the identification of potential zones of interest (see Chapters 6 and 7).

3.2 Methodological aspects

In this thesis, we use several methods and strategies to analyze populations.

3.2.1 Construction of statistical models

As we mentioned in Chapter 2, a large number models were proposed in the field of statistical and machine learning. The key idea is simple: it consists in predicting some information about a patient given the knowledge of previously observed patients. A classical example is the classification one, where for example one can aim to predict the diagnosis given the observation of the image of the patient. However, several challenges occur when one tries to apply these techniques on datasets of medical images (curse of dimensionality, curse of big data, skewed classes, etc). In this thesis, we study several strategies to face these challenges.

3.2.2 Definitions of descriptors and distances

The use of statistical and learning models requires

1. the definition of features (i.e. descriptors),
2. the definition of an appropriate distance.

In this thesis, we study descriptors encoding information at several levels: at the voxel level (i.e. at the anatomical neighborhood), at the patient level (i.e. at a single time point), and the evolution of the patient level.

In the case where descriptors are at the patient level, no matter if it is a cross-sectional or longitudinal study, it is possible to measure distances between patients by using an Euclidean distance, or by building more complex distances, based on deformations. We show that using a "simple" Euclidean distance already enable us to obtain insight on the populations. In the literature, many models were introduced in the field of *computational anatomy*. We study in particular algorithms for atlas creation, which enable us to represent patients in the same space, and the creation of descriptors from the large deformation diffeomorphic metric mapping (LDDMM) framework.

3.2.3 Number of parameters and dimensionality reduction

Dimensionality reduction is a popular technique to reduce the dimensionality of a system. In the case of image processing, it can be used to

- facilitate the visualization in order to do exploratory data analysis,
- regularize learning algorithms.

In this thesis, we use non-linear dimensionality reduction for visualization and EDA (see Chapters 6 and 7). We also realized experiments using Laplacian support vector machines (SVM) [Belkin 2004]. The results were interesting on synthetic data, though as our tests on real data did not outperform the ones obtained using SVM, they are not presented here.

3.2.4 The importance of regularizations

Many medical imaging problems (registration, segmentation, learning, etc) can be formulated with a variational form. The solution is expressed as a minimizer (in a certain space) of a functional. To control the smoothness of the solution, regularizations are introduced.

Moreover, as we saw in Section 2.3.3, problems can be *ill-posed*, by opposition of the notion of *well-posed* problem of Hadamard. This is typically the case when one wants to optimize a number of parameters that is higher than the number of observations. We saw that finding the solution $\mathbf{w} \in \mathbb{R}^p$ of the linear inverse problem

$$\mathbf{y} = \mathbf{X}\mathbf{w}\,, \tag{3.1}$$

where $\mathbf{y} \in \mathbb{R}^n$ and $\mathbf{X} \in \mathbb{R}^{n \times p}$, can be ill-posed.

In Chapter 10, we use a similar linear prediction model

$$\mathbf{y} \overset{\text{def.}}{=} F(\mathbf{X}\mathbf{w} + b) \ , \tag{3.2}$$

where $\mathbf{y} \in \{\pm 1\}^n$ is the behavioral variable, $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the design matrix containing n observations of dimension p, F is the prediction function and $(\mathbf{w}, b) \in \mathbb{R}^p \times \mathbb{R}$ are the parameters to estimate. We use this model to build disease progression classifiers from initial momenta encoding the evolutions of patients. To solve the model efficiently, we evaluate standard regularizations and also introduce several spatial regularizations.

3.3 Contributions of this thesis

3.3.1 Objectives

In Chapter 1, we introduced the clinical context of AD and explained why medical imaging can be beneficial for large-scale population studies. However, as mentioned in Chapter 2, the use of statistical learning algorithms in medical imaging raise several challenges.

In this thesis, we aim at building models to analyze both cross-sectional and longitudinal populations. In particular, we want to answer questions raised in several areas:

- **White matter hyper-intensities (WMH) segmentation**:
  - How to define a pipeline for WMH segmentation?
  - Is it worth acquiring several modalities?
  - Which local features should be used?
  - How to define an appropriate region of interest (ROI)?
- **Manifold learning**:
  - Can it be used to find shape/appearance trends in populations?
  - How to improve the low-dimension representation of a patient using clinical information?
- **Longitudinal population analysis**:
  - Can initial momenta from the LDDMM framework be used to model patient evolutions?
  - Do these local descriptors outperforms global descriptors such as volume variation?
  - What template algorithm can/should be used?
  - How to transport the tangent information?

  - What are good classification strategies (in terms of model, regularization, optimization, etc.)?

In the next paragraph, we present the different contributions.

3.3.2 Contributions

In Part II, we focus on cross-sectional population analysis. In this setting, a population of patients is under study, considering *one temporal point by patient* (which means one image by patient).

In Chapter 4, we review several machine learning algorithms and various applications in medical imaging such as image registration, image segmentation, population analysis and machine learning.

In Chapter 5, we address the WMH segmentation problem by building classifiers from local features. Our contributions in this chapter are:

- the introduction of a new segmentation pipeline based on SVM, including in particular the definition of a new ROI,
- the validation on a large database from the Australian imaging biomarker and lifestyle (AIBL) study,
- the evaluation of the relative contribution of each modality,
- the evaluation of the classification performance for different feature types,
- the comparison with other supervised classification algorithms.

In Chapter 6, we consider the use of non linear dimensionality reduction (NLDR) algorithms for population analysis. Our contributions is this chapter are:

- the review of several NLDR methods,
- the application in medical imaging for single- and multi-modality images to visualize trends in populations.

In Chapter 7, we address the question of using clinical data to improve low-dimensional patient representation. Our contributions in this chapter are:

- the introduction of a new dimensionality reduction (DR) method combining image and clinical data (extension of the Laplacian eigenmaps (LEM) algorithm),
- the validation of this method on a large dataset from the Alzheimer's disease neuroimaging initiative (ADNI) study, with both continuous and discrete clinical data.

In Part III, we focus on *longitudinal population analysis*. In this setting, we have not one but several images by patient, screened at different time points. The goal is now to analyze the evolutions of the patients.

In Chapter 8, we introduce the field of *computational anatomy*. In this domain, several registration techniques have been introduced, able to estimate biologically realistic deformations. We review several transformations models, as well as several template building algorithms, and introduce the question of transport.

In Chapter 9, we use the theory of *large deformation diffeomorphic metric mappings (LDDMMs)*, which offers a Riemannian framework and the possibility to encode the deformation information via tangent vectors. We present the use of *initial momenta* to build disease progression classifiers. Our contributions in this chapter are:

- the introduction of a new method to separate stable from progressive patients from the use of initial momenta encoding the hippocampal shape evolutions,
- the introduction of two extensions of the Karcher algorithm able to (1) compute the average shape of a population up to rigid transformations and (2) avoid the smoothing due to successive resamplings,
- the validation of the proposed method on a set of patients from ADNI,
- a proof-of-concept on the use of subregions for the classification, enabling local representations to outperform the classification performance of global representations.

In Chapter 10, we address the question of spatial regularization for the classification of AD progression and the detection of hippocampus deformation related to the disease. Our contributions in this chapter are: ce chapitre sont :

- the introduction of a logistic classification model with spatial regularization,
- the comparison of standard regularizations (Ridge, LASSO and ElasticNet) and spatial regularizations (Sobolev, Total Variation and fused-LASSO) for the classification and the detection of biomarkers,
- the validation on a dataset from the ADNI study.

3.3.3 List of publications

Journal papers

[1] Jean-Baptiste Fiot, Laurent D Cohen, Parnesh Raniga, and Jurgen Fripp. Efficient brain lesion segmentation using multi-modality tissue-based feature selection and support vector machines. *International Journal for Numerical Methods in Biomedical Engineering*, Jan 2013.

[2] Jean-Baptiste Fiot, Hugo Raguet, Laurent Risser, Laurent D. Cohen, Jurgen Fripp, and François-Xavier Vialard. Longitudinal deformation models, spatial regularizations and learning strategies to quantify Alzheimer's disease progression. *Submitted*, 2013.

Conference proceedings

[1] Jean-Baptiste Fiot, Laurent Risser, Laurent D. Cohen, Jurgen Fripp, and François-Xavier Vialard. Local vs global descriptors of hippocampus shape evolution for Alzheimer's longitudinal population analysis. In *2nd International MICCAI Workshop on Spatiotemporal Image Analysis for Longitudinal and Time-Series Image Data (STIA'12)*, volume 7570, pages 13–24, Nice, France, October 2012. 201, 202

[2] Jean-Baptiste Fiot, Jurgen Fripp, and Laurent D. Cohen. Combining imaging and clinical data in manifold learning: distance-based and graph-based extensions of Laplacian Eigenmaps. In *Proc. IEEE International Symposium on Biomedical Imaging 2012*, 2012.

[3] Jean-Baptiste Fiot, Laurent D. Cohen, Parnesh Raniga, and Jurgen Fripp. Efficient Lesion Segmentation using Support Vector Machines. In *Computational Vision and Medical Image Processing: VipIMAGE 2011*, page ISBN 9780415683951, Olhão, Portugal, September 2011. 102

[4] Jean-Baptiste Fiot, Laurent D. Cohen, Pierrick Bourgeat, Parnesh Raniga, Oscar Acosta, Victor Villemagne, Olivier Salvado, and Jurgen Fripp. Multimodality Imaging Population Analysis using Manifold Learning. In *Computational Vision and Medical Image Processing: VipIMAGE 2011*, page ISBN 9780415683951, Olhão, Portugal, September 2011.

3.3.4 List of oral communications

Talks

- Mathematical methods for medical image population analysis. *Université Paris Dauphine, Feb 2013.*
- Overview of my PhD research directions. *"Mathematical methods for Imaging" group meeting, Université Paris Dauphine, Oct 2012.*
- Local vs global descriptors of hippocampus shape evolution for Alzheimer's longitudinal population analysis. *2nd International MICCAI Workshop on Spatiotemporal Image Analysis for Longitudinal and Time-Series Image Data (STIA'12), Nice, France, Oct 2012.*
- Local vs global descriptors of hippocampus shape evolution for Alzheimer's longitudinal population analysis. *PhD student seminar, Université Paris Dauphine, France, Sept 2012.*

- Introduction to mathematical methods for medical image population analysis. *Samsung Advanced Insitute of Technology, South Korea, Aug 2012.*
- Mathematical methods for medical image population analysis. *Seoul National University, South Korea, July 27th, 2012.*
- Analyse de population d'images par apprentissage de variétés. *CNRS - GdR ISIS workshop, Télécom Paris Tech, France, May 2012.*
- Image population analysis using non-linear dimensionality reduction. *PhD student seminar, Université Paris Dauphine, April 2012.*
- Image population analysis using non-linear dimensionality reduction. *"Mathematical methods for Imaging" group meeting, Université Paris Dauphine, Dec 2011.*
- Overview of my PhD research projects. *NatImage'11 workshop, IHP Paris, France, Nov 2011.*
- *Efficient Lesion Segmentation using Support Vector Machines. VipIMAGE 2011, Olhão, Portugal, Oct 2011.*
- Overview of my PhD research projects. *"Mathematical methods for Imaging" group meeting, Université Paris Dauphine, Sep 2011.*
- Efficient Lesion Segmentation using Support Vector Machines. *Centre for Medical Image Computing, London, UK, Nov 2010.*
- Multimodality Imaging Population Analysis using Manifold Learning. *Centre for Medical Image Computing, London, UK, Nov 2010.*

Poster presentations

- Local vs global descriptors of hippocampus shape evolution for Alzheimer's longitudinal population analysis. *Workshop SIGMA'2012 Signal, Image, Geometry, Modeling, Approximation, Marseille, France, Nov 2012.*
- Combining imaging and clinical data in manifold learning: distance-based and graph-based extensions of Laplacian eigenmaps. *IEEE International Symposium on Biomedical Imaging (ISBI), Barcelona, Spain, May 2012.*
- Multimodality Imaging Population Analysis using Manifold Learning. *VipIMAGE conference, Olhão, Portugal, Oct 2011.*

Part II

Cross-sectional population analysis

Chapter 4

State of the art

Contents

Résumé

Dans le cadre de l'analyse à grande échelle de collections d'images, alors que chaque image peut être considérée comme un point dans un espace de grande dimension, un ensemble d'images peut être représenté par une variété (ou une collection de variétés) de dimension intrinsèque beaucoup plus faible. Dans ce chapitre, nous présentons différentes méthodes d'apprentissage de variétés utilisées dans la littérature. Ensuite, nous présentons comment ces algorithmes peuvent être utilisés dans diverses applications telles que le recalage, la segmentation, l'analyse de population et l'apprentissage statistique.

Mots clés : Apprentissage de variétés, recalage, segmentation, analyse transversale de population, régularisation

Abstract

In large scale image database analysis, whereas each image can be seen a point in a high-dimensional space, a set of images can be assumed to be well-represented by a manifold (or a collection of manifolds) of much lower intrinsic dimension. In this chapter, we will first review some

popular manifold learning algorithms. Second, we will see how such algorithms can be used in several applications such as image registration, image segmentation, population analysis and machine learning.

Keywords: Manifold learning, registration, segmentation, cross-sectional population analysis, manifold regularization

4.1 Manifold learning

4.1.1 Definition and algorithms

Definition 4.1.1 (Dimensionality reduction (DR) problem)**.** Given a set of n vectors $\left\{\mathbf{x}_i \in \mathbb{R}^d; \quad 1 \leq i \leq n\right\}$ and a *target dimension* $\delta < d$, the dimensionality reduction consists in finding n corresponding vectors $\left\{\tilde{\mathbf{x}}_i \in \mathbb{R}^\delta; \quad 1 \leq i \leq n\right\}$ which are optimal in some sense.

Remark. In the literature, another formulation is sometimes used, where the dimensionality reduction (DR) algorithm does not take a target dimension as input, but instead finds an "optimal" one (in a sense to be defined). This will be discussed further on in Section 4.1.2.

In the literature, a large number of DR algorithms have been proposed, and applied in a wide range of applications. Several reviews can be found in [Cayton 2005, Lee 2007, van der Maaten 2007]. These algorithms can be categorized into

- Information-based methods
  - maximum of variance: principal component analysis (PCA), kernel principal component analysis (kPCA),
  - entropy measure [Lawrence 2011],
- Geometry-based methods
  - global: multi dimension scaling (MDS), isometric mapping (ISOMAP),
  - local: local linear embeddings (LLE), diffusion maps (DM), local tangent space alignment (LTSA), Hessian eigenmaps (HEM), Laplacian eigenmaps (LEM).

Let us now briefly describe the key ideas of these algorithms.

Principal component analysis (PCA) [Jolliffe 1986] PCA is a popular and widely used linear DR technique. It computes low-dimensional coordinates encoding as much variance as possible. For example, the direction of maximum variance can be found by solving

$$\max_{\mathbf{w}} \sum_{i=1}^{n} \langle \mathbf{x}_i, \mathbf{w} \rangle. \tag{4.1}$$

In practice the PCA is solved by performing a singular value decomposition (SVD) of the covariance matrix.

Kernel principal component analysis (kPCA) [Schölkopf 1996] . kPCA is an extension of PCA using the kernel trick to compute non-linear low-dimensional coordinates.

Multidimensional scaling (MDS) [Cox 1994] MDS is a linear technique aiming at conserving the pairwise distance. It minimizes the functional

$$\mathcal{L}(\tilde{\mathbf{X}}) = \sum_{i,j=1}^{n} \left(d(\mathbf{x}_i, \mathbf{x}_j) - \|\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j\|\right)^2 , \tag{4.2}$$

where $d(\mathbf{x}_i, \mathbf{x}_j)$ is a distance between $\mathbf{x}_i$ and $\mathbf{x}_j$.

Isometric mapping (ISOMAP) [Tenenbaum 2000] ISOMAP is a technique that builds upon the MDS approach, using approximation of geodesic distances for the $d(\mathbf{x}_i, \mathbf{x}_j)$. It builds a weighted neighborhood graph (usually a k nearest neighbours (kNN) graph), then computes the weights between all pairs of points using shortest paths on graphs, and finally constructs the low-dimensional embedding via an eigenvalue problem.

Remark. ISOMAP is considered being a global method because it constructs an embedding derived from the geodesic distance between all pairs of points. The neighborhood graph is only built to approximate the geodesic distance by a path on the graph.

Local linear embeddings (LLE) [Roweis 2000] LLE builds a kNN graph, then computes the optimal weights minimizing the sum of the errors of linear reconstructions in the high dimensional space, and finally solve an eigenvalue problem to map to embedded coordinates.

Remark. LLE is considered a local method because the cost function that is used to construct the embedding only considers the placement of each point with respect to its neighbors. Similarly, LEM and the derivatives of LLE are local methods.

Laplacian Eigenmaps (LEM) [Belkin 2003] LEM first builds a weighted adjacency graph and then solves an eigenvalue optimization problem based on the Laplacian operator.

Remark. The details of this algorithm are provided in Chapter 6.

Hessian eigenmaps (HEM, also called Hessian-based LLE) [Donoho 2003] HEM identifies the kNN, obtains tangent coordinates by SVD, and then computes the embedding coordinates using the Hessian operator and eigen-analysis.

Diffusion maps (DM) [Coifman 2006] In this paper, the authors introduce a framework based upon diffusion processes. They use spectral analysis of Markov matrices to compute the low dimension coordinates of the dataset. A Markov matrix represent a graph where the edge weights are transition probabilities. In that setting, the distance between two connected nodes is represented by the probability of a random walk to go from one node to the second one.

Algorithm	Number of parameters
MDS	1 (δ)
PCA	
kPCA	1 (δ) + kernel params
ISOMAP	2 (δ, k)
LLE	
LEM	
HEM	
LTSA	
DM	3 (δ, σ, α)

Table 4.1: Number of parameters of DR algorithms.

Remark. The notion of a cluster from a random walk point of view is a region in which the probability of escaping this region is low.

Local tangent space alignment (LTSA) [Zhang 2003] LTSA uses the tangent space in the neighbourhood of a data point (typically the kNN) to represent the local geometry. The optimization of the alignment of those tangent spaces is used to construct the low-dimension coordinates.

4.1.2 Parameter selection

Parameters Many DR algorithms share common parameters such as the target dimension δ, the number of nearest neighbors k, the width σ of the Gaussian kernel, the normalization control α ($\alpha = 0$: Graph Laplacian, $\alpha = 1/2$: Fokker-Plank operator, $\alpha = 1$: Laplace-Beltrami operator). Table 4.1 summarizes the parameters of the various algorithms. Let us now mention limitations and possible considerations with regard to the number of nearest neighbors and the target dimension.

Number of nearest neighbors As illustrated in Fig. 4.1, an inappropriate number of nearest neighbors may cause the DR algorithm to fail.

Remark (Manifold ranking). In [Wei 2008], instead of using a ranking based on the euclidean distance, a geometric ranking called *manifold ranking* [Zhou 2004] is used. This method should fix the problems arising from an unreasonably high k value.

Remark (Adaptative Neighborhood). Chosing a global value for k is not necessarily accurate. We might want to have a local way to define the neighborhood. In [Zhan 2009], the neighborhoods are defined based on local linearities (via the decrease of the eigenvalues).

Target dimension To select the target dimension δ (i.e. the reduced dimension of the representation of the data), two point of views can be considered:

1. try to estimate the "intrinsic dimension" from the dataset,
2. consider as an input dimension of the DR algorithm, and optimize δ as an hyper-parameter (for example using classification performance).

Several methods have been proposed to find the intrinsic dimension from a finite dataset. They can be classified in 4 categories:

1. projection methods (Local-PCA [Fukunaga 1971]),
2. topological methods (Topology Representing Network [Martinetz 1994]),
3. trial-and-error methods (e.g. based on reconstruction error),
4. geometric methods (fractal-based [Camastra 2002, Fan 2009], Packing Numbers [Kégl 2003], Maximum-likelihood [Levina 2005]).

Reviews of these methods can be found in [Camastra 2003] and in the 3rd chapter of [Lee 2007].

Remark (Strategy in medical imaging). In this setting, the second point of view is often preferred. Indeed, in such area, building low-representation of the data is a tool used towards a more general medical objective such as disease detection. In that sense, the optimal target dimension in terms of performance of the final goal should be selected.

4.1.3 Toy examples

In Fig. 4.1, ISOMAP embeddings are computed for the Swiss roll dataset. This dataset is composed of a set of points in $\mathbb{R}^3$ that lie on a $2D$ manifold. Successful and unsuccessful examples are given for different sets of parameters.

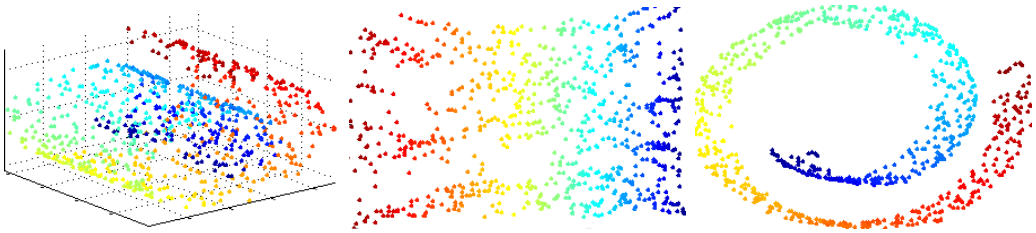
In Fig. 4.2, ISOMAP and HEM embeddings are computed for the Swiss hole dataset. This dataset is similar to the Swiss roll dataset, except that a subset of points are removed so that a hole is created on the manifold. This example shows that ISOMAP creates distortions in the case of non-convex data (the convexity is actually one assumption of ISOMAP [Donoho 2003]), whereas HEM is able to unwrap the manifold properly.

In Fig. 4.3, LTSA is able to uncover the intrinsic parameters of the dataset (face rotation and illumination).

4.2 Applications in medical imaging

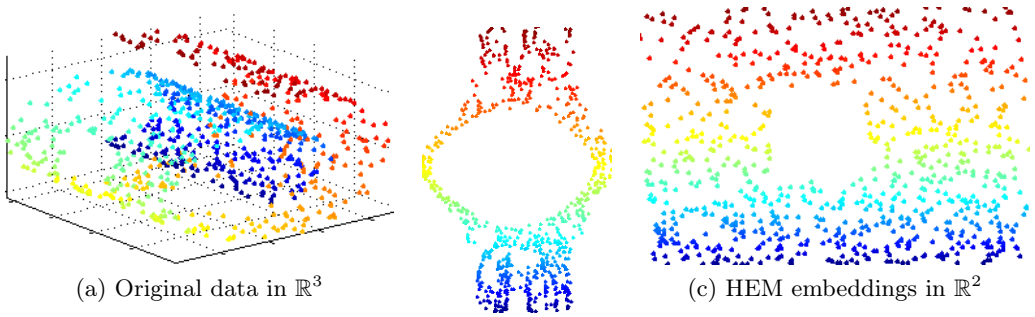
4.2.1 Image registration

This section does not intend to review the state of the art in image registration, but instead only illustrates how manifold learning can be used for image registration. Image registration is a process that is used to align two images to facilitate their comparison. The usual way to perform registration is to look for a function ϕ to



(a) Original data in $\mathbb{R}^3$ (b) $\delta = 2, k = 8$ (c) $\delta = 2, k = 200$

Figure 4.1: ISOMAP embeddings in $\mathbb{R}^2$ of the Swiss roll toy example. In 4.1b, the Swiss roll is properly unwrapped. However when the k parameter gets too high, the neighborhood graph "jumps" between layers, and the the Swiss roll is not properly unwrapped (4.1c).



(a) Original data in $\mathbb{R}^3$ (b) ISOMAP embeddings in $\mathbb{R}^2$ (c) HEM embeddings in $\mathbb{R}^2$

Figure 4.2: Embeddings in $\mathbb{R}^2$ of the Swiss hole toy example (i.e. Swiss roll with an extra hole). Since the manifold is not intrinsically convex, ISOMAP creates distortions (4.2b). HEM is able to deal with such manifolds (4.2c).

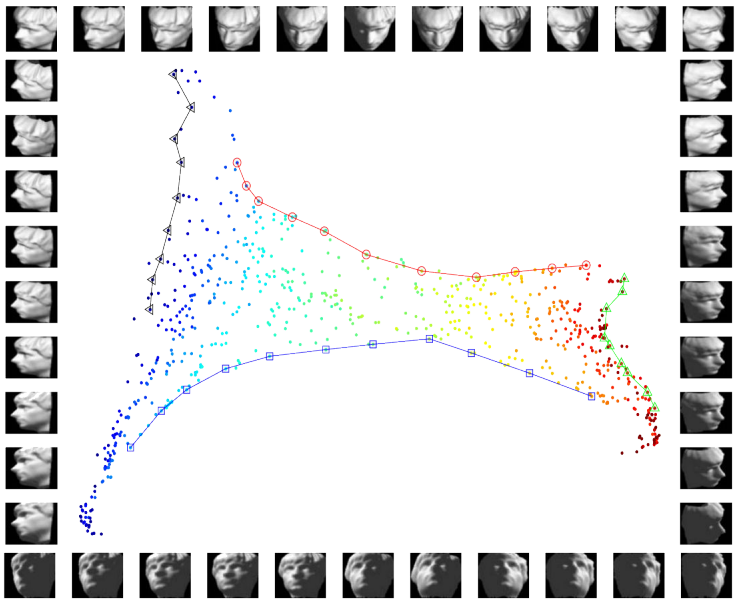


Figure 4.3: LTSA embeddings in $\mathbb{R}^2$ of a set of images. The algorithm is able to uncover the intrinsic parameters of the dataset (face rotation and illumination).

deform a source image $I_s \colon \Omega \subset \mathbb{R}^n \to \mathbb{R}$ towards a target image $I_t \colon \Omega \to \mathbb{R}$. The quality of the registration is often measured by the fitting term

$$\mathcal{L}_{fit}(\phi) \overset{\text{def.}}{=} \|I_t - I_s \circ \phi^{-1}\|^2. \tag{4.3}$$

The optimal transformation is the one minimizing the matching term in Equation (4.3). Registration is often categorized as *rigid* or *non-rigid*. A rigid registration (RR) does not deform locally an image whereas non-rigid registration (NRR) can do it. When computing a NRR, a regularization term is usually added to ensure that the optimal transformation ϕ is smooth enough. A non-rigid registration is therefore usually computed via the variational problem

$$\underset{\phi}{\operatorname{argmin}} \|I_t - I_s \circ \phi^{-1}\|^2 + \lambda J(\phi), \tag{4.4}$$

where J is a regularization function and λ a weighting coefficient between matching and smoothness.

Now let us see how manifold learning can be used for image registration. In [Hamm 2010], the authors introduced the geodesic registration on anatomical manifold (GRAM) framework based on an idea from ISOMAP [Tenenbaum 2000], where the geodesic path of the analytical manifold is replaced by the shortest path on a kNN graph that approximates the metric structure of the empirical manifold. The motivation is that the classic registration between two very different shapes might be difficult, and therefore an alternative strategy is to register two "far away" images via the composition of simpler registrations (see Fig. 4.4). The authors claim

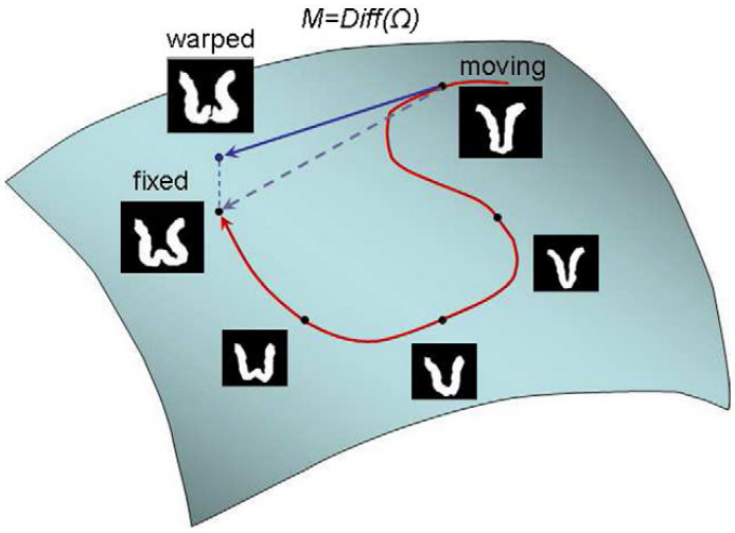


Figure 4.4: Motivation for the GRAM framework: registering two images "far away" from each other might be difficult (blue path). An alternative approach is to replace a difficult registration by the composition of simpler registrations (red path). Source: [Hamm 2010].

the following properties

1. learning of anatomical manifolds,
2. computational efficiency,
3. visualization and automatic template selection.

Remark (Models for large deformations and computational efficiency). In the field of computational anatomy, several registration frameworks have been introduced in order to deal with large deformations (see Chapter 8, in particular Section 8.2). Because each intermediate registration in the GRAM framework is assumed to be "simple", it can be performed by a simpler and less computationally intensive algorithm.

Remark (Large deformations and registration accuracy). As mentioned in [Aljabar 2012], a limitation of the GRAM framework is that the composition of a large number of transformations might lead to a compounding of smaller registration errors into larger ones.

4.2.2 Image segmentation

The segmentation of $I\colon \Omega \subset \mathbb{R}^n \to \mathbb{R}$ consists in assigning a *label* to each voxel $\omega \in \Omega$. Formally speaking, it consists in building a function $S\colon \Omega \to \mathcal{S}$ where $\mathcal{S}$ is a (finite) discrete set of labels. For example, these labels can represents various organs, the presence/absence of lesions, etc. In this section, we will see how manifold

learning can be used to improve the accuracy of atlas-based segmentation. The concept of *atlas segmentation propagation* (or *atlas-based segmentation*) consists in segmenting an image given the segmentations of a single or several reference images called atlas(es). The *single atlas segmentation propagation* technique uses only a single image as atlas. To compute the segmentation of the new image, first the atlas is registered to the new image. Then the computed deformation is used to deform the segmentation of the atlas, which gives the segmentation of the new image. The segmentation of the atlas is said to be *propagated*.

Remark (Segmentation accuracy). The choice of the registration algorithm is the key choice here, and an accurate registration is required to obtain an accurate segmentation.

The *multi atlas segmentation propagation* technique [Heckemann 2006] is an extension of the single atlas segmentation propagation technique. In this setting, several segmented atlases are used, instead of just one in the previous technique. All of them are registered to the new image to segment, and the computed transformations are applied to the known segmentations. The deformed segmentations are finally combined (usually using a voting scheme) to get the segmentation of the new image.

Remark (Combination scheme). Besides the choice of the registration algorithm, the way to combine the different segmentations in the voting scheme has to be defined. The easiest choice is to give the same weight to all the votes. More elaborate techniques tries to give more importance to the "closest" atlases (e.g. in terms of normalized mutual information (NMI), quantity of displacement in the registration, etc). When some weights are set to zero if the distance is too high, the atlases associated to non-zero weights are said to be *selected*. In [Aljabar 2009], the authors studied the choice of an atlas selection strategy and its effect on accuracy.

Another class of approaches is *probabilistic atlas-based segmentation*, where atlases contain tissue probabilities [Van Leemput 2009].

Remark (Review). In [Cabezas 2011], a large number of atlas-based segmentation methods are reviewed for the case of magnetic resonance (MR) brain images.

Improving segmentation accuracy using manifold learning. In [Wolz 2009], the authors introduced the learning embeddings for atlas propagation (LEAP) algorithm to improve the performance of atlas-based segmentation. The key idea is that the segmentation accuracy could be higher if the propagation is done progressively on a dataset: from segmented atlas towards close images, and iteratively until all images are segmented. This would avoid the direct segmentation of an image too "far" from the atlases. Their algorithm (see Fig. 4.6) is composed of the following steps

1. compute the embeddings,
2. select a subset of images to segment (based on the Euclidean distance on the manifold coordinates),

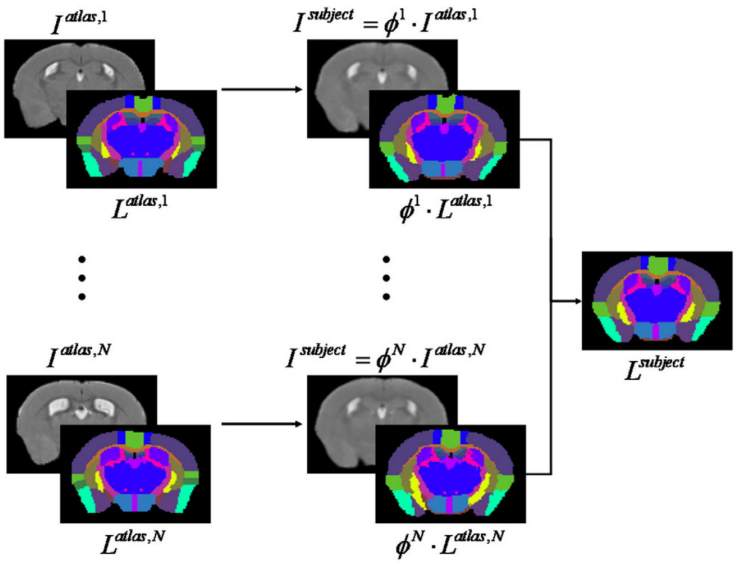


Figure 4.5: Multi-atlas segmentation-propagation. $(I^{atlas,i}, L^{atlas,i})_{1\leq i\leq n}$ are the labeled atlases. $(\phi^i)_{1\leq i\leq n}$ are the transformations registering the images of the atlas towards the images of the subject. The segmentation $L^{subject}$ is obtained via the combination of the deformed segmentations $(\phi^i \cdot L^{atlas,i})_{1\leq i\leq n}$. Source: [Bai 2012]

3. for each of these, select a subset of atlases which are going to vote for the segmentation,
4. propagate the segmentation (do the registrations and voting),
5. iterate step 2-4 until all images are segmented.

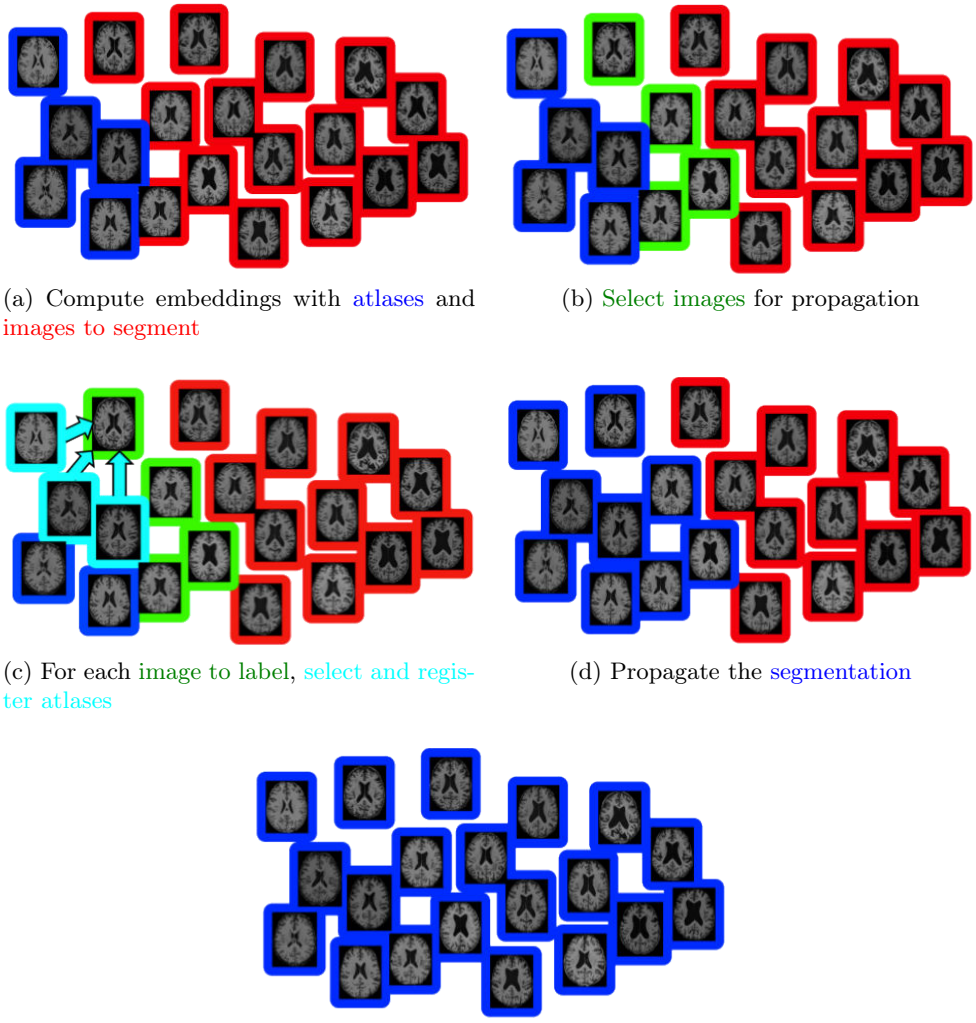
Remark (Segmentation accuracy in LEAP). In the LEAP setting, accuracy might be limited similarly as in the GRAM framework: a large number of iterations might compound small segmentation errors into larger ones.

4.2.3 Population analysis

In this section, we describe how manifold learning can be used to find trends and modes of variation in a population.

Generative models In [Gerber 2009, Gerber 2010], the authors introduced a generative model to describe a population. The main assumption is that all the images in the population derive from a small number of brains. In their setting, new brain images can be projected onto the manifold. Figure 4.7 illustrates the result of their method on the OASIS[1] dataset. Via a manifold kernel regression [Davis 2007] on the manifold coordinates, the authors are also able to reconstruct images from low-dimension coordinates (see Fig. 4.8).

[1] www.oasisbrains.org



(a) Compute embeddings with atlases and images to segment

(b) Select images for propagation

(c) For each image to label, select and register atlases

(d) Propagate the segmentation

(e) Iterate until all images are segmented

Figure 4.6: LEAP algorithm for the segmentation of images by iterative segmentation-propagation of atlases. Source: adapted from [Aljabar 2012].

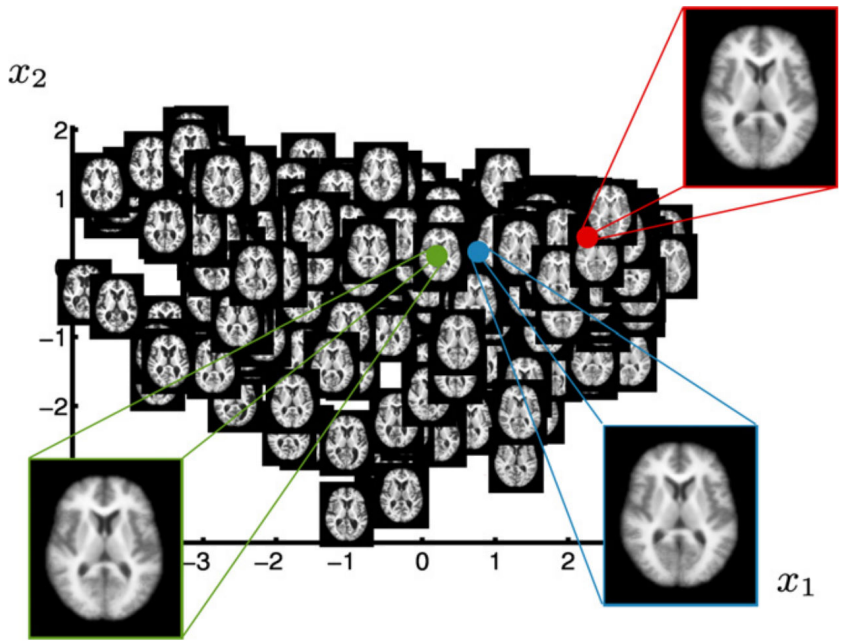


Figure 4.7: 2D parametrization of OASIS brain MRI dataset. The insets show the mean (green), median (blue) and mode[2](red) of the learned manifold and the corresponding reconstructed images. Source [Gerber 2010].

In [Sabuncu 2009], the authors introduced the iCluster algorithm that clusters a set of images while co-registering them using a parameterized, nonlinear transformation model. This algorithm is based on a generative model of an image population as a mixture of deformable templates. The computed template images represent different modes in a population.

Atlas stratification In [Blezek 2007], the authors use the mean shift algorithm [Fukunaga 1975] to identify modes in a populations. Their method builds multiple atlases, each from a subset of the population. The various clusters are visualized using MDS (see Fig. 4.9).

4.2.4 Machine learning

Manifold regularization In the context of semi-supervised learning, a geometric method is presented in [Belkin 2004] to improve the results of classification algorithms by using the geometry of unlabeled data. As illustrated in Fig. 4.10, it can be reasonable to adjust the class prior belief according to the geometry of unlabeled data. An extra term based on the Laplacian operator is added to the cost function of the learning algorithm, to force the labeling function to be smooth according

[2]In statistics, the mode of a random variable $\mathbf{X}$ is the most probable value the variable can take. For a discrete variable, it is the x maximizing $\mathbb{P}(\mathbf{X} = x)$. For a continuous variable, it is the x maximizing the density of probability $f(x)$.

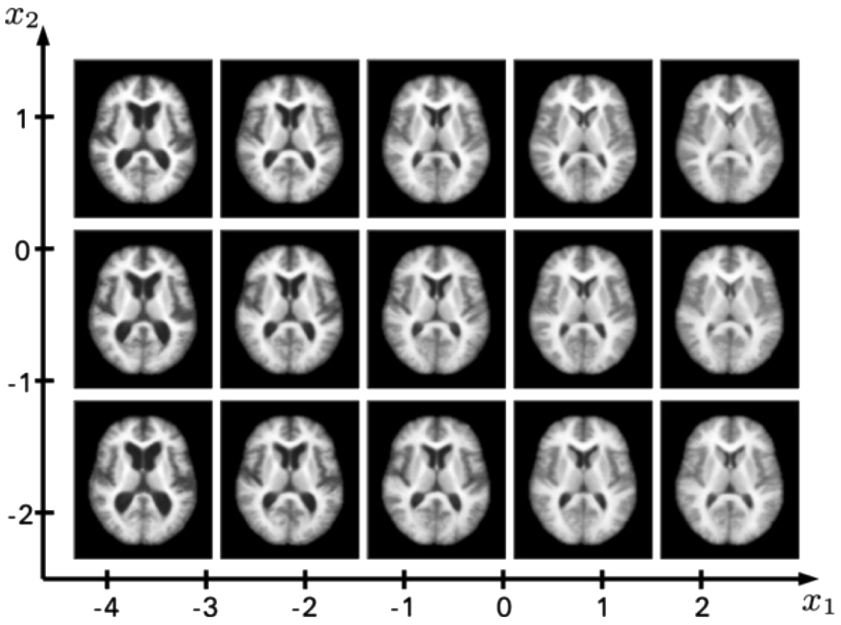
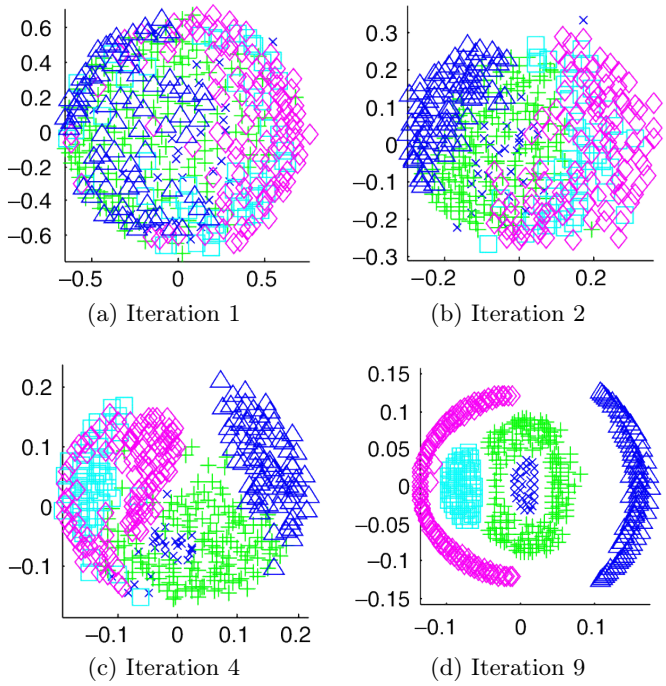


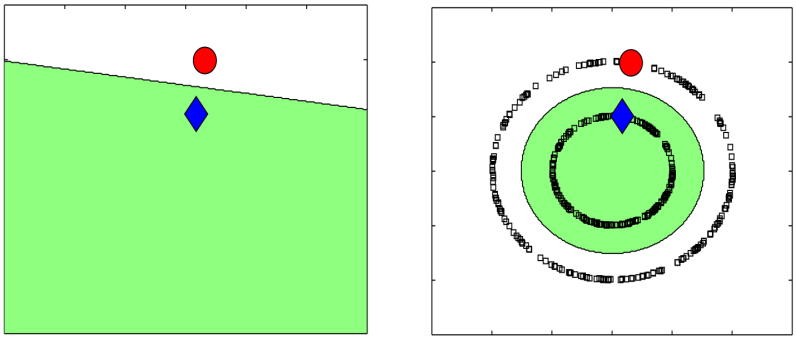
Figure 4.8: Reconstructions from the manifold coordinates to the space of images. Source [Gerber 2010].



(a) Iteration 1 (b) Iteration 2

(c) Iteration 4 (d) Iteration 9

Figure 4.9: Atlas stratification identifying five modes in a population (2D visualization using MDS). Source: [Blezek 2007].



(a) With one observation from each class

(b) With extra unlabeled points

Figure 4.10: In green is represented the prior belief of the class to which the blue diamond belongs. The red circle represent an example from the other class. On the left, the decision boundary is expected to be linear. On the right, the geometry of unlabeled data makes a disk prior reasonable. Source: [Belkin 2004].

to the geometry of the data. This technique has been applied to regularized least squares (RLS) and support vector machines (SVM).

Example 4.2.1 (Disease manifolds)**.** In [Batmanghelich 2008], manifold regularization is used to compute disease manifolds. The authors assume that tissue deterioration related to the disease can be viewed as a continuous change from healthy to disease, and hence can be modeled by a non-linear manifold. Using the semi-supervised data (see Fig. 4.11) from several modalities, they apply Laplacian SVM at the voxel level to build tissue abnormality maps. It is important to notice that their Laplacian operator contains a parameter to weight the spatial regularization between pairs of unlabeled voxels (see Fig. 4.12).

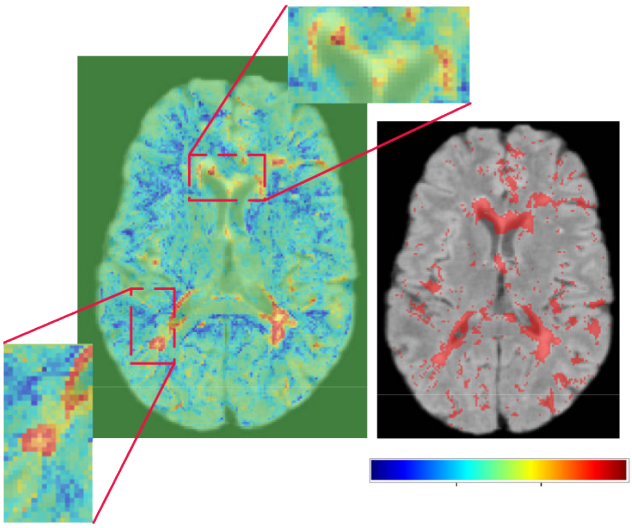
4.3 Conclusion

In this section, we have reviewed some DR algorithms (PCA, kPCA, ISOMAP, LLE, LEM, HEM, LTSA, and DM). We have seen that these algorithms have been used in medical imaging in a wide range of applications such as image registration, image segmentation, population analysis and machine learning.

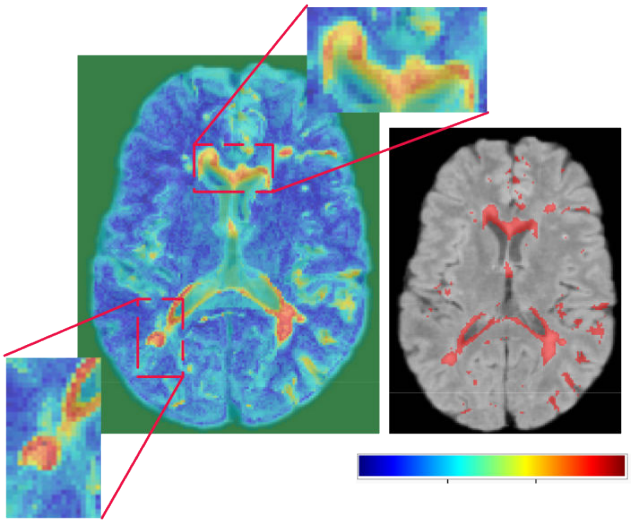
In Chapter 6, these algorithms are used to find trends of shape and appearance in population of MR and positron emission tomography (PET) images. In Chapter 7, we describe how image and clinical informations can be combined in manifold learning to build low-dimensional representations of patients.

(a) FLAIR image (b) Lesion region (c) Healthy tissue

Figure 4.11: Example of semi-supervised data used in [Batmanghelich 2008]. The red voxels correspond to lesion, the green ones correspond to healthy tissue, and the remaining ones are unlabeled.



(a) Low spatial regularization



(b) High spatial regularization

Figure 4.12: Abnormality maps computed in [Batmanghelich 2008] with different spatial regularizations.

Chapter 5

Lesion segmentation using Support Vector Machines

Contents

Résumé

Les séparateurs à vaste marge (SVM) sont une technique d'apprentissage statistique qui a été utilisée dans la littérature pour la segmentation et la classification d'images médicales, en particulier pour la segmentation d'hyperintensités de la matière blanche (HMB). Les approches actuelles utilisant les SVM pour la segmentation d'HMB extraient des descripteurs du cerveau, les classifient, et utilisent des techniques de post-traitement complexes pour supprimer les faux positifs. La méthode présentée dans ce chapitre combine l'utilisation d'information spatiale, un pré-traitement avancé (en utilisant la propagation de segmentation de tissus d'un atlas) et une classification SVM permettant de segmenter les HMB de manière efficace et avec des performances élevées. A partir d'une base de données de 125 patients, des descripteurs combinant jusqu'à quatre modalités (T1-w, T2-w, PD et FLAIR), différentes tailles de voisinage, et l'utilisation de descripteurs multi-échelles ont été évalués. Nos résultats montrent que même si la combinaison des quatre modalités donne la meilleure performance (scores Dice moyens de

0.54 ± 0.12, 0.72 ± 0.06 et 0.82 ± 0.06 pour des charges de lésion respectivement faibles, modérées, et élevées, avec des descripteurs d'intensité de taille $3 \times 3 \times 3$), cette performance n'est pas statistiquement meilleure ($p = 0.50$) que celle obtenue si l'on utilise seulement les modalités T1-w et FLAIR (scores Dice moyens de 0.52 ± 0.13, 0.71 ± 0.08 et 0.81 ± 0.07 pour les mêmes charges de lésion et type de descripteur). De plus, la différence de performance est négligeable ($p = 0.93$) entre l'utilisation de descripteurs de taille $5 \times 5 \times 5$ et ceux de taille $3 \times 3 \times 3$. Finalement, nous montrons qu'une sélection appropriée de descripteurs et des techniques de pré-traitement non seulement réduit les besoins de stockage et de puissance de calcul, mais permet aussi d'obtenir une classification plus efficace, dont les performances dépassent celles obtenues si l'on entraîne sur tous les descripteurs et que la classification est suivie d'un post-traitement.

Mots clés : Imagerie cérébrale, Segmentation de lésion, Classification, Séparateurs à vaste marge

Abstract

Support vector machines (SVM) are a machine learning technique that have been used for segmentation and classification of medical images, including segmentation of white matter hyper-intensities (WMH). Current approaches using SVM for WMH segmentation extract features from the brain and classify these followed by complex post-processing steps to remove false positives. The method presented in this chapter combines the use of domain knowledge, advanced pre-processing (based on tissue segmentation and atlas propagation) and SVM classification to obtain efficient and accurate WMH segmentation. Features from a dataset of 125 patients, generated from up to four MR modalities (T1-w, T2-w, PD and FLAIR), differing neighbourhood sizes and the use of multi-scale features were compared. We found that although using all four modalities gave the best overall classification (average Dice scores of 0.54 ± 0.12, 0.72 ± 0.06 and 0.82 ± 0.06 respectively for small, moderate and severe lesion loads, using $3 \times 3 \times 3$ neighborhood intensity features); this was not significantly different ($p = 0.50$) from using just T1-w and FLAIR sequences (Dice scores of 0.52 ± 0.13, 0.71 ± 0.08 and 0.81 ± 0.07 for the same lesion loads and feature type). Furthermore, there was a negligible difference ($p = 0.93$) between using $5 \times 5 \times 5$ and $3 \times 3 \times 3$ features. Finally, we show that careful consideration of features and pre-processing techniques not only saves storage space and computation time but also leads to more efficient classification which outperforms the one based on all features with post-processing.

Keywords: Brain imaging, Multi-modality, Lesion segmentation, Classification, Support vector machines

5.1 Introduction

White matter hyper-intensities (WMH) are regions in the brain white matter (WM) that appear with bright signal on T2-weighted (T2-w) and fluid attenuated inversion recovery (FLAIR) magnetic resonance (MR) modalities. They are a possible risk factor for Alzheimer's disease (AD) and vascular dementia, with progression associated with vascular factors and cognitive decline [Lao 2008]. To quantify these changes in large scale population studies, it is desirable to have fully automatic and accurate segmentation methods to avoid time-consuming, costly and non-reproducible manual segmentations. However, WMH segmentation using a single modality is challenging because their signal intensity range overlaps with that of normal tissue: in T1-weighted (T1-w) images, WMH have intensities similar to grey matter (GM), and in T2-w and proton density (PD) images, WMH look similar to cerebro-spinal fluid (CSF). FLAIR images have been shown to be most sensitive to WMH [Anbeek 2004], but can also present hyper-intensity artifacts that can lead to false positives. To improve the WMH segmentation performance, additional discriminative information is extracted from multiple MR modalities. An alternative strategy can be found in [Samaille 2012], where the proposed method is not based on intensities but instead relies on contrast.

Now the most successful lesion segmentation methods in the literature have been developed for the detection of multiple sclerosis lesions, with a recent grand challenge comparing the performance of various techniques [Styner 2008]. It is also worth mentioning algorithms for brain metastasis segmentation [Ambrosini 2010, Farjam 2012] and a review of algorithms for brain tumor segmentation [Bauer 2013]. Lesion segmentation algorithms can be categorized into unsupervised clustering or (semi-)supervised voxel-wise classification [Llad'o 2012]. Unsupervised methods suffer from the issue of model selection. Supervised methods such as neural networks [Dyrby 2008], k nearest neighbours (kNN) [Anbeek 2004], Naive Bayes classifier [Scully 2010] and Parzen windows [Sajja 2006, Datta 2006] have been proposed. Neural networks can be efficient but designing an appropriate network architecture and setting the parameters are difficult steps to obtain high performance. A recent review [Klöppel 2011] compares several approaches (including SVM), however this study uses a very limited dataset of only 20 patients and only FLAIR and T1-w images.

We present an SVM based segmentation scheme whose preliminary results were presented as a conference paper in [Fiot 2011], and inspired by the work in [Lao 2008, Zacharaki 2008]. Lao et al. applied four steps: pre-processing (co-registration, skull-stripping, intensity normalisation and inhomogeneity correction), SVM training with Adaboost, segmentation and elimination of false positives. Our implementation utilises a similar but more advanced pre-processing pipeline and a simpler training procedure. As one of the primary causes of errors in other approaches is false positive cortical regions, we use information from multiple modalities to define a mask of potential WMH. This mask is built from patient specific tissue segmentation and atlas based population tissue priors. It leads to three main advantages compared

to existing techniques. First, such careful feature selection enables to have a more accurate model without the use of boosting. Second it limits the areas where the classification is performed on the training set, which means a faster overall brain classification. Third, it reduces the false positive regions that are usually found with naive classifiers, so the advanced post processing required by other techniques [Lao 2008] are not necessary. We also evaluated the relative value of each MR acquisition protocol for WM lesion segmentation. This scheme is quantitatively validated on a significantly larger dataset including healthy aging, mild cognitive impairment and AD subjects. These results were compared with other supervised classification algorithms such as kNN, Naive Bayes, Parzen windows and decision tree.

5.2 Methods

5.2.1 Global pipeline

The proposed algorithm uses the standard supervised classification design for segmentation: given images and corresponding segmentations, the goal is to build a classifier to segment new images (Fig. 5.1). To obtain good performance, adequate preprocessing, mask and feature type have to be defined. This application-specific part is followed by a machine learning process. The steps of the proposed algorithm are summarized in Fig. 5.2.

In this chapter, we use features containing *only local information*. In particular, such features do not contain any spatial information. Indeed as the training sets available in this application are generally limited in size, we do not want the algorithm to be able to predict lesion voxels only in areas where it has "observed" some in the training set. However, using purely local feature (i.e. not containing any spatial information whatsoever) could lead to many false positives. To avoid this side effect, we define a mask to limit the area of interest. This mask is used twice: first in the training step training features are taken inside this mask, and second in the segmentation step, voxels outside the mask are directly labeled as non-lesion by our algorithm.

Section 5.2.2 defines the mask. Section 5.2.3 defines the local features, detail the support vector machine classifier, and list other classifiers and indicators used to benchmark our pipeline.

5.2.2 Mask creation

To improve the performance of our lesion segmentation procedure, the definition of a proper region of interest (ROI) is critical. In this section, we define a mask to improve the lesion segmentation accuracy. The whole mask creation process is summarized in Fig. 5.9, and here we present the formal definitions.

First, let us recall that a global threshold on FLAIR images provides a high sensitivity, but poor specificity, which means it can only be used to define areas of

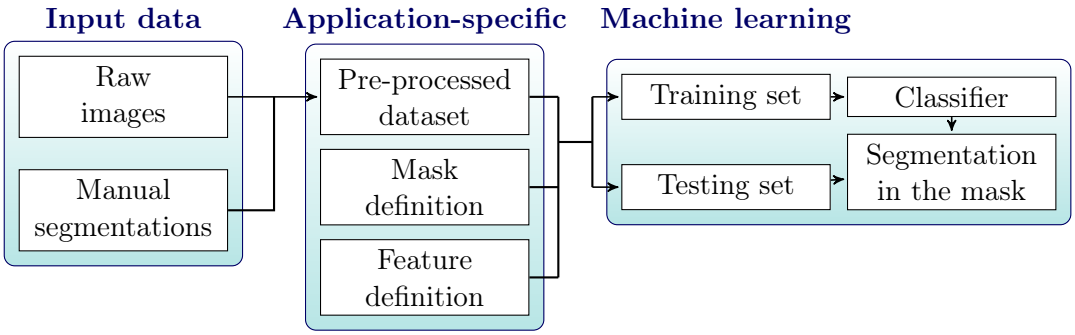


Figure 5.1: Supervised classification algorithms for segmentation aim to build a classifier from images and corresponding segmentations. To obtain good performance, adequate pre-processing, mask and feature definitions have to be used. This application-specific part is followed by a machine learning process, where a classifier is built from training examples and then used to segment new (testing) images.

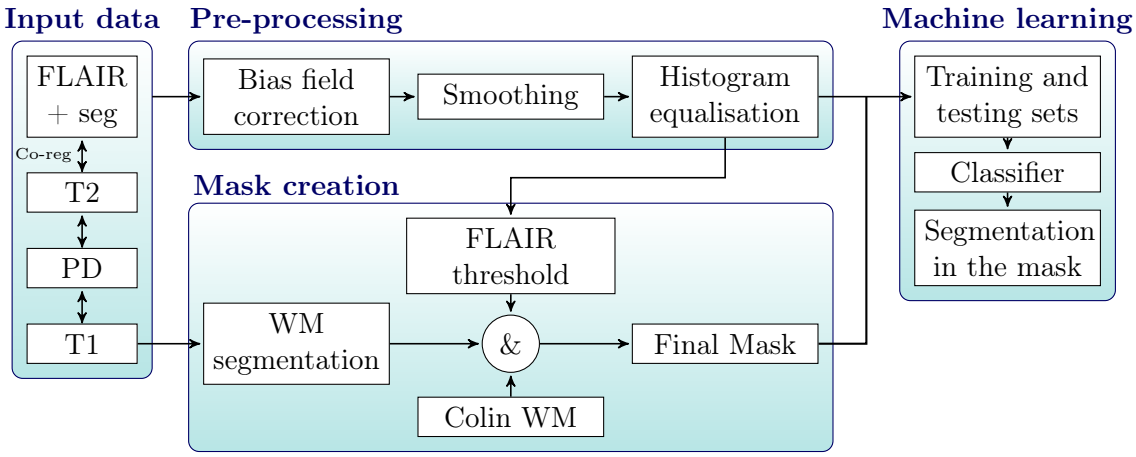


Figure 5.2: The proposed WMH segmentation pipeline is composed of 3 mains steps: pre-processing, mask creation and machine learning.

interest. To further refine the areas of interest, we define the region $\Omega_W \subset \Omega$ as the intersection of the dilated Colin WM mask [Collins 1998] (which was registered rigidly [Ourselin 2001] then non-rigidly [Rueckert 1999] to the subject) and the WM mask (from the tissue segmentation in patient space, see [Acosta 2009]).

Definition 5.2.1 (white matter domain)**.** We introduce the notion of WM domain as

$$\Omega_W \overset{\text{def.}}{=} \text{WM} \cap \text{ColinWM}, \tag{5.1}$$

where WM and ColinWM are obtained as described in the above references.

So far we have selected a potential area for white matter. This is a first step as we are interested in white matter hyper-intensities. Now let us refine even more to select hyperintensities. Within this white matter domain Ω_W, we associate a score to each voxel to measure how far above or under is the intensity of this voxel compared to the mean intensity in Ω_W.

Definition 5.2.2 (Map of interest)**.** We define $I \colon \Omega \to \mathbb{R} \cup \{-\infty\}$ by

$$\forall \omega \in \Omega, \quad I(\omega) \overset{\text{def.}}{=} \begin{cases} \frac{FLAIR(\omega) - \mu_W}{\sigma_W} & \text{if } \omega \in \Omega_W, \\ -\infty & \text{otherwise,} \end{cases} \tag{5.2}$$

where $\mu_W \overset{\text{def.}}{=} \text{mean}\{FLAIR(\omega); \quad \omega \in \Omega_W\}$ and $\sigma_W \overset{\text{def.}}{=} \text{std}\{FLAIR(\omega); \quad \omega \in \Omega_W\}$ are the mean and standard deviation of the FLAIR intensities on Ω_W

Remark. The FLAIR image and Ω_W are assumed such that $\sigma_W \neq 0$.

Up to now, we have obtained a map of potential WMH. To obtain the final mask, we threshold this map of coefficients.

Definition 5.2.3 (Final mask)**.** Given a threshold parameter $\tau \in \mathbb{R}$, we define the mask $M_\tau \colon \Omega \to \mathbb{R}$ by

$$\forall \omega \in \Omega, \quad M_\tau(\omega) \overset{\text{def.}}{=} \begin{cases} 1 & \text{if } I(\omega) > \tau, \\ 0 & \text{otherwise,} \end{cases} \tag{5.3}$$

where $I \colon \Omega \to \mathbb{R}$ is the map of interest defined in (5.2).

This mask is a binary image which depends on a threshold τ. To understand better the impact of this parameter, it is interesting to see the limit cases. By definition, the mask can contain ones only in some parts of the white matter domain Ω_W. The higher the τ, the more selective is the mask, and the less ones it contains. At the contrary, the lower the τ, the less selective is the mask and the more voxels in the white matter domain Ω_W are selected. The following property evaluates if the limits of the mask when τ goes to 0 or ∞. Its proof is given in Appendix A.2.1.

Proposition 5.2.1 (Limits of the mask)**.** *Assuming the FLAIR image to be bounded on Ω_W (this is always the case in practice because the number of voxels is finite), we have*

$$\lim_{\tau \to -\infty} M_\tau = \mathbb{1}|_{\Omega_W}, \tag{5.4}$$

$$\lim_{\tau \to \infty} M_\tau = 0, \tag{5.5}$$

where $\mathbb{1}|_S$ denotes the charateristic function of a set S.

5.2.3 Classification

5.2.3.1 Feature definitions

As mentioned in section 5.2.1, we evaluated supervised lesion segmentation in the case of features containing only local information. Below are the definitions of neighborhood intensity and pyramidal features evaluated in this study.

Definition 5.2.4 (Neighborhood intensity feature)**.** Given an image $I\colon \Omega \subset \mathbb{R}^3 \to \mathbb{R}$, a voxel $\omega \in \Omega$ and an odd integer $p \in \mathbb{N}$, we define the neighbourhood intensity feature of size $p \times p \times p$ by

$$\mathbf{x}(\omega) \overset{\text{def.}}{=} I|_{B(\omega, \frac{p-1}{2})}, \tag{5.6}$$

where $B(\omega, \frac{p-1}{2}) = \left\{ \omega' \in \Omega; \quad \|\omega - \omega'\|_\infty \leq \frac{p-1}{2} \right\}$ is the ball centered on ω and of radius $\frac{p-1}{2}$ for the uniform norm.

Definition 5.2.5 (Pyramidal feature)**.** Given an image $I\colon \Omega \subset \mathbb{R}^3 \to \mathbb{R}$ and a set of scales $0 = \sigma_1 < \cdots < \sigma_p$, the pyramidal features with p levels are evaluated via the operator

$$\mathbf{x} \overset{\text{def.}}{=} (k_{\sigma_i} \star I)_{1 \leq i \leq p} : \Omega \to \mathbb{R}^p, \tag{5.7}$$

where $k_\sigma \star$ represents the convolution by the Gaussian kernel of scale σ, i.e. $\forall \omega \in \Omega, \quad (k_\sigma \star I)(\omega) = \frac{1}{(\sqrt{2\pi}\sigma)^3} \int_\Omega \exp(-\frac{\|\omega - u\|^2}{2\sigma^2}) I(u) du$. By convention, $k_0 \star I = I$ (i.e. k_0 is the Dirac operator).

5.2.3.2 Support vector machine theory

Lesion segmentation can be formulated as a binary classification problem. The SVM technique [Schölkopf 2001] solves it in a supervised way: given a training set, it builds a function $f\colon \mathcal{X} \to \mathbb{R}$ such that $\text{sign}(f)$ is an optimal labeling function. To build f, we consider a Reproducing Kernel Hilbert Space $\mathcal{H}_K$ of functions $\mathcal{X} \to \mathbb{R}$, of associated Mercer Kernel $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. As mentioned in Chapter 2, the performance of a predictive model is often evaluated via a loss function (see Definition 2.1.2). Let us see the various terms in the SVM loss function.

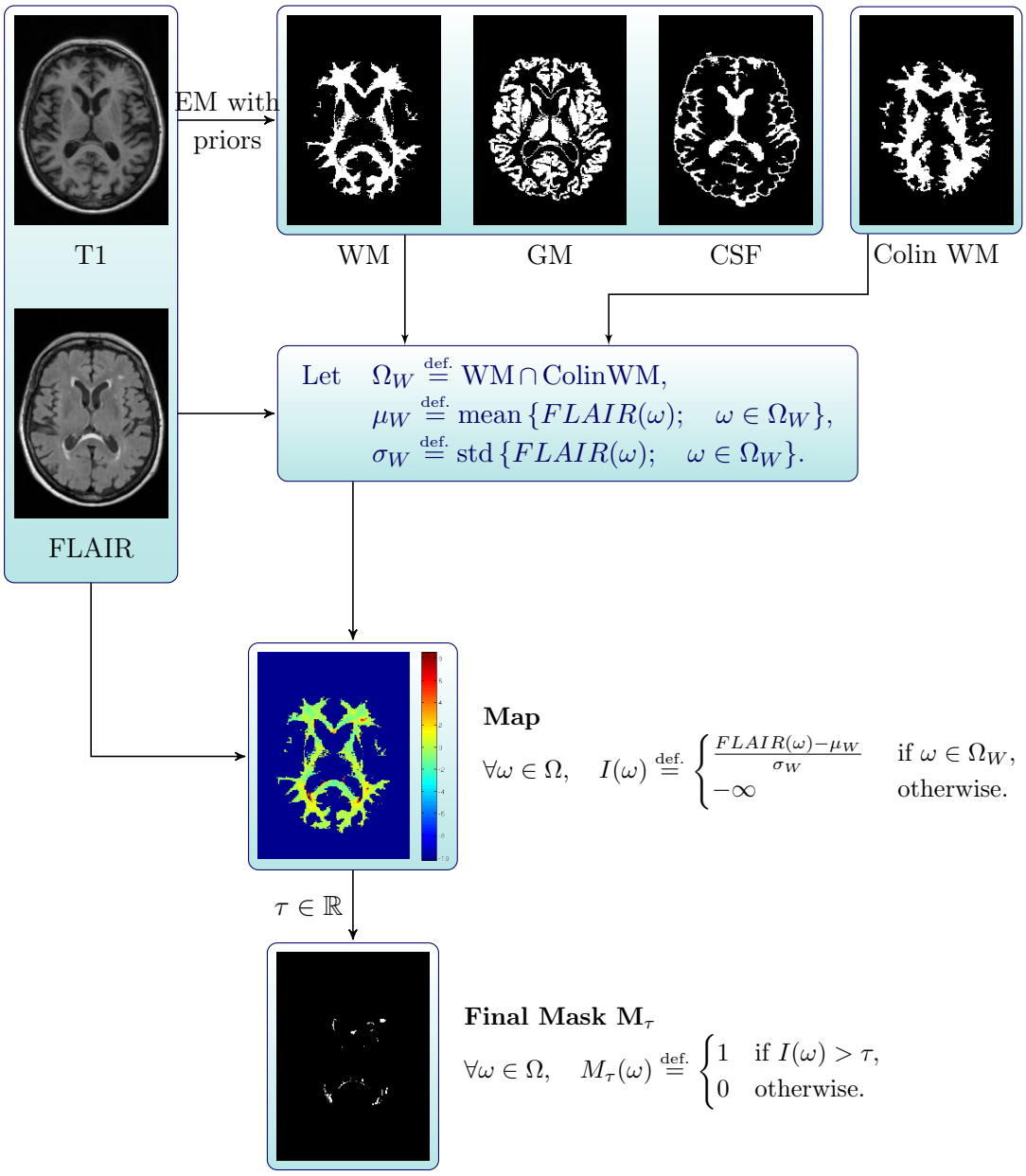


Figure 5.3: The mask creation uses both FLAIR and T1-w modalities to combine intensity-based and tissue-based properties. First, an expectation-maximisation (EM) technique on the T1-w is used to generate WM/GM/CSF segmentation. On the intersection Ω_W of the patient WM and the registered Colin WM, a normalized scalar map is computed from the FLAIR intensities. A final threshold $\tau \in \mathbb{R}$ on this map provides the mask M_τ.

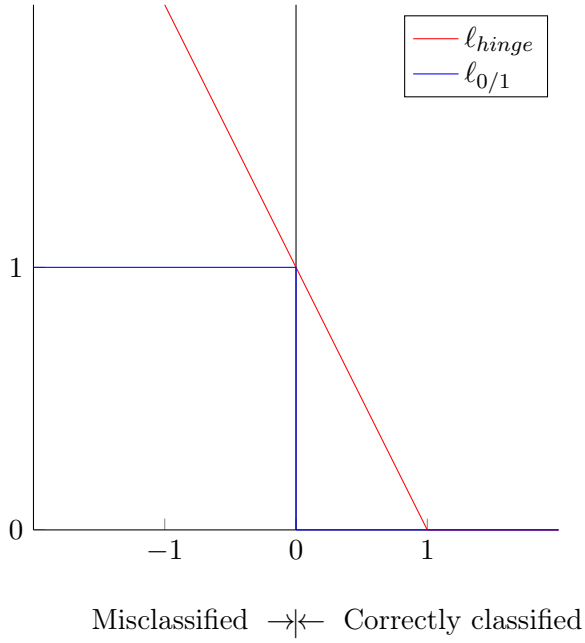


Figure 5.4: Hinge loss $u \overset{\text{def.}}{=} y \times f(\mathbf{x}) \mapsto \ell_{hinge}(f(\mathbf{x}), y)$ and misclassification loss $u \mapsto \ell_{0/1}(f(\mathbf{x}), y)$.

Definition 5.2.6 (SVM data fitting loss)**.** Given a training set of n labeled features $(\mathbf{x}_i, y_i)_{1 \leq i \leq n} \in \mathcal{X} \times \{-1, 1\}$, the *SVM data fitting loss* is defined by

$$\mathcal{L}_{fit}(f) \overset{\text{def.}}{=} \frac{1}{n} \sum_{i=1}^{n} \ell_{hinge}(f(\mathbf{x}_i), y_i), \tag{5.8}$$

where the hinge loss is defined as follows.

Definition 5.2.7 (Hinge loss)**.** The hinge loss $\ell_{hinge} \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is defined as

$$\forall (y_1, y_2) \in \mathbb{R} \times \mathbb{R}, \quad \ell_{hinge}(y_1, y_2) \overset{\text{def.}}{=} \max\{0, 1 - y_1 y_2\}. \tag{5.9}$$

Remark. Figure 5.4 compares the hinge and misclassification losses (seen as functions of $y \times f(\mathbf{x})$).

Then to avoid overfitting (see Chapter 2, in particular Section 2.1.2), a regularization term is defined.

Definition 5.2.8 (SVM regularization loss)**.**

$$\mathcal{L}_{reg}(f) \overset{\text{def.}}{=} \|f\|_K^2 \,, \tag{5.10}$$

where $\| \ \|_K$ is the norm associated to K.

Combining both terms, we get the final optimization problem.

Definition 5.2.9 (SVM optimization problem)**.**

$$f^* \overset{\text{def.}}{=} \underset{f\in\mathcal{H}_K}{\operatorname{argmin}}\, \mathcal{L}_{fit}(f) + \gamma\mathcal{L}_{reg}(f), \tag{5.11}$$

$$= \underset{f\in\mathcal{H}_K}{\operatorname{argmin}}\, \frac{1}{n}\sum_{i=1}^{n} \ell_{hinge}(f(\mathbf{x}_i), y_i) + \lambda\|f\|_K^2, \tag{5.12}$$

where $\lambda \geq 0$ is a parameter weighting the first term which controls the labeling performance and the second term which controls the smoothness of the solution.

Remark. In the literature, the regularization/weighting parameter λ is often called γ. Nonetheless, we use the notation λ in this thesis to have similar notations in all chapters.

This optimization problem can be of infinite dimension (and therefore hard to optimize). The Riesz representation theorem shows that this optimization problem can be written as a finite optimization problem. Its proof is given in Appendix A.2.2

Theorem 5.2.2 (Riesz representation theorem)**.** *The solution of* (5.11) *exists in* $\mathcal{H}_K$ *and*

$$\exists(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n, \quad \forall \mathbf{x} \in \mathcal{X}, \quad f^*(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i K(\mathbf{x}, \mathbf{x}_i)\,. \tag{5.13}$$

The optimization problem is convex because of the convexity of the hinge loss function. However as the objective function is not differentiable, the problem is reformulated with additional slack variables $\xi_1, \ldots, \xi_n \in \mathbb{R}$.

Definition 5.2.10 (SVM formulation with slack variables)**.**

$$f^* = \underset{\substack{f\in\mathcal{H}_K \\ \xi_1,\ldots,\xi_n\in\mathbb{R}}}{\operatorname{argmin}}\, \frac{1}{n}\sum_{i=1}^{n} \xi_i + \lambda\|f\|_K^2\,, \tag{5.14}$$

$$\text{subject to } \forall i \in [\![1, n]\!], \quad \xi_i \geq \ell_{hinge}(f(\mathbf{x}_i), y_i).$$

Corollary 5.2.3 (Finite dimension SVM formulation with slack variables)**.** *By plugging the expansion of* f *from* (5.13) *in* (5.14)*, the optimization problem becomes a finite dimension optimization problem. Let the matrix* $\mathbf{K}$ *be defined as* $\forall (i,j) \in [\![1, n]\!]^2, \quad \mathbf{K}_{i,j} \overset{\text{def.}}{=} K(\mathbf{x}_i, \mathbf{x}_j)$*. The optimization problem is now*

$$\min_{\substack{\alpha_1,\ldots,\alpha_n\in\mathbb{R} \\ \xi_1,\ldots,\xi_n\in\mathbb{R}}} \frac{1}{n}\sum_{i=1}^{n} \xi_i + \lambda\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha}\,, \tag{5.15}$$

$$\textit{subject to } \forall i \in [\![1, n]\!], \quad \begin{cases} \xi_i - 1 + y_i \sum_{j=1}^{n} \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \geq 0, \\ \xi_i \geq 0. \end{cases}$$

Definition 5.2.11 (SVM Lagrangian)**.** Let $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathbb{R}^n$ be the Lagrangian multipli-

ers. The Lagrangian of this problem is

$$L(\boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\nu}, \boldsymbol{\mu}) = \frac{1}{n}\sum_{i=1}^{n}\xi_i + \lambda \boldsymbol{\alpha}^T \mathbf{K}\boldsymbol{\alpha} - \sum_{i=1}^{n}\mu_i\left(\xi_i - 1 + y_i\sum_{j=1}^{n}\alpha_j K(\mathbf{x}_i, \mathbf{x}_j)\right) - \sum_{i=1}^{n}\nu_i\xi_i\,. \tag{5.16}$$

First, let us study the optimality conditions with regard to $\boldsymbol{\alpha}$ and $\boldsymbol{\xi}$. The proofs of the following propositions are given in Appendix A.2.3 and A.2.4.

Proposition 5.2.4 (Optimality condition with regard to $\boldsymbol{\alpha}$)**.** *Solving* $\nabla_{\boldsymbol{\alpha}} L = 0$ *leads to*

$$\forall i \in [\![1, n]\!], \quad \alpha_i^*(\boldsymbol{\mu}, \boldsymbol{\nu}) = \frac{y_i \mu_i}{2\lambda}. \tag{5.17}$$

Proposition 5.2.5 (Optimality condition with regard to $\boldsymbol{\xi}$)**.** *Solving* $\nabla_{\boldsymbol{\xi}} L = 0$ *leads to*

$$\forall i \in [\![1, n]\!], \quad \mu_i + \nu_i = \frac{1}{n}. \tag{5.18}$$

Definition 5.2.12 (Lagrange dual function)**.** The Lagrange dual function is

$$q(\boldsymbol{\mu}, \boldsymbol{\nu}) \overset{\text{def.}}{=} \inf_{\boldsymbol{\alpha}, \boldsymbol{\xi} \in \mathbb{R}^n} L(\boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\nu}, \boldsymbol{\mu}) \tag{5.19}$$

Now the next proposition gives the formulation of the SVM Lagrange dual function. Its proof is given in Appendix A.2.5.

Proposition 5.2.6 (SVM Lagrange dual function)**.** *The SVM Lagrange dual function reads*

$$q(\boldsymbol{\mu}, \boldsymbol{\nu}) = \begin{cases} \sum_{i=1}^{n}\mu_i - \frac{1}{4\lambda}\sum_{i,j=1}^{n} y_i y_j \mu_i \mu_j K(\mathbf{x}_i, \mathbf{x}_j) & \text{if } \mu_i + \nu_i = \frac{1}{n}, \\ -\infty & \text{otherwise.} \end{cases} \tag{5.20}$$

Now, let us define the SVM dual problem, and give its formulation (proof is given in Appendix A.2.6).

Definition 5.2.13 (SVM dual problem)**.** The dual problem consists in maximizing $q(\boldsymbol{\mu}, \boldsymbol{\nu})$ subject to $\begin{cases} \boldsymbol{\mu} \geq 0 \\ \boldsymbol{\nu} \geq 0 \end{cases}$.

Proposition 5.2.7. *The SVM dual problem is equivalent to*

$$\max_{0 \leq \boldsymbol{\mu} \leq \frac{1}{n}} \sum_{i=1}^{n}\mu_i - \frac{1}{4\lambda}\sum_{i,j=1}^{n} y_i y_j \mu_i \mu_j K(\mathbf{x}_i, \mathbf{x}_j). \tag{5.21}$$

As the prediction function can easily be computed using the α_i (see Theorem 5.2.2), it is in practice convenient to solve an optimization problem on $\boldsymbol{\alpha}$. Using the Propositions 5.2.4, 5.2.7 and the fact that $\frac{1}{y_i} = y_i$ (since $y_i \in \{-1, 1\}$), we get the following corollary.

Corollary 5.2.8. *The problem that* $\boldsymbol{\alpha}$ *must solve is*

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^n} 2\sum_{i=1}^{n}\alpha_i y_i - \sum_{i,j=1}^{n}\alpha_i\alpha_j K(\mathbf{x}_i,\mathbf{x}_j) \tag{5.22}$$

$$= \max_{\boldsymbol{\alpha}\in\mathbb{R}^n} 2\boldsymbol{\alpha}^T\mathbf{y} - \boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha}. \tag{5.23}$$

Definition 5.2.14 (Support vectors)**.** The training vectors associated with non-null values in $\boldsymbol{\alpha}$ are called the *support vectors*. We note SV the set of support vectors

$$SV \overset{\text{def.}}{=} \{\mathbf{x}_i; \quad \alpha_i \neq 0\}. \tag{5.24}$$

A geometrical interpretation of the SVM optimization is related to the notion of *margin* (see [Schölkopf 2001]).

Definition 5.2.15 (Margin)**.** The *margin* is defined as the distance between the decision boundary and the support vectors.

Remark (Maximization of the margin [Rosset 2003]). Margin maximization properties can be interesting for both the theoretical point of view (e.g. in terms of generalization error analysis) and the geometric interpretation. The SVM optimization maximizes the margin.

5.2.3.3 Other algorithms used for benchmark

Definition 5.2.16 (kNN classifier)**.** The kNN classifier classifies a feature $\mathbf{x} \in \mathcal{X}$ to the class with the highest cardinality among the k nearest neighbours of $\mathbf{x}$ in the training set.

$$\forall(\mathbf{x},y)\in\mathcal{X}\times\mathcal{Y}, \quad f_{kNN}(\mathbf{x}) \overset{\text{def.}}{=} \underset{y}{\text{argmax}} \ \text{card}\{\mathbf{x}_i \in kNN(\mathbf{x}); \quad y_i = y\}, \tag{5.25}$$

where $kNN(\mathbf{x})$ is the set of k nearest neighbours of $\mathbf{x}$.

Definition 5.2.17 (Naive Bayes classifier)**.** The Naive Bayes method computes the posterior probability of a feature $\mathbf{x}$ belonging to each class y, and classifies according the largest posterior probability (see Eq. (5.26)). To compute the parameters of the probability density of feature $\mathbf{x}$ given class y, the features are assumed conditionally independent given the class.

$$\forall(\mathbf{x},y)\in\mathcal{X}\times\mathcal{Y}, \quad f_{Bayes}(\mathbf{x}) \overset{\text{def.}}{=} \underset{y}{\text{argmax}}\, P(y|\mathbf{x}) \tag{5.26}$$

$$= \underset{y}{\text{argmax}}\, \frac{P(\mathbf{x}|y)P(y)}{P(\mathbf{x})} \tag{5.27}$$

$$= \underset{y}{\text{argmax}}\, P(\mathbf{x}|y)P(y). \tag{5.28}$$

Definition 5.2.18 (Parzen window classifier)**.** In the case of the Parzen window, the prior probability is estimated as

$$\forall(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}, \quad P(\mathbf{x}|y) \overset{\text{def.}}{=} \frac{1}{\operatorname{card}\{i; \quad y_i = y\}} \sum_{i, y_i = y} K(\mathbf{x}, \mathbf{x}_i), \tag{5.29}$$

where K is for example the Gaussian kernel. The Parzen window classifier is then found using the Bayes rule

$$f_{Parzen}(\mathbf{x}) \overset{\text{def.}}{=} \operatorname*{argmax}_y P(y|\mathbf{x}) \tag{5.30}$$

$$= \operatorname*{argmax}_y \frac{\sum_{i, y_i = y} K(\mathbf{x}, \mathbf{x}_i)}{\sum_i K(\mathbf{x}, \mathbf{x}_i)}. \tag{5.31}$$

5.2.3.4 Performance indicators

Model performances were compared using the indicators defined below. Statistical significance was also analysed via the p-values of paired t-tests [Ott 2008].

Definition 5.2.19 (Dice score [Dice 1945])**.** The Dice score (DSC) is defined as

$$\text{DSC} \overset{\text{def.}}{=} 2\frac{\text{Vol}(S \cap GT)}{\text{Vol}(S) + \text{Vol}(GT)}, \tag{5.32}$$

where S the computed segmentation, GT the ground truth and Vol an operator counting the number of voxels in a volume.

Remark. The Dice score ranges from 0 (no overlap between the predicted segmentation and the ground truth) to 1 (perfect match).

Definition 5.2.20 ((Number of) true/false positives/negatives (TP, FP, TN, FN) voxels)**.**

$$\text{TP} \overset{\text{def.}}{=} \operatorname{card}\{\omega \in \Omega; \quad S(\omega) = GT(\omega) = 1\}, \tag{5.33}$$

$$\text{TN} \overset{\text{def.}}{=} \operatorname{card}\{\omega \in \Omega; \quad S(\omega) = GT(\omega) = 0\}, \tag{5.34}$$

$$\text{FP} \overset{\text{def.}}{=} \operatorname{card}\{\omega \in \Omega; \quad S(\omega) = 1 \neq GT(\omega) = 0\}, \tag{5.35}$$

$$\text{FN} \overset{\text{def.}}{=} \operatorname{card}\{\omega \in \Omega; \quad S(\omega) = 0 \neq GT(\omega) = 1\}, \tag{5.36}$$

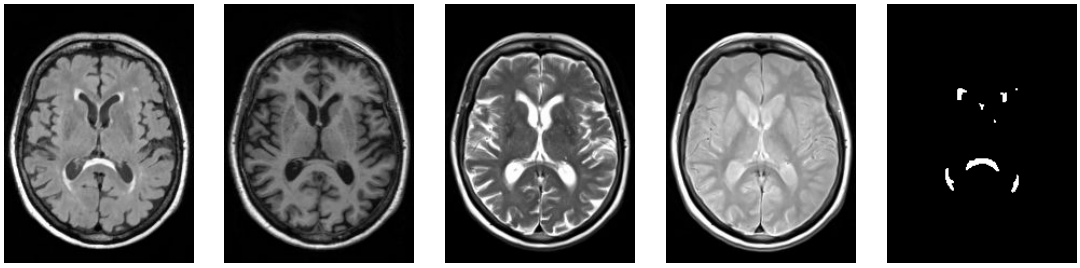
where S is computed segmentation and GT is the true segmentation.

Definition 5.2.21 (Specificity)**.** The specificity is defined as

$$\text{Spec} \overset{\text{def.}}{=} \frac{\text{TN}}{\text{TN} + \text{FP}}, \tag{5.37}$$

where TN and FP are the number of true negatives and false positives.

Remark. The specificity ranges from 0 to 1. An algorithm providing a high specificity is said to be *specific*. It does not generate many false positives, and therefore if a



(a) FLAIR (b) T1-w (c) T2-w (d) PD (e) Manual segmentation

Figure 5.5: Axial slices from one subject illustrating the different MR modalities and manual segmentation. Lesions can be seen in the FLAIR and T2-w as a bright signal.

feature is labeled as positive, there is a "high" chance that the real label is indeed positive.

Definition 5.2.22 (Sensitivity)**.** The sensitivity is defined as

$$\text{Sens} \overset{\text{def.}}{=} \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{5.38}$$

where TP and FN are the number of true positives and false negatives.

Remark. The sensitivity ranges from 0 to 1. An algorithm providing a high sensitivity is said to be *sensitive*. It does not generate many false negative, and therefore is able to "detect well" the positive features.

Remark (Recall)*.* The sensitivity is also called *recall*.

Remark. Higher is better for DSC, TP, TN, sensitivity and specificity. Lower is better for FP and FN.

5.3 Material and Results

5.3.1 Data

The dataset used in this paper comes from the Australian imaging biomarker and lifestyle (AIBL) study [Ellis 2009], where T1-w (resolution $160 \times 240 \times 256$, spacing $1.2 \times 1 \times 1$ mm in the sagittal, coronal and axial directions, TR = 2300 ms, TE = 2.98 ms, flip angle = 9°), FLAIR ($176 \times 240 \times 256$, $0.90 \times 0.98 \times 0.98$ mm, TR = 6000 ms, TE = 421 ms, flip angle = 120°, TI = 2100 ms), T2-w ($228 \times 256 \times 48$, $0.94 \times 0.94 \times 3$, TR = 3000 ms, TE = 101 ms, flip angle = 150°) and PD ($228 \times 256 \times 48$, $0.94 \times 0.94 \times 3$, TR = 3000 ms, TE = 11 ms, flip angle = 150°) images were acquired for 125 subjects. WM lesions were manually segmented by one of the authors (PR), reviewed by a neuro-radiologist and used as ground truth in the classification (Fig. 5.5).

5.3.2 Experiments

Preprocessing: Images were rigidly co-registered [Ourselin 2001], bias-field corrected [Salvado 2006], smoothed using anisotropic diffusion and histogram equalised to a reference subject. T1-w images were segmented into WM, GM, CSF using an EM approach with priors [Acosta 2009]. For each modality, features were extracted within the mask defined below, and scaled to [0, 1]. Multi-modality features were created by concatenation of single modality features. Neighbourhood intensities features ($3 \times 3 \times 3$ and $5 \times 5 \times 5$ sizes) and pyramidal features (with 4 levels, $(\sigma_1, \ldots, \sigma_4) = (0, 0.5, 1, 1.5)$) were examined.

Mask creation: The influence of the threshold τ was studied in term of classification performance bounds (see Section 5.3.3.1). Then the mask M_τ with $\tau = 2$ was computed and used for all experiments.

Machine learning: A subset of 10 000 features, with half belonging to the lesion class, the other half belonging to the non-lesion class, randomly selected and equally distributed among the training samples was used to generate the classifiers. A Matlab implementation solving SVM in its primal formulation was used [Melacci 2011, Melacci 2009]. The chosen kernel was the (Gaussian) radial basis function. The width of the kernel and the regularization weight were selected via a 10-fold cross validation. Then each image in the test set was segmented within the patient mask created. Pixels outside this region were set to the non-lesion class. As post-processing, all the connected components segmented as lesion with less than 10 voxels were removed.

Performance evaluation: The dataset was randomly split equally into training and test sets. A classifier was built using the training set, and then used to segment the test set. Then training set and test set were swapped, another classifier was built, and the rest of the segmentations were computed. Results were then merged. We performed experiments to test the influence of the combination of modalities, the influence of the feature type and the influence of using the mask in pre-processing instead of in the post-processing. The performance of the SVM classifier was compared to the performance of other supervized classification algorithms. As the overall lesion load impacts the segmentation performance, as previously reported in [Anbeek 2004], results are displayed for low (<3mL), moderate (3-10mL) and severe (>10mL) lesion loads.

5.3.3 Results

5.3.3.1 Performance bounds from mask parameter setting

The influence of the τ parameter in the mask creation on the segmentation performance was evaluated in terms of performance bounds (Fig. 5.6). Those bounds

are independent of the modality combination and feature type used later in the classification.

First, we studied the positive effects of increasing τ. As the mask was originally created to limit the number of FP, we evaluate the positive effects via the use of the "lesion-everywhere" classifier (worst-case scenario in terms of FP, FN and specificity). We notice that increasing τ decreases the upper bound of FP, increases the lower bounds of TN and specificity.

To evaluate the negative effects of increasing τ in terms of performance bounds, we use the optimal classifier within the mask (best case scenario in terms of TP, FN, sensitivity and Dice). We notice that increasing τ decreases the upper bounds of TP, sensitivity and Dice score, and increases the lower bound of FN.

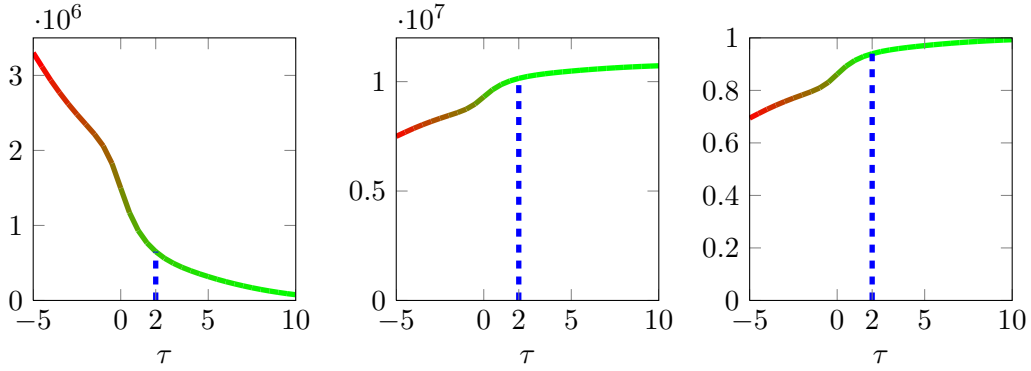
These graphs give insight on the impact of the τ parameter. If τ is too high, the final segmentation performance will be low, no matter how good is the classifier segmenting inside M_τ. If τ is too low, the algorithm has a high risk of FP, which is a known drawback in WM lesion segmentation. Setting τ therefore involves a trade-off between the use of tissue-information to reduce the risk of FP and a high upper bound performance. The value $\tau = 2$, which was selected for all experiments, is in the range of acceptable values decreasing the FP upper bound without decreasing too much the Dice upper bound.

5.3.3.2 Performance with regard to modality combinations

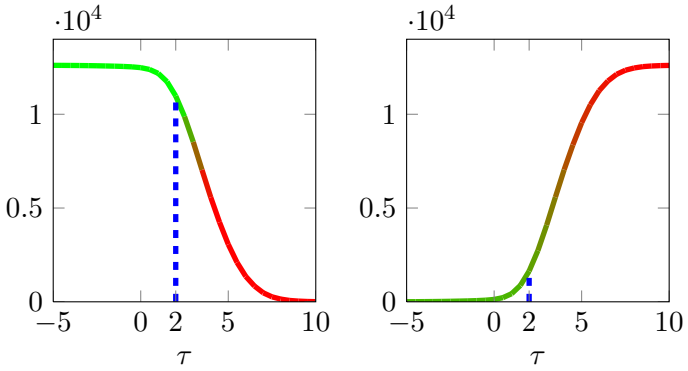
The segmentation performance was evaluated for various combinations of modalities (using $3 \times 3 \times 3$ neighourhood features). Figure 5.7 shows the values of the previously defined performance indicators for four single-modality and four multi-modality features. When using one modality, FLAIR gives the best performance. However, combining several modalities reduces FP and FN and increases TP. Table 5.1 indicates that on low and moderate lesion loads the T1-w + FLAIR combination performs statistically better than FLAIR (T2-w + FLAIR does not). On the overall dataset, the T1-w + T2-w + FLAIR combination performs statistically better than FLAIR. The model with the four modalities performs the best (Fig. 5.7), but not significantly better than T1-w + FLAIR ($p = 0.50$).

5.3.3.3 Performance with regard to feature type

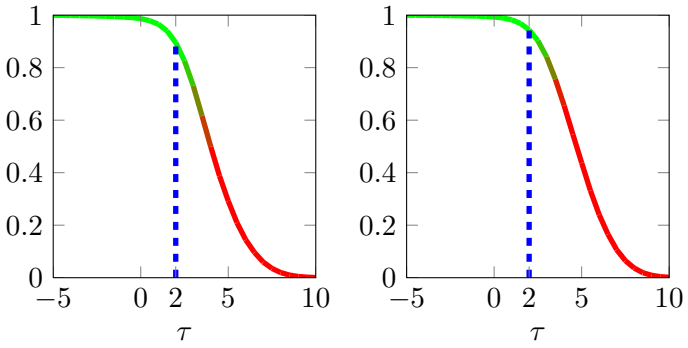
Using the four modalities, the segmentation performance was evaluated for various feature types (Fig. 5.8). With neighbourhood intensity features, a $5 \times 5 \times 5$ size slightly increased the DSC compared to $3 \times 3 \times 3$, but the difference was not statistically significant ($p = 0.93$). Pyramidal features with 4 levels did not perform as well as neighbourhood intensity features, but the DSC difference was not statistically significant ($p = 0.21$ when compared with $3 \times 3 \times 3$ features, $p = 0.18$ with $5 \times 5 \times 5$ features).



(a) FP upper bound (b) TN lower bound (c) Specificity lower bound

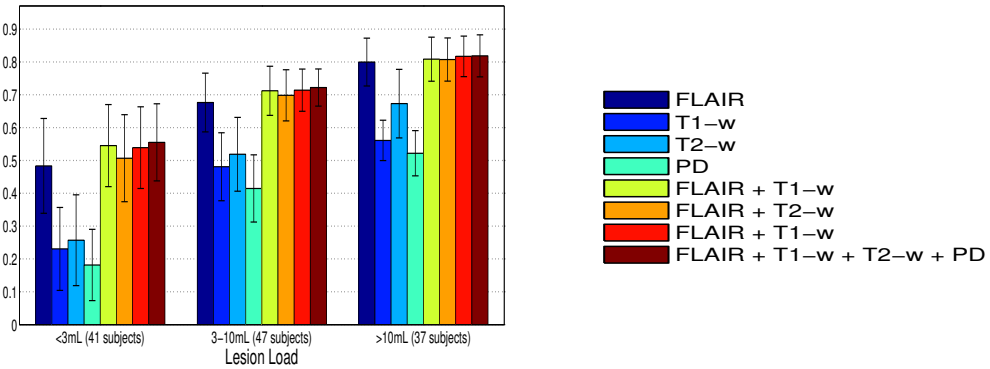


(d) TP upper bound (e) FN lower bound

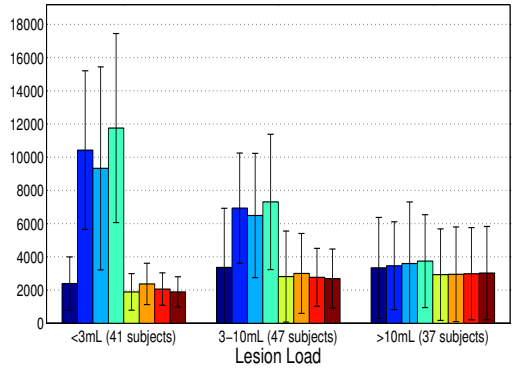


(f) Sensitivity upper bound (g) Dice upper bound

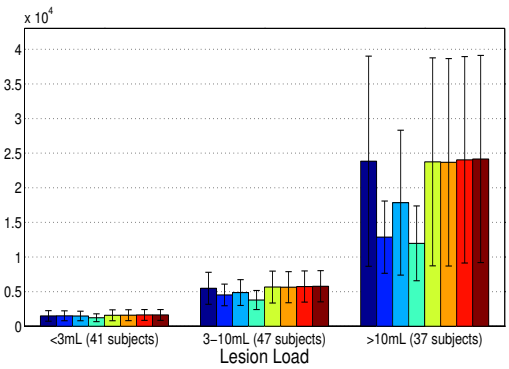
Figure 5.6: Performance bounds due to the threshold τ in the mask creation. The first line illustrates the positive effects of increasing τ: it decreases the upper bound of FP (a), increases the lower bounds of TN (b) and specificity (c). The second and third lines illustrate its negative effects: it decreases the upper bounds of TP (d), sensitivity (f) and Dice score (g), and increases the lower bound of FN (e). The value $\tau = 2$ was the value selected for all experiments.



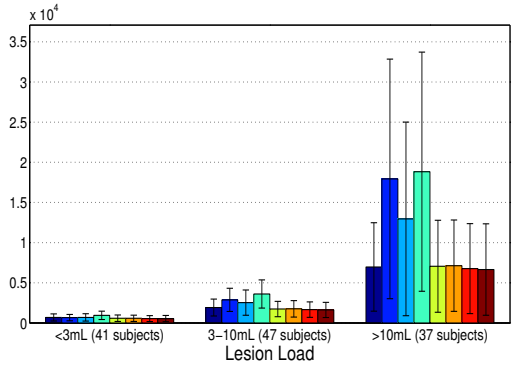
(a) Dice



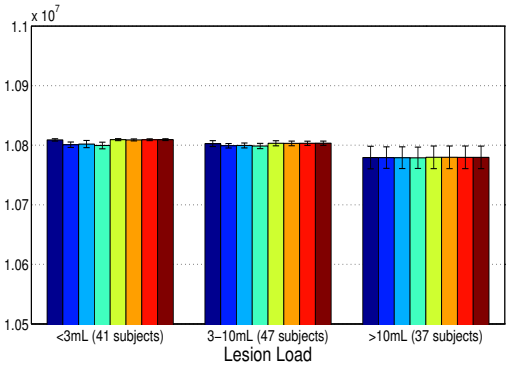
(b) False positive



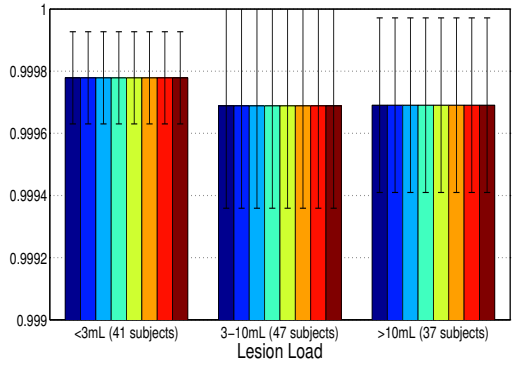
(c) True positive



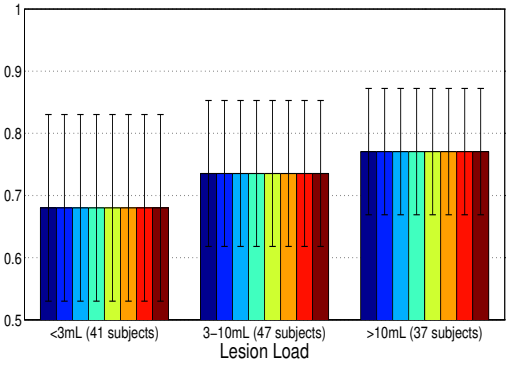
(d) False negative



(e) True negative



(f) Specificity

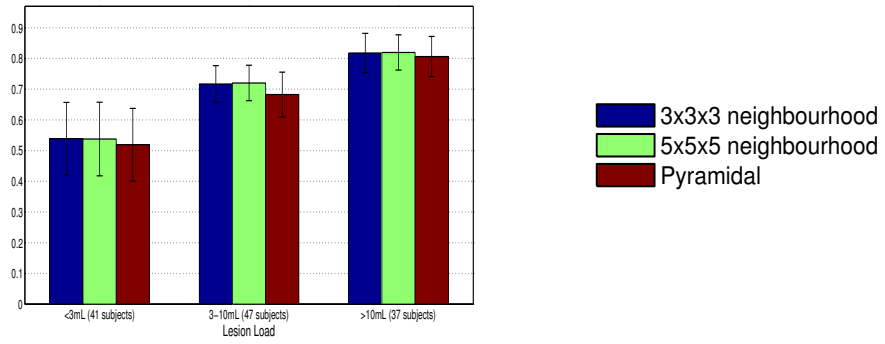


(g) Sensitivity

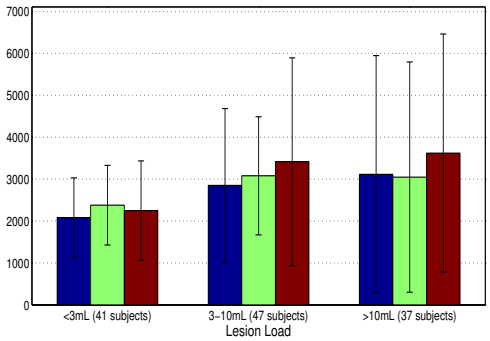
Figure 5.7: Segmentation performance with different modality combinations (using the $3 \times 3 \times 3$ neighbourhood intensity feature type).

Table 5.1: p-values of paired t-tests using $3 \times 3 \times 3$ features. Statistically significant differences ($p < \alpha = 0.05$) in bold green.

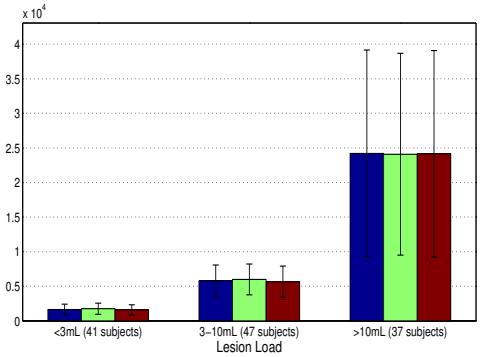
Modalities		**p-values of t-tests for lesion load in mL (number of subjects)**			
Model 1	**Model 2**	< 3 (35)	3-10 (47)	>10 (43)	Any (125)
FLAIR	FLAIR, T2-w	0.47	0.19	0.62	0.38
FLAIR	FLAIR, T1-w	**0.047**	**0.032**	0.57	0.070
FLAIR	FLAIR, T1-w, T2-w	**0.048**	**0.014**	0.26	**0.047**
FLAIR	FLAIR, T1-w, T2-w, PD	**0.011**	**0.002**	0.23	**0.014**
FLAIR, T1-w	FLAIR, T1-w, T2-w, PD	0.59	0.41	0.51	0.50
FLAIR, T1-w, T2-w	FLAIR, T1-w, T2-w, PD	0.56	0.54	0.93	0.64



(a) Dice

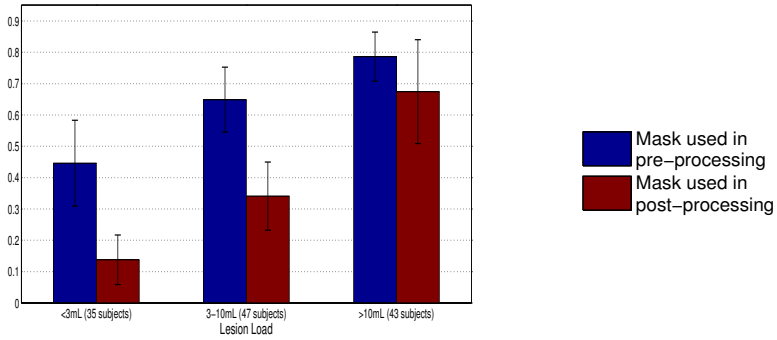


(b) False positive

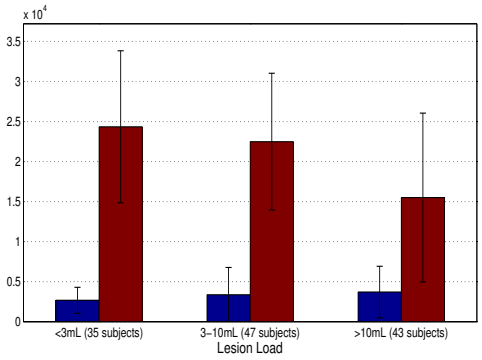


(c) True positive

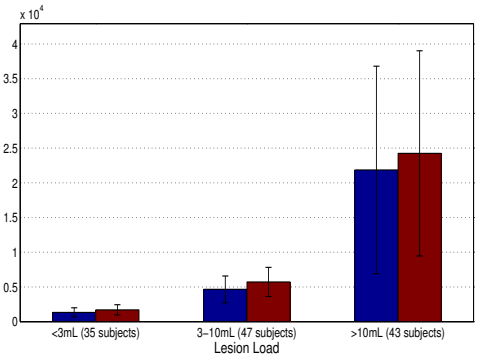
Figure 5.8: Segmentation performance with different feature types (using the 4 modalities).



(a) Dice



(b) False positive



(c) True positive

Figure 5.9: Using our mask M_τ in the pre-processing gives better results than using it only as a post-processing step.

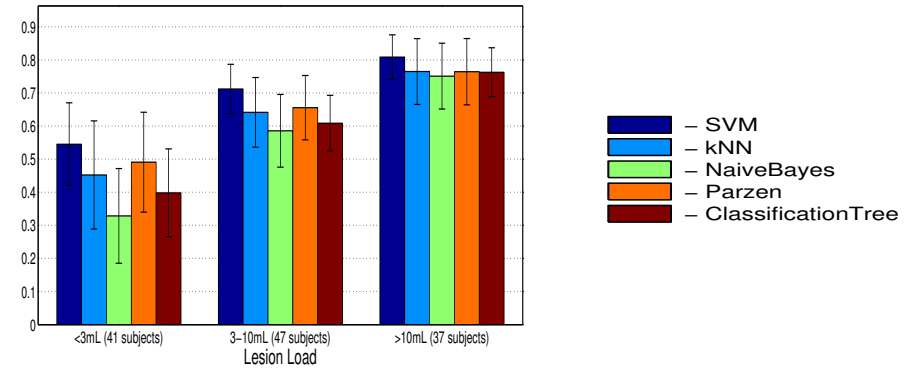
5.3.3.4 Performance with regard to the mask use

Using the FLAIR modality and $3 \times 3 \times 3$ neighbourhood feature type, the impact of the mask on the segmentation performance has been evaluated. The use of the mask in the pre-processing instead of post-processing significantly decreased FP and led to a much better DSC (Fig. 5.9). The computation time in the prediction step being linear in the number of features to label, computing predictions for a significantly lower number of features (only within the mask) reduced the computation time (41 times computation speed-up on our dataset with $\tau = 2$).

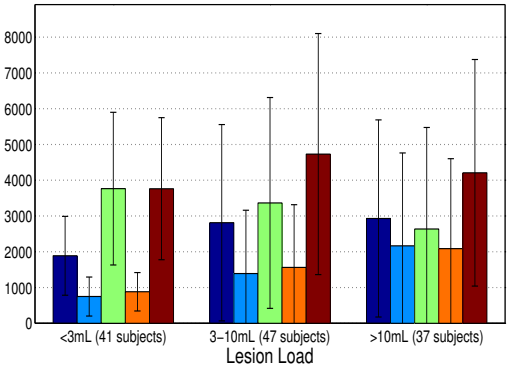
5.3.3.5 Performance comparison with other supervised classification algorithms

Using the FLAIR + T1-w combination and $3 \times 3 \times 3$ feature type, the performance of the SVM classifier was compared with several other supervised classifiers: kNN (with k=100 as in [Anbeek 2004]), Naives Bayes, Parzen window and decision tree.

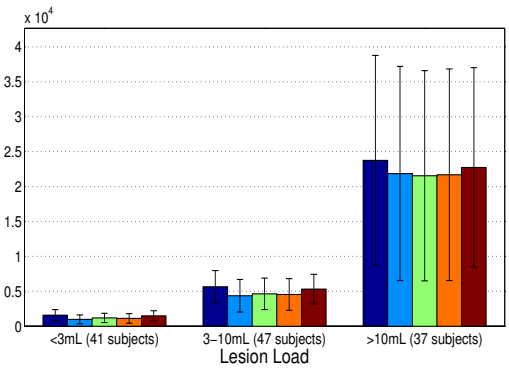
On this dataset, using combined $3 \times 3 \times 3$ features from FLAIR and T1-w



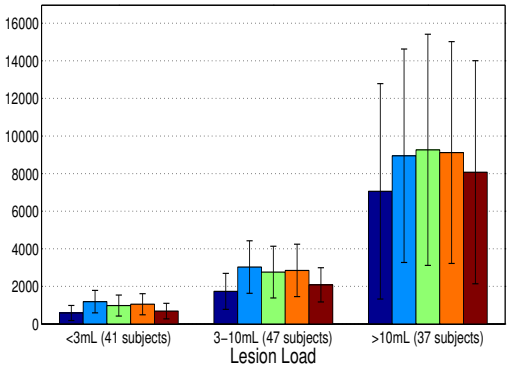
(a) Dice



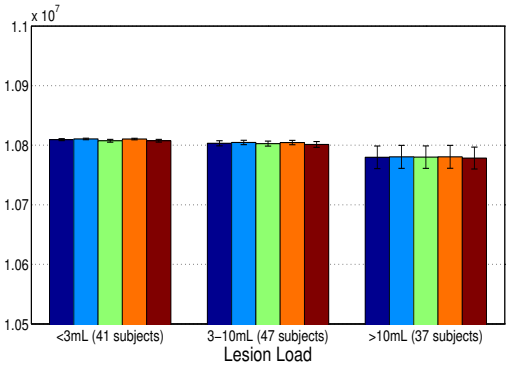
(b) False positive



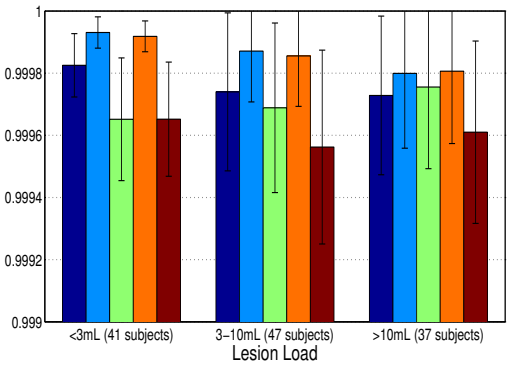
(c) True positive



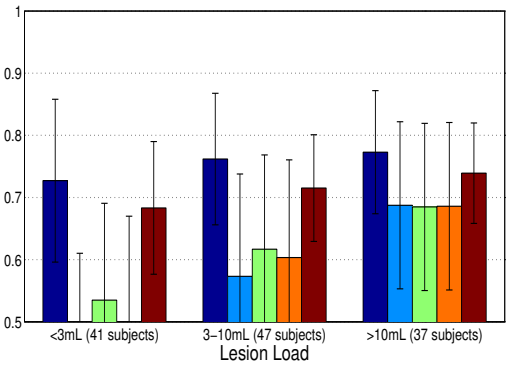
(d) False negative



(e) True negative



(f) Specificity



(g) Sensitivity

Figure 5.10: Segmentation performance with different algorithms (using the $3\times3\times3$ neighbourhood intensity feature type, FLAIR and T1-w modalities).

modalities, SVM obtained the best DSC results, followed by Parzen windows, kNN, decision tree and Naive Bayes (Fig. 5.10). In terms of FP, kNN, Parzen window and SVM provide the best results. These tree classifiers also provide the best specifities. However, kNN and Parzen window classifier have a quite-low sensitivity, whereas SVM is the most sensitive (followed by classification tree).

5.3.3.6 Performance comparison with the literature

Even though it is difficult to compare results with the literature that are obtained with different datasets or evaluated with different indicators, let us try to comment the performance results.

First, we notice that [Klöppel 2011] and [Samaille 2012] also reported higher performance of SVM compared to kNN. In our study, we only compared supervised methods. In [Samaille 2012], the authors also evaluated unsupervised methods, and reported they had lower performance on their datasets.

Second, the Dice scores we obtain are similar to the ones in [Anbeek 2004] (Dice scores ranged from 0.5 to 0.85 using up to five modalities: compared to us, they also have inversion recovery images); similar to the ones in [Samaille 2012] (average Dice score if 0.72); and better than the ones in [Klöppel 2011] (Dice scores of 0.5 and below using FLAIR and T1-w and different methods).

Finally, it is worth mentioning that [Samaille 2012] evaluated the robustness of their method in a multicenter setup, i.e. using data from different datasets. Perspectives of our work include the evaluation of our method in such setting.

5.4 Conclusion

We have presented a machine learning scheme applied to the WMH segmentation problem. Our approach is inspired by the previous work on SVM but has a number of differences. It combines the use of tissue segmentation, atlas propagation techniques and SVM classification to get efficient and accurate segmentation results. Using our pipeline and our dataset, SVM has a higher classification performance than other supervised algorithms such as kNN, Naive Bayes, Parzen window and decision tree.

In this work we also quantified the relative performance variations with regard to different modalities or feature types. Regarding the modalities, our results confirm that using all of the four modalities adds discriminative information and improves the segmentation results, as reported in [Lao 2008]. However, our quantitative results show that using only FLAIR and T1-w modalities can give similar performance at a lower cost. One reason could be the lower axial resolution of our T2-w and PD images. Regarding the feature types, there is a trade off between complexity, storage place and computation time versus the performance.

As other important contribution of this work, the mask we define and use in the pre-processing has several positive impacts. First, it improves the classifier performance as the training features are selected in regions of interest, which leads to better classifiers. We have given insight on the trade-off related to the threshold

parameter selection in the mask creation. Increasing the threshold decreases the upper bound of FP (and therefore the potential risk of final FP). However, a low FP upper bound comes at a price as increasing the threshold increases the FN lower bound and decreases the Dice upper bound, which means a threshold too high would cause poor final performance no matter how good is the classifier. Second, computation time and storage space required are significantly lower (41 times lower on our dataset with the chosen threshold) as features and predictions are computed in a restricted area. Finally, using our mask in the pre-processing makes most of the complex post-processing steps required in current state-of-art methods redundant.

Chapter 6

Image and shape analysis via manifold learning

Contents

Résumé

L'identification à grande échelle des variations anatomiques et des propriétés des tissus dans des populations constitue un challenge en imagerie médicale. Diverses méthodes d'analyse statistique, de réduction de dimension et de partitionnement ont été introduites pour apporter des éléments de réponse à ce problème. Ces techniques permettent de mieux comprendre les effets démographiques ou les facteurs génétiques des évolutions de maladie. Elles peuvent également être utilisées pour améliorer la précision ou supprimer les biais de divers algorithmes de recalage ou de segmentation. Dans ce chapitre, nous évaluons la capacité de techniques de réduction de dimension à établir des marqueurs simples du vieillissement et de la maladie d'Alzheimer à partir d'images multi-modales (IRM et PET) de 128 patients, sains, atteints de troubles cognitifs légers, ou malades. En appliquant les *Laplacian Eigenmaps* sur des images T1-w d'IRM, nous montrons que la variation principale dans cette population est la taille des ventricules. À partir d'images PiB PET, nous construisons des variétés montrant une transition suivant la rétention du radio-traceur PiB. La combinaison des deux modalités donne

des variétés dont différents endroits correspondent à différentes tailles de ventricule et différentes charges de bêta-amyloïde.

Mots clés : Analyse de population, réduction de population non linéaire, apprentissage de variétés, imagerie cérébrale

Abstract

Characterizing the variations in anatomy and tissue properties in large populations is a challenging problem in medical imaging. Various statistical analysis, dimension reduction and clustering techniques have been developed to reach this goal. These techniques can provide insight into the effects of demographic and genetic factors on disease progression. They can also be used to improve the accuracy and remove biases in various image segmentation and registration algorithms. In this chapter we explore the potential of some non linear dimensionality reduction (NLDR) techniques to establish simple imaging indicators of aging and Alzheimer's disease (AD) on a large population of multimodality brain images (magnetic resonance (MR) and Pittsburgh compound B marker (PiB) positron emission tomography (PET)) composed of 218 patients including healthy control, mild cognitive impairment and AD. Using T1-weighted (T1-w) MR images, we found using Laplacian eigenmaps (LEM) that the main variation across this population was the size of the ventricles. For the grey matter signal in PiB PET images, we built manifolds that showed transition from low to high PiB retention. The combination of the two modalities generated a manifold with different areas that corresponded to different ventricle sizes and beta-amyloid loads.

Keywords: Population Analysis, Non Linear Dimensionality Reduction, Manifold Learning, Brain Imaging

6.1 Introduction

Analyzing trends and modes in a population, as well as computing meaningful regressions, are challenges in the field of medical imaging. A considerable amount of work has been done to simplify the use of medical images for clinicians, and summarizing the information in just few imaging biomarkers, that would for example quantify and easily allow the interpretation of disease evolution. This is of great interest not only for clinical diagnosis, but also for clinical studies and to stratify cohorts during clinical trials.

Large medical databases challenge manual analysis of a population. Unbiased atlases can be used to describe a population [Lorenzen 2005]. [Blezek 2007] introduced the atlas stratification technique, discovering modes of variation in a population using a mean shift algorithm. [Sabuncu 2009] introduced iCluster, a clustering algorithm computing multiple templates that represent different modes in the population. [Davis 2007] demonstrated the use of manifold kernel regression to regress the images with regard to a known parameter, such as age. [Wolz 2009] introduced the Learning Embeddings for Atlas Propagation technique, and showed that the use of manifold learning can improve the segmentation results compared to the simple use of image similarity in multi-atlas segmentation techniques. [Gerber 2010] developed a generative model to describe the population of brain images, under the assumption that the whole population derive from a small number of brains. These techniques usually rely on computations of diffeomorphisms or transformations to compute distances between images. Alternatively it is also possible to use dimensionality reduction techniques directly on the image pixels intensities [Wolz 2009], as we propose in this paper. Most dimensionality reduction techniques rely either on information theory or geometry. Information-based assumptions can be related to the maximum of variance (principal component analysis (PCA), kernel principal component analysis (kPCA)), entropy measure, etc. Geometric assumptions are either global (multi dimension scaling (MDS), isometric mapping (ISOMAP)), or local (local linear embeddings (LLE), LEM, Hessian eigenmaps (HEM), diffusion maps (DM), local tangent space alignment (LTSA)). References to these algorithms can be found in [van der Maaten 2007].

In this publication, we examine the use of NLDR techniques to analyse multimodality brain images. Alzheimer's disease (AD) is associated with the deposition in the brain of amyloid plaques, which can be imaged with PET using the PiB, and with brain atrophy, which can be imaged with MR T1-w images. We are investigating the use of manifold learning techniques for studying PET-PiB and T1-w.

6.2 Methods

6.2.1 Global pipeline

The proposed algorithm, summarized in Fig. 6.1, consists of the following steps:

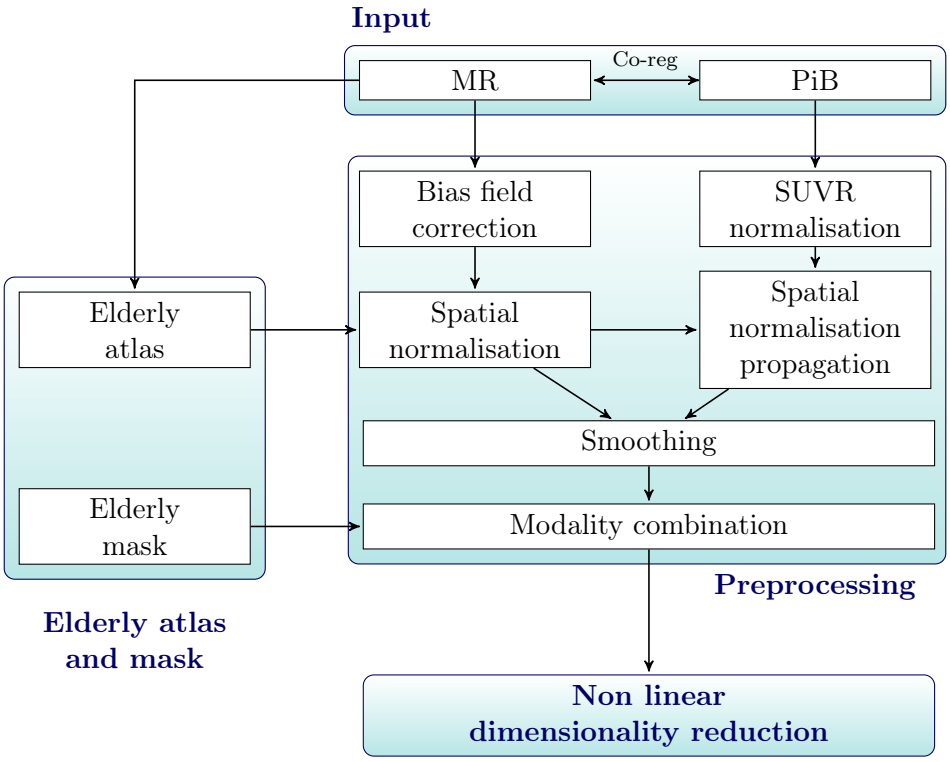


Figure 6.1: Overview of the algorithmic pipeline

Pre-processing: PiB and MR images were affinely co-registered. All PiB Images were Standardised Uptake Value Ratio normalised to the mean uptake in the cerebellum crus region [Raniga 2008]. T1-w images were bias-field corrected in the mask creation process. T1-w images were then spatially normalised using an elderly brain atlas using affine and then non-rigid transformations. These transformations were then propagated to the PiB images. Noise was reduced in T1-w images using anisotropic diffusion, and in PiB images using a 2mm Gaussian convolution.

Mask creation: using a subset of 98 MR images, an average elderly brain atlas and its associated probabilistic tissue priors (white matter (WM), grey matter (GM) and cerebro-spinal fluid (CSF)) is created from the segmentations obtained using [Acosta 2009] and a voting method. The segmentation of the atlas is used to create the mask used in the NLDR step (whole brain (union of WM, GM and CSF) or GM only).

Dimensionality reduction (dimensionality reduction (DR)): this last step is described in details in the next section.

6.2.2 Dimensionality reduction

Definition 6.2.1 (DR problem)**.** Given n vectors $\{\mathbf{x}_1, \dots, \mathbf{x}_n\} \in \mathbb{R}^d$ and a *target dimension* $\delta < d$, the dimensionality reduction consists in finding n corresponding vectors $\{\widetilde{\mathbf{x}_1}, \dots, \widetilde{\mathbf{x}_n}\} \in \mathbb{R}^\delta$ optimal in some sense.

Remark. When designing or using a DR algorithm, ones needs to answer several questions:

- How to define the notion of "optimality" of the low dimension coordinates?
- What are the assumptions on the data coming from this definition?
- Is it optional / desirable / important / critical to be able to be able to compute the low-dimension representation of a new vector $\mathbf{x} \in \mathbb{R}^d$ without recomputing everything?
- Is it optional / desirable / important / critical to be able to reconstruct the high dimension representation from the low dimension one?

Remark. In the case of medical imaging, the $\mathbf{x}_i$ are typically scalar, vectorial or tensor images. Their dimension d is generally the number of voxels in the image (possibly masked), or a multiple of it (so d~10^4 to 10^7). The number of images / patients n ranges from 10^2 to 10^4. Finally the target dimension δ depends on the final application, but usually ranges from 10^0 to 10^2.

In the literature, several NLDR strategies have been proposed. LEM first builds a weighted adjacency graph and then solves an eigenvalue optimisation problem based on the Laplacian operator. The weighted adjacency graph is usually a graph of k nearest neighbours (kNN). In this graph, each image defines one vertex, and every image is connected with an edge to its kNN. Edges are bidirectional, and weighted based on distances between images, usually using the heat kernel. ISOMAP builds a weighted neighbourhood graph (usually kNN), then computes the weights between all pairs of points using shortest paths on graphs, and finally constructs the low-dimensional embedding via an eigenvalue problem. LLE builds a kNN graph, then computes the optimal weights minimising the sum of the errors of linear reconstructions in the high dimensional space, and finally solve an eigenvalue problem to map to embedded coordinates. HEM identifies the kNN, obtains tangent coordinates by singular value decomposition, and then computes the embedding coordinates using the Hessian operator and eigen-analysis. LTSA uses the tangent space in the neighbourhood of a data point (typically the kNN) to represent the local geometry, and then align those tangent spaces to construct the global coordinate system for the nonlinear manifold by minimizing the alignment error for the global coordinate learning.

Laplacian eigenmaps LEM [Belkin 2003] is a distance-based dimensionality reduction algorithm. It aims at minimizing a weighted sum of the distances in the final space.

Definition 6.2.2 (LEM loss function)**.** The LEM loss function is defined as

$$\mathcal{L}(\widetilde{\mathbf{x}_1}, \dots, \widetilde{\mathbf{x}_n}) \overset{\text{def.}}{=} \sum_{i,j=1}^{n} w_{ij} ||\widetilde{\mathbf{x}_i} - \widetilde{\mathbf{x}_j}||^2, \tag{6.1}$$

where $\{w_{ij} \in \mathbb{R}; \quad (i,j) \in [\![1,n]\!]\}$ is a set of weights such that w_{ij} is "high" if $\mathbf{x}_i$ and $\mathbf{x}_j$ are "close".

Remark. Appropriate constraints must be added to prevent the solution to be the trivial one.

Now let us see how the weights are defined. First, a graph is built with edges connecting nearby points to each other. There are 2 variants: ε-graph and kNN graph.

Definition 6.2.3 (Distance matrix)**.** Given n objects $\mathbf{x}_1, \dots, \mathbf{x}_n$, the *distance matrix* is defined by

$$\Delta \overset{\text{def.}}{=} (d(\mathbf{x}_i, \mathbf{x}_j))_{1 \leq i,j \leq n}, \tag{6.2}$$

where $d(\mathbf{x}_i, \mathbf{x}_j)$ is a distance between $\mathbf{x}_i$ and $\mathbf{x}_j$.

Definition 6.2.4 (ε-graph)**.** Given n objects $\mathbf{x}_1, \dots, \mathbf{x}_n$ and their distance matrix Δ, the *ε-graph* is defined as a set of nodes $(n_i)_{1 \leq i \leq n}$, and edges connecting the nodes corresponding close enough objects

$$\forall (i,j) \in [\![1,n]\!]^2, \quad \eta_i \sim \eta_j \quad \text{if } d(\mathbf{x}_i, \mathbf{x}_j)^2 \leq \varepsilon. \tag{6.3}$$

Definition 6.2.5 (Graph connected)**.** A graph of nodes $(n_i)_{1 \leq i \leq n}$ is *connected* if

$$\forall i \in [\![1,n]\!], \quad \exists j \in [\![1,n]\!] \setminus \{i\}, \quad n_i \sim n_j. \tag{6.4}$$

Remark. Given n objects $\mathbf{x}_1, \dots, \mathbf{x}_n$, the ε-graph is not connected if ε is too small. (For example, one can take $\varepsilon = \frac{1}{2} \min \left\{ d(\mathbf{x}_i, \mathbf{x}_j); \quad (i,j) \in [\![1,n]\!]^2 \right\}$).

Definition 6.2.6 (Set of kNN)**.** Given n objects $\mathbf{x}_1, \dots, \mathbf{x}_n$ and their distance matrix Δ, we define the *set of k nearest neighbours* ($k \in \mathbb{N}$) by

$$kNN(\mathbf{x}_i) \overset{\text{def.}}{=} \{\mathbf{x}_{\sigma_i}; \quad j \in [\![1,k]\!]\}, \tag{6.5}$$

where $\{\sigma_j; \quad j \in [\![1, n-1]\!]\} = [\![1,n]\!] \setminus \{i\}$ and $d(\mathbf{x}_i, \mathbf{x}_{\sigma_1}) \leq \dots \leq d(\mathbf{x}_i, \mathbf{x}_{\sigma_{n-1}})$.

Definition 6.2.7 (kNN graph)**.** Given n objects $\mathbf{x}_1, \dots, \mathbf{x}_n$ and their distance matrix Δ, the *kNN-graph* is defined as a set of nodes $(\eta_i)_{1 \leq i \leq n}$, and edges connecting each object to its k nearest neighbours

$$\forall (i,j) \in [\![1,n]\!]^2, \quad \eta_i \sim \eta_j \quad \text{if } \begin{cases} \mathbf{x}_i \in kNN(\mathbf{x}_j) \text{ or,} \\ \mathbf{x}_j \in kNN(\mathbf{x}_i). \end{cases} \tag{6.6}$$

Definition 6.2.8 (Heat kernel)**.** The weights of the graph with the *heat kernel* are defined as

$$w_{ij} \overset{\text{def.}}{=} \begin{cases} e^{\frac{-d(\mathbf{x}_i,\mathbf{x}_j)^2}{\sigma}} & \text{if } \eta_i \sim \eta_j, \\ 0 & \text{otherwise,} \end{cases} \tag{6.7}$$

where $\sigma \in \mathbb{R}$ is a scaling coefficient.

Definition 6.2.9 (Simple-minded kernel)**.** The weights of the graph with the *simple-minded kernel* are defined as

$$w_{ij} \overset{\text{def.}}{=} \begin{cases} 1 & \text{if } \eta_i \sim n_j, \\ 0 & \text{otherwise.} \end{cases} \tag{6.8}$$

Remark. The simple-minded kernel is equivalent to a heat kernel with $\sigma = \infty$.

Definition 6.2.10 (Degree matrix $\mathbf{D}$ of $\mathbf{W}$)**.** Given a matrix $\mathbf{W} \in \mathbb{R}^{n\times n}$ containing the weights of a graph, we define the *degree matrix* as a diagonal matrix

$$\begin{cases} \forall i \in [\![1, n]\!], & d_{ii} = \sum_j w_{ij}, \\ \forall (i, j) \in [\![1, n]\!]^2 \text{ such that } i \neq j, & d_{ij} = 0. \end{cases} \tag{6.9}$$

Definition 6.2.11 (Graph Laplacian)**.** The graph Laplacian is a matrix in $\mathbb{R}^{n\times n}$ defined by

$$\mathbf{L} \overset{\text{def.}}{=} \mathbf{D} - \mathbf{W}, \tag{6.10}$$

where $\mathbf{D}$ and $\mathbf{W}$ are respectively the degree and weight matrices corresponding to a graph.

Proposition 6.2.1. *Using the previous definitions, the LEM loss function can be re-written as*

$$\mathcal{L}(\widetilde{\mathbf{X}}) = 2\,\mathrm{Tr}\left(\widetilde{\mathbf{X}}^T \mathbf{L} \widetilde{\mathbf{X}}\right), \tag{6.11}$$

where $\widetilde{\mathbf{X}} = (\widetilde{\mathbf{x}_1}, \ldots, \widetilde{\mathbf{x}_n})^T \in \mathbb{R}^{n\times\delta}$.

Remark. The proof is given in Appendix A.3.1.

Remark. To avoid the trivial solution, the minimization problem is solved under the constraint $\widetilde{\mathbf{X}}^T \widetilde{\mathbf{X}} = \mathbf{I}$.

Finally one can establish the LEM solution using [Lütkepohl 1996].

Proposition 6.2.2 (LEM solution)**.** *The low dimensional representation can therefore be found by solving the eigenvalue problem*

$$\textit{Find } (\mathbf{v}, \lambda) \in \mathbb{R}^n \times \mathbb{R} \textit{ such that } \mathbf{L}\mathbf{v} = \lambda \mathbf{v}. \tag{6.12}$$

Then the δ eigenvectors $(v_i)_{1\leq i\leq \delta}$ corresponding to the smallest nonzero eigenvalues form the low-dimensional data representation $\widetilde{\mathbf{X}}$, i.e.

$$\forall i \in [\![1, n]\!], \quad \tilde{\mathbf{x}}_i = \left(v_1^{(i)}, \ldots, v_\delta^{(i)}\right) . \tag{6.13}$$

6.3 Material and Results

6.3.1 Data

The dataset is composed of 218 patients from the Australian imaging biomarker and lifestyle (AIBL) study [Ellis 2009]. T1-w (image matrix $60 \times 240 \times 256$, image spacing of $1.2 \times 1 \times 1$ mm in the sagittal, coronal and axial directions, TR=2300ms, TR=2.98ms, TI=900ms, flip angle=9°) and PiB (reconstructed image matrix $28 \times 128 \times 90$, $2 \times 2 \times 2$ mm spacing) scans were acquired. Extra clinical information available in the database (such as age, clinical score, diagnosis) was also used in the visualization (but not in the creation of the low-dimension representations of the patients). Finally a preprocessed set of 3D hippocampi represented by point clouds was used.

6.3.2 Experiments

Initially LLE, LEM, HEM, and LTSA were investigated for multi modality brain imaging population analysis. Although we initially investigated several algorithms, we only report the LEM results as it was the only method we found to give stable manifold structures and that did not lead to numerical issues. In particular, HEM was found to have a prohibitive processing time. On our data, LLE had numerical stability problems that resulted from nearly-singular matrices (some eigenvalues being close to zero). LTSA did not reveal any meaningful manifold structures on our data. Moreover, several target dimensions were initially investigated, however we only report the results of 2D dimensional manifolds within, as they provided more stable and meaningful structures.

The NLDR was performed on the middle 2D slice using the mani Matlab implementation available at [Wittman 2005]. The LEM version using a kNN graph was used. The default k parameter from [Wittman 2005] ($k = 8$) was used. The robustness of the manifold with regard to k was also analysed. To compute the edge weights, we used the simple-minded version.

Low dimension representation of population of T1-w, PiB and combined T1-w/PiB were then studied. As the AIBL also provide clinical information (diagnosis, age, cognitive information), we have also computed some low-dimension representation from the images and displayed the clinical information on top of it. Finally, we have computed LEM embeddings of a preprocessed set of 3D hippocampi available in the AIBL database.

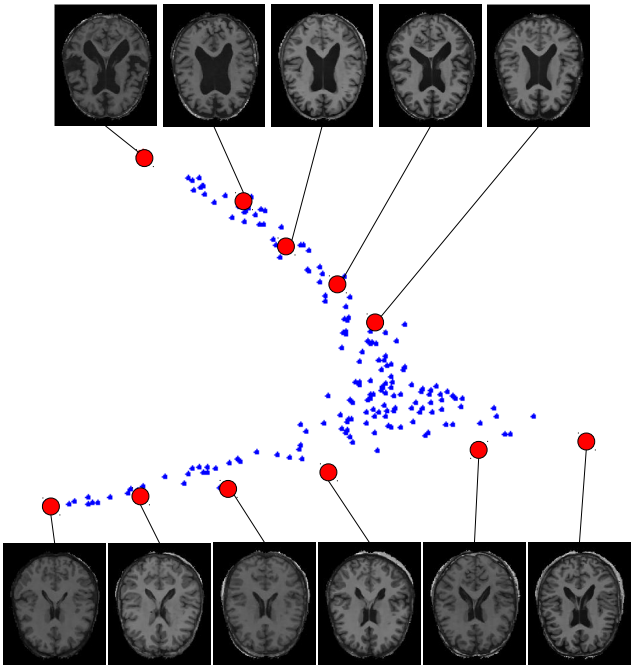


Figure 6.2: LEM embeddings using MR images registered with affine transformations and a global brain mask (218 images, input dimension: 23346, target dimension: 2). Several examples of corresponding images are also plotted showing increased ventricle size from bottom to top.

6.3.3 Results

The enlargement of the ventricles is one of the most obvious changes seen in MR images of the brain as one ages. Figure 6.2 shows the LEM embeddings (i.e. the low dimension representation of the data) in dimension 2 with the MR images using a global tissue mask corresponding to the whole brain. A structure with two branches appears. The top branch corresponds to images with large ventricles, whereas the lower branch corresponds to smaller ventricles. Figure 6.3 shows that if only the central part of the brain image is used as input data (by eroding the mask), the structure of the manifold is conserved, with the same separation of ventricle sizes.

Amyloid load as observed using PET PiB is known to be related to AD. Figure 6.4 shows LEM embeddings in dimension 2 with PiB images. When using a global brain mask (input dimension: 23346) and images registered with affine transformations (Fig. 6.4a), the point cloud obtained has a similar structure as the one with MR images (Fig. 6.2) with two branches. The images in the bottom branch have increased PiB retention compared to the ones in the top branch. With a GM mask (input dimension: 12212) and images registered with affine transformations (Fig. 6.4b), the structure with two branches disappears. However, from top to bottom, the PiB retention increases. If the images are registered non-rigidly and a GM mask is used (Fig. 6.4c), there is a structure with 2 branches, the top branch with

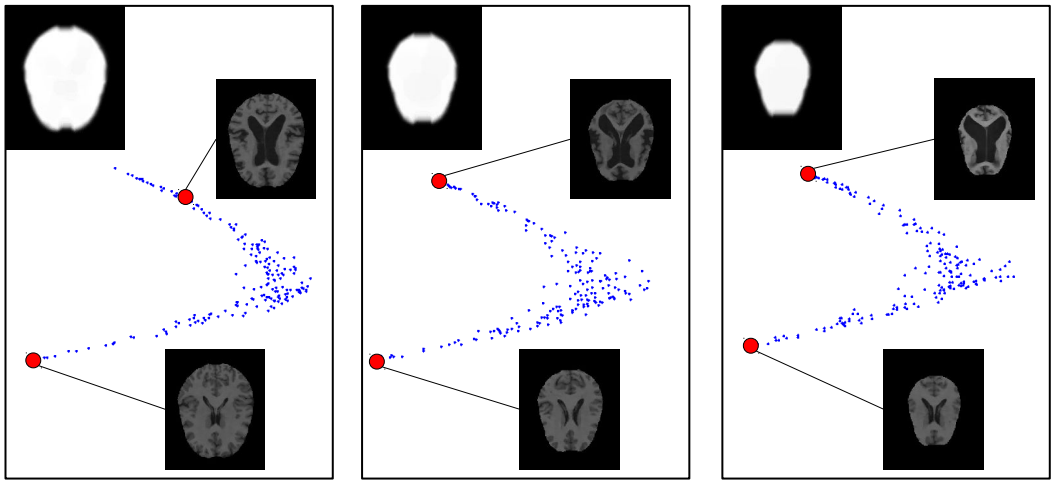


Figure 6.3: LEM embeddings using MR images (registered using affine transformations) and global brain masks more and more eroded (218 images, target dimension: 2). The structure with two branches is conserved.

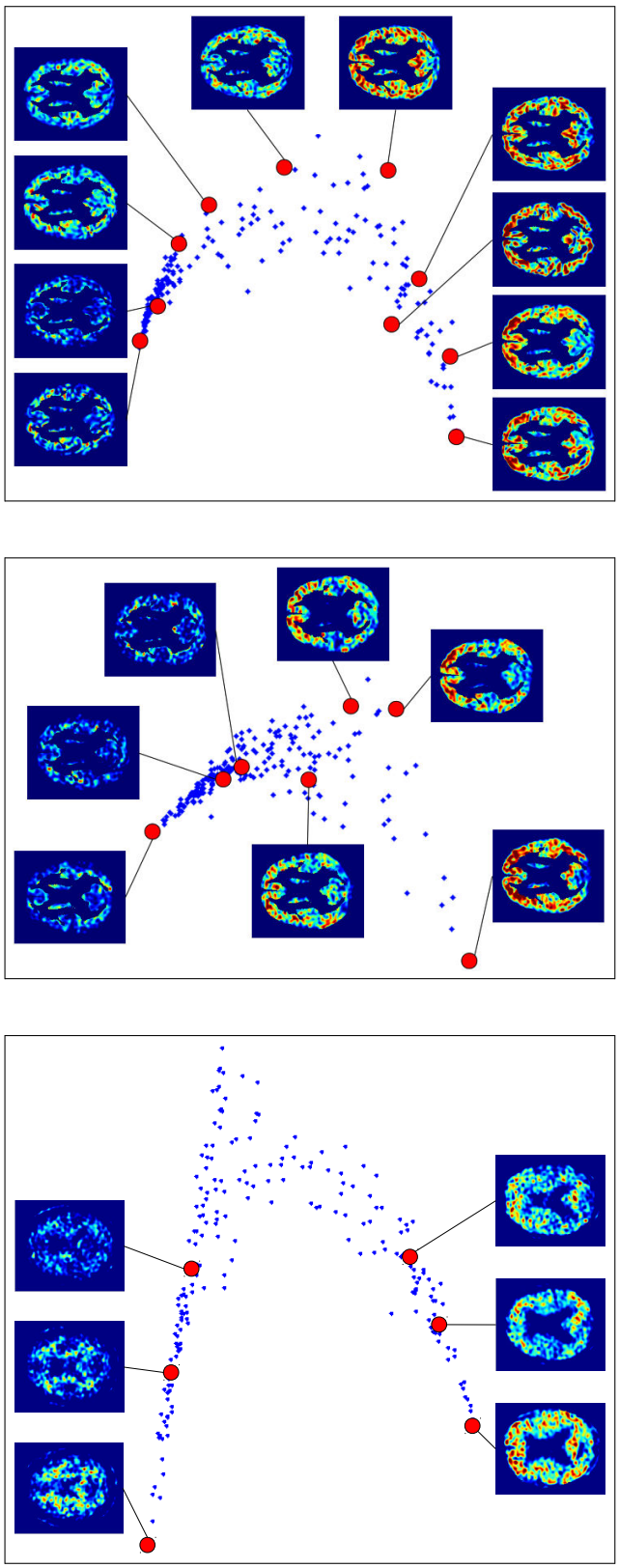
a low PiB retention, the other one with high PiB retention.

Figure 6.5a shows the LEM embeddings in dimension 2 of the data when combining the MR and PiB modalities, registered using affine transformations. Top left images have large ventricles, and bottom right images have a higher PiB retention. When images are registered non-rigidly, the structure with 2 branches appears again, and the PiB retention increases from top to bottom (Fig. 6.5b).

Figure 6.6 illustrates the robustness with regard to the number of nearest neighbours k used in the neighbourhood graph. If k is too low or too high, the structure with two branches is destroyed. A value of k too high leads to jumps between different parts on the manifold.

In Fig. 6.7, we displayed several indicators on top of the manifolds. This exploratory view can be used to identify potential correlations between different types of clinical data. It would be also interesting to see how changing the metric would impact such manifolds. Finally it motivates the NLDR methods presented in Chapter 7, which are able to combine different types of data.

In Fig. 6.8, 2D LEM embeddings from 3D hippocampi are represented. Each hippocampus is represented by a group of points on its surface. As all the hippocampi have the same number of points and point repartition on the surface, it is possible to define a distance between hippocampi by summing the distances between all their corresponding points. However, it is hard to identify a clear pattern from these LEM embeddings. To try to obtain a more meaningful structure, one can modify the distance. In Part III, we investigate other models based on deformations, which might be better for hippocampus shape analysis.

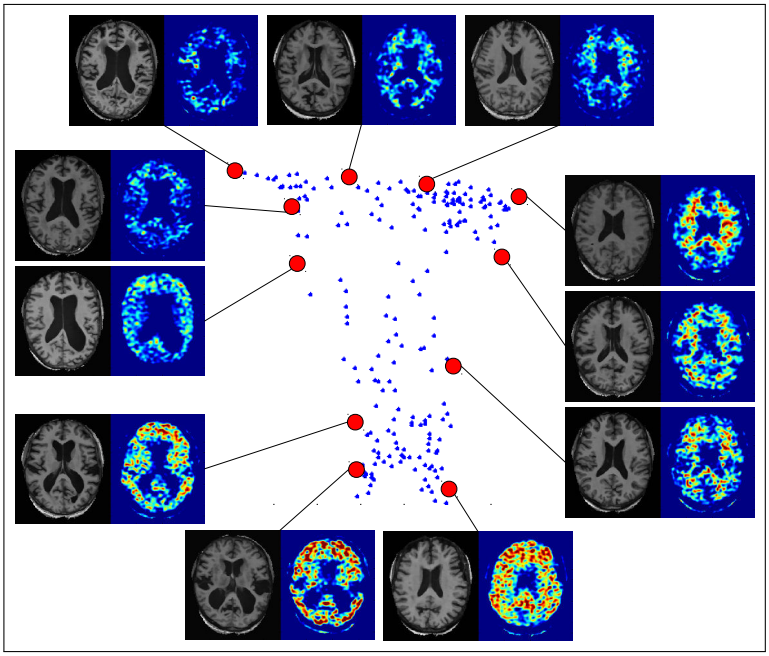


(a) Global brain mask, affine registrations

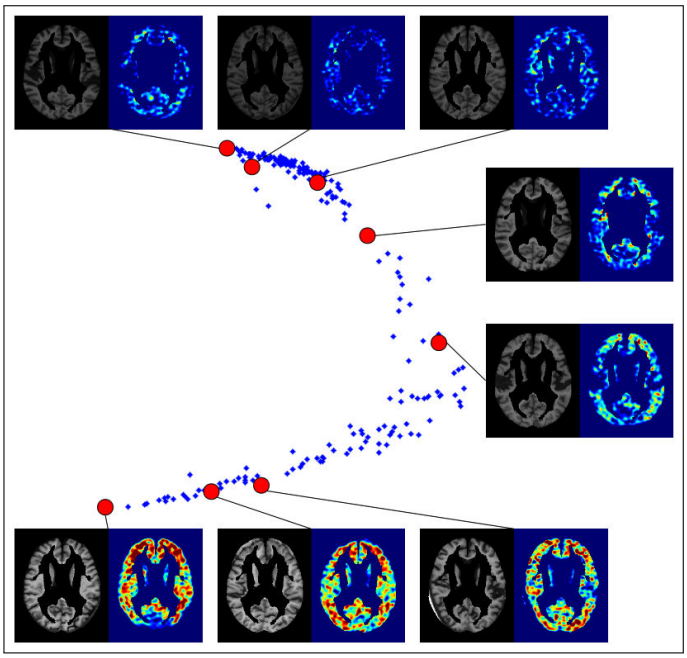
(b) GM Mask, affine registrations

(c) GM Mask, NRR registrations

Figure 6.4: LEM embeddings in dimension 2 using PiB images.

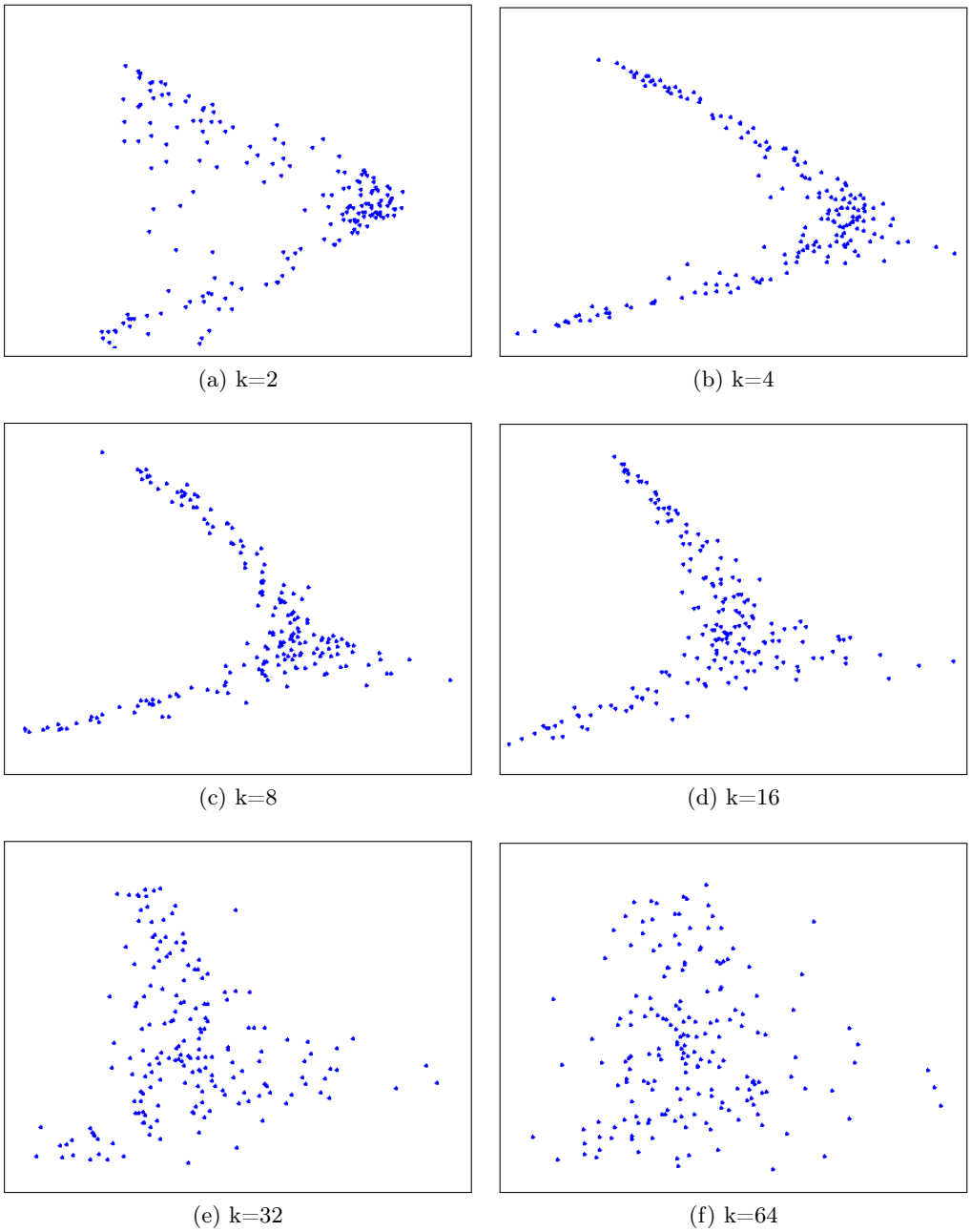


(a) With a global brain mask (input dimension: 46692)



(b) With a GM mask (input dimension: 24424)

Figure 6.5: LEM embeddings in 2D using the combination MR + PiB (registered using affine transformations).

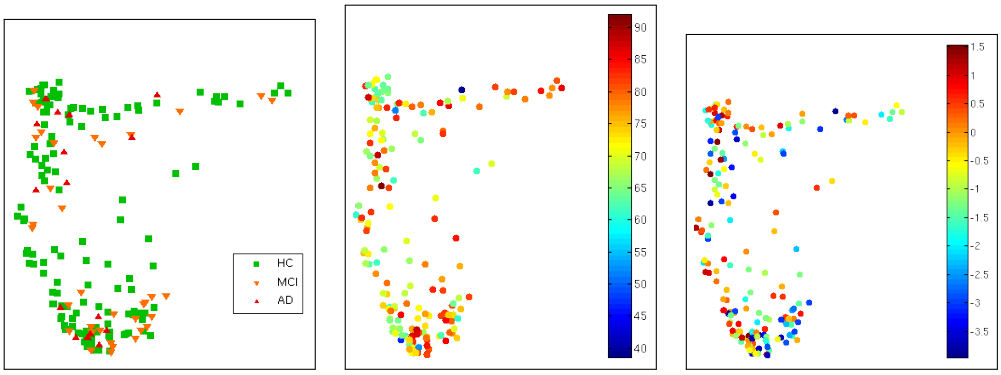


(a) k=2 (b) k=4

(c) k=8 (d) k=16

(e) k=32 (f) k=64

Figure 6.6: Test of robustness of LEM embeddings in dimension 2 with regard to K (number of Nearest Neighbours in the graph creation), using MR images and a global brain mask. If K is too low or too high, the structure with two branches gets destroyed.



(a) Diagnosis (b) Age (c) Cognitive score

(d) Diagnosis (e) Age (f) Cognitive score

Figure 6.7: Clinical information displayed on top of 2D LEM embeddings. In the first line, the embeddings were computed from MR images aligned to a template, and with a mask on the hippocampal area. In the second line, the embeddings are computed using PET-PiB images aligned non-rigidly to a template, and with a GM mask.

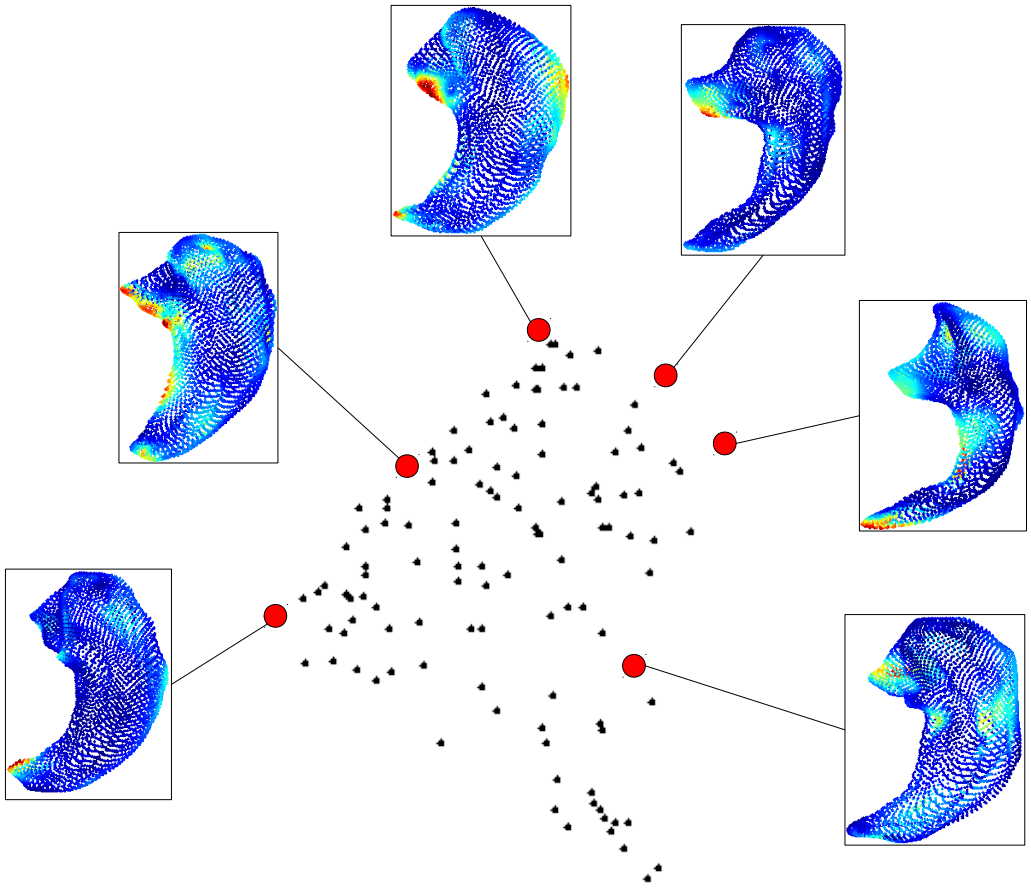


Figure 6.8: Two-dimensional LEM embeddings of 3D hippocampi. The shape of each hippocampus is represented as a set of points in $\mathbb{R}^3$. All the hippocampi have the same number of points, and the color represent the distance to the mean position in the population.

6.4 Conclusion

In this paper, we investigated the use of LEM to model PET-PiB and MR-T1-w to characterize the shape and appearance of images in a large clinical Alzheimer study. This can be particularly useful in atlas selection techniques, but can be applied in other areas. As far as shape analysis is concerned, NLDR techniques revealed that the ventricle size was the main variation in this population of brain images. The structure of the resulting 2D manifold with two branches was conserved when the cortical details were masked, leaving only the ventricles. This was expected as we used a L2 distance and many voxels were strongly affected with ventricle enlargement associated with the disease and ageing. To avoid biases from the ventricles (Fig. 6.4a), we examined only the GM voxels when studying PiB intensity (Fig. 6.4b and Fig. 6.4c).

Many studies advice to use an image metric based on deformations to analyze population of images [Gerber 2010]. Nonetheless, we have shown that a simple Euclidean distance in LEM allowed identifying a low dimensional manifold structure corresponding to some anatomical and/or intensity variations. It is expected that using L2 distance would be less computationally expensive than deformation based approaches, such as diffeomorphic or elastic registrations. This could offer faster processing especially for large databases.

Chapter 7

Manifold learning combining imaging and clinical data

Contents

Résumé

Les techniques d'apprentissage de variétés sont très populaires dans la littérature pour représenter en faible dimension les images par résonance magnétique (IRM) de cerveaux de patients. Entrainés sur ces coordonnées, des classifieurs de diagnostique visent à séparer les patients sains, les patients ayant des troubles cognitifs légers et les patients atteints de la maladie d'Alzheimer. La performance de ces classifieurs peut être améliorée en incorporant des données cliniques, disponibles dans la plupart des études à grande échelle. Cependant, les algorithmes standards de réduction de dimension non-linéaire ne peuvent pas être appliqués directement à la combinaison d'images et de données cliniques. Dans ce chapitre, nous présentons une nouvelle extension des *Laplacian Eigenmaps* permettant l'apprentissage de variétés en combinant des images et des données cliniques. Dans le cas de données cliniques continues,

cette méthode, basée sur la distance, est plus appropriée que la méthode existante basée sur une extension du graphe, qui convient dans le cas discret. Ces méthodes sont évaluées en termes de performance de classification sur une base de données provenant de l'étude ADNI et constituée d'images IRM et de données cliniques (génotypes ApoE, concentrations de $A\beta_{42}$, et scores cognitifs MMSE) de 288 patients.

Mots clés : Apprentissage de variétés, analyse de population, traitement d'image, données cliniques, maladie d'Alzheimer

Abstract

Manifold learning techniques have been widely used to produce low-dimensional representations of patient brain MR images. Diagnosis classifiers trained on these coordinates attempt to separate healthy, mild cognitive impairment and Alzheimer's disease patients. The performance of such classifiers can be improved by incorporating clinical data available in most large-scale clinical studies. However, the standard non-linear dimensionality reduction algorithms cannot be applied directly to imaging and clinical data. In this chapter, we introduce a novel extension of *Laplacian Eigenmaps* that allows the computation of manifolds while combining imaging and clinical data. This method is a distance-based extension that suits better continuous clinical variables than the existing graph-based extension, which is suitable for clinical variables in finite discrete spaces. These methods were evaluated in terms of classification accuracy using 288 MR images and clinical data (ApoE genotypes, $A\beta_{42}$ concentrations and mini-mental state exam (MMSE) cognitive scores) of patients enrolled in the Alzheimer's disease neuroimaging initiative (ADNI) study.

Keywords: Manifold learning, population analysis, image processing, clinical data, Alzheimer's disease

7.1 Introduction

Large scale population studies aim to improve the understanding of the causes of diseases, define biomarkers for early diagnosis, and develop preventive treatments. In the context of the AD, imaging biomarkers, blood biomarkers, cognitive tests, lifestyle and diet biomarkers are all potential sources of information to diagnose the disease as early as possible.

Manifold learning techniques have been used to analyze trends in populations and describe the space of brain images by a low-dimensional non-linear manifold [Gerber 2010, Wolz 2011]. These studies attempt to describe the space of brain images via a low-dimensional manifold while capturing relevant information with regard to disease diagnosis. Diagnosis classifiers trained on the low-dimension coordinates evaluate the ability to capture this information and separate healthy, mild cognitive impairment (MCI) and AD patients [Wolz 2011].

In large-scale clinical studies, an efficient representation for imaging data that captures the population variability is useful for comparison, data exploration and prediction. To build more informative manifolds, the use of non-imaging information can illuminate otherwise hidden relations. However, as the imaging and clinical data are in different spaces, the non-linear dimensionality reduction cannot be applied directly and must be adapted. We introduce a distance-based extension and compare it theoretically to an existing graph-based extension [Wolz 2011]. We also evaluate their numerical classification performances on a large dataset from the ADNI study [Mueller 2005].

7.2 Methods

7.2.1 Population analysis and diagnosis classification from manifold learning

It has been shown that the space of brain images in $\mathbb{R}^d$ can be described by a non-linear manifold of intrinsic dimension $\delta \ll d$ [Gerber 2010]. LEM [Belkin 2003] can be used to compute the low-dimensional representation of the data (Fig. 7.1). Given a matrix $\Delta_{img} \in \mathbb{R}^{n\times n}$ of pairwise distances between n images and a number of kNN $k \in \mathbb{N}$, an adjacency-graph is computed. Each node n_i represents an image, and weighted edges connecting each image to its kNN are created. From the weight matrix $\mathbf{W} = (w_{ij})_{1\leq i,j\leq n}$, a diagonal matrix $\mathbf{D}$ is computed with $d_{ii} \overset{\text{def.}}{=} \sum_j w_{ij}$. The graph Laplacian is given by $\mathbf{L} \overset{\text{def.}}{=} \mathbf{D}-\mathbf{W}$. Its eigenvectors $\{\mathbf{v}_j \in \mathbb{R}^n\}_{1\leq j\leq \delta}$ associated to the δ smallest non-zero eigenvalues provide the low-dimension coordinates $\{\tilde{\mathbf{x}}_i = (\mathbf{v}_1^i, \ldots, \mathbf{v}_\delta^i) \in \mathbb{R}^\delta\}_{1\leq i\leq n}$. Noting $\widetilde{\mathbf{X}} \overset{\text{def.}}{=} (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n)^T$, these coordinates are the solutions of the optimization problem

Definition 7.2.1 (Standard LEM optimization problem)**.**

$$\underset{\tilde{\mathbf{X}}^T\mathbf{D}\tilde{\mathbf{X}}=\mathbf{I}}{\operatorname{argmin}} \sum_{ij} w_{ij}||\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j||^2 . \tag{7.1}$$

$$\underset{\Delta \in \mathbb{R}^{n\times n}}{\text{Distance matrix}} \rightarrow \underset{\mathbf{W} \in \mathbb{R}^{n\times n}}{\text{k-NN graph}} \rightarrow \underset{\mathbf{L} \in \mathbb{R}^{n\times n}}{\text{Laplacian}} \rightarrow \underset{\widetilde{\mathbf{X}} \in \mathbb{R}^{n\times \delta}}{\text{Low dimension coordinates}}$$

Figure 7.1: Standard LEM pipeline to compute low-dimension coordinates $\widetilde{\mathbf{X}}$. The distance-based extension modifies Δ, whereas the graph-based extension modifies $\mathbf{W}$.

To evaluate how well the representation captures the disease progression we compute the classification performance of the low dimensional parameterization.

7.2.2 Extended Laplacian eigenmaps based on distance matrix combination

To combine imaging and clinical data in the manifold learning process, one can define a distance on the clinical data, combine linearly the image-based and clinical-based distance matrices, and apply the standard LEM algorithm. This extension adds two constraints to the original algorithm:

1. the need for a distance on the clinical data and,
2. the need to define a weight for the clinical data.

Definition 7.2.2 (Combined image and clinical distance)**.**

$$\Delta \overset{\text{def.}}{=} \Delta_{img} + \gamma \Delta_{clinical}, \tag{7.2}$$

where $\gamma \geq 0$ is weighting the importance of the clinical data.

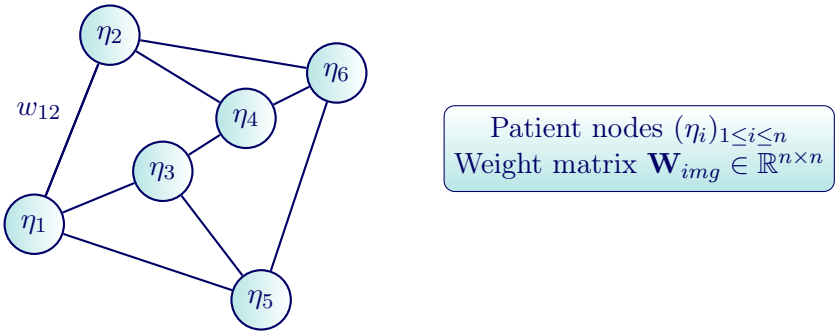
Remark. Combining the two distance matrices and applying LEM creates a graph with the same nodes but different edges and weights $\hat{w}_{ij}$ (Fig. 7.2a and 7.2b). Let us note $\hat{\mathbf{W}} = (\hat{w}_{ij})_{1\leq i,j\leq n}$ and $\hat{\mathbf{D}}$ the weight and degree matrices associated with Δ.

Proposition 7.2.1 (Distance-based LEM extension optimization problem)**.** *Using the previous notations, the optimization problem becomes*

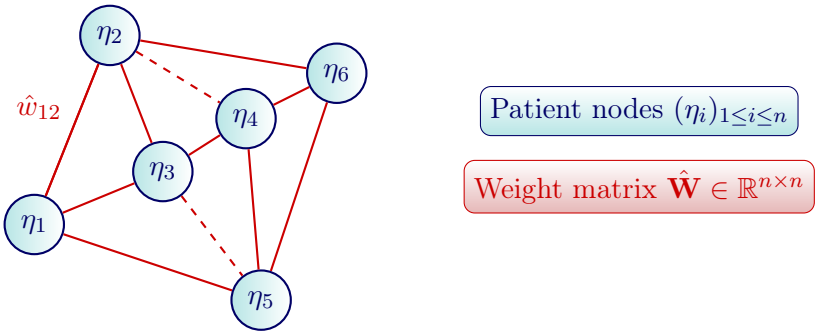
$$\underset{\widetilde{\mathbf{X}}^T\hat{\mathbf{D}}\widetilde{\mathbf{X}}=\mathbf{I}}{\operatorname{argmin}} \sum_{i,j=1}^{n} \hat{w}_{ij} \, ||\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j||^2 . \tag{7.3}$$

7.2.3 Extended Laplacian eigenmaps based on adjacency graph extension

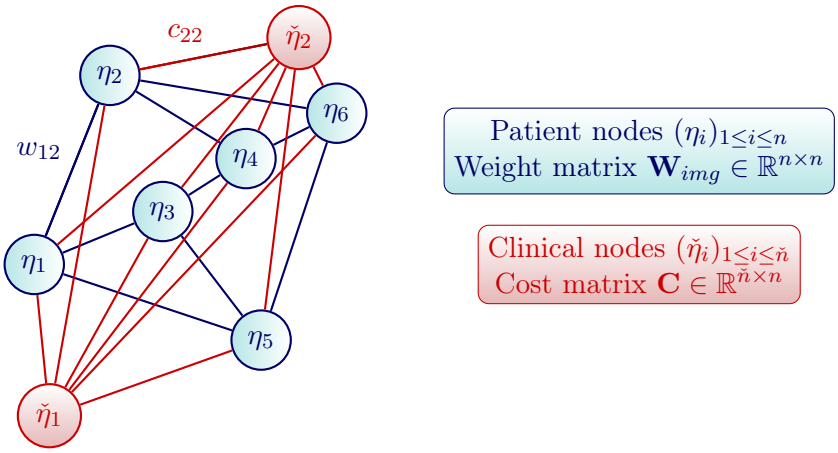
An alternative method to combine imaging and clinical data is to extend the adjacency graph by adding extra nodes and edges. One such technique has been presented in [Wolz 2011]. This extension also adds two constraints to the original algorithm:



(a) Standard LEM



(b) Distance-based LEM extension



(c) Graph-based LEM extension

Figure 7.2: Comparison of the graphs in the standard LEM algorithm and in the two extensions. When combining distances matrices, one gets a graph as in 7.2b with the same nodes as the standard LEM 7.2a but different edges and different weights. In the graph-based LEM extension, the graph 7.2c is built from the graph of the standard LEM 7.2a, then extra new nodes and weights are added.

1. a set of rules to extend the graph (extra nodes and extra weights),
2. the need to define a weight for the clinical data.

Let us assume we have a set of images $(\mathbf{x}_i)_{1\leq i\leq n}$ and a matrix of pairwise distances $\Delta_{img} \in \mathbb{R}^{n\times n}$. Let us note $(\eta_i)_{1\leq i\leq n}$ the nodes of the adjacency graph and $W_{img} \in \mathbb{R}^{n\times n}$ the weights built in the standard LEM algorithm. Now let us assume we have a set of clinical variables $(z_i)_{1\leq i\leq n} \in Z^n$. Let us see first how the graph is extended when the clinical data lies in a *discrete* space and then when it lies in a *continuous* space.

Extended graph in the discrete case: If $\operatorname{card} Z = \check{n} < \infty$, let us denote $Z = \{\overline{z}_1, \dots, \overline{z}_{\check{n}}\}$. Now let us define the extra nodes, edges and weights.

Definition 7.2.3 (Extra nodes in the discrete case)**.** In this case, a set of extra nodes is added: $(\check{\eta}_k)_{1\leq k\leq \check{n}}$.

Definition 7.2.4 (Extra edges in the discrete case)**.** Extra edges are added to connect each patient to the node of their clinical state

$$\forall (i,k) \in [\![1,n]\!] \times [\![1,\check{n}]\!], \quad \eta_i \sim \check{\eta}_k \quad \text{if } z_i = \overline{z}_k. \tag{7.4}$$

The matrix $\mathbf{C} \in \mathbb{R}^{n\times\check{n}}$ contains the extra weights

$$\forall (i,k) \in [\![1,n]\!] \times [\![1,\check{n}]\!], \quad c_{ik} \overset{\text{def.}}{=} \begin{cases} 1 & \text{if } z_i = \overline{z}_k, \\ 0 & \text{otherwise.} \end{cases} \tag{7.5}$$

Remark (Example: apolipoprotein E (ApoE) genotype). In this case,

$$\begin{aligned} Z = \{&\overline{z}_1 = (\varepsilon_2, \varepsilon_2), \overline{z}_2 = (\varepsilon_2, \varepsilon_3), \overline{z}_3 = (\varepsilon_2, \varepsilon_4), \\ &\overline{z}_4 = (\varepsilon_3, \varepsilon_3), \overline{z}_5 = (\varepsilon_3, \varepsilon_4), \overline{z}_6 = (\varepsilon_4, \varepsilon_4)\}. \end{aligned} \tag{7.6}$$

One extra node is created for each ApoE genotype, and then all patients are connected to the node of their genotype.

Extended graph in the continuous case Now let us consider the case $\operatorname{card} Z = \infty$ (typically $Z = \mathbb{R}^{\check{d}}$). In this case, Wolz et al. proposed to partition this continuous space and set the weights as the fuzzy probabilities of belonging to each partition.

Definition 7.2.5 (Partition of the clinical space)**.** Let us assume the space Z is partitioned into $\check{n} \in \mathbb{N}$ parts.

$$Z = \cup_{i=1}^{\check{n}} Z_i, \tag{7.7}$$

such that $\forall (i,j) \in [\![1,\check{n}]\!]^2$ with $i \neq j, \quad Z_i \cap Z_j = \emptyset$.

Definition 7.2.6 (Extra nodes in the continuous case)**.** In this case, a set of extra nodes is added: $(\check{\eta}_k)_{1\leq k\leq \check{n}}$, where $\check{n}$ is the number of parts in the partition.

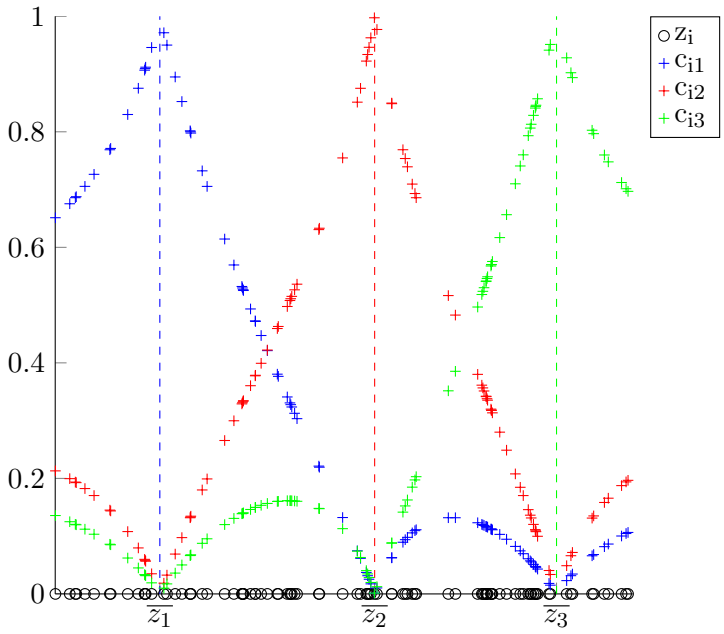


Figure 7.3: Numerical simulation: given a random set $\{z_i \in \mathbb{R}; \quad i \in [\![1, n]\!]\}$ with $n = 100$, the coefficients $\{c_{ik}; \quad (i, k) \in [\![1, n]\!] \times [\![1, \check{n}]\!]\}$ ($\check{n} = 3$) are computed according to the equation (7.9).

Definition 7.2.7 (Extra edges in the continuous case)**.** Extra edges are added to connect each patient to the nodes of all parts of Z

$$\forall (i, k) \in [\![1, n]\!] \times [\![1, \check{n}]\!], \quad \eta_i \sim \check{\eta}_k \,. \tag{7.8}$$

The matrix $\mathbf{C} \in \mathbb{R}^{n \times \check{n}}$ contains the extra weights designed as fuzzy probabilities of being in the parts of Z

$$\forall (i, k) \in [\![1, n]\!] \times [\![1, \check{n}]\!], \quad c_{ik} \overset{\text{def.}}{=} \frac{\frac{1}{d(z_i, \overline{z_k})}}{\sum_{k=1}^{\check{n}} \frac{1}{d(z_i, \overline{z_k})}} \,, \tag{7.9}$$

where d is a distance on Z, and $\overline{z_k} \in Z_k$ is a chosen "center" of the partition Z_k.

Remark (Case $Z = \mathbb{R}$). Wolz et al. proposed to use $\check{n} = 3$ partition defined using the minimum, the 33% and 67% percentiles, and the maximum computed empirically on the dataset.

Remark (Numerical simulation). Figure 7.3 illustrates the computation of the coefficients $\{c_{ik}; \quad (i, k) \in [\![1, n]\!] \times [\![1, \check{n}]\!]\}$ from a random set $\{z_i \in \mathbb{R}; \quad i \in [\![1, n]\!]\}$, with $n = 100$ and $\check{n} = 3$.

Remark (Examples). $A\beta_{42}$ concentration or MMSE clinical score are examples of continuous clinical variables.

Now let us see the final weight matrix and cost function corresponding to the extended graph.

Definition 7.2.8 (Weight matrix of the extended graph)**.** The weight matrix of the extended graph reads

$$\check{\mathbf{W}} \overset{\text{def.}}{=} \begin{pmatrix} \mathbf{I} & \frac{\gamma}{2}\mathbf{C}^T \\ \frac{\gamma}{2}\mathbf{C} & \mathbf{W}_{img} \end{pmatrix}, \tag{7.10}$$

where $\mathbf{I} \in \mathbb{R}^{\check{n}\times\check{n}}$ is the identity matrix, $\mathbf{W}_{img} \in \mathbb{R}^{n\times n}$ is the weight matrix of the standard LEM on images, $\mathbf{C} \in \mathbb{R}^{n\times\check{n}}$ contains the weights of the extra-edges, and $\gamma \geq 0$ is weighting the clinical data versus the imaging data.

Remark. Figure 7.2c represents the extended graph.

Proposition 7.2.2 (Solution of the extended LEM with graph extension)**.** *When extending the graph by $\check{n}$ nodes, we are now looking for $\check{\mathbf{X}} = (\check{\mathbf{x}}_1, \ldots, \check{\mathbf{x}}_{\check{n}}, \tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n) \in \mathbb{R}^{(\check{n}+n)\times\delta}$ as a solution of the optimization problem*

$$\underset{\check{\mathbf{X}}^T\check{\mathbf{D}}\check{\mathbf{X}}=\mathbf{I}}{\operatorname{argmin}} \sum_{ij} w_{ij} \, ||\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j||^2 + \gamma \sum_{ik} c_{ik} \, ||\tilde{\mathbf{x}}_i - \check{\mathbf{x}}_k||^2 . \tag{7.11}$$

7.3 Material and Results

7.3.1 Data

A dataset of 288 MR images from 101 patients enrolled in the ADNI study[1] [Mueller 2005] has been used to compare the diagnosis classification performances of the standard LEM algorithm and its two extensions.

As clinical data, ADNI provides the ApoE genotype. Three ApoE alleles exist $(\varepsilon_2, \varepsilon_3, \varepsilon_4)$, and since each individual carries two alleles, six ApoE genotypes are possible. The ε_4 allele has been shown to increase the risk of developing AD, whereas ε_2 decreases this risk [Macdonald 2000]. Moreover an $A\beta_{42}$ protein analysis of CSF is provided. A decrease in the concentration of this protein has been shown to be associated with a development of AD [Mueller 2005]. Table 7.2 summarizes the clinical information for the various diagnostics in the dataset.

7.3.2 Experiments

The 288 images were intensity normalized by histogram equalization to the ICBM152 atlas [Mazziotta 2001] used as template. All the images were then rigidly registered to the atlas using [Ourselin 2001]. The image distance matrix was computed using the Euclidean distance on the hippocampus area. The ApoE genotype was used considering all possible pairs of alleles and considering ApoE carriers as in [Wolz 2011], respectively leading to 6 and 3 extra nodes in the graph-based extension (see Tab. 7.1). For the graph-based extension with the continuous clinical

[1] http://www.loni.ucla.edu/ADNI

Table 7.1: Table of correspondences between ApoE genotypes and ApoE carriers from [Wolz 2011].

ApoE genotype	ApoE carriers
$(\varepsilon_3, \varepsilon_3)$	Standard carrier
$(\varepsilon_2, \varepsilon_2)$ $(\varepsilon_2, \varepsilon_3)$	ε_2-carrier
$(\varepsilon_3, \varepsilon_4)$ $(\varepsilon_4, \varepsilon_4)$	ε_4-carrier
$(\varepsilon_2, \varepsilon_4)$	*Undefined*

variables ($A\beta_{42}$ and MMSE), 3 extra nodes were added as in [Wolz 2011]. Adjacency graphs were 100-nearest neighbor graphs with edges weights computed using the Gaussian kernel with a kernel width equal to the standard deviation of the distance matrix coefficients. LEM was applied with target dimension $\delta \in [\![2, 100]\!]$. The classifiers used were 50-nearest neighbor classifiers. Training set and test sets were built using a leave-5%-out scheme. The optimal target dimension in LEM and optimal λ (resp. γ) were automatically selected from a 20-cross validation on the training set on $\{1, \ldots, 100\} \times \{0.1, 1, 2, 5, 10, 25\}$ (resp. $\{1, \ldots, 100\} \times \{0.001, 0.01, 0.1, 0.5, 1, 2, 5, 10, 25\}$).

7.3.3 Results

Table 10.1 presents the classification performance of the standard LEM algorithm using the imaging data, and the two extensions using the combined imaging and clinical data. Using clinical data combined with imaging data improves classification results for both methods compared to the standard LEM on only imaging data. For the discrete clinical variable ApoE genotype, the two extensions have similar performance on this dataset. For the continuous clinical variables $A\beta_{24}$ CSF concentration and MMSE cognitive score, the distance-based extension performs better than the graph-based extension.

7.4 Discussion

We have presented two extensions of LEM able to perform non-linear dimensionality reduction with data from different spaces, such as imaging and clinical data. Both methods come with two additional constraints. In particular, they both need to set an extra parameter to balance how much weight is given to the clinical information versus the weight of the imaging information.

From a theoretical point of view, the graph-based extension seems more natural when the clinical variable is in a finite discrete space, whereas the distance based extension seems more natural when the clinical data lives in a continuous space. First, when the clinical variable's space is a finite discrete space, it is easy to add one node per possible value and edges with weights equal to one for class mem-

Table 7.2: Number of patients, ApoE genotypes, mean and standard deviation of $A\beta_{42}$ concentration in CSF and MMSE cognitive scores are shown for the normal controls (NC), mild cognitive impairment (MCI) and Alzheimer's disease (AD) patients.

Diagnosis	N	ApoE genotype					$A\beta_{42}$	MMSE cognitive score
		$(\varepsilon_2, \varepsilon_2)$	$(\varepsilon_2, \varepsilon_3)$	$(\varepsilon_2, \varepsilon_4)$	$(\varepsilon_3, \varepsilon_3)$	$(\varepsilon_3, \varepsilon_4)$	$(\varepsilon_4, \varepsilon_4)$	
NC	94	0	12	0	65	17	0	210.15 ± 58.15	29.28 ± 1.02
MCI	114	0	2	0	58	46	8	160.48 ± 43.50	26.62 ± 1.92
AD	80	0	0	3	26	37	14	137.53 ± 24.54	21.53 ± 4.74

Table 7.3: Diagnosis classification accuracy (%) from low-dimension coordinates from the standard LEM algorithm or its extensions.

Data	DR algorithm	NC - MCI	NC - AD	MCI - AD
Imaging	LEM	65.6	63.3	61.9
Imaging & ApoE carriers	Distance-based LEM extension	66.7	71.8	66.1
	Graph-based LEM extension	65.8	73.0	64.8
Imaging & ApoE pairs	Distance-based LEM extension	62.3	62.5	66.0
	Graph-based LEM extension	63.8	65.8	65.3
Imaging & $A\beta_{42}$	Distance-based LEM extension	70.7	75.5	67.1
	Graph-based LEM extension	65.2	70.7	65.5
Imaging & MMSE	Distance-based LEM extension	83.2	93.1	67.1
	Graph-based LEM extension	65.8	75.3	68.8

berships. However, using the distance-based extension when the clinical variable is in a discrete space requires to define a distance on that space. Depending on the problem, this can raise difficult questions. In the case of the ApoE genotype, we can for example wonder if creating a distance being equal to one between all pairs of different genotypes is really optimal. Having $d((\varepsilon_2, \varepsilon_2), (\varepsilon_4, \varepsilon_4))$ higher than $d((\varepsilon_2, \varepsilon_2), (\varepsilon_2, \varepsilon_3))$ would not be absurd given the known biological impact of the ApoE alleles [Macdonald 2000]. This example illustrates that the distance-based extension is not necessarily well suited for discrete clinical variables. On the other hand, when the clinical variable lives in a continuous space such as $\mathbb{R}^n$, many distances are commonly associated (e.g. distances from l^p norms $\|x\|_p \overset{\text{def.}}{=} (\sum_i {x_i}^p)^{1/p}$). However, if one wants to use the graph extension technique, it is obviously impossible to add an infinite number of nodes. So the continous space has to be discretized into a finite number of subparts. At this point, using memberships to these subparts would mean that each z value would be considered as being one of the $\overline{z_k}$. To avoid this huge loss of information, Wolz et al. introduced fuzzy memberships. Nonetheless, there is no natural way to select the number of elements of the partition. In their paper, Wolz et al. have a clinical variable in $z \in \mathbb{R}$, they add $\check{n} = 3$ extra nodes, and the weights were defined by the minimum of z, its 33% and 67% percentiles and its maximum value, but this choice is rather arbitrary.

From a numerical point of view, when the graph-based LEM extension is used with a continuous clinical variable, the divisions in the c_{ik} can be sources of numerical instability.

7.5 Conclusion

In this chapter, we have introduced a novel extension of LEM able to perform non linear dimensionality reduction while combining imaging data and clinical data which are in different spaces. This distance-based extension leads to a graph with the same nodes as from the standard LEM but with different edges and weights, whereas the previously existing graph-based extension leads to a graph where all the nodes and edges from the standard LEM are kept and extra ones are created. This new distance-based extension is better suited for a continuous clinical data than the graph-based which is well-suited when the clinical variable lives in a finite discrete space.

We have shown that both extensions improve the numerical classification performance compared to the original LEM on a large dataset from ADNI. Performances of both extensions are similar with the discrete ApoE genotype clinical value, and our new distance-based extension have higher classification accuracy with the continuous clinical variables $A\beta_{42}$ CSF concentrations and MMSE clinical scores. This performance increase indicates a better representation of the data with regard to disease progression.

In terms of generalization of the two extensions to other dimensionality reduction algorithms, the existing graph-based extension can potentially be adapted only if

the dimensionality reduction process is based on a graph. Our new distance-based extension is more general and can be directly used in any dimensionality reduction algorithm that requires a distance of pairwise distances between all objects as input.

Perspectives of this work also include the investigation of other clinical information, such as phosphorylated tau (ptau). And to go even further, the presented methods could be used to combine images and several clinical indicators at the same time. This is particularly interesting as it would be a way to also explore possible *interactions* between different variables. A recent study [Manczak 2013] identified that the interactions of $A\beta_{42}$ and ptau may be implied to synaptic dysfunction and neuronal damage, and our method could be way to further investigate these results.

Part III

Longitudinal population analysis

Chapter 8

State of the art

Contents

Résumé

Au début du 20ème siècle, D'arcy Thompson a introduit la notion d'étude de la variabilité biologique d'un point de vue mathématique [Thompson 1917]. Dans ses travaux, il a comparé différentes espèces de poissons via des transformations géométriques. Avec le développement des technologies d'imagerie médicale, le domaine de l'*anatomie numérique* a généré un vif intérêt dans les communautés scientifiques, tant du point de vue théorique que appliqué.

Pour analyser les variations anatomiques, le recalage est un élément clé. Dans la littérature, ce problème est souvent formalisé comme un problème variationnel comprenant un terme de similarité et un terme de régularisation, et dont une solution est une transformation optimale du problème. Dans ce cadre variationnel, de nombreux algorithmes ont

été proposés dans la littérature, avec diverses classes de transformation, divers termes de similarité, divers termes de régularisation, et diverses stratégies d'optimisation. Dans ce chapitre, nous présentons différents modèles de déformations difféomorphiques. Ensuite, nous présentons différents algorithmes permettant de calculer un atlas de population. Finalement, nous introduisons la notion de transport suivant une déformation et différents méthodes de transport. A la fin de chaque section, nous mentionnons les modèles utilisés dans les chapitres suivants.

Mots clés : Anatomie numérique, forme, déformation, analyse longitudinale de population, recalage, atlas, transport

Abstract

Back to the early 20^{th} century, Sir D'arcy Thompson introduced the idea of studying the biological variability from a mathematical point of view [Thompson 1917]. In his work, he compared different fish species via geometrical transformations. With the development of medical imaging technologies, the field of *computational anatomy* aroused a lot of interest in the scientific communities, both from theoretical and applied points of view.

To analyze anatomical variations, the registration problem plays a key role. In the literature, this problem is usually formalized as a variational problem, where an optimal transformation is computed, balancing a matching term and a regularization term. Within this variational framework, a large number of algorithms have been proposed, with different classes of transformations, different matching terms, different regularization terms, and different optimization strategies. In this chapter, we present several diffeomorphic image deformation models, Then, we present several algorithms to compute a population atlas. Finally, we introduce the notion of transport according to a deformation and several transport methods. At the end of each section, we mention which models are used in the following chapters.

Keywords: Computational anatomy, shape, deformation, longitudinal population analysis, registration, template, transport

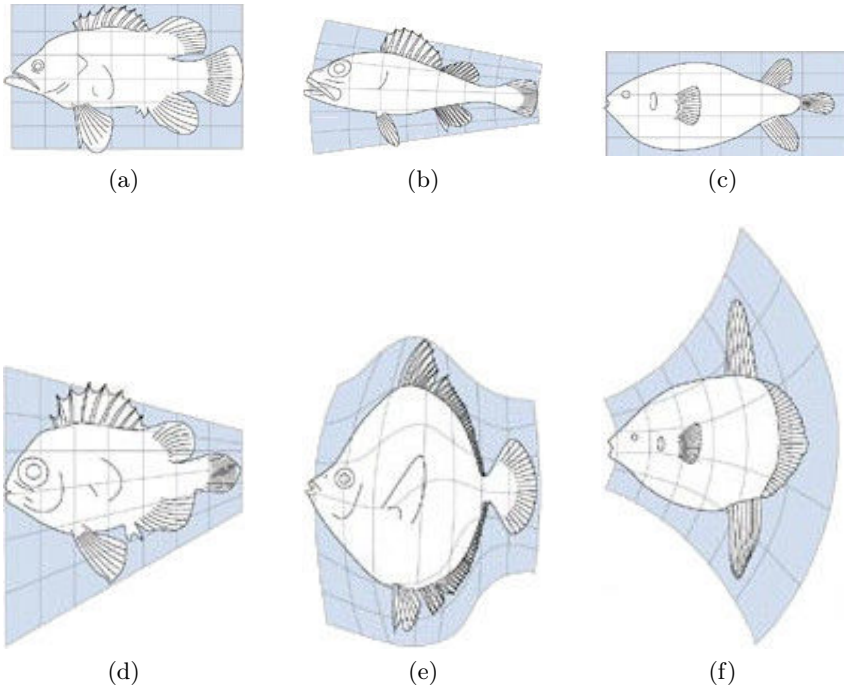


Figure 8.1: Transformations relating different fish to each other. Each fish on the second line is obtained by deforming the corresponding fish above.

8.1 Computational anatomy

The term *anatomy* refers to the "science of the shape and structure of organisms and their parts"[1]. *Computational anatomy* consists in studying the anatomy via the numerical analysis of images. Its aims range from the construction of representative atlases of populations (i.e. anatomical models) to the estimation of the variability of organs across species, the identification of biomarkers of disease progression, the modeling of longitudinal evolution, the detection of correlation with other genetic or phenotype information, etc.

The idea of analyzing the biological variability of shapes from a mathematical point of view goes back to the early 20^{th} century, where Sir D'arcy Thomson related different fish species via geometrical transformations [Thompson 1917] (see Fig. 8.1). This key idea is the starting point of many deformation models, where ones builds a metric on shapes via a metric on the space of deformations.

Computational anatomy is a discipline at the interface of mathematics, geometry and statistics [Grenander 1998, Miller 2004]. Several types approaches voxel-based morphometry (VBM), deformation-based morphometry (DBM), tensor-based morphometry (TBM) have been introduced and widely studied in the literature [Frackowiak 2004, Chapter 36]. In VBM [Ashburner 2000, Mechelli 2005], voxel-

[1] http://www.thefreedictionary.com/anatomy

wise statistics such as tissue probabilities are built. In DBM, the position of anatomical structures are studied via statistitics on the deformation fields. In TBM, local structural changes (e.g. local surface or volume change) are studied from the gradients of deformation fields.

8.2 Deformation models

In the analysis of shapes, non-rigid registration plays a key role. Indeed, as mentioned before, the comparison of two shapes can be performed by the analysis of the deformation bringing one shape to another one. In the case of population analysis, the construction of a template and the estimation of the variability of a population can be studied via deformation models. In this section, we review several well-established registration methods and list their pros and cons. In the following, let us focus on the case of images. We denote by $I\colon \Omega \subset \mathbb{R}^3 \to \mathbb{R}$ the moving image, $J\colon \Omega \subset \mathbb{R}^3 \to \mathbb{R}$ the target image, and $\phi\colon \Omega \to \Omega$ a transformation deforming the source image towards the moving image.

Remark (The small deformation framework and its limitations). In the case of small deformations, the deformation is written in the form

$$\phi \overset{\text{def.}}{=} \mathrm{Id} + u, \tag{8.1}$$

where $u\colon \Omega \to \Omega$ is a displacement field. In that setting, the inverse transformation is sometimes approximated using the subtraction of the displacement field, and the composition approximated by the voxel-wise sum of displacement fields. In [Ashburner 2007], some numerical examples are given to illustrate the limitation of this approach. In medical imaging, given the large variety of anatomies, it is often preferable to take advantage of more involved models. In this section, we present several deformation models that have been widely used in the literature in the case of images.

8.2.1 Free-forms

The free-form deformations were introduced in [Rueckert 1999]. The basic idea of free-form deformations (FFDs) is to deform an object by manipulating an underlying mesh of control points. Computed deformations are based on B-splines and under certain conditions are smooth enough to be diffeomorphisms. Statistics can be performed on the parameters encoding such deformations.

Definition 8.2.1 (Free-form deformation (FFD))**.** A FFD is a deformation $u\colon \Omega \to \Omega$ of the form

$$\forall \omega \in \Omega, \quad u(\omega) \overset{\text{def.}}{=} u_{global}(\omega) + u_{local}(\omega)\,, \tag{8.2}$$

where $u_{global}\colon \omega \in \Omega \mapsto \mathbf{R}\ \omega + \mathbf{t}$ with $\mathbf{R} \in \mathbb{R}^{3\times 3}$ a rotation matrix and $\mathbf{t} \in \mathbb{R}^3$ a

translation vector and

$$\forall \omega \in \Omega, \quad u_{local}(\omega) \overset{\text{def.}}{=} \sum_{0 \leq p_1,p_2,p_3 \leq 3} B_{p_1}(\hat{\omega}_1) B_{p_2}(\hat{\omega}_3) B_{p_3}(\hat{\omega}_3) \phi_{\tilde{\omega}_1+p_1,\tilde{\omega}_2+p_2,\tilde{\omega}_3+p_3} \quad (8.3)$$

where (ϕ_{ijk}) are the control points of a $n_1 \times n_2 \times n_3$ mesh with uniform spacing, $\forall i \in [\![1,3]\!], \quad \hat{\omega}_i = \frac{\omega_i}{n_i} - \lfloor \frac{\omega_1}{n_1} \rfloor$ and $\tilde{\omega}_i = \lfloor \frac{\omega_i}{n_i} \rfloor - 1$, B_p represents the p-th basis function of the B-spline i.e.

$$B_0(\hat{\omega}) \overset{\text{def.}}{=} (1-\hat{\omega})^3/6\,, \quad (8.4)$$

$$B_1(\hat{\omega}) \overset{\text{def.}}{=} (3\hat{\omega}^3 - 6\hat{\omega}^2 + 4)/6\,, \quad (8.5)$$

$$B_2(\hat{\omega}) \overset{\text{def.}}{=} (-3\hat{\omega}^3 + 3\hat{\omega}^2 + 3\hat{\omega} + 1)/6\,, \quad (8.6)$$

$$B_3(\hat{\omega}) \overset{\text{def.}}{=} \hat{\omega}^3/6\,. \quad (8.7)$$

Definition 8.2.2 (Free-form registration)**.** The free-form registration aims at minimizing the following cost function

$$\mathcal{L}(u) = \mathcal{L}_{fit}(u(I), J) + \lambda \mathcal{L}_{reg}(u)\,, \quad (8.8)$$

where the fitting term is for example the normalized mutual information (NMI) between the deformed source image and the target image $\mathcal{L}_{fit}(u(I), J) = \frac{H(u(I))+H(J)}{H(u(I),J)}$ (i.e. the sum of their marginal entropies divided by their joint entropy) [Rueckert 1999], and the regularization term $\mathcal{L}_{reg}(u) = \frac{1}{\text{Vol}\,\Omega} \int_\Omega \sum_{1 \leq i,j \leq 3} \left(\frac{\partial^2 u}{\partial \omega_i \partial \omega_j} \right)^2$.

In [Rueckert 2006], the authors introduced conditions for the FFD to be diffeomorphic: the maximum displacement along all axes cannot exceed a constant times the grid spacing. The value of the constant was found in [Choi 2000].

Proposition 8.2.1 (Injectivity condition)**.** *A FFD based on cubic B-splines is locally injective over all the domain if*

$$\max_{\omega \in \Omega} \|u(\omega)\|_\infty \leq \frac{1}{K} \times \Delta x, \quad (8.9)$$

where Δx is the grid spacing (assumed to be uniform) and $K \approx 2.48$. So in practice the maximum displacement along each axis is forced to be lower than $0.4 \times \Delta x$ for the FFD to be diffeomorphic.

Remark (Pros and cons of FFD).

+ simple vector statistics,

- dependency on the underlying mesh,

- inconsistency with group properties.

8.2.2 Large deformation diffeomorphic metric mapping (LDDMM)

The LDDMM framework [Dupuis 1998] is a framework that was widely studied and used in the literature, as it is able two register two objects that are "far-away" from each other. This framework has been applied to various shape representations. One of the first settings was landmark matching [Joshi 2000]. Later on, it was used for images (see below), currents [Durrleman 2010], etc.

Remark (Currents). The idea of the currents is to define a metric between curves and surfaces which does not assume point correspondence between structures [Vaillant 2005, Glaunès 2005, Durrleman 2008, Durrleman 2009]. They can be seen as a generalizations of distributions to vectors: whereas distributions are known though their action on smooth test functions, currents integrate smooth vector fields (they measure flux along lines or though surfaces). An interesting extension is the notion of functional currents [Charon 2013].

Now let us get back to LDDMM in the case of images.

Definition 8.2.3 (LDDMM registration)**.** The LDDMM framework [Beg 2005] is a Riemannian framework that introduces the following functional

$$\mathcal{L}(v) \overset{\text{def.}}{=} \frac{1}{2}\|I \circ \phi_{0,1}^{-1} - J\|_{L^2}^2 + \lambda \int_0^1 \|v_t\|_K^2 \, dt \,, \tag{8.10}$$

where $v\colon (t,\omega) \in [0,1] \times \Omega \subset \mathbb{R}^3 \to \Omega$ is a time dependent velocity field that belongs to a reproducing kernel Hilbert space $\mathcal{H}_K$ ([Schölkopf 2001]) of smooth enough vector fields defined on Ω, and of associated kernel K and norm $\| \ \|_K$. For $(t,\omega) \in [0,1] \times \Omega$, we note $v_t(\omega) = v(t,\omega)$.

The deformation $\phi\colon [0,1]^2 \times \Omega \subset \mathbb{R}^3 \to \Omega$ is given by the flow of v_t, i.e.

$$\forall (t,\omega) \in [0,1] \times \Omega, \quad \begin{cases} \frac{\partial \phi_{0,t}}{\partial t}(\omega) = v_t \circ \phi_{0,t}(\omega) \\ \phi_{t,t}(\omega) = \omega \,, \end{cases} \tag{8.11}$$

where ϕ_{t_1,t_2} is the deformation from $t = t_1$ to $t = t_2$.

Remark. More details about the LDDMM framework are given in Chapter 9.

Remark (Pros and cons).

+ solid mathematical foundations,

- computationally intensive for images.

Remark (Sparse extension). In [Durrleman 2013], the authors introduced a new parametrization that can be forced to be sparse in terms of control points.

Remark (Kernel bundle extension [Sommer 2013]). One of the key feature that made LDDMM very popular in its ability to handle large deformations. The ability to detect small scale deformations is also desirable in many applications, for example

in longitudinal studies where ones wants to find correlations between shape evolution and disease progression. Several framework were proposed in the literature to extend LDDMM to *multi-scale*. For example, the technique of [Risser 2011b] is used in Chapters 9 and 10. This method adds kernels of different scales, though such parameters need to be selected beforehand and are applied for all locations in the image. The approach developed in [Sommer 2011, Sommer 2013] differs, and is designed to allow a sparse deformation description across space and scales. It also removes the need for classical selection of scales. The authors implemented and illustrated their method for landmark matching, and state that extending it to images using a control point formulation poses no conceptual problem.

8.2.3 Log-demons

The log-euclidean framework has been introduced in [Arsigny 2006] and used in different registration settings [Ashburner 2007, Bossa 2007, Bossa 2008, Vercauteren 2008, Modat 2011].

Definition 8.2.4 (Flow)**.** A *flow on* $\Omega \subset \mathbb{R}^3$ is a mapping $\phi\colon [0,1] \times \Omega \to \Omega$ such that

$$\forall (t_1, t_2, \omega) \in [0,1]^2 \times \Omega, \quad \begin{cases} \phi(0, \omega) = \omega, \\ \phi(t_2, \phi(t_1, \omega)) = \phi(t_1 + t_2, \omega)\,. \end{cases} \tag{8.12}$$

For $(t, \omega) \in [0,1]^2 \times \Omega$, we note $\phi_t(\omega) = \phi(t, \omega)$.

Definition 8.2.5 (Stationary velocity field)**.** A *stationary velocity field (SVF)* is a velocity (i.e. vector) field $v\colon \Omega \subset \mathbb{R}^3 \to \Omega$ that does not depend on the time. In the literature, a SVF is sometimes called *steady velocity field*.

Remark. The use of SVF is the key difference with the LDDMM setting, where the velocity-fields are time-dependent.

Definition 8.2.6 (Log-demon deformation)**.** In the log-demon framework, the deformation belongs to the subset of diffeomorphisms generated by the flow of SVFs.

$$\forall (t, \omega) \in [0,1]^2 \times \Omega, \quad \frac{\partial \phi(\omega, t)}{\partial t} = v(\phi(t, \omega))\,, \tag{8.13}$$

with the conditions (8.12).

Definition 8.2.7 (Exponential of a SVF)**.** The flow at time one of the ODE (8.13) is called the *exponential of the SVF* v, and we note $\exp(v) = \phi_1$.

Definition 8.2.8 (Logarithm of a diffeomorphism)**.** The *logarithm of a diffeomorphism* ϕ is the unique vector field v such that $\exp(v) = \phi$. We note $log(\phi) = v$.

Definition 8.2.9 (Log-demon registration)**.**

$$\mathcal{L}(v,\hat{v}) \overset{\text{def.}}{=} \frac{1}{\sigma_i^2}\|J - I\circ\exp(\hat{v})\|_{L_2}^2 + \frac{1}{\sigma_x^2}\|\log\left(\exp(-v)\circ\exp(\hat{v})\right)\|_{L_2}^2 + \frac{1}{\sigma_T^2}\mathcal{L}_{reg}(v)\,, \tag{8.14}$$

where v is the SVF of the transformation, $\hat{v}$ an auxiliary correspondence field, $\sigma_i, \sigma_x, \sigma_T$ are weighting coefficients.

Remark (Solving the Log-demon registration). The equation (8.14) can be minimized alternatively with regard to v and $\hat{v}$.

Remark (Fast implementations). The *scaling and squaring property* $\exp(v) = \exp(v/2)\circ\exp(v/2)$ allows efficient numerical computation [Arsigny 2006]. More recently, a new algorithm has been proposed, based on a series in terms of the group exponential and the Baker-Campbell-Hausdorff formula [Bossa 2008].

Remark (Pros and cons).

+ efficient numerical methods,

- some mathematical properties are not guaranteed, for instance whether the one-parameter subgroups are still geodesics or if the space is complete ([Lorenzi 2012b]).

8.2.4 Other models

Metamorphosis The framework of *metamorphosis* [Trouvé 2005, Holm 2008] is similar to the LDDMM framework in the sense that deformations are also built from smooth enough time-dependent velocity fields. However, the key difference is that to deform a source image towards a target image, the metamorphosis does not only deform the space, but also change intensity values.

Definition 8.2.10 (Metamorphosis registration)**.** The *metamorphosis* framework introduces the following functional

$$\mathcal{L}(v) \overset{\text{def.}}{=} \int_0^1 \frac{1}{2}\|I_t + \langle \operatorname{grad} I_t, v_t\rangle\|_{L^2}^2 + \lambda\|v_t\|_K^2\, dt\,, \tag{8.15}$$

with the same notations as in the LDDMM framework (see Section 8.2.2).

Remark. The choice between the LDDMM and the metamorphosis frameworks relies on the desired properties of the registration algorithm. On one hand, in an application where topological changes can occur, the metamorphosis might be better. It is indeed able to deal with topological changes where the LDDMM could lead to unrealistic deformations. On the other hand, when studying a set of shapes with common topology, the LDDMM is probably more suited. For example, on a set on binary hippocampus immages, the topology is the same for all images, and ones wants to capture transform a source image by a deformation but not by changing voxel intensities.

8.2.5 Choice of a deformation model

In the beginning of Section 8.2, we have listed several deformation models that have been successfully applied in medical imaging. When choosing a deformation model and a registration algorithm (i.e. a particular implementation solving the deformation model), one needs to answer several questions:

- What is the purpose of the deformation model? Is it "only" for registration? Is it to build shape statistics?
- What shape representation will it work on? Lines? Surfaces? 2D/3D scalar images? Vector fields? Tensor images?
- What properties are vital/desirable for the computed deformations?
- What are the computational speed constraints? What is the dimension of the data that will be processed? How many runs of the algorithm will be necessary?

Model used in the following chapters In Chapters 9 and 10, the LDDMM model will be used. Our goal will be to build disease classifiers based on descriptors of longitudinal evolutions. In particular, we aim to study the potential of *initial momenta* of the LDDMM framework for classification of disease progression and identification of biomarkers. Even though SVF-based approaches appear as good alternatives to the LDDMM framework for many applications, they have not been designed to estimate geodesic transformations. Finally, the collaboration with François-Xavier Vialard and Laurent Rissert helped both in terms of understanding of the LDDMM deformation model and in terms of implementation[2] and parameter optimization.

8.3 Population template

As explained in [Thompson 2000], a population template is a powerful research tool with a wide range of applications. In the following section, we review several algorithms that have been introduced in the context of brain imaging. In the whole section, let us note $(I^i)_{1\leq i\leq n}$ the population images.

Remark (Groupwise registration). In the literature, when an algorithm outputs both a population template and the registered images of the population towards this template, it is sometimes referred as a *groupwise registration* algorithm.

8.3.1 Deterministic approaches

Notion of intrinsic mean [Pennec 1996, Fletcher 2004, Pennec 2006] In these articles, the authors introduced the notion of intrinsic mean, which is the basis

[2] http://sourceforge.net/projects/utilzreg/

of several deterministic approaches. Assuming the images of the population lie on some manifold, a *Fréchet mean* F [Fréchet 1948] is defined as a minimizer

$$F \in \underset{\hat{F}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} d(\hat{F}, I^i)^2 \,, \tag{8.16}$$

where d is a geodesic distance on the manifold of the images. Such a minimizer is not necessarily unique.

Remark. This definition is analogue to the notion of barycenter in vectorial spaces.

In practice, a solution is found via an optimization procedure. Therefore, a global minimum is not necessarily reached. When one aims to compute local minima, the solutions are called *Karcher means* or *Riemannian centers of mass* [Karcher 1977, Pennec 1996, Pennec 1999, Pennec 2006]. These local minima can be identified by a null gradient and a positive definite Hessian, and computed via iterative procedures.

[Beg 2006] In this paper, the authors introduced an algorithm to compute an average anatomical atlas using LDDMM and geodesic shooting. The key idea is to used informations about the deformations registering the current template estimate towards all the images of the population. As mentioned by the authors, averaging the deformations is not valid since the space of transformations is not a vector space. It would be reasonable in first approximation in the case of small deformations. As the LDDMM setting is designed to handle large deformations, the authors instead propose to average inital vector fields, and then use the conservation of the momentum to update the template estimate. Algorithm 1 provides the details of the procedure. Figure 8.2 illustrates the result of this procedure on heart images from a (healthy) population.

Algorithm 1: Template estimation from [Beg 2006].

Input $(I^i)_{1 \leq i \leq n}$ set of images.
Output Template T.
Initialize Set T_0 as one of the input images and compute rigid registrations $(R^i)_{1 \leq i \leq n}$ from all $(I^i)_{1 \leq i \leq n}$ towards T_0.
repeat
  compute LDDMM registration from T_k to each $I^i \circ \left(R^i\right)^{-1}$,
  average corresponding initial velocity field $v_0^{mean} = \frac{1}{n} \sum_{i=1}^{n} v_0^i$,
  update template estimate from T_k and v_0^{mean}.
until *convergence*;

[Risser 2011a] In this paper, the author introduced an algorithm that shares similarities with the one of [Beg 2006]: they are both based on LDDMM, and compute the template via an iterative procedure. In both algorithms, the computed template remains on the orbit of the initialization. However, they differ in two ways:

(a) Heart images

(b) Template

Figure 8.2: Heart images of a (healthy) population and estimated template. Source: [Beg 2006].

1. the LDDMM registration algorithm differs (see [Beg 2005] and [Vialard 2012]),

2. the computation of the update is based on averaging the initial velocity fields for [Beg 2006] and the initial momenta for [Risser 2011a].

Algorithm 2 provides the details of the procedure.

Algorithm 2: Template estimation from [Risser 2011a].

Input $(I^i)_{1\leq i\leq n}$ set of images.
Output Template T.
Initialize Set T_0 and compute rigid registrations $(R^i)_{1\leq i\leq n}$ from all $(I^i)_{1\leq i\leq n}$ towards T_0.
repeat
  compute LDDMM registration from T_k to each $I^i \circ \left(R^i\right)^{-1}$,
  average corresponding initial momenta $P_0^{mean} = \frac{1}{n}\sum_{i=1}^{n} P_0^i$,
  update template estimate from T_k and P_0^{mean}.
until *convergence*;

Remark (Extensions). Several extensions of this algorithm are introduced in Chapter 9.

[Avants 2004] In this paper, the authors use the diffeomorphic framework to flow an initial template estimate along the geodesic path towards the centroid of the population. Algorithm 3 provides the details of the procedure.

8.3.2 Probabilistic approaches

Several methods have been introduced to build a population template from a probabilistic point of view. For example, several algorithms have been proposed in the Bayesian framework.

Algorithm 3: Template estimation from [Avants 2004].

Input $(I^i)_{1\leq i\leq n}$ set of images, step size ε.
Output Template T.
Initialize Set T_0 as one of the input images, $\phi^{mean} = Id$.
repeat

1. for each time t, initialize $v^{mean} = Id$,
2. for time t, for each $i \in [\![1, n]\!]$, use the constant arc length estimation method to minimize the following functional

$$\mathcal{L}(v_t^i) = \frac{1}{2}\|I^i \circ (\phi^{mean})^{-1} - T_t \circ \phi^{-1}\| + \lambda \|v_t^i\|_L^2, \tag{8.17}$$

   with $\|v_t^i\|_L = \varepsilon$,
3. set $v_t^{mean} = \frac{1}{n}\sum_{i=1}^n v_t^i$,
4. set $\phi_{t+\varepsilon}^{mean} = \phi_t^{mean} \circ (Id + v_t^{mean})$,

until ϕ^{mean} *converges and all images are registered*;

[Allassonnière 2008] In this paper, the authors use the statistical framework introduced in [Allassonnière 2007] to address the problem of population average and estimation of the underlying geometrical variability as a MAP computation problem.

[Ma 2008] In this paper, the authors assume the populations images $(I_i)_{1\leq i\leq n}$ are generated by shooting the template though Gaussian distributed random initial momenta. The template is modeled as a deformation from a given hypertemplate.

8.3.3 Mixed approaches

[Joshi 2004] In this paper, the authors propose a method for building an atlas using intensity voxel averaging and diffeomorphic registrations in the LDDMM setting. Formally the template is estimated via the optimization of the functional

$$\mathcal{L}(T, \phi^1, \dots, \phi^n) \overset{\text{def.}}{=} \sum_{i=1}^n \|T - I^i \circ (\phi^i)^{-1}\|_{L_2}^2 + \int_0^1 \|v_t^i\|_L^2 dt\,, \tag{8.18}$$

where $(\phi^i)_{1\leq i\leq n}$ are deformations, L is a partial differential operator and $\|v_t^i\|_L^2 = \|Lv_t^i\|_{L_2}^2$. In practice, L is the Navier-Stokes operator $L \overset{\text{def.}}{=} \alpha\Delta + \beta\,\text{div} + \gamma\,\text{Id}$. This optimization problem is solved iteratively and alternatively with regard to T and $(\phi^i)_{1\leq i\leq n}$, as detailed in Algorithm 4.

Remark. As mentioned in [Risser 2011a], such method does not preserve the shape topology, since it mixes two averaging strategies: intrinsic and extrinsinc means

Algorithm 4: Template estimation from [Joshi 2004].

Input $(I^i)_{1\leq i\leq n}$ set of images, ε step size.
Output Template T.
Initialize $\phi_0^i = Id, v_0^i = 0$.
repeat
  update the transformations $\phi_{k+1}^i = \phi_k^i \circ (Id + \varepsilon v_k^i)$,
  compute $I_{k+1}^i = I^i \circ (\phi_{k+1}^i)^{-1}$,
  update the template estimate $T_{k+1} = \frac{1}{n}\sum_{i=1}^n I_{k+1}^i$,
  compute force functions $F_{k+1}^i = -\left(I_{k+1}^i - T_{k+1}\right) \operatorname{grad} I_{k+1}^i$,
  compute the velocity fields $v_{k+1}^i = L^{-1} F_{k+1}^i$,
until *convergence*;

[Fletcher 2004]. Besides, as mentioned in [Jia 2010], such method could provide a blurry group mean image from a population of sharp images (i.e. with clear anatomical structures).

[Seghers 2004] In this paper, each image is deformed by the average deformation field of all deformation fields from this image towards all the other ones. The atlas is built by averaging all the deformed images.

Remark. This method can be computationally expensive as the number of registration is proportional to the square of the number of images.

[Bhatia 2004] In this paper, the authors formalize the group-wise registration problem as an optimization problem under constraints. Each image is deformed using a FFD (see Section 8.2.1). The functional to maximize is based on the NMI, with the constraint that the sum of all the deformations must be equal to zero. The optimization is solved via a steepest gradient descent method with projection. The atlas is obtained by averaging all the deformed images.

[Jia 2010] In this paper, the authors introduce an interesting approach. One of their key motivations is that it might be difficult to register each towards all the other ones when anatomical variations are large. Therefore, they propose that in each iteration, each image is modified via the average of the deformation fields computed from the registrations towards some of its neighbors (and not towards all the other images). In the end, the template is obtained by averaging all the registered images.

Remark. The authors state they "can average all registered images to obtain the atlas since the mean image of a well-aligned dataset is sharp and keeps all major anatomical structures". Their simulation on a synthetic dataset works well in that regard. Nonetheless, this final averaging classifies this algorithm as a "mixed" approach, as there is no guarantee that on some dataset the alignment would not fail for an outlier and thus lead to a blurry template.

8.3.4 Choice of a population template model

In the beginning of Section 8.3, we have listed several template models that have been successfully applied in medical imaging. When choosing a template algorithm, one needs to answer several questions:

- Is the template representing a single anatomical structure or a collection of structures?
- Are there topological variations within the population?
- What type of shape representation is used?
- What are the complexity and computational speed constraints?

Model used in the following chapters In the Chapters 9 and 10, an extended version of [Risser 2011a] will be introduced. In these chapters, the atlas represents a single anatomical structure (the hippocampus), and the population consists of a collection of binary images. In that setting, a template with sharp boundaries is also desirable.

8.4 Transport

8.4.1 Examples of transport methods

Given a deformation $\phi\colon \Omega \to \Omega$ built according to a chosen model, the question of transport consists in modifying (we say *transporting*) an application f defined on Ω. The transport itself depends on the nature of the quantity transported, or in other words, on the image space of f. In this section, we give the definitions of a few transport methods, then examine $2D$ examples, and finally list other methods from the literature.

Definition 8.4.1 (Image transport)**.** The standard transport of an image $I\colon \Omega \subset \mathbb{R}^3 \to \mathbb{R}$ by the action of a diffeomorphism $\phi\colon \Omega \to \Omega$ is defined by

$$\forall \omega \in \Omega, \quad \widetilde{I}(\omega) \overset{\text{def.}}{=} I \circ \phi^{-1}(\omega)\,. \tag{8.19}$$

Definition 8.4.2 (Transport as a density)**.** The standard transport for a density $n\colon \Omega \subset \mathbb{R}^3 \to \mathbb{R}$ is defined by

$$\forall \omega \in \Omega, \quad \widetilde{n}(\omega) \overset{\text{def.}}{=} \det(\mathrm{Jac}_{\phi^{-1}}(\omega))\, n \circ \phi^{-1}(\omega)\,, \tag{8.20}$$

where det is the notation for the determinant.

Definition 8.4.3 (Vector field transport)**.** The transport via the standard conjugation of a velocity field $v\colon \Omega \subset \mathbb{R}^3 \to \mathbb{R}^3$ is defined by

$$\forall \omega \in \Omega, \quad \widetilde{v}(\omega) \overset{\text{def.}}{=} \mathrm{Jac}_{\phi}(\omega)\; v \circ \phi^{-1}(\omega)\,, \tag{8.21}$$

where $\mathrm{Jac}_\phi(\omega) \in \mathbb{R}^{3\times 3}$ is the Jacobian matrix defined by $\mathrm{Jac}_\phi(\omega) \overset{\text{def.}}{=} \left(\frac{\partial \phi_i}{\partial \omega_j}\right)_{1\leq i,j\leq 3}$.

Example 8.4.1 (Transport examples in 2D)**.** Let us consider a simple example, where a deformation composed of a scaling, rotation and translation, i.e. of the form

$$\phi\colon \begin{pmatrix} x \\ y \end{pmatrix} \in \Omega \subset \mathbb{R}^2 \mapsto \alpha \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \end{pmatrix} \tag{8.22}$$

where $(\alpha, \theta, t_x, t_y) \in \mathbb{R}_+^* \times [0, 2\pi] \times \mathbb{R}^2$. For convenience, we note $\mathbf{R}_\theta \overset{\text{def.}}{=} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$, $\mathbf{t} \overset{\text{def.}}{=} \begin{pmatrix} t_x \\ t_y \end{pmatrix}$. Such a transformation is illustrated in Fig. 8.3a.

Let us consider we have a scalar field $I\colon \Omega \to \mathbb{R}$ that we want to transport. In Fig. 8.3b, the image transport is used. One should note that the image on the right is a deformed version on the one on the left. Now let us study the transport as a density. The inverse transformation is $\phi^{-1} = \frac{1}{\alpha}\mathbf{R}_{-\theta}(\mathrm{Id} - \mathbf{t})$. Finally, as the determinant of a 2D rotation is equal to one, the transported image is $\frac{1}{\alpha} I \circ \phi^{-1}$. In particular, we notice that when the image is scaled by a factor α, the values are multiplied by $\frac{1}{\alpha}$, and we notice that the integral of I over the domain Ω is conserved (see Fig. 8.3c).

Now let us consider we have a scalar field $v\colon \Omega \to \mathbb{R}^2$ that we want to transport according to a transformation ϕ (see Fig. 8.4a). If the vector field is transported component by component as an image, the orientation of the vector field is not aligned anymore with the shape (see Fig. 8.4b). When using the conjugate action $Jac_\phi \circ v \circ \phi^{-1} = \alpha \mathbf{R}_\theta \circ v \circ \phi^{-1}$, the orientation gets respected (see Fig. 8.4c).

Parallel transport This method has been introduced to transport tangent vectors and applied in the Alzheimer's disease (AD) context [Younes 2007, Younes 2008, Younes 2009, Qiu 2008, Lorenzi 2011].

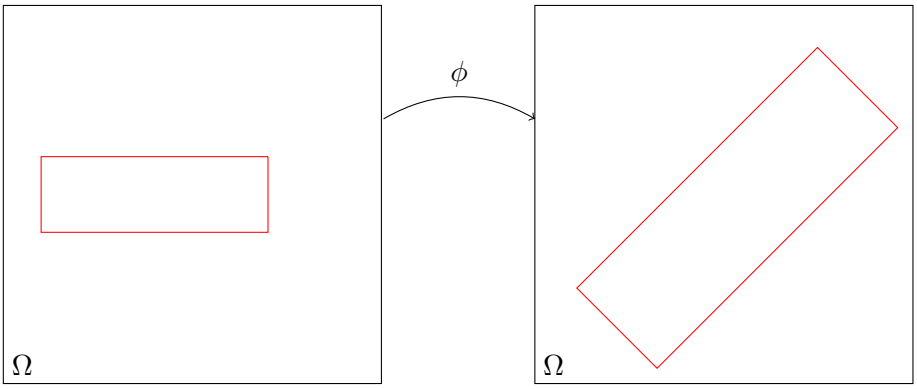
Remark. The parallel transport of a vector depends on the trajectory (see Fig. 8.5).

8.4.2 Choice of a transport method

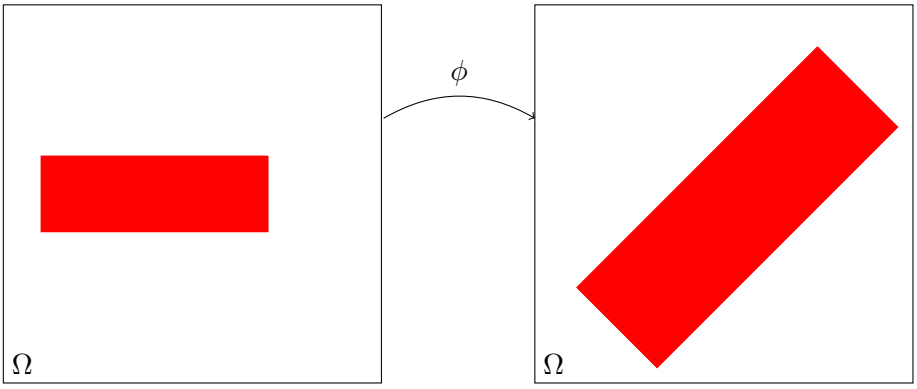
In the beginning of Section 8.4, we presented the notion of transport and mentioned several methods. When choosing a transport method, one needs to answer several questions:

- What is the quantity to be transported? (scalar field, vector field, ...)
- Which properties are desirable for the transport in a specific application?

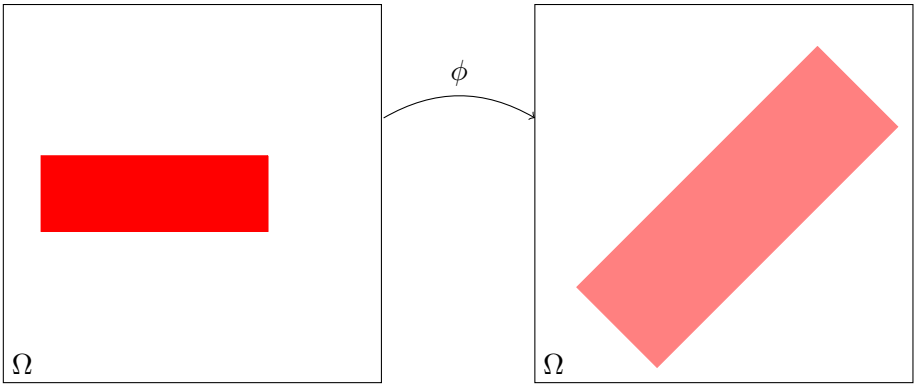
Methods used in the following chapters In Chapter 9, different models will be tested and compared in terms of classification. In Chapter 10, only one method will be used for the sake of simplicity. However, we consider the question of transport as still open.



(a) Left: original shape. Right: Shape deformed by ϕ.

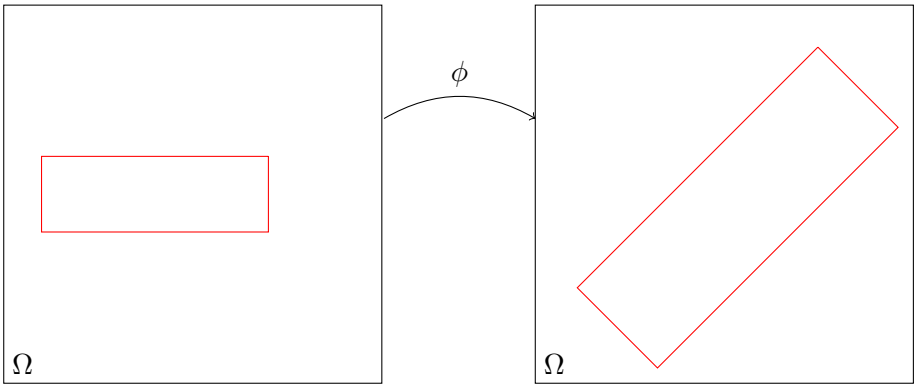


(b) Using image transport, the values are kept between the original image and the deformed image.

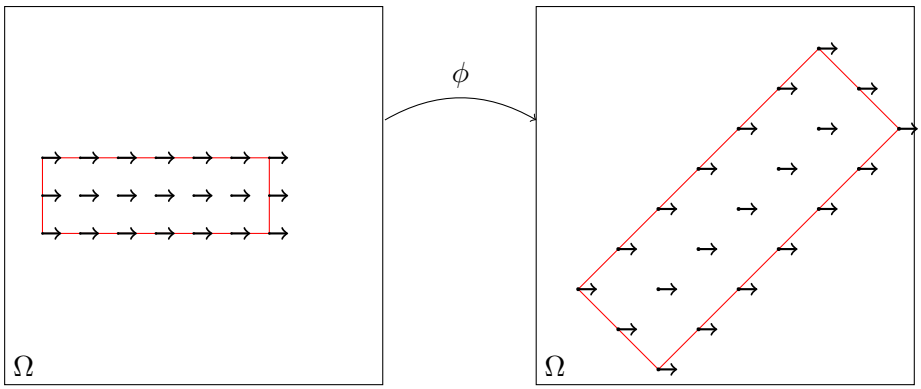


(c) Using tranport as a density, the integral over the domain Ω is conserved.

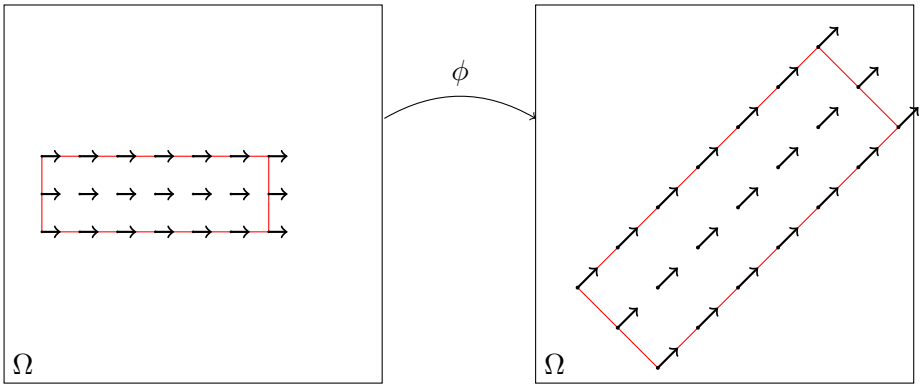
Figure 8.3: Illustration of the image transport and transport as a density for scalar fields in 2D.



(a) Left: original shape. Right: Shape deformed by ϕ.



(b) Vector field transported via the image transport of each component.



(c) Vector transport respecting the orientation with regard to the shape.

Figure 8.4: Illustration of the two vector field transports in 2D.

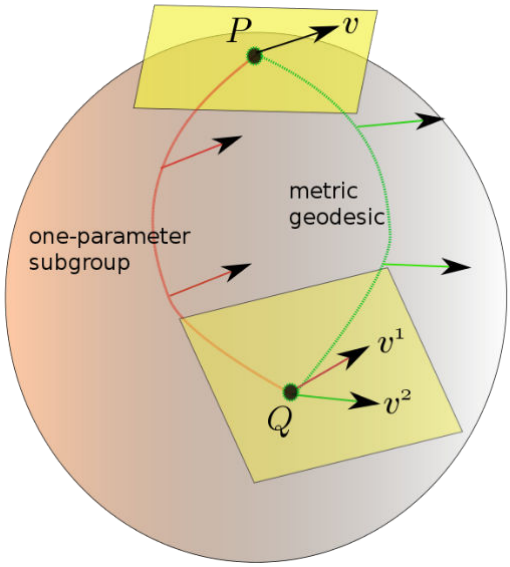


Figure 8.5: The parallel transport of a vector v closely depends on the chosen trajectory, and generally transporting along different curves leads to different parallel vectors. Image source: [Lorenzi 2012a].

8.5 Conclusion

In this section, we have introduced the field of *computational anatomy*. We have reviewed several choices that can be key steps to analyze a population of medical images: a deformation model, a template model, and a transport method.

In Chapter 9, we investigate the use of the LDDMM framework and initial momenta to build descriptors of hippocampus shape evolutions. In Chapter 10, we introduce several spatial regularizations in a logistic classification framework.

Chapter 9

Longitudinal hippocampus shape analysis via geodesic shootings

Contents

Résumé

Dans le contexte de la maladie d'Alzheimer, les méthodes de l'état de l'art permettent de séparer les patients sains des patients malades ou les patients sains des patients progressifs (patients ayant des troubles cognitifs légers puis développant la maladie) avec des taux de classification corrects. Cependant, leurs taux de classification restent très faibles lorsqu'il s'agit de séparer les patients ayant des troubles cognitifs légers et stables (c'est-à-dire ne développant pas la maladie) des patients progressifs. Au lieu d'utiliser des descripteurs provenant d'un seul point temporel, nous résolvons ce problème en utilisant des descripteurs représentant les évolutions des hippocampes des patients entre deux points temporels. Pour encoder ces transformations, nous utilisons le cadre des larges déformations par diffémorphismes, qui permet de calculer des évolutions géodésiques. Pour effectuer des statistiques sur ces descripteurs dans un système de coordonnées commun,

nous présentons une nouvelle extension de l'algorithme de la moyenne de Karcher, que nous définissons à une transformation rigide près, et un critère d'initialisation permettant un meilleur recalage des patients vers l'atlas. Finalement, puisque les descripteurs locaux ne permettent pas directement d'obtenir une meilleure performance que les descripteurs globaux (comme la variation de volume), nous proposons une nouvelle stratégie combinant l'utilisation de tirs géodésiques, d'une nouvelle version de l'algorithme de Karcher, du transport par densité et de l'intégration sur une sous-région de l'hippocampe. Cette nouvelle stratégie permet d'obtenir des performances de classification plus élevées que celles obtenues avec des descripteurs locaux.

Mots clés : Imagerie cérébrale, analyse de population, maladie d'Alzheimer, tir géodésique, moyenne de Karcher

Abstract

In the context of Alzheimer's disease (AD), state-of-the-art methods separating normal control (NC) from AD patients or NC from progressive MCI (mild cognitive impairment patients converting to AD) achieve decent classification rates. However, they all perform poorly at separating stable MCI (MCI patients not converting to AD) and progressive MCI. Instead of using features extracted from a single temporal point, we address this problem using descriptors of the hippocampus evolutions between two time points. To encode the transformation, we use the framework of *large deformations by diffeomorphisms* that provides geodesic evolutions. To perform statistics on those local features in a common coordinate system, we introduce an extension of the Karcher mean algorithm that defines the template modulo rigid registrations, and an initialization criterion that provides a final template leading to better matching with the patients. Finally, as local descriptors transported to this template do not directly perform as well as global descriptors (e.g. volume difference), we propose a novel strategy combining the use of initial momentum from geodesic shooting, extended Karcher algorithm, density transport and integration on a hippocampus subregion, which is able to outperform global descriptors.

Keywords: Brain imaging, population analysis, Alzheimer's disease, geodesic shooting, time-series image data, Karcher mean

9.1 Introduction

Large scale population studies aim to improve the understanding of the causes of diseases, define biomarkers for early diagnosis, and develop preventive treatments. For Alzheimer's disease, several classification strategies have been proposed to separate patients according to their diagnosis. These methods can be split into three categories: voxel-based [Klöppel 2008, Vemuri 2008, Lao 2004, Magnin 2009, Fan 2007, Fan 2008a, Fan 2008b], cortical-thickness-based [Klöppel 2008, Querbes 2009, Desikan 2009] and hippocampus-based [Chupin 2007, Chupin 2009, Gerardin 2009] methods. While decent classification rates can be achieved to separate AD from NC or NC from p-MCI (progressive mild cognitive impairment patients, i.e. converting to AD), all methods perform poorly at separating s-MCI (stable mild cognitive impairment patients, i.e. non converting to AD) and progressive mild cognitive impairment (p-MCI). A recent review comparing these methods can be found in [Cuingnet 2011b].

In this paper, we investigate the use of longitudinal evolution quantifiers either local or global to separate between stable MCI and progressive MCI. To extract information between two successive time-points, we use a one-to-one deformation mapping the first image onto the second one. Different registration algorithms are available to compute plausible deformations in this context. However, only one, the *large deformation diffeomorphic metric mapping (LDDMM)* [Beg 2005], provides a Riemannian setting that enables to represent the deformations using tangent vectors: initial velocity fields or equivalently initial momenta. This can be used in practice to retrieve local information and to perform statistics on it as presented in [Vaillant 2004, Wang 2007]. In this direction, it is worth mentioning paper [Singh 2010] which shows the correlation between principal modes of deformations and diagnosis. In order to compare this information among the population, we need to define a common coordinate system. This implies (1) the definition of a template and (2) a methodology for the transport of the tangent vector information.

In the literature, point (1) is addressed via different methods [Ma 2009, Fletcher 2004]. Combining geodesic shooting algorithm presented in [Vialard 2012], we chose to develop a Karcher method to average a set of shapes. A first approach has been presented in [Risser 2011a]. A natural requirement on the Karcher average is that it could be invariant with respect to rigid transformations of each subject of the population. However, this is not the case in [Risser 2011a]. One of the contribution of the present paper is to propose a methodology to define such invariant Karcher averages. We also use a finer strategy to update the deformations at each iteration of the algorithm. Point (2) benefited in each different settings from various contributions that go beyond the standard transport actions. The key-point in our application is that inter-subject variability is much higher than the longitudinal variation so that one expects the statistical results to be strongly influenced by the choice of the transport. To address this issue, parallel transport has been proposed in the LDDMM setting in [Younes 2007] and it has been applied to longitudinal data discrimination, very similar to our problem, in [Qiu 2008]. Note that parallel

transport preserves the norm of the velocity field and since this norm is not invariant with respect to rescaling, the population variability is still contained in the parallel transport of the tangent information. For other frameworks, such as Log-demons, Schild's ladder approach has been introduced in [Lorenzi 2011] to extend parallel transport to their Lie-group setting. In any case, we consider the question of transporting tangent information as still open and this motivates us to compare different transport strategies in the classification step.

Section 9.2 introduces the global pipeline used to perform statistics from local descriptors of hippocampus deformations. Section 9.3 presents the data used and the numerical results. Section 9.4 concludes the paper.

9.2 Methods

9.2.1 Global pipeline

Let us assume we have a population of patients and the *binary* segmentations of their hippocampus at two different time points, called *screening* and *follow-up*.

Remark. The fact that we are using binary images (and not gray-scale) is important for the methodological choices. Such binary images will be used in all this chapter and in Chapter 10.

Let us also assume that all patients have the same diagnosis at the first time point, and only a part of them have converted to another diagnosis at the second time point. Our goal is to compare patient evolutions, and classify them with regard to disease progression, i.e. stable diagnosis versus progressive diagnosis.

We use the pipeline summarized in Fig. 9.1. First, the evolution descriptors is computed locally for each patient (independently). To be able to compare these descriptors, one needs to transport them in a common space. To do so, a population template is computed, towards which all the local descriptors are transported. Finally, classification is performed to separate progressive from stable patients. Several local descriptors were tested: initial momentum and initial velocity field of geodesic shooting. The use of subregion and integration were also introduced. As for global shape deformation descriptors, volume variation and relative volume variation were tested.

9.2.2 Geodesic shooting

To register a source image $I\colon \Omega \subset \mathbb{R}^3 \to \mathbb{R}$ towards a target image $J\colon \Omega \subset \mathbb{R}^3 \to \mathbb{R}$, the LDDMM framework introduces the following minimization problem

$$\underset{v \in L^2([0,1],\mathcal{H}_K)}{\operatorname{argmin}} \ \frac{1}{2}\|I \circ \phi_{0,1}^{-1} - J\|_{L^2}^2 + \lambda \int_0^1 \|v_t\|_K^2 \, dt \,, \tag{9.1}$$

where $v\colon (t,\omega) \in [0,1] \times \Omega \subset \mathbb{R}^3 \to \Omega$ is a time dependent velocity field that belongs to a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ of smooth enough vector fields

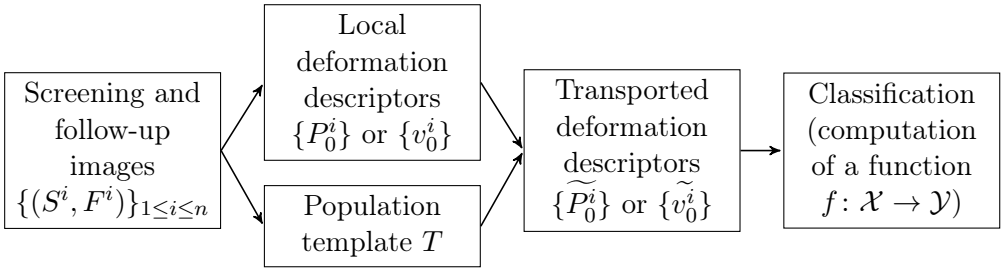


Figure 9.1: Four steps are needed to classify patient evolutions using local descriptors of shape deformations: (1) the local descriptors are computed for each patient independently, (2) a population template is computed, (3) all local shape deformation descriptors are transported towards this template, (4) classification is performed.

defined on Ω, and of associated kernel K and norm $\|\ \|_K$. For $(t,\omega) \in [0,1]\times\Omega$, we note $v_t(\omega) = v(t,\omega)$. The deformation $\phi\colon [0,1]^2\times\Omega \subset \mathbb{R}^3 \to \Omega$ is given by the flow of v_t

$$\forall (t,\omega)\in[0,1]\times\Omega, \quad \begin{cases} \frac{\partial \phi_{0,t}}{\partial t}(\omega) = v_t \circ \phi_{0,t}(\omega) \\ \phi_{t,t}(\omega) = \omega\,, \end{cases} \tag{9.2}$$

where ϕ_{t_1,t_2} is the deformation from $t=t_1$ to $t=t_2$. Such approach induces a right-invariant metric on the group of diffeomorphisms as well as a Riemannian metric on the orbit of I, i.e. the set of all deformed images by the registration algorithm [Miller 2006]. The first term in formula (9.1) is a similarity term controlling the matching quality whereas the second one is a smoothing term controlling the deformation regularity. Now noting $I_t \overset{\text{def.}}{=} I\circ\phi_{0,t}^{-1}$ and $J_t \overset{\text{def.}}{=} J\circ\phi_{t,1}$, the Euler-Lagrange equation asssociated with (9.1) reads

$$\forall (t,\omega)\in[0,1]\times\Omega, \quad v_t(\omega) = K\star\big(\operatorname{grad} I_t(\omega)\operatorname{Jac}_{\phi_{t,1}}(\omega)(I_t(\omega)-J_t(\omega))\big)\big) \tag{9.3}$$

where K the translation-invariant kernel of the reproducing kernel Hilbert space, $\star$ the convolution operator, grad the image gradient in space and Jac_ϕ the Jacobian of ϕ.

Definition 9.2.1 (Momentum)**.** Let us define the momentum $P\colon [0,1]\times\Omega\to\mathbb{R}$ by

$$\forall (t,\omega)\in[0,1]\times\Omega, \quad P(t,\omega) \overset{\text{def.}}{=} \operatorname{Jac}_{\phi_{t,1}}(\omega)(I_t(\omega)-J_t(\omega))\,, \tag{9.4}$$

and note $P_t(\omega) = P(t,\omega)$.

The Euler-Lagrange equation (9.3) can be rewritten as a set of geodesic shooting equations

$$\forall (t,\omega)\in[0,1]\times\Omega, \quad \begin{cases} \frac{\partial I_t}{\partial t}(\omega) + \langle \operatorname{grad} I(\omega), v_t(\omega)\rangle = 0\,, \\ \frac{\partial P_t}{\partial t}(\omega) + \operatorname{div}\left(P_t(\omega)v_t(\omega)\right) = 0\,, \\ v_t(\omega) + K\star\operatorname{grad} I_t(\omega)\ P_t(\omega) = 0\,, \end{cases} \tag{9.5}$$

where div is the divergence operator.

Remark (Geodesic shooting). Given a initial image I_0 and an initial momentum P_0, one can integrate the system (9.5). Such a resolution is called *geodesic shooting*. We say that *we shoot from* I_0 *using* P_0.

The minimization problem (9.1) can be reformulated using a shooting formulation on the initial momentum P_0

$$\underset{P_0}{\operatorname{argmin}} \frac{1}{2}\|I \circ \phi_{0,1}^{-1} - J\|_{L^2}^2 + \lambda \langle \operatorname{grad} I_0 P_0, K \star \operatorname{grad} I_0 P_0 \rangle_{L^2} \tag{9.6}$$

subject to the shooting system (9.5).

Theorem 9.2.1 ([Vialard 2012])**.** *The gradient of the functional in* (9.6) *is given by:*

$$\forall \omega \in \Omega, \quad \nabla_{P_0} \mathcal{L}(\omega) = -\hat{P}_t(\omega) + \lambda \langle \operatorname{grad} I_t(\omega), K \star (P_0(\omega) \operatorname{grad} I_0(\omega)) \rangle \tag{9.7}$$

where $\hat{P}_0$ *is given by the solution of the following PDE solved backward in time*

$$\forall (t, \omega) \in [0,1] \times \Omega, \quad \begin{cases} \frac{\partial \hat{I}_t}{\partial t}(\omega) + \operatorname{div}(v_t(\omega)\hat{I}_t(\omega) + \operatorname{div}(P_t(\omega)\hat{v}_t(\omega)) = 0\,, \\ \frac{\partial \hat{P}_t}{\partial t}(\omega) + \langle v_t(\omega), \operatorname{grad} \hat{P}_t(\omega) \rangle - \langle \operatorname{grad} I_t(\omega), \hat{v}_t(\omega) \rangle = 0\,, \\ \hat{v}_t(\omega) + K \star \left(\hat{I}_t(\omega) \operatorname{grad} I_t(\omega) - P_t(\omega) \operatorname{grad} \hat{P}_t(\omega) \right)\,, \end{cases} \tag{9.8}$$

subject to the initial conditions $\hat{I}_1 = J - I_1$ *and* $\hat{P}_0 = 0$ *and that* P_t, I_t *are the solution of the shooting system* (9.5) *for the initial conditions* I_0, P_0.

In order to solve the new optimization problem (9.6), we use the methodology described in [Vialard 2012]. Note that the choice of the kernel matters for retrieving plausible deformations and we refer to [Risser 2011b] for an extensive discussion on the parameter choices.

For each patient, a two-step process was performed to encode the deformations of the hippocampus shape evolution between the screening image S (scanned at the first time point $t = t_0$) to the follow-up image F (scanned at the second time point $t = t_0 + 12$ months), as described in Fig. 9.2. First F was rigidly registered back to S. We note $R \colon \Omega \subset \mathbb{R}^3 \to \Omega$ the rigid transformation obtained. Second, the geodesic shooting was performed with the screening image as source image ($I = S$) point towards the registered second time point as target image ($J = F \circ R^{-1}$). Inital momenta and initial velocity fields from different patients are local descriptors that were used to compare hippocampus evolutions.

9.2.3 Population template

Need for a template: As mentioned in section 9.2.1, local descriptors of hippocampus evolutions need to be transported in a common space prior to any statistical analysis. One way to obtain spatial correspondences between local descriptors

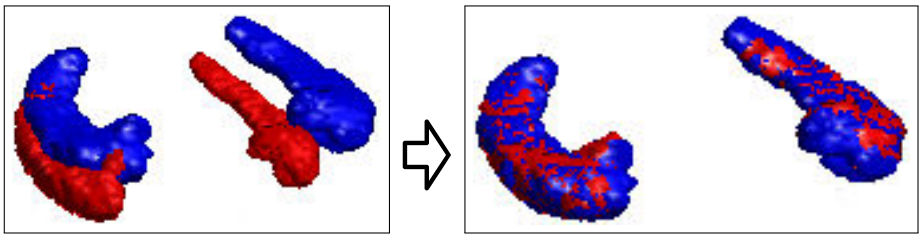
Step 1. Rigid registration of F towards S

Input:

- source image: follow-up image F of the patient, scanned at $t = t_0 + 12$ months,
- target image: screening image S of the patient, scanned $t = t_0$.

Output:

- rigid transformation R.



Step 2. Geodesic shooting from S towards $F \circ R^{-1}$

Input:

- source image: $I = S$, screening image,
- target image: $J = F \circ R^{-1}$, follow-up image rigidly registered back to the screening image.

Output:

- initial momentum P_0.

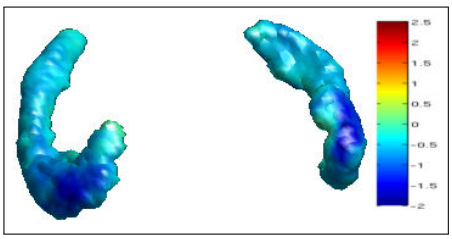


Figure 9.2: For each patient, the initial momentum encoding the hippocampus evolution is computed in a two-step process. First, the follow-up image F (i.e. second time point, $t = t_0 + 12$ months) is rigidly registered to the scanning image S (i.e. first time point, $t = t_0$). Second, the geodesic shooting is computed from the screening image S to the previously rigidly registered follow-up image $F \circ R^{-1}$.

of different patients consist in building a population template and then aligning these descriptors on the template.

Notions of Fréchet and Karcher means: In the Riemannian framework used for the geodesic shooting, a Fréchet mean [Fréchet 1948] can be used to define an average shape from a population [Pennec 1999, Fletcher 2004, Pennec 2006]. Given n images $\{S^i \colon \Omega \subset \mathbb{R}^3 \to \mathbb{R}\}_{1 \leq i \leq n}$ and d a Riemannian metric on the space of images, the Fréchet mean $T \colon \Omega \subset \mathbb{R}^3 \to \mathbb{R}$ is defined as a minimizer of the sum of the geodesic distances to all images

$$\mathcal{L}^{SK}(T) \overset{\text{def.}}{=} \sum_{i=1}^{n} d(T, S^i)^2. \tag{9.9}$$

As the solution is found via an optimization procedure, there is not necessarily existence (the global minimum is not necessarily reached), and the uniqueness is not guaranteed either [Pennec 1999]. Besides, in practice (9.9) is often solved via an optimization procedure looking for local minima (not global). In that case, one refers to the found solutions as *Karcher means* [Karcher 1977].

Remark. Before going any further, let us recall that in our study we want to build a population template from *binary* images. Moreover, all the patient images have the same topology.

Given our images, the solution of this optimization problem can be computed using a gradient descent procedure [Risser 2011a].

Modifications of the algorithm: We introduce two modifications from the algorithm in [Risser 2011a]: (1) the population template is computed up to rigid transformations and (2) the template is regenerated from a reference image at every iteration. The first modification involves the definition of the model, whereas the second modification involves the numerical resolution of the model.

The first modification is introduced because current implementations of LDDMM algorithms are not invariant by the action of the group of orthogonal linear transformations, so the resulting template reflects the orientation variability of the population. This means a minimizer of (9.9) could depend on the orientations of the input images S^i. In other words, if S^i is replaced by $S^i \circ R^{-1}$ where $i \in [\![1, n]\!]$ and $R \colon \Omega \to \Omega$ is a rigid transformation, the solution of the optimization could be different. In practice, a preprocessing step where all images are rigidly registered towards the initialization T_0 is usually performed. The idea of this step is to align roughly the images to avoid convergence issues of finer non-rigid registrations during the template computation, or at least to save computation time. Not only it is an arbitrary way to set the orientations of the population images S^i, but it also has the side effect of biasing all the result towards T_0. To circumvent this issue, we propose

the following optimization problem:

$$\mathcal{L}^{IK}(T) \overset{\text{def.}}{=} \sum_{i=1}^{n} d\left(T, \mathcal{R}^i\right)^2 , \tag{9.10}$$

where $\mathcal{R}^i$ is the orbit of S^i under the group action of the rigid transformations, i.e. $\mathcal{R}^i = \{S^i \circ R^{-1}, R\colon \Omega \to \Omega \text{ rigid transformation}\}$, and the d defined as $d\left(T, \mathcal{R}^i\right) \overset{\text{def.}}{=} \inf\{d(T, S^i \circ R^{-1}), R\colon \Omega \to \Omega \text{ rigid transformation}\}$. As the previous infima are minima, one can rewrites the functional

$$\mathcal{L}^{IK}(T, R_T^1, \ldots, R_T^n) = \sum_{i=1}^{n} d\left(T, S^i \circ \left(R_T^i\right)^{-1}\right)^2 , \tag{9.11}$$

where $\{R_T^i\colon \Omega \to \Omega\}_{1 \leq i \leq n}$ are the optimal rigid transformations registering the S^i towards the image T. Now let us study the resolution of the optimization problem (9.11). The proposed approach is an iterative procedure inspired by the work of [Risser 2011a]. The difference with their approach is that here we alternatively update T with R_T^i fixed and R_T^i with T fixed. For this reason, we drop the dependency on T and simply note R^i for the rigid transformations. Now, since the functional from (9.1) does not give a geodesic distance between two images - but between a source image and the deformed image, we approximate (9.11) by:

$$\mathcal{L}^{IK}(T, R^1, \ldots, R^n) \approx \sum_{i=1}^{n} d\left(T, J_1^k\right)^2 , \tag{9.12}$$

where J_1^k is the result of the shooting equations for the initial conditions $I = T$ and P_0^k where P_0^k is a minimizer of (9.6) with $J = S^i \circ \left(R^i\right)^{-1}$. In this case, each term of the sum is equal to $\langle \operatorname{grad} IP_0, K \star (\operatorname{grad} IP_0)\rangle_{L^2}$, and the gradient with regard to T is:

$$\forall \omega \in \Omega, \quad \nabla_T \mathcal{L}^{IK}(T, R^1, \ldots, R^n)(\omega) \approx -\frac{1}{n} \sum_{i=1}^{n} K \star \operatorname{grad} TP_0^i(\omega) , \tag{9.13}$$

where P_0^i is the initial momentum matching T on $S^k \circ (R^i)^{-1}$ via the shooting system (9.5). Here comes the second modification of the standard Karcher algorithm, which is as stated before purely motivated by numerical considerations. The standard formulation [Risser 2011a] smoothes the template estimate at each iteration (as T_{k+1} is computed by shooting from T_K). In our algorithm, at every iteration the new Karcher estimate T_{k+1} is computed from a reference image (typically the initialization T_0):

$$\forall \omega \in \Omega, \quad T_{k+1}(\omega) = T_0\left(u_{k+1}(\omega)\right) , \tag{9.14}$$

where $u_{k+1}\colon \Omega \to \Omega$ is a deformation field. It is initialized as $u_0 = Id$, and updated at every iteration by composition $u_{k+1} = u_{mean} \circ u_k$ where u_{mean} is the deformation field associated with the geodesic shooting from T_k using P_0^{mean}. Such update

procedures allows the template estimate to keep sharp boundaries. Altogether, each Karcher iteration is composed of four steps, as described in Fig. 9.3.

9.2.4 Tangent information and associated transport

The local descriptors computed for each patient as explained in section 9.2.2 need to be transported in a common coordinate space: the space of the Karcher average defined in section 9.2.3. We chose transport rules that only depend on the final deformation (it does not include parallel transport which depends on the chosen path). A two-step process was then used to transport local descriptors of hippocampus evolutions to the template space (Fig. 9.4). First, the screening hippocampus was registered towards the template rigidly [Ourselin 2001] then non-rigidly [Modat 2010]. The resulting deformation is denoted by ϕ. Second, this transformation was used to transport the local descriptors of hippocampus deformations towards the template. The transport itself depends on the nature of the quantity transported. For instance, we call *image transport* the standard transformation of an image $I\colon \Omega \to \mathbb{R}$ by the action of a diffeomorphism $\phi\colon \Omega \subset \mathbb{R}^3 \to \Omega$

$$\forall \omega \in \Omega, \quad \widetilde{I}(\omega) \overset{\text{def.}}{=} I \circ \phi^{-1}(\omega)\,. \tag{9.15}$$

From the mathematical point of view, the momentum is an adjoint variable to the image. As a consequence, it is transported by the adjoint action of the group which reduces to the standard transport for a density $n\colon \Omega \subset \mathbb{R}^3 \to \mathbb{R}$, defined by

$$\forall \omega \in \Omega, \quad \widetilde{n}(\omega) \overset{\text{def.}}{=} \det(\mathrm{Jac}_{\phi^{-1}}(\omega))\, n \circ \phi^{-1}(\omega)\,, \tag{9.16}$$

where det is the notation for the determinant. Note that this action preserves the global integration of the density by a simple change of variable. Last, we present the transport via the standard conjugation of a velocity field $v\colon \Omega \subset \mathbb{R}^3 \to \mathbb{R}^3$ defined by

$$\forall \omega \in \Omega, \quad \widetilde{v}(\omega) \overset{\text{def.}}{=} d\phi(\omega) \circ v \circ \phi^{-1}(\omega)\,, \tag{9.17}$$

where for $\omega = (\omega_1, \omega_2, \omega_3) \in \mathbb{R}^3$, if we note $\phi(\omega) = (\phi_1(\omega), \phi_2(\omega), \phi_3(\omega)))$, $d\phi(\omega)$ is a 3×3 matrix (i.e. an operator from $\mathbb{R}^3$ to $\mathbb{R}^3$) defined by $(d\phi(\omega))_{ij} = \frac{\partial \phi_i}{\partial \omega_j}$, i.e. $d\phi(\omega) = (\mathrm{grad}\,\phi_1(\omega), \mathrm{grad}\,\phi_2(\omega), \mathrm{grad}\,\phi_3(\omega))^T \in \mathbb{R}^{3\times 3}$.

All those transport methods were tested in the classification step. We did not include parallel transport in this study since no public implementation is available and its implementation is rather involved, especially in the case of images.

9.2.5 Classification

The final step in the proposed pipeline is the classification step, as described in the section 9.2.1. The local descriptors of shape evolution have been computed for each patient independently (Section 9.2.2) and transported in a common space: the space of the template (sections 9.2.3 and 9.2.4). Let us first give a brief reminder on the

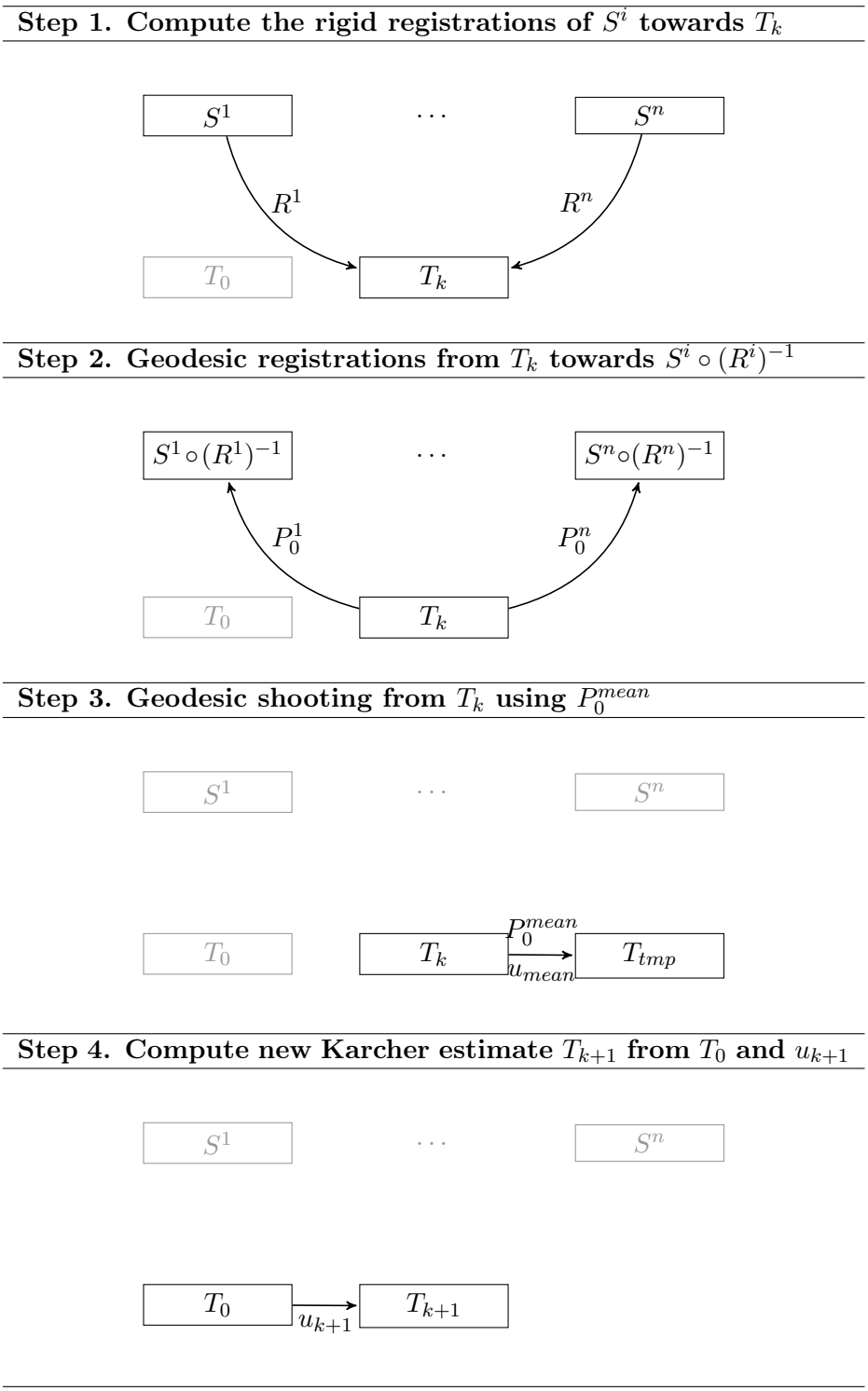


Figure 9.3: Each Karcher iteration is composed of four steps: (1) the images S^i are rigidly aligned towards the current Karcher mean estimate T_k, (2) geodesic shootings from the current Karcher estimate T_k towards all the registered images $S^i \circ (R^i)^{-1}$ are computed (3) geodesic shooting from T_k using $P_0^{mean} = \frac{1}{n}\sum_i P_0^i$ generates a deformation field u_{mean}, and (4) the composed deformation field $u_{k+1} = u_{mean} \circ u_k$ is used to compute the updated estimate from the reference image.

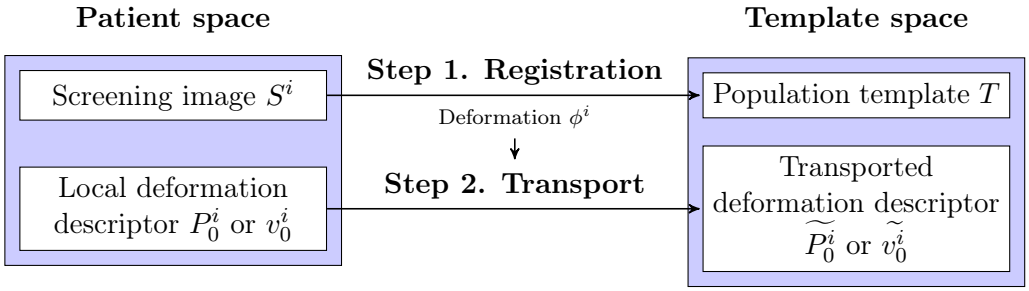


Figure 9.4: Local descriptors of hippocampus evolutions are transported to the template in a two-step process. First the deformation field from the patient space to the population template. Second, this deformation field is used to transport the local descriptors.

support vector machines (SVM) technique and then describe the various features evaluated.

Support Vector Machines: SVMs [Schölkopf 2001] are a supervised classification method. Given n labelled features $(\mathbf{x}_i, y_i) \in \mathcal{X} \times \{-1, 1\}$, it aims at building a function $f^* \colon \mathcal{X} \to \mathbb{R}$ such that $y = \text{sign}(f^*)$ is an optimal labeling function. The function f is a solution of the optimization problem

$$f^* \stackrel{\text{def.}}{=} \underset{f \in \mathcal{H}_K}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell_{hinge}(f(\mathbf{x}_i), y_i) + \gamma \|f\|_K^2 \tag{9.18}$$

where $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a Mercer kernel, $\mathcal{H}_K$ its associated Reproducing Kernel Hilbert Space of functions $\mathcal{X} \to \mathbb{R}$ and its corresponding norm $\| \ \|_K$, and ℓ_{hinge} is the hinge loss defined as $\ell_{hinge}(f(\mathbf{x}), y) = \max\{0, 1 - y \times f(\mathbf{x})\}$. The loss function ℓ_{hinge} controls the labeling performance, and the second term controls the smoothness of the solution.

Remark (Choice of the SVM kernel). In this study, we are using the Gaussian kernel defined as $K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)$

Local vs global features: In the case on AD, several global descriptors of shape evolution such as (relative) volume difference seem to be related to the disease progression [Braak 1998, Gosche 2002]. In order to both (1) compare the relevance of the local descriptors with regard to disease progression and (2) assess the extra spatial information, several features were computed from the local descriptors. The definitions of the various derived features are summarized in Table 9.1.

Table 9.1: Definitions of the various features derived from the local descriptors of hippocampus shape evolutions. From $\widetilde{P_0^i}\colon \Omega \to \mathbb{R}$, four derived features are defined: local, local restricted to a subregion $\Omega_r \subset \Omega$, local integrated on the whole domain, and local integrated on a subregion.

Feature type	Definition	Dimension
Local	$\mathbf{x}_i = \widetilde{P_0^i}$	$d \sim 10^5$ to 10^6
Local restricted to a subregion	$\mathbf{x}_i = \widetilde{P_0^i}\vert_{\Omega_r}$	$d \sim 10^4$ to 10^5
Local integrated on the whole domain	$\mathbf{x}_i = \int_{\Omega} \widetilde{P_0^i}(\omega) d\omega$	1
Local integrated on a subregion	$\mathbf{x}_i = \int_{\Omega_r} \widetilde{P_0^i}(\omega) d\omega$	1

9.3 Material and Results

9.3.1 Data

A dataset of 206 hippocampus binary segmentations from 103 patients enrolled in Alzheimer's disease neuroimaging initiative (ADNI)[1] [Mueller 2005] has been used to estimate the efficiency of local and global descriptors of hippocampus evolution with regard to disease progression. For each patient, 'screening' and 'month 12' follow-up were the two time points selected. All patients were mild cognitive impairment (MCI) at the screening point, 19 became AD by month 12, and the remaining 84 stayed MCI.

9.3.2 Experiments

First, all screening images were resampled to a common isotropic voxel size $1.0 \times 1.0 \times 1.0$ mm, similar to their original size. Rigid transformations aligning the month 12 hippocampus towards the screening ones were computed using [Ourselin 2001]. The geodesic shootings [Vialard 2012] were performed[2] using a sum of three kernels (sizes $1, 3$ and 6 mm, with respective weights $2, 1$ and 1), and 200 gradient descent iterations.

To compute the template, a subset of 20 images was used. This subset and the initialization was based on a shape volume criterion. Four Karcher iterations were performed, with respectively $200, 150, 150$ and 100 gradient descent iterations in the geodesic shootings. To compute the transformations ϕ^i from the screening hippocampi towards the template (Fig. 9.4), rigid [Ourselin 2001] then non-rigid [Modat 2010] registration algorithms were applied with their default parameters.

To classify from local descriptors, a mask computed by dilating the template was used. To compute classification on subregions, each hippocampus (left and right)

[1] http://www.loni.ucla.edu/ADNI
[2] http://sourceforge.net/projects/utilzreg/

from the template was dilated. The bounding box was cut equally in thirds along the longest axis, and intersections were used as masks (Fig. 9.7).

Using a leave-10%-out scheme, training and test sets were created. With training features equally distributed among classes, SVM classifiers were computed (the Matlab functions from the Bioinformatics Toolbox were used). All the patients were then classified. The Gaussian kernel was used and 20 kernel widths were tested. This procedure was repeated 50 times and classification accuracy averaged. From the numbers of true/false positives/negatives (TP, FP, TN, FN), four indicators were used to measure classification accuracy: specificity $Spec \overset{\text{def.}}{=} \frac{TN}{TN+FP}$, sensitivity $Sens \overset{\text{def.}}{=} \frac{TP}{TP+FN}$, negative predictive value $NPV \overset{\text{def.}}{=} \frac{TN}{TN+FN}$, and positive predictive value $PPV \overset{\text{def.}}{=} \frac{TP}{TP+FP}$.

9.3.3 Results

To validate the proposed pipeline, several quality checks were performed. To check the quality of the geodesic shooting computed for each patient i (second step in 9.2), the evolution of the Dice score DSC between S_t^i which is the deformed screening image at time t and the target image $F^i \circ (R^i)^{-1}$ was computed, and the average final DSC is 0.94 ± 0.01 (Fig. 9.5).

Using the modified Karcher mean algorithm and the criterion mentioned above, the average Dice score between the 103 registered patients and the template was 0.87 ± 0.02, whereas it was only 0.44 ± 0.11 when matching to a template computed using a criterion based on the distance to the L2 mean.

To check the quality of the registration ϕ^i computed to transport the local descriptor of the patient i (first step in 9.4), the Dice score was computed between the rigidly registered screening image and the template (i.e. $\text{DSC}(S \circ (R^i)^{-1}, T)$) and between the final registered screening image and the template (i.e. $\text{DSC}(S \circ (\phi^i)^{-1}, T)$), see Fig. 9.6.

Now regarding descriptors of hippocampus evolutions, the local descriptors did not perform as well as global descriptors, when used directly as input features (Fig. 9.8). However, when integrated on the whole domain, the performances were similar. When integrated on some subregion, they can outperform the global descriptors. Detailed results are displayed in Fig. 9.8, and Table 10.1 displays the four unbiased indicators when the sum of specificity and sensitivity is maximized.

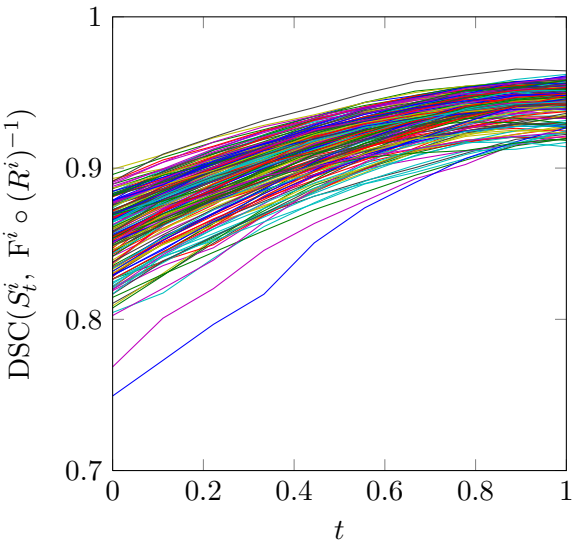


Figure 9.5: To check the quality of the geodesic shooting computed for each patient i (second step in 9.2), the evolution of the Dice score DSC between S_t^i which is the deformed screening image at time t and the target image $F^i \circ (R^i)^{-1}$ was computed. The average final DSC is 0.94 ± 0.01 .

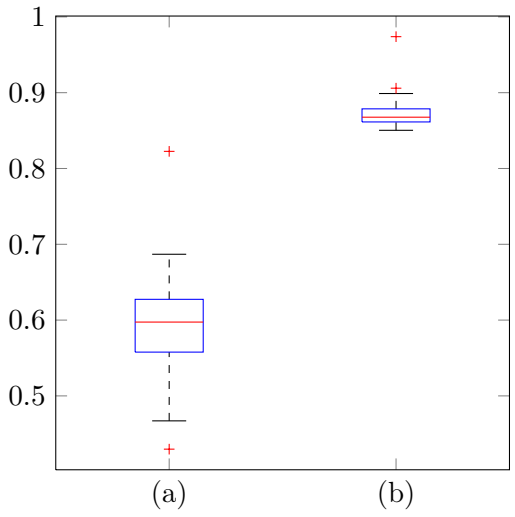


Figure 9.6: To check the quality of the registration ϕ^i computed to transport the local descriptor of the patient i (first step in 9.4), the Dice score was computed between the rigidly registered screening image and the template (i.e. DSC$(S \circ (R^i)^{-1}, T)$, see (a)) and between the final registered screening image and the template (i.e. DSC$(S \circ (\phi^i)^{-1}, T)$, see (b)).

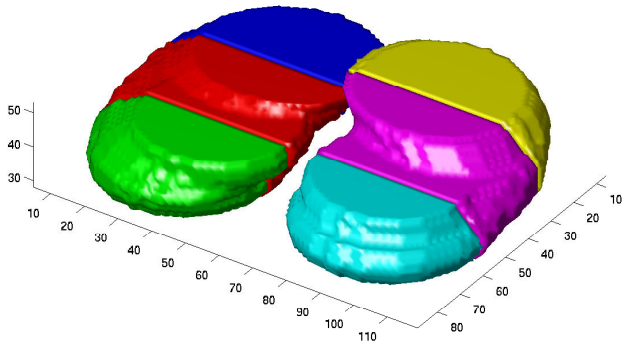
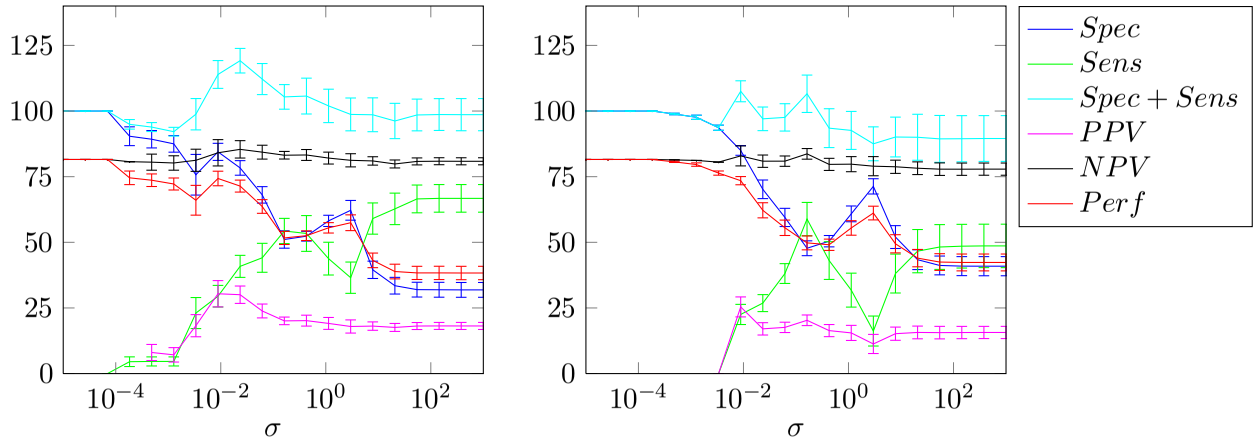
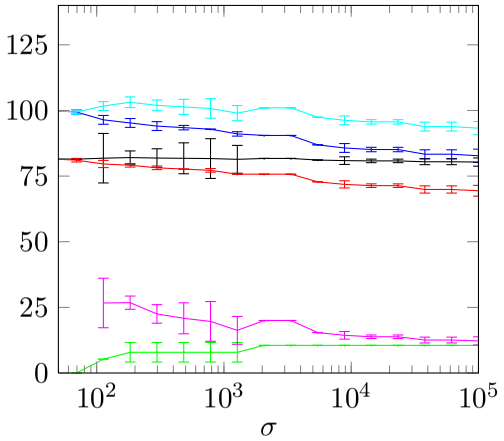


Figure 9.7: Subregions $\{\Omega_i\}_{1\leq i\leq 6} \subset \Omega$ of the hippocampus used as proof-of-concept in the classification step. Each hippocampus was dilated and then cut in thirds along the longest axis.

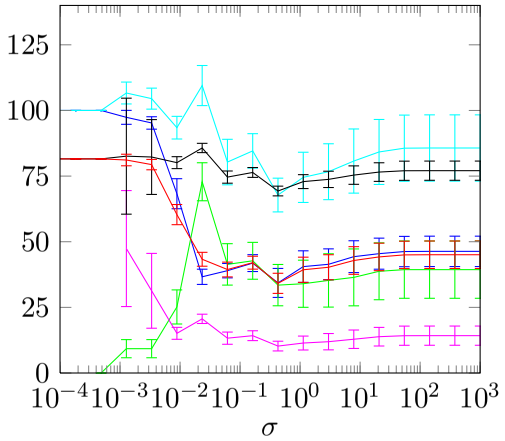


(a) Volume difference

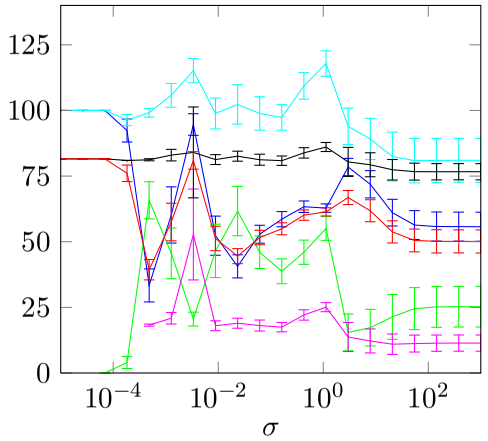
(b) Relative volume difference



(c) Initial momentum with image transport

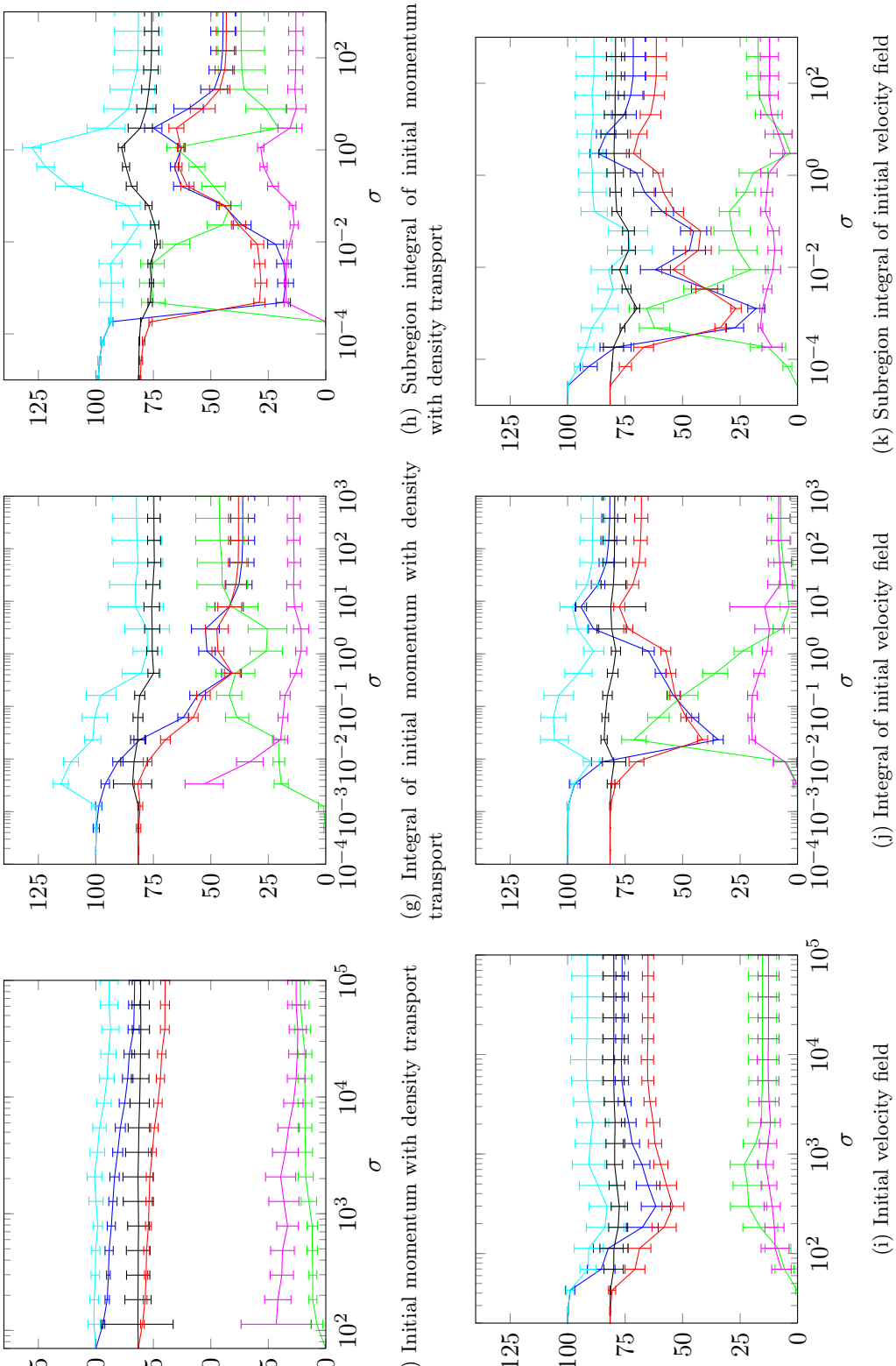


(d) Integral of initial momentum with image transport



(e) Subregion integral of initial momentum with image transport

Figure 9.8: Figure to be continued on page 192



(f) Initial momentum with density transport

(g) Integral of initial momentum with density transport

(h) Subregion integral of initial momentum with density transport

(i) Initial velocity field

(j) Integral of initial velocity field

(k) Subregion integral of initial velocity field

Figure 9.8: Classification performance (depending on the SVM Gaussian kernel width σ) for global descriptors (9.8a, 9.8b), local (9.8c, 9.8f, 9.8i), local integrated on the whole image (9.8d, 9.8g, 9.8j) and local integrated on a subregion (9.8e, 9.8h, 9.8k). Higher for $Spec + Sens$ (in cyan blue) is better.

9.4 Conclusion

We have studied the use of global, semi-local and local descriptors of hippocampus evolutions to predict AD conversion for MCI patients, using a dataset of binary segmentations provided by ADNI. This study focuses on shape evolutions between two time points, whereas (to the best of our knowledge), studies in this application field usually extract features from a single time point and perform diagnosis classification.

The proposed extension of the Karcher mean algorithm with a subpopulation and initialization criterion based on shape volume improved the matching quality to the template (average Dice of 0.87 ± 0.02 instead of 0.44 ± 0.11) without the need of modifying the default registration parameters.

In our experiments, the local descriptors did not perform as well as global descriptors such as volume difference when they were directly used as input features of the SVM classifiers. However, when integrated over the whole domain, classification performances were similar. When integrated on a subregion, they could even outperform the global descriptors. The method we propose combines (1) the use of initial momentum of geodesic shooting, (2) an extended version of the Karcher mean algorithm, (3) the use of density transport and (4) the integration on a subregion. On our dataset, this method was the only one able to outperform the global descriptors. It should be noted that in our study the definition of the subregion was sub-optimal and used as a proof-of-concept. The most promising perspectives are (1) developing a strategy to define subregions maximizing the classification results and (2) adding more time-points to the study using the geodesic regression method introduced in [Niethammer 2011] or cubic spline interpolation in [Trouvé 2012].

Table 9.2: Performance indicators for various descriptors of the hippocampus evolutions. These indicators are computed using a SVM classifier, with a Gaussian kernel. Kernel width is such that the sum of specificity and sensitivity is maximized. The proposed method is the only one with $Spec + Sens$ outperforming the same sum for the volume difference global descriptor.

Global / Local	**Deformation descriptor**	**Spec+ Sens**	**Spec**	**Sens**	**NPV**	**PPV**
Global	Volume difference	1.19	0.78	0.41	0.85	0.30
	Relative volume difference	1.08	0.85	0.23	0.83	0.25
Local integrated on the whole domain	Initial momentum, image transport	1.10	0.37	0.73	0.86	0.21
	Initial momentum, density transport	1.15	0.96	0.19	0.84	0.53
	Initial velocity field	1.07	0.46	0.61	0.84	0.20
Local integrated on a subregion	Initial momentum, image transport	1.18	0.63	0.55	0.86	0.25
	Initial momentum, density transport	**1.28**	**0.63**	**0.65**	**0.89**	**0.28**
	Initial velocity field	0.92	0.79	0.13	0.80	0.11
Local	Initial momentum, image transport	1.01	0.96	0.05	0.82	0.27
	Initial momentum, density transport	1.01	0.95	0.06	0.82	0.21
	Initial velocity field	0.92	0.77	0.15	0.80	0.13
Local restricted to a subregion	Initial momentum, image transport	1.10	0.68	0.42	0.84	0.23
	Initial momentum, density transport	1.17	0.68	0.49	0.85	0.26
	Initial velocity field	0.98	0.38	0.60	0.81	0.18

Chapter 10

Spatial regularizations for the classification of AD progression and detection of related hippocampus deformations

Contents

Résumé

Dans le contexte de la maladie d'Alzheimer, (1) la prédiction de l'apparition de la maladie chez les patients atteints de troubles cognitifs légers et (2) la caractérisation des changements locaux de l'hippocampe liés à la progression de la maladie sont des problèmes difficiles. Dans ce chapitre, nous étudions l'emploi d'un modèle de classification logistique pour la résolution simultanée de ces deux problèmes. Étant donné les observations de l'hippocampe d'un patient à deux points temporels, nous quantifions les déformations à l 'aide du modèle des larges déformations par difféomorphismes. Puisque les transformations sont a-priori structurées en espace, nous introduisons plusieurs *régularisations spatiales*. Nous comparons des régularisations générant des cartes de coefficients

lisses (Sobolev), constantes par morceaux (variation totale), ou parcimonieuses (fused LASSO) aux régularisations classiques (LASSO, ridge et ElasticNet). Les performances de classification sont évaluées sur une base de données composée de 103 patients provenant d'ADNI.

Mots clés : Maladie d'Alzheimer, imagerie cérébrale, évolution de la maladie, classification logistique, régularisation spatiale, carte de coefficients

Abstract

In the context of Alzheimer's disease, both (1) the identification of mild-cognitive impairment patients likely to convert and (2) the characterization of local hippocampal changes specific to disease progression are challenging issues. In this chapter, we investigate the use of a logistic classification model to address both problems simultaneously. Given the observations of the hippocampus of a patient at two time points, we quantify its deformation using the framework of large deformations by diffeomorphisms. Since the deformations are expected to be spatially structured, we introduce several *spatial regularizations*. We compare regularizations enforcing coefficient maps that are smooth (Sobolev), piecewise constant (total variation) or sparse (fused LASSO) to standards regularizations (LASSO, ridge and ElasticNet). Their performances are evaluated on a dataset of 103 patients from ADNI.

Keywords: Alzheimer's disease, brain imaging, disease progression, logistic classification, spatial regularization, coefficient map

10.1 Introduction

Large scale population studies aim to improve the understanding of the causes of diseases, define biomarkers for early diagnosis, and develop preventive treatments. An important challenge for medical imaging is to analyze the variability in magnetic resonance (MR) acquisitions of normal control (NC), mild cognitive impairment (MCI), and Alzheimer's disease (AD) patients.

As seen in Chapter 9, the *large deformation diffeomorphic metric mapping (LDDMM)* framework [Beg 2005] has proven useful to describe shape variations within a population. In the case of longitudinal analysis, it is not anymore the shapes that are compared but their evolutions in time. In [Qiu 2008] the authors estimate the typical deformation of several clinical groups from the deformations between baseline and follow-up hippocampus surfaces.

Quality of shape descriptors regards to the disease is often evaluated though statistical significance tests or classification performance. In order to deal with the latter we use the same pipeline as in Chapter 9, and evaluate descriptors on a logistic classification task. In spite of its simplicity, it has the advantage of providing a map of coefficients weighting the relevance of each voxel. Such map could help to localize the AD-related hippocampus deformations. However, the dimensionality of the problem being much higher than the number of observations ($p \sim 10^6 \gg n \sim 10^2$), the problem requires proper regularization. Now standard regularization methods such as *ridge* [Hoerl 1970], *LASSO* [Tibshirani 1994] and *elastic net* [Zou 2005] do not take into account any spatial structure of the coefficients. In contrast, total variation was used to regularize a logistic regression on functional MR data [Michel 2011]. Total variation promotes coefficient maps with spatially homogeneous clusters. Similar ideas can be found in [Cuingnet 2012] where the authors defined the notion of spatial proximity to regularize a linear support vector machines (SVM) classifier. In [Durrleman 2013], the authors introduce sparse parametrization of the diffeomorphisms in the LDDMM framework. Our goal is different: we want spatial properties (smoothness, sparsity, etc.) to be found *across* the population (i.e. on the common template) and we want this coherence to be driven by the disease progression. In this chapter, we investigate the use of total variation, Sobolev and fused LASSO regularizations. Compared to total variation, Sobolev enforces smoothness of the coefficient map, whereas fused LASSO adds a sparsity constraint.

The deformation model used to assess longitudinal evolutions in the population was presented in Chapter 9. The model of logistic classification with spatial regularization is described in Section 10.2.1. Data used and numerical results are presented in Section 10.3.3. We illustrate that initial momenta capture information related to AD progression, and that spatial regularizations significantly increase classification performance.

10.2 Methods

10.2.1 Logistic Classification with Spatial Regularization

Let us define a predictive model that reads

$$\mathbf{y} \overset{\text{def.}}{=} F(\mathbf{X}\mathbf{w} + b) \ , \tag{10.1}$$

where $\mathbf{y} \in \{\pm 1\}^n$ is the behavioral variable (i.e. the patient disease progression), $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the design matrix containing n observations of dimension p (i.e. the initial momenta representing the deformations of the patient hippocampus), F is the prediction function and $(\mathbf{w}, b) \in \mathbb{R}^p \times \mathbb{R}$ are the parameters to estimate. The binary logistic classification model defines the probability of observing y_i given the data $\mathbf{x}_i$ as

$$p\left(y_i \mid \mathbf{x}_i, \mathbf{w}, b\right) \overset{\text{def.}}{=} \frac{1}{1 + \exp\left(-y_i \ \left(\mathbf{x}_i^T \mathbf{w} + b\right)\right)} \ . \tag{10.2}$$

Given parameters $(\hat{\mathbf{w}}, \hat{b})$ and a new data point $\mathbf{x}$ the prediction is the maximum likelihood, i.e. $\text{class}(\mathbf{x}) = \text{argmax}_{y \in \{\pm 1\}} p\big(y \mid \mathbf{x}, \hat{\mathbf{w}}, \hat{b}\big) = \text{sign}\left(\mathbf{x}^T \hat{\mathbf{w}} + \hat{b}\right)$. Accordingly the parameters are estimated as minimizers of the opposite log likelihood of the observations, considered as independent

$$\mathcal{L}(\mathbf{w}, b) \overset{\text{def.}}{=} \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i (\mathbf{x}_i^T \mathbf{w} + b)\right)\right) \ . \tag{10.3}$$

Since the number of observations is much smaller than the dimension of the problem ($n \ll p$) minimizing directly the loss (10.3) leads to overfitting, and proper regularization is required. This is commonly performed by introducing a regularization function J and the final problem becomes

$$\text{Find } (\hat{\mathbf{w}}, \hat{b}) \text{ in } \underset{\mathbf{w}, b}{\text{argmin}} \ \mathcal{L}(\mathbf{w}, b) + \lambda J(\mathbf{w}) \ , \tag{10.4}$$

where λ is a coefficient tuning the balance between loss and regularization.

The standard *elastic net* regularization [Zou 2005] uses a combined ℓ_1 and squared ℓ_2 penalization $\lambda EN(\mathbf{w}) \overset{\text{def.}}{=} \lambda_1 ||\mathbf{w}||_1 + \lambda_2 ||\mathbf{w}||_2^2 = \sum_{j=1}^{p} \lambda_1 |w_j| + \lambda_2 w_j^2$, with the limit cases $\lambda_2 = 0$ referred to as *LASSO* [Tibshirani 1994] and $\lambda_1 = 0$ referred to as *ridge* [Hoerl 1970]. However as mentioned in [Michel 2011], one drawback of such methods is that they do not take into account any geometrical structure of $\mathbf{w}$. Since coefficients are expected to be locally correlated in space, we investigate the following regularizations

$$\text{Sobolev semi-norm} \qquad \text{SB}(\mathbf{w}) \overset{\text{def.}}{=} \sum_{\omega \in \Omega} \| \operatorname{grad} \mathbf{w}(\omega) \|_2^2 \ , \tag{10.5}$$

$$\text{Total Variation semi-norm} \qquad \text{TV}(\mathbf{w}) \overset{\text{def.}}{=} \sum_{\omega \in \Omega} \| \operatorname{grad} \mathbf{w}(\omega) \|_2 \ , \tag{10.6}$$

$$\text{Fused LASSO} \qquad \text{FL}(\mathbf{w}) \overset{\text{def.}}{=} \text{TV}(\mathbf{w}) + \|\mathbf{w}\|_1 \ . \tag{10.7}$$

The above sums go over all voxels ω in the domain $\Omega \subset \mathbb{R}^3$, and grad is a linear operator implementing the image gradient by finite differences. By indexing each voxel ω by integer coordinates on a 3D lattice, we define grad by

$$\operatorname{grad} \mathbf{w}(\omega_{ijk}) \stackrel{\text{def.}}{=} \begin{pmatrix} \Delta\mathbf{w}\left(\omega_{ijk}, \omega_{(i+1)jk}\right) \\ \Delta\mathbf{w}\left(\omega_{ijk}, \omega_{i(j+1)k}\right) \\ \Delta\mathbf{w}\left(\omega_{ijk}, \omega_{ij(k+1)}\right) \end{pmatrix} , \tag{10.8}$$

where $\Delta\mathbf{w}(\omega_1, \omega_2) \stackrel{\text{def.}}{=} \begin{cases} \mathbf{w}(\omega_2) - \mathbf{w}(\omega_1) & \text{if } (\omega_1, \omega_2) \in \Omega^2 \\ 0 & \text{otherwise} \end{cases}$. This definition allows to restrain Ω to any region of interest and boundaries of the domain are not penalized. Rationals and differences for those regularizations are discussed in Section 10.3.3.

10.2.2 Solving the Model

Let us first study differentiability and convexity of the objective function in problem (10.4). For convenience, we define $\tilde{\mathbf{w}} \stackrel{\text{def.}}{=} (\mathbf{w}^T, b)^T$ and for all i, $\tilde{\mathbf{x}}_i \stackrel{\text{def.}}{=} (\mathbf{x}_i^T, 1)^T$, with associated data matrix $\tilde{\mathbf{X}} \stackrel{\text{def.}}{=} (\tilde{\mathbf{x}}_{ij})_{\substack{1 \le i \le n \\ 1 \le j \le p+1}} \in \mathbb{R}^{n \times (p+1)}$. Then (10.3) becomes

$$\mathcal{L}(\tilde{\mathbf{w}}) = \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \exp(-y_i \tilde{\mathbf{x}}_i^T \tilde{\mathbf{w}})\right) . \tag{10.9}$$

It is twice differentiable and its first and second order partial derivatives read for all j, k in $[1..p+1]$

$$\frac{\partial \mathcal{L}}{\partial \tilde{\mathbf{w}}_j}(\tilde{\mathbf{w}}) = \frac{1}{n} \sum_{i=1}^{n} \frac{-y_i \tilde{\mathbf{x}}_{ij} \exp(-y_i \tilde{\mathbf{x}}_i^T \tilde{\mathbf{w}})}{1 + \exp(-y_i \tilde{\mathbf{x}}_i^T \tilde{\mathbf{w}})} = -\frac{1}{n} \sum_{i=1}^{n} \frac{y_i \tilde{\mathbf{x}}_{ij}}{1 + \exp(y_i \tilde{\mathbf{x}}_i^T \tilde{\mathbf{w}})} ,$$
$$\frac{\partial^2 \mathcal{L}}{\partial \tilde{\mathbf{w}}_j \partial \tilde{\mathbf{w}}_k}(\tilde{\mathbf{w}}) = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i^2 \tilde{\mathbf{x}}_{ij} \tilde{\mathbf{x}}_{ik} \exp(y_i \tilde{\mathbf{x}}_i^T \tilde{\mathbf{w}})}{\left(1 + \exp(y_i \tilde{\mathbf{x}}_i^T \tilde{\mathbf{w}})\right)^2} .$$

Denoting $e_i \stackrel{\text{def.}}{=} \exp(y_i \tilde{\mathbf{x}}_i^T \tilde{\mathbf{w}})$ we deduce the expression of the gradient and Hessian

$$\nabla \mathcal{L}(\tilde{\mathbf{w}}) = -\frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{1 + e_i} \tilde{\mathbf{x}}_i \; ; \quad \nabla^2 \mathcal{L}(\tilde{\mathbf{w}}) = \frac{1}{n} \sum_{i=1}^{n} \frac{e_i}{(1 + e_i)^2} \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T .$$

For all i, $\tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T$ is a nonnegative matrix and $\frac{e_i}{(1+e_i)^2} > 0$, so that $\nabla^2 \mathcal{L}(\tilde{\mathbf{w}})$ is nonnegative. This establishes the convexity of the loss function.

When the regularization J is also convex and twice differentiable the reference optimization algorithms include quasi-Newton methods; in particular for large-scale problems the limited memory Broyden-Fletcher-Goldfarb-Shanno (LM-BFGS) is very popular. For instance the Sobolev regularization defined in (10.5) is quadratic. It is easy to derive $\nabla \mathrm{SB}(\mathbf{w}) = -2 \operatorname{div}(\operatorname{grad}(\mathbf{w}))$, where div is the usual notation for the opposite of the adjoint of the linear operator grad.

However for non-differentiable regularizations such as total variation and fused LASSO optimization raises theoretical difficulties. Proximal methods such as monotonous fast iterative shrinkage thresholding algorithm (M-FISTA) [Beck 2009] and generalized forward-backward (GFB) [Raguet 2011] have been considered. Unfortunately their low convergence rates are prohibitive for extensive investigation of the classification scheme (parameter λ, domain Ω, training design matrix $\mathbf{X}$). Therefore we used the hybrid algorithm for non-smooth optimization (HANSO) [Lewis 2012] which is a LM-BFGS algorithm with weak Wolfe conditions line search. It is reported to be efficient for minimizing functions that are almost everywhere differentiable; in particular termination results are proven for Lipschitz-continuous or semi-algebraic functions [Lewis 2012]. This comprises both the total variation semi-norm and the ℓ_1-norm, with almost everywhere

$$\begin{aligned}\nabla \operatorname{TV}(\mathbf{w}) &= -\operatorname{div}\left(\left(\|\operatorname{grad}\mathbf{w}(\omega)\|_2^{-1}\operatorname{grad}\mathbf{w}(\omega)\right)_{\omega\in\Omega}\right)\ ,\\ \nabla\|\mathbf{w}\|_1 &= \left(\operatorname{sign}\left(\mathbf{w}(\omega)\right)\right)_{\omega\in\Omega}\ .\end{aligned}$$

10.2.3 Weighted Loss Function

In supervised learning, classifiers trained with observations not equally distributed between classes can be biased in favor of the majority class. Several strategies can be used to alleviate this. One strategy is to restrict the training set to be equally distributed among classes. To use the full training set, an alternative strategy is to introduce weights $(q_i)_{i\in[\![1,n]\!]}$ in the loss function as follows

$$\mathcal{L}_q(\tilde{\mathbf{w}}) \overset{\text{def.}}{=} \frac{1}{n}\sum_{i=1}^{n} q_i \log\left(1+\exp\left(-y_i\tilde{\mathbf{x}}^T\tilde{\mathbf{w}}\right)\right) \tag{10.10}$$

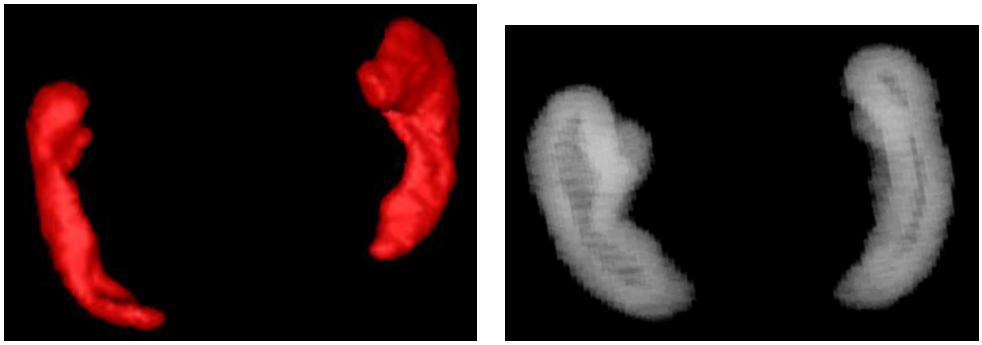
where $q_i \overset{\text{def.}}{=} n/\left(n_c \times \operatorname{card}\{j\in[1..n]\,|\,y_j=y_i\}\right)$, n_c being the number of classes (2 in our case). When the observations are equally distributed among classes $q_i=1$ for all i and one retrieves (10.9), whereas $q_i<1$ (respectively $q_i>1$) when the class of observation i is overrepresented (respectively underrepresented) in the training set.

10.3 Material and Results

10.3.1 Data

A dataset of 206 hippocampus binary segmentations from 103 patients enrolled in ADNI[1] [Mueller 2005] has been used. For each patient, 'screening' and 'month 12' were the two time points selected. All patients were MCI at the screening point, 19 converted to AD by month 12, and the remaining 84 stayed MCI.

[1] http://www.loni.ucla.edu/ADNI



(a) Template T (b) Region of interest Ω

Figure 10.1: The region of interest Ω (visualized with transparency) is computed by difference of the dilated template minus the eroded template.

10.3.2 Experiments

Computation of initial momenta Preprocessing, computation of the initial momenta[2], of the template of the population, and transport by density were performed following the methodology from [Fiot 2012] (or Chapter 9).

Computation of the region of interest Ω The region of interest Ω was restricted around the surface of the template (see Fig. 10.1), where lie the high values of the initial momenta. Moreover, this allows greater differences of coefficient values from one side to the other when using Sobolev regularization.

Logistic classification optimization Since stable and progressive classes in the dataset are unbalanced, the weighted version of the loss function defined in Section 10.2.3 was used. Solution of the optimization problems was computed via HANSO[3] with a maximum of 20 iterations.

Performance evaluation First, the effect of spatial regularizations was compared. The spatial regularizations introduced in Section 10.2.1 aim at enforcing local correlations between the coefficients in $\mathbf{w}$. Using the whole dataset, the effect of the various regularizations were compared. Second, the model was evaluated in terms of classification of AD progression. All patients were classified using a leave-10%-out scheme. This procedure was repeated for $N = 50$ random draws and classification results were averaged. From the numbers of true/false positives/negatives (TP, FP, TN, FN), four indicators were used to measure classification accuracy: specificity $Spec \overset{\text{def.}}{=} \frac{TN}{TN+FP}$, sensitivity $Sens \overset{\text{def.}}{=} \frac{TP}{TP+FN}$, negative predictive value $NPV \overset{\text{def.}}{=} \frac{TN}{TN+FN}$, and positive predictive value $PPV \overset{\text{def.}}{=} \frac{TP}{TP+FP}$.

[2] http://sourceforge.net/projects/utilzreg/
[3] http://www.cs.nyu.edu/overton/software/hanso

10.3.3 Results

10.3.3.1 Effect of spatial regularizations

When using standard regularizations, increasing the regularization does not lead to any spatial coherence (Fig. 10.2a, 10.2b and 10.2c). In contrast, the higher the spatial regularization, the more structured are the coefficients. Note that delimited areas are coherent across different spatial regularizations. Sobolev regularization leads to smooth coefficient maps (Fig. 10.2d) whereas total variation tends to a piecewise constant maps (Fig. 10.2e). Finally, fused LASSO adds sparsity by zeroing out the lowest coefficients (Fig. 10.2f).

Remark (Effective degrees of freedom (DOF)). When solving equation (10.4), it is interesting to evaluate the effective DOF of the system [Efron 1986]. In the literature, the degrees of freedom were studied in the case of univariate multiple regression analysis [Kruggel 2002], LASSO [Zou 2007, Dossal 2011], group-LASSO [Vaiter 2013], etc. Evaluation of the DOF is a potential research directions to evaluate the statistical significance of the methods presented in this chapter.

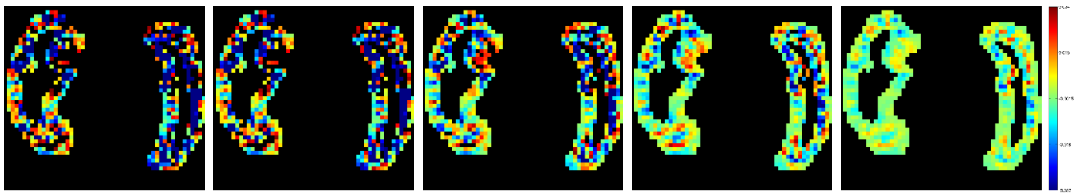
10.3.3.2 Classification of Alzheimer's disease progression.

Table 10.1 displays the classification performance indicators for the logistic classification model with various regularizations. Without any regularization, the resulting classifier always predicts the same class. All regularizations improve significantly the classification performance, the top 3 being the three spatial regularizations. On this dataset, total variation is the one providing the best results, similar to the performance from [Fiot 2012] (and Chapter 9). Nonetheless, note that even though fused LASSO (resp. ElasticNet) is the sum of two different regularizations we scaled them with the same coefficient λ. Optimizing over two different regularization coefficients should lead to better results than total variation (resp. ridge) which is only a limit case.

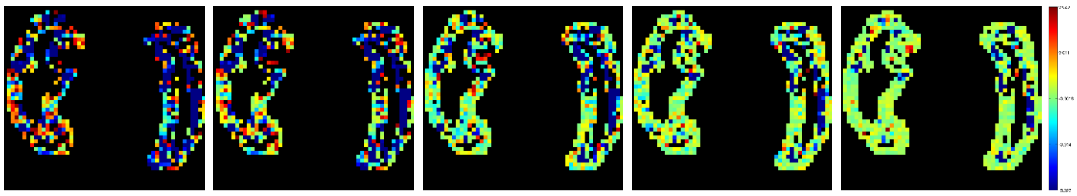
10.4 Conclusion

We investigated the use of logistic classification with spatial regularization in the context of Alzheimer's disease. Results indicate that initial momenta of hippocampus deformations are able to capture information relevant to the progression of the disease. Another contribution of this paper is the joint use of a simple linear classifier with complex spatial regularizations. Achieving results as good as in [Fiot 2012] which uses non-linear SVM classifier, our method also provides coefficient maps with direct anatomical interpretation. Moreover, we compared Sobolev, total variation and fused LASSO regularizations. While they all successfully enforce different priors (respectively smooth, piecewise constant and sparse), their resulting coefficient maps are coherent one to the other.

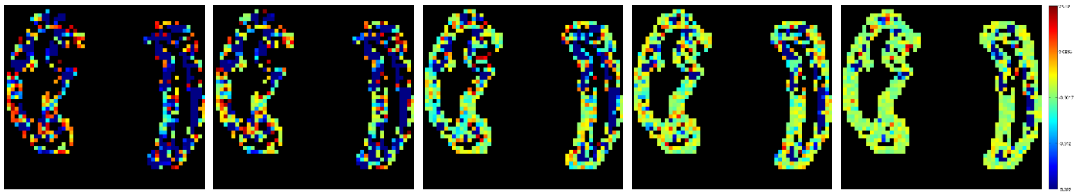
Those promising results pave the way to the design of better regularizations such as group sparsity. Other perspectives include the adaptation of the whole



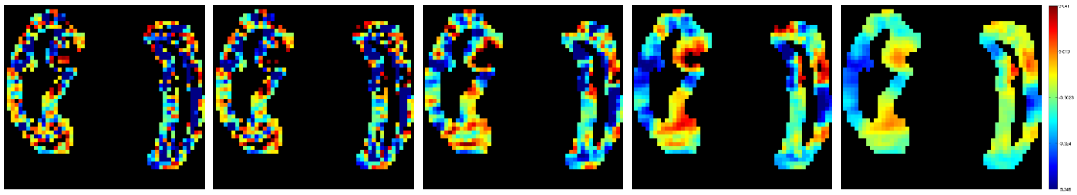
(a) Ridge



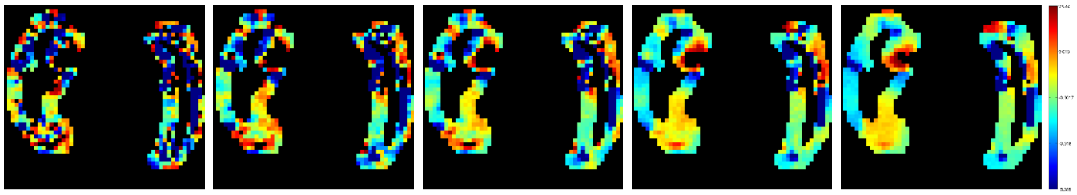
(b) LASSO



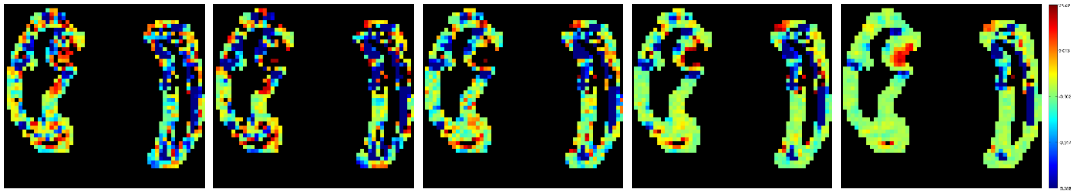
(c) ElasticNet



(d) Sobolev



(e) Total variation



(f) Fused LASSO

Figure 10.2: Effects of various regularizations on the coefficients of $\mathbf{w}$. Each small image represent the coefficients of one slice of $\mathbf{w}$, solution of the optimization problem with spatial regularization. On each row, the regularization is increasing from left to right. Fig. 10.2a and 10.2b show standard regularizations whereas Fig. 10.2d, 10.2e and 10.2f show spatial regularizations. Spatial regularizations provide more structured coefficients.

framework to surface representation of shapes, since initial momenta lie on surfaces in this application.

Table 10.1: Prediction accuracy of MCI patients' progression. Results are averaged over 50 random draws of training and testing sets. The sum of specificity and sensitivity is given as mean ± standard deviation.

Regularization		λ **range**	$\hat{\lambda}$	**Spec+ Sens**	**Spec**	**Sens**	**NPV**	**PPV**
None		0	0	1.00 ± 0.00	0.00	1.00	NaN	0.18
Standard	Ridge	$[10^{-9}, 10^{0}]$	0.001	1.10 ± 0.06	0.92	0.17	0.83	0.35
	LASSO	$[10^{-9}, 10^{0}]$	0.01	1.06 ± 0.03	0.23	0.83	0.85	0.20
	ElasticNet	$[10^{-9}, 10^{0}]$	0.01	1.06 ± 0.03	0.23	0.83	0.86	0.20
Spatial	Sobolev	$[10^{-9}, 10^{4}]$	$2.4 * 10^{3}$	1.15 ± 0.05	0.55	0.60	0.86	0.23
	Total Variation	$[10^{-9}, 10^{0}]$	0.01	1.26 ± 0.03	0.47	0.79	0.91	0.25
	Fused LASSO	$[10^{-9}, 10^{0}]$	0.006	1.18 ± 0.04	0.40	0.78	0.89	0.23

Conclusion

Thanks to the development and generalization of imaging technologies over the last decades, larger databases of observations of patients are progressively available. With an increased resolution of the images, increased number of patients, and increased number of observations per patient, population analysis can be more accurate and statistically meaningful. The potential benefits of such studies include better understanding of diseases, identification of risk factors, development of preventive and curative treatments. In the end, patients are progressively getting higher quality medical care with improved follow-up and improved quality of life.

⋆
⋆ ⋆

In this thesis, we studied mathematical methods of image analysis for both cross-sectional and longitudinal studies. In this conclusion, we first analyze the contributions with a transverse point of view, and then present potential perspectives of this thesis.

Revisited contributions

Definition of features for population analysis

In machine learning, defining the representation of the data is a vital step to obtain high numerical performance. This holds true in many applications, and in particular for population analysis via medical imaging.

First, we performed studies where each feature represents *an anatomical neighborhood*, i.e. it contains information at the voxel level (it can be related to the anatomy in magnetic resonance (MR) images, activity in functional MR images,

etc). In particular, we evaluated several local features for the segmentation of white matter hyper-intensities (WMH). These local features were built using up to four modalities, and the relative contribution of each modality was evaluated. Our study showed a numerical trade-off between the quantity of information and performance vs numerical efficiency.

Second, we performed studies where each feature represents *the snapshot of a patient*, i.e. that contains information of the patient at a specific time point. We evaluated single and multi-modality representations of patients for manifold learning. We saw that results can depend on the pre-processing (e.g. registration) or the brain area considered.

Third, we performed studies where each feature represents *the evolution of a patient*, i.e. that contains information of the patient between several time points. We evaluated the use of local descriptors from the large deformation diffeomorphic metric mapping (LDDMM) framework. We studied how local descriptors such as initial momenta can be used in disease progression classification, and what could be done so that they outperform global descriptors such as (relative) volume variation.

Building a distance between features

Once we have defined the features, the next question is: how to measure the distance between them? When the features represent anatomical neighborhoods, the Euclidean distance is commonly used. However, the answer is not as straightforward when it comes to features representing patients. One way to address the problem is to register the features towards a common template and compute Euclidean distance once the features are transported to this common space. Using this strategy, the results depends on (1) the computation of the template, (2) the quality and properties of the registration algorithm and (3) the choice of the transport.

With regard to the first point, we introduced extensions of the Karcher algorithm to define a template up-to rigid transformations, that keeps sharp boundaries by avoiding consecutive smoothings. For an efficient computation, we only selected a subpart of the population and used an interesting selection criterion to ensure proper final matching between the images and the final template.

With regard to the second point, we showed that having the choice of the registration algorithm can be helpful in some situations. For example, a non-rigid registration can help to remove shape effects when ones wants to focus his/her study on intensity. Please note that we neither quantified the sensibility of our pipelines to this step nor evaluated template-free methods, though these ideas can be interesting perspectives.

With regard to the third point, we evaluated several transport strategies in Chapter 9. Please note that the results are still preliminary and we consider the question of transport as still open.

Optimization strategies and regularizations

In this thesis, we built predictive models in several challenging settings.

First, we built classification models in studies with *many observations containing few information*. This was the case in Chapter 5, where the design and use of an appropriate region of interest (ROI) was vital to obtain accurate lesion segmentation.

Second, we built classification models in studies with *few observations in high dimension*. That was the case in Chapters 7, 9, 10. In that setting, an appropriate choice of the distance between features and the introduction of regularizations can improve the results. For example, the introduction of spatial regularizations in Chapter 10 led to more meaningful maps of coefficients.

Finally, we built classification models with *skewed classes* in all studies (Chapters 5, 7, 9, 10), i.e. the datasets did not contain as many observations for all classes. This is a common issue in medical imaging, due to the nature of the observations. In this thesis, we used two strategies: either we forced the training sets to be balanced, or used weighted loss functions.

Discovering and quantifying trends in populations

Being able to predict the evolution of a patient is one of the major goals of medical research. It helps the medical staff to provide the most appropriate treatments to patients. Understanding disease progression or uncovering modes of variation or trends in a population are likely to help to achieve this goal.

First, we used non-linear dimensionality reduction (DR) algorithms to build low-dimensional manifolds. Such manifolds enable the *visualization* of trends in populations. Several pipelines can be used in order to find intensity or shape trends.

Second, we also introduced an extension of Laplacian Eigenmaps able to process combined imaging and clinical data. Such extensions were shown to be able to improve the diagnosis *classification* performance. We also proposed a pipeline based on the LDDMM framework for the classification performance for disease progression.

Third, we first did a proof of concept on the use of hippocampus subregions to improve classification results. Then we evaluated a logistic classification models, and showed that spatial regularizations can be introduced to discover hippocampus areas related to disease progression. This is an important step for the *identification of biomarkers*.

⋆

⋆ ⋆

Perspectives

Here is a non-exhaustive lists of perspectives of the work presented in this thesis.

Discovering and quantifying trends in populations

In this thesis, we built manifolds for cross-sectional studies with simple Euclidean distances. Despite the simplicity of this distance, these manifolds provide interesting insight of the trends of the studied populations. It would be interesting to build some using deformation-based distances, for both cross-sectional and longitudinal studies.

Definition of features and distances for population analysis

We have seen in this thesis that defining appropriate features is an important step for population analysis.

A promising perspective is the use of *surface models*, for example for the longitudinal analysis of hippocampi. Indeed, the high values in the initial momenta lie close to the surface boundary, hence surface models seem appropriate.

Another perspective is to work on the definition of features encoding the evolutions of patients with *more than two time points*, for example based on geodesic regression [Niethammer 2011] or cubic spline interpolation [Trouvé 2012].

As mentioned earlier, methods based on local comparisons can depend on the quality of the template and/or registration algorithm. Instead of building features that need to be transported and compared in a common space, structural approaches could be interesting alternatives [Mangin 2004]. Such strategies have been used in the literature in the field of computer vision.

Optimization strategies and regularizations

We have seen that optimization strategies and regularizations are important given the challenging settings in population analysis. Several perspectives and further studies would require new regularizations.

In Chapter 10, we introduced spatial regularizations in 3D. As mentioned earlier, one perspective of our work is to work on surface models to encode hippocampus shape evolutions. It would be particularly interesting to study in the effect of *spatial regularizations on surfaces* for predictive models.

We also introduced the idea of working with spatio-temporal data with more than two time points. In that case, the introduction of *spatio-temporal regularization* could be relevant.

Appendices

Appendix A

Proofs

A.1 Proofs of Chapter 2

In this section, the proofs of Chapter 2 are presented. Notations are the same as in Chapter 2.

A.1.1 Proof of Theorem 2.1.1

Proof. Let us assume that for all $\mathbf{x} \in \mathcal{X}$ the infimum $\inf_{y \in \mathcal{Y}} \mathbb{E}_{p_{\mathbf{XY}}(dY|\mathbf{x})} \ell(Y, y)$ is reached. Because the infimum is reached, we can define the function $f^* \colon \mathcal{X} \to \mathcal{Y}$ such that $\forall \mathbf{x} \in \mathcal{X}, f^*(x) \in \operatorname{argmin}_{y \in \mathcal{Y}} \mathbb{E}_{p_{\mathbf{XY}}(dY|\mathbf{x})} \ell(Y, y)$. The theorem states that f^* is a target function, i.e. a prediction function minimizing the risk. Let us consider a prediction function $f \in \mathcal{F}(\mathcal{X}, \mathcal{Y})$ and show that the risk of f is higher or equal than the one of f^*. The definition of the risk of f is

$$R(f) = \mathbb{E}\left[\ell(f(\mathbf{X}), \mathbf{Y})\right]. \tag{A.1}$$

Using conditional probabilities

$$R(f) = \mathbb{E}_{p_{\mathbf{XY}}(d\mathbf{X})}\left[\mathbb{E}_{p_{\mathbf{XY}}(d\mathbf{Y}|\mathbf{X})}\left[\ell(f(\mathbf{X}), \mathbf{Y})\right]\right]. \tag{A.2}$$

By definition of the infimum we get

$$R(f) \geq \mathbb{E}_{p_{\mathbf{XY}}(d\mathbf{X})}\left[\inf_{y \in \mathcal{Y}} \mathbb{E}_{p_{\mathbf{XY}}(d\mathbf{Y}|\mathbf{X})}\left[\ell(y, \mathbf{Y})\right]\right]. \tag{A.3}$$

By definition of f^* we get

$$R(f) \geq \mathbb{E}_{p_{\mathbf{XY}}(d\mathbf{X})}\left[\mathbb{E}_{p_{\mathbf{XY}}(d\mathbf{Y}|\mathbf{X})}\left[\ell(f^*(\mathbf{X}), \mathbf{Y})\right]\right]. \tag{A.4}$$

Removing the conditional probabilities we get

$$R(f) \geq \mathbb{E}\left[\ell(f^*(\mathbf{X}), \mathbf{Y})\right]. \tag{A.5}$$

So by definition of the risk of f^*

$$R(f) \geq R(f^*) \tag{A.6}$$

□

A.1.2 Proof of Theorem 2.1.2

Proof. To establish this result, we are going to use the following lemma.

Lemma A.1.1 (Variance decomposition)**.** *Let* $\mathbf{W}$ *be a random variable and* $a \in \mathbb{R}$*, then*

$$\mathbb{E}\left[(\mathbf{W}-a)^2\right] = \mathbb{E}\left[(\mathbf{W}-\mathbb{E}\mathbf{W})^2\right] + (\mathbb{E}\mathbf{W}-a)^2. \tag{A.7}$$

Proof of Lemma A.1.1. Let $V_a \overset{\text{def.}}{=} \mathbb{E}\left[(\mathbf{W}-a)^2\right]$ for any $a \in \mathbb{R}$. Expanding the square gives

$$V_a = \mathbb{E}\left[\mathbf{W}^2 - 2a\mathbf{W} + a^2\right]. \tag{A.8}$$

By linearity of the expected values we have

$$V_a = \mathbb{E}\left[\mathbf{W}^2\right] - 2a\mathbb{E}\mathbf{W} + a^2, \tag{A.9}$$
$$= \mathbb{E}\left[\mathbf{W}^2\right] - 2a\mathbb{E}\mathbf{W} + a^2 + (\mathbb{E}\mathbf{W})^2 - (\mathbb{E}\mathbf{W})^2. \tag{A.10}$$

Hence by reordering

$$V_a = \mathbb{E}\left[\mathbf{W}^2\right] - (\mathbb{E}\mathbf{W})^2 + (\mathbb{E}\mathbf{W}-a)^2 \tag{A.11}$$

In particular $V_{\mathbb{E}\mathbf{W}} = \mathbb{E}\left[\mathbf{W}^2\right] - (\mathbb{E}\mathbf{W})^2$. Therefore

$$V_a = V_{\mathbb{E}\mathbf{W}} + (\mathbb{E}\mathbf{W}-a)^2 \tag{A.12}$$

□

Now let us get back to the proof of Theorem 2.1.2. We apply the Lemma A.1.1 with $\mathbb{E} = \mathbb{E}_{p_{\mathbf{XY}}(d\mathbf{Y}|\mathbf{X}=\mathbf{x})}$, $\mathbf{W} = \mathbf{Y}$ and $a = y$.

$$\begin{aligned}\mathbb{E}_{p_{\mathbf{XY}}(d\mathbf{Y}|\mathbf{X}=\mathbf{x})}\ell(\mathbf{Y},y) &= \mathbb{E}_{p_{\mathbf{XY}}(d\mathbf{Y}|\mathbf{X}=\mathbf{x})}(\mathbf{Y}-y)^2 \\ &= \mathbb{E}_{p_{\mathbf{XY}}(d\mathbf{Y}|\mathbf{X}=\mathbf{x})}(\mathbf{Y}-\mathbb{E}_{p_{\mathbf{XY}}(d\mathbf{Y}|\mathbf{X}=\mathbf{x})}\mathbf{Y})^2 + (\mathbb{E}_{p_{\mathbf{XY}}(d\mathbf{Y}|\mathbf{X}=\mathbf{x})}\mathbf{Y}-y)^2\end{aligned}$$

The first term does not depend on y, so the infimum of $\mathbb{E}_{p_{\mathbf{XY}}(d\mathbf{Y}|\mathbf{X}=\mathbf{x})}\ell(\mathbf{Y},y)$ for $y \in \mathcal{Y}$ is obtained for $y = \mathbb{E}(\mathbf{Y}|\mathbf{X}=\mathbf{x})$. Therefore,

$$\eta^*(\mathbf{x}) \overset{\text{def.}}{=} \mathbb{E}(\mathbf{Y}|\mathbf{X}=\mathbf{x}) \tag{A.13}$$

is a target function. Now we want to quantify the excess of risk of $\eta\colon \mathcal{X} \to \mathbb{R}$. The risk of η is

$$R(\eta) = \mathbb{E}(\mathbf{Y}-\eta(\mathbf{X}))^2. \tag{A.14}$$

Using conditional probabilities we get

$$R(\eta) = \mathbb{E}_{p_{\mathbf{XY}}(d\mathbf{X})}\left[\mathbb{E}_{p_{\mathbf{XY}}(d\mathbf{Y}|\mathbf{X}=\mathbf{x})}(\mathbf{Y}-\eta(\mathbf{x}))^2\right]. \tag{A.15}$$

Now we apply the Lemma A.1.1 with $\mathbb{E} = \mathbb{E}_{p_{\mathbf{XY}}(d\mathbf{Y}|\mathbf{X}=\mathbf{x})}$, $\mathbf{W} = \mathbf{Y}$ and $a = \eta(\mathbf{x})$

$$R(\eta) = \mathbb{E}_{p_{\mathbf{XY}}(d\mathbf{X})}\left[\mathbb{E}_{p_{\mathbf{XY}}(d\mathbf{Y}|\mathbf{X}=\mathbf{x})}(\mathbf{Y} - \mathbb{E}(\mathbf{Y}|\mathbf{X}))^2 + (\mathbb{E}(\mathbf{Y}|\mathbf{X}) - \eta(\mathbf{x}))^2\right]. \quad \text{(A.16)}$$

By definition of η^*

$$R(\eta) = \mathbb{E}_{p_{\mathbf{XY}}(d\mathbf{X})}\left[\mathbb{E}_{p_{\mathbf{XY}}(d\mathbf{Y}|\mathbf{X}=\mathbf{x})}(\mathbf{Y} - \eta^*(\mathbf{x}))^2 + (\eta^*(\mathbf{x}) - \eta(\mathbf{x}))^2\right]. \quad \text{(A.17)}$$

Finally by linearity of the expected values and definition of the risk, we get

$$R(\eta) = R(\eta^*) + \mathbb{E}_{p_{\mathbf{XY}}(d\mathbf{X})}\left[(\eta^*(\mathbf{X}) - \eta(\mathbf{X}))^2\right], \quad \text{(A.18)}$$

$$= R(\eta^*) + \mathbb{E}(\eta - \eta^*)^2. \quad \text{(A.19)}$$

□

A.1.3 Proof of Theorem 2.1.3

Proof. The Theorem 2.1.3 claims the form of the target functions in a classification problem with the $\ell^{0/1}$ loss. First, since $\operatorname{card}\mathcal{Y} < \infty$, for all $\mathbf{x} \in \mathcal{X}$ the infimum $\inf_{y\in\mathcal{Y}} \mathbb{E}_{p_{\mathbf{XY}}(dY|\mathbf{X}=\mathbf{x})}\ell(Y, y)$ is reached. Therefore by Theorem 2.1.1 we can define f^* such that

$$\forall \mathbf{x} \in \mathcal{X}, \quad f^*(\mathbf{x}) \in \operatorname*{argmin}_{y\in\mathcal{Y}} \mathbb{E}_{p_{\mathbf{XY}}(d\mathbf{Y}|\mathbf{X}=\mathbf{x})}\ell(Y, y). \quad \text{(A.20)}$$

With $\ell = \ell^{0/1}$, we get

$$\forall \mathbf{x} \in \mathcal{X}, \quad f^*(\mathbf{x}) \in \operatorname*{argmin}_{y\in\mathcal{Y}} \mathbb{E}_{p_{\mathbf{XY}}(d\mathbf{Y}|\mathbf{X}=\mathbf{x})}(\mathbb{1}_{\mathbf{Y}\neq y}). \quad \text{(A.21)}$$

By definition of the expected value of the characteristic function, we get

$$\forall \mathbf{x} \in \mathcal{X}, \quad f^*(\mathbf{x}) \in \operatorname*{argmin}_{y\in\mathcal{Y}} p_{\mathbf{XY}}(\mathbf{Y} \neq y|\mathbf{X} = \mathbf{x}), \quad \text{(A.22)}$$

or equivalently

$$\forall \mathbf{x} \in \mathcal{X}, \quad f^*(\mathbf{x}) \in \operatorname*{argmax}_{y\in\mathcal{Y}} p_{\mathbf{XY}}(\mathbf{Y} = y|\mathbf{X} = \mathbf{x}). \quad \text{(A.23)}$$

Now when $\mathcal{Y} = \{-1, +1\}$, $\eta^*(\mathbf{x})$ can be decomposed using the definition of the expected value

$$\begin{aligned}
\eta^*(\mathbf{x}) &= \mathbb{E}(\mathbf{Y}|\mathbf{X} = \mathbf{x}),\\
&= p_{\mathbf{XY}}(\mathbf{Y} = 1|\mathbf{X} = \mathbf{x}) \times 1 + p_{\mathbf{XY}}(\mathbf{Y} = -1|\mathbf{X} = \mathbf{x}) \times (-1),\\
&= p_{\mathbf{XY}}(\mathbf{Y} = 1|\mathbf{X} = \mathbf{x}) \times 1 + (1 - p_{\mathbf{XY}}(\mathbf{Y} = 1|\mathbf{X} = \mathbf{x})) \times (-1),\\
&= 2 \times p_{\mathbf{XY}}(\mathbf{Y} = 1|\mathbf{X} = \mathbf{x}) - 1.
\end{aligned}$$

Therefore

$$\begin{aligned}\text{sign}(\eta^*(\mathbf{x})) = 1 \Leftrightarrow \eta^*(\mathbf{x}) \geq 0 &\Leftrightarrow p_{\mathbf{XY}}(\mathbf{Y}=1|\mathbf{X}=\mathbf{x}) \geq \frac{1}{2},\\ &\Leftrightarrow p_{\mathbf{XY}}(\mathbf{Y}=1|\mathbf{X}=\mathbf{x}) \geq p_{\mathbf{XY}}(\mathbf{Y}=-1|\mathbf{X}=\mathbf{x}),\\ &\Leftrightarrow 1 \in \underset{y\in\{0,1\}}{\text{argmax}}\, p_{\mathbf{XY}}(\mathbf{Y}=y|\mathbf{X}=\mathbf{x}).\end{aligned}$$

□

A.1.4 Proof of Proposition 2.1.4

Proof. We want to bound $R(f_{ERM}) - R(\hat{f})$. Let us introduce some terms in the expression and then bound the terms by pairs.

$$\begin{aligned}R(f_{ERM}) - R(\hat{f}) &= R(f_{ERM}) - r(f_{ERM}) + \underbrace{r(f_{ERM}) - r(\hat{f})}_{\leq 0} + r(\hat{f}) - R(\hat{f}),\\ &\leq \sup_{f\in\hat{\mathcal{F}}} |R(f) - r(f)| + 0 + \sup_{f\in\hat{\mathcal{F}}} |r(f) - R(f)|,\\ &\leq 2 \times \sup_{f\in\hat{\mathcal{F}}} |R(f) - r(f)|\,.\end{aligned}$$

□

A.1.5 Proof of the inequalities for the bias-variance trade-off

Proof. First, f_{ERM} is defined as a minimizer of r on $\hat{\mathcal{F}}$ whereas $\hat{f}$ is defined as a minimizer of R on $\hat{\mathcal{F}}$. Therefore, we have $R(f_{ERM}) \geq R(\hat{f})$.

□

A.2 Proofs of Chapter 5 [1]

In this section, the proofs of Chapter 5 are presented. Notations are the same as in Chapter 5.

A.2.1 Proof of the Proposition 5.2.1

Proof. Because the fluid attenuated inversion recovery (FLAIR) image is bounded on Ω_W, the minimum (resp. maximum) $m = \min\{I(\omega); \quad \omega \in \Omega_W\}$ (resp. $M = \max\{I(\omega); \quad \omega \in \Omega_W\}$) is well defined.

1. $\forall \tau < m$ and $\omega \in \Omega_W$, $I(\omega) \geq m > \tau$ and so $M_\tau(\omega) = 1$. For $\omega \in \Omega \setminus \Omega_W, I(\omega) = -\infty$, so $I(\omega) > \tau$ is false and $M_\tau(\omega) = 0$.

[1] The proofs in this section are inspired by the class "Kernel methods" by Jean-Philippe Vert from the MSc "Math, Vision, Learning" of ENS Cachan, France.

2. $\forall \tau > M$ and $\omega \in \Omega$, $I(\omega) > \tau > M$ is false by definition of M and then $M_\tau(\omega) = 0$.

□

A.2.2 Proof of Theorem 5.2.2

Proof. Let us note $\mathcal{H}_K^s$ the linear span in $\mathcal{H}_K$ of the vectors $K_{\mathbf{x}_i}$

$$\mathcal{H}_K^s \overset{\text{def.}}{=} \left\{ f \in \mathcal{H}_K \colon f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i K(\mathbf{x}, \mathbf{x}_i); \quad (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n \right\}. \tag{A.24}$$

$\mathcal{H}_K^s$ is a finite-dimensional subspace, therefore any function $f \in \mathcal{H}_K$ can be uniquely decomposed as

$$f = f^s + f^\perp, \tag{A.25}$$

with $f^s \in \mathcal{H}_K^s$ and $f^\perp \in (\mathcal{H}_K^s)^\perp$.

Since $\mathcal{H}_K$ is a RKHS, $\forall i \in [\![1, n]\!], \quad f^\perp(\mathbf{x}_i) = \langle f^\perp, K(\mathbf{x}_i, \cdot) \rangle_{\mathcal{H}_K}$. Therefore

$$\forall i \in [\![1, n]\!], \quad f(\mathbf{x}_i) = f^s(\mathbf{x}_i). \tag{A.26}$$

Then Pythagora's theorem gives

$$\|f\|_K^2 = \|f^s\|_K^2 + \|f^\perp\|_K^2. \tag{A.27}$$

From (A.26) and (A.27) the value functional in (5.11) is therefore higher or equal for f than for f^s, with equality when $f^\perp = 0$. The value of the minimum is therefore in $\mathcal{H}_K^s$. □

A.2.3 Proof of Proposition 5.2.4

Proof. L has a quadratic term and a linear term in $\boldsymbol{\alpha}$, so it gives us

$$\nabla_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\nu}, \boldsymbol{\mu}) = 2\gamma K\boldsymbol{\alpha} - K\mathbf{Y}\boldsymbol{\mu}, \tag{A.28}$$

$$= K(2\gamma\boldsymbol{\alpha} - \mathbf{Y}\boldsymbol{\mu}), \tag{A.29}$$

where $\mathbf{Y}$ is the diagonal matrix with entries $\mathbf{Y}_{ii} = y_i$. So solving $\nabla_{\boldsymbol{\alpha}} L = 0$ leads to

$$\boldsymbol{\alpha} = \frac{\mathbf{Y}\boldsymbol{\mu}}{2\gamma} + \varepsilon \tag{A.30}$$

with ε such that $K\varepsilon = 0$. However, ε does not change f, so we can chose $\varepsilon = 0$, hence the result. □

A.2.4 Proof of Proposition 5.2.5

Proof. To compute $\nabla_{\boldsymbol{\xi}} L$ (gradient of L with regard to ξ), we compute the partial derivatives $\frac{\partial L}{\partial \xi_i}(\boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\nu}, \boldsymbol{\mu})$. Since L is linear in ξ_i, we get

$$\forall i \in [\![1, n]\!], \quad \frac{\partial L}{\partial \xi_i}(\boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\nu}, \boldsymbol{\mu}) = \frac{1}{n} - \mu_i - \nu_i. \tag{A.31}$$

When the gradient of L with regard to $\boldsymbol{\xi}$ is null, all the partial derivatives of L with regard to ξ_i are null, which leads to

$$\forall i \in [\![1, n]\!], \quad 0 = \frac{1}{n} - \mu_i - \nu_i. \tag{A.32}$$

□

A.2.5 Proof of Proposition 5.2.6

Proof. In Appendix A.2.4, we have seen that L is linear in ξ_i with the coefficient $\frac{1}{n} - \mu_i - \nu_i$. Therefore if $\exists i$ such that $\frac{1}{n} - \mu_i - \nu_i \neq 0$, then $\inf_{\xi_i} L(\boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\nu}, \boldsymbol{\mu}) = -\infty$ and $q(\boldsymbol{\mu}, \boldsymbol{\nu}) = -\infty$.

Now if $\forall i \in [\![1, n]\!], \quad 0 = \frac{1}{n} - \mu_i - \nu_i$, we get

$$L(\boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\nu}, \boldsymbol{\mu}) = \gamma \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} - \sum_{i=1}^{n} \mu_i \left(-1 + y_i \sum_{j=1}^{n} \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \right), \tag{A.33}$$

$$= \gamma \sum_{i,j=1}^{n} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^{n} \mu_i - \sum_{i,j=1}^{n} \mu_i y_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j), \tag{A.34}$$

$$= \sum_{i=1}^{n} \mu_i + \sum_{i,j=1}^{n} \left(\gamma \alpha_i \alpha_j - \mu_i y_i \alpha_j \right) K(\mathbf{x}_i, \mathbf{x}_j). \tag{A.35}$$

Now using the optimality condition of Proposition 5.2.4, we get

$$q(\boldsymbol{\mu}, \boldsymbol{\nu}) = \sum_{i=1}^{n} \mu_i + \sum_{i,j=1}^{n} \left(\gamma \frac{y_i \mu_i}{2\gamma} \frac{y_j \mu_j}{2\gamma} - \mu_i y_i \frac{y_j \mu_j}{2\gamma} \right) K(\mathbf{x}_i, \mathbf{x}_j), \tag{A.36}$$

$$= \sum_{i=1}^{n} \mu_i - \frac{1}{4\gamma} \sum_{i,j=1}^{n} y_i y_j \mu_i \mu_j K(\mathbf{x}_i, \mathbf{x}_j). \tag{A.37}$$

□

A.2.6 Proof of Proposition 5.2.7

Proof. The dual problem consists in maximizing $q(\boldsymbol{\mu}, \boldsymbol{\nu})$ subject to $\begin{cases} \boldsymbol{\mu} \geq 0 \\ \boldsymbol{\nu} \geq 0 \end{cases}$.

However, if $\exists i$ such that $\mu_i > \frac{1}{n}$: as $\nu_i \geq 0$, $\mu_i + \nu_i \neq \frac{1}{n}$ and $q(\boldsymbol{\mu}, \boldsymbol{\nu}) = -\infty$.

If $\forall i,\ 0 \leq \mu_i \leq \frac{1}{n}$, the dual function takes finite values when $\nu_i = \frac{1}{n} - \mu_i$.
In that case, the dual problem can be written as an optimization problem on the μ_i.
□

A.3 Proofs of Chapter 6

A.3.1 Proof of Proposition 6.2.1

.

Proof.

$$\begin{aligned}
\sum_{i,j=1}^{n} w_{ij}||\widetilde{\mathbf{x}_i} - \widetilde{\mathbf{x}_j}||^2 &= \sum_{i,j} w_{ij}\Big(||\widetilde{\mathbf{x}_i}||^2 - 2 < \widetilde{\mathbf{x}_i}, \widetilde{\mathbf{x}_j} > + ||\widetilde{\mathbf{x}_j}||^2\Big) \\
&= 2\sum_{i} \underbrace{\Big(\sum_{j} w_{ij}\Big)}_{d_{ii}} ||\widetilde{\mathbf{x}_i}||^2 - 2\sum_{i,j} w_{ij} < \widetilde{\mathbf{x}_i}, \widetilde{\mathbf{x}_j} > \\
&= 2\,\mathrm{Tr}\left(\widetilde{\mathbf{X}}^T \mathbf{D} \widetilde{\mathbf{X}}\right) - 2\,\mathrm{Tr}\left(\widetilde{\mathbf{X}}^T \mathbf{W} \widetilde{\mathbf{X}}\right) \\
&= 2\,\mathrm{Tr}\left(\widetilde{\mathbf{X}}^T \mathbf{L} \widetilde{\mathbf{X}}\right)
\end{aligned}$$

□

Bibliography

[Acosta 2009] Oscar Acosta, Pierrick Bourgeat, Maria A. Zuluaga, Jurgen Fripp, Olivier Salvado and Sébastien Ourselin. *Automated voxel-based 3D cortical thickness measurement in a combined Lagrangian-Eulerian PDE approach using partial volume maps.* Medical Image Analysis, vol. 13, no. 5, pages 730 – 743, 2009. 105, 114, 128

[Aljabar 2009] P. Aljabar, R.A. Heckemann, A. Hammers, J.V. Hajnal and D. Rueckert. *Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy.* NeuroImage, vol. 46, no. 3, pages 726–738, 2009. 90

[Aljabar 2012] P. Aljabar, R. Wolz and D. Rueckert. Machine learning in computer-aided diagnosis: Medical imaging intelligence and analysis, chapitre Manifold Learning for Medical Image Registration, Segmentation, and Classification. IGI Global, 2012. 18, 89, 92

[Allassonnière 2007] S. Allassonnière, Y. Amit and A. Trouvé. *Towards a coherent statistical framework for dense deformable template estimation.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 69, no. 1, pages 3–29, 2007. 168

[Allassonnière 2008] Stéphanie Allassonnière, Estelle Kuhn and Alain Trouvé. *MAP Estimation of Statistical Deformable Templates Via Nonlinear Mixed Effects Models : Deterministic and Stochastic Approaches.* In Xavier Pennec, editeur, 2nd MICCAI Workshop on Mathematical Foundations of Computational Anatomy, pages 80–91, New-York, United States, October 2008. 168

[Ambrosini 2010] Robert D. Ambrosini, Peng Wang and Walter G. O'Dell. *Computer-aided detection of metastatic brain tumors using automated three-dimensional template matching.* Journal of Magnetic Resonance Imaging, vol. 31, no. 1, pages 85–93, 2010. 102

[Amieva 2007] Hélène Amieva, Sandrine Andrieu, Claudine Berr and et al. *Maladie d'Alzheimer : enjeux scientifiques, médicaux et sociétaux.* Bilan, INSERM, 2007. XV - 654 pages, illustrations, figures. 44

[Anbeek 2004] Petronella Anbeek, Koen L. Vincken, Matthias J. P. van Osch, Robertus H. C. Bisschops and Jeroen van der Grond. *Probabilistic segmentation of white matter lesions in MR imaging.* NeuroImage, vol. 21, no. 3, pages 1037 – 1044, 2004. 102, 114, 120, 122

[Arsigny 2006] Vincent Arsigny, Olivier Commowick, Xavier Pennec and Nicholas Ayache. *A log-Euclidean framework for statistics on diffeomorphisms.* Med

Image Comput Comput Assist Interv, vol. 9, no. Pt 1, pages 924–931, 2006. 163, 164

[Ashburner 2000] J. Ashburner and K. J. Friston. *Voxel-based morphometry–the methods.* Neuroimage, vol. 11, no. 6 Pt 1, pages 805–821, Jun 2000. 159

[Ashburner 2007] John Ashburner. *A fast diffeomorphic image registration algorithm.* Neuroimage, vol. 38, no. 1, pages 95–113, Oct 2007. 160, 163

[Avants 2004] Brian Avants and James C. Gee. *Geodesic estimation for large deformation anatomical shape averaging and interpolation.* NeuroImage, vol. 23, no. Supplement 1, pages S139 – S150, 2004. Mathematics in Brain Imaging. 167, 168

[Bai 2012] Jordan Bai, Thi Lan Huong Trinh, Kai-Hsiang Chuang and Anqi Qiu. *Atlas-based automatic mouse brain image segmentation revisited: model complexity vs. image registration.* vol. 30, 2012. 18, 91

[Batmanghelich 2008] Verma R. Batmanghelich N. *On non-linear characterization of tissue abnormality by constructing disease manifolds.* pages 1 –8, June 2008. 18, 95, 96, 97

[Bauer 2013] Stefan Bauer, Roland Wiest, Lutz-P Nolte and Mauricio Reyes. *A survey of MRI-based medical image analysis for brain tumor studies.* Physics in Medicine and Biology, vol. 58, no. 13, page R97, 2013. 102

[Beck 2009] A. Beck and M. Teboulle. *Fast Gradient-Based Algorithms for Constrained Total Variation Image Denoising and Deblurring Problems.* Image Processing, IEEE Transactions on, vol. 18, no. 11, pages 2419 –2434, nov. 2009. 200

[Beg 2005] M. Faisal Beg, Michael I. Miller, Alain Trouvé and Laurent Younes. *Computing Large Deformation Metric Mappings via Geodesic Flows of Diffeomorphisms.* Int. J. Comput. Vision, vol. 61, no. 2, pages 139–157, 2005. 162, 167, 177, 197

[Beg 2006] M. F. Beg and A. Khan. *Computing an average anatomical atlas using LDDMM and geodesic shooting.* In Biomedical Imaging: Nano to Macro, 2006. 3rd IEEE International Symposium on, pages 1116–1119. IEEE, April 2006. 20, 166, 167

[Belkin 2003] Mikhail Belkin and Partha Niyogi. *Laplacian Eigenmaps for Dimensionality Reduction and Data Representation.* Neural Comput., vol. 15, 2003. 84, 129, 143

[Belkin 2004] Mikhail Belkin, Partha Niyogi, Vikas Sindhwani and Peter Bartlett. *Manifold Regularization: A Geometric Framework for Learning from Examples.* Technical report, 2004. 18, 36, 72, 93, 95

[Bhatia 2004] K. K. Bhatia, J. V. Hajnal, B. K. Puri, A. D. Edwards and D. Rueckert. *Consistent groupwise non-rigid registration for atlas construction.* In Biomedical Imaging: Nano to Macro, 2004. IEEE International Symposium on, pages 908–911 Vol. 1, 2004. 169

[Bishop 2007] Christopher M. Bishop. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer, 1st ed. 2006. corr. 2nd printing 2011 édition, October 2007. 17, 59, 60, 61

[Blezek 2007] Daniel J. Blezek and James V. Miller. *Atlas stratification.* Medical Image Analysis, vol. 11, 2007. 18, 93, 94, 127

[Bossa 2007] Matias Bossa, Monica Hernandez and Salvador Olmos. *Contributions to 3D diffeomorphic atlas estimation: application to brain images.* Med Image Comput Comput Assist Interv, vol. 10, no. Pt 1, pages 667–674, 2007. 163

[Bossa 2008] Matias Bossa and Salvador Olmos. *A New Algorithm for the Computation of the Group Logarithm of Diffeomorphisms.* In Xavier Pennec and Sarang Joshi, editeurs, Second International Workshop on Mathematical Foundations of Computational Anatomy - Geometrical and Statistical Methods for Modelling Biological Shape Variability, New York, USA, 2008. 163, 164

[Braak 1998] H. Braak and E. Braak. *Evolution of neuronal changes in the course of Alzheimer's disease.* In K. Jellinger, F. Fazekas and M. Windisch, editeurs, Ageing and Dementia, volume 53 of *Journal of Neural Transmission. Supplementa*, pages 127–140. Springer Vienna, 1998. 186

[Cabezas 2011] Mariano Cabezas, Arnau Oliver, Xavier Lladó, Jordi Freixenet and Meritxell Bach Cuadra. *A review of atlas-based segmentation for magnetic resonance brain images.* Comput. Methods Prog. Biomed., vol. 104, no. 3, pages e158–e177, December 2011. 90

[Camastra 2002] Francesco Camastra and Alessandro Vinciarelli. *Estimating the Intrinsic Dimension of Data with a Fractal-Based Method.* IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 10, pages 1404–1407, 2002. 86

[Camastra 2003] F. Camastra. *Data dimensionality estimation methods: a survey.* Pattern Recognition, vol. 36, no. 12, pages 2945–2954, December 2003. 86

[Cayton 2005] Lawrence Cayton. Algorithms for manifold learning. June 2005. 83

[Charon 2013] Nicolas Charon and Alain Trouvé. *Functional Currents: A New Mathematical Tool to Model and Analyse Functional Shapes.* Journal of Mathematical Imaging and Vision, pages 1–19, 2013. 162

[Choi 2000] Yongchoel Choi and Seungyong Lee. *Injectivity Conditions of 2D and 3D Uniform Cubic B-Spline Functions.* Graphical Models, vol. 62, no. 6, pages 411 – 427, 2000. 161

[Chupin 2007] Marie Chupin, A. Romain Mukuna-Bantumbakulu, Dominique Hasboun, Eric Bardinet, Sylvain Baillet, Serge Kinkingnéhun, Louis Lemieux, Bruno Dubois and Line Garnero. *Anatomically constrained region deformation for the automated segmentation of the hippocampus and the amygdala: Method and validation on controls and patients with Alzheimer's disease.* Neuroimage, vol. 34, no. 3, pages 996–1019, Feb 2007. 177

[Chupin 2009] Marie Chupin, Emilie Gérardin, Rémi Cuingnet, Claire Boutet, Louis Lemieux, Stéphane Lehéricy, Habib Benali, Line Garnero, Olivier Colliot and 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, vol. 19, no. 6, pages 579–587, Jun 2009. 177

[Ciarlet 1988] Philippe G. Ciarlet and Jacques-Louis Lions. Introduction à l'analyse numérique matricielle et à l'optimisation. 1988. 63

[Coifman 2006] Ronald R. Coifman and Stephane Lafon. *Diffusion maps.* Applied and Computational Harmonic Analysis, vol. 21, no. 1, pages 5–30, July 2006. 84

[Collins 1998] D.L. Collins, A.P. Zijdenbos, V. Kollokian, J.G. Sled, N.J. Kabani, C.J. Holmes and A.C. Evans. *Design and construction of a realistic digital brain phantom.* Medical Imaging, IEEE Transactions on, vol. 17, no. 3, pages 463 –468, june 1998. 105

[Cox 1994] T. Cox and M. Cox. Multidimensional scaling. Chapman and Hall, 1994. 84

[Cuingnet 2011a] Rémi Cuingnet. *Contributions to statistical learning for structural neuroimaging data - Contributions à l'apprentissage automatique pour l'analyse d'images cérébrales anatomiques.* PhD thesis, University of Paris Sud 11, 2011. 43

[Cuingnet 2011b] Rémi Cuingnet, Emilie Gerardin, Jérôme Tessieras, Guillaume Auzias, Stéphane Lehéricy, Marie-Odile Habert, Marie Chupin, Habib Benali and Olivier Colliot. *Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database.* NeuroImage, vol. 56, no. 2, pages 766 – 781, 2011. 177

[Cuingnet 2012] Rémi Cuingnet, Joan Alexis Glaunès, Marie Chupin, Habib Benali and Olivier Colliot. *Spatial and Anatomical Regularization of SVM: A General Framework for Neuroimaging Data.* IEEE Trans Pattern Anal Mach Intell, Jun 2012. 197

[Datta 2006] Sushmita Datta, Balasrinivasa Rao Sajja, Renjie He, Jerry S. Wolinsky, Rakesh K. Gupta and Ponnada A. Narayana. *Segmentation and quan-*

tification of black holes in multiple sclerosis. NeuroImage, vol. 29, no. 2, pages 467 – 474, 2006. 102

[Davis 2007] B.C. Davis, P.T. Fletcher, E. Bullitt and S. Joshi. *Population Shape Regression From Random Design Data.* In ICCV, pages 1–7, Oct. 2007. 91, 127

[Desikan 2009] Rahul S Desikan, Howard J Cabral, Christopher P Hess, William P Dillon, Christine M Glastonbury, Michael W Weiner, Nicholas J Schmansky, Douglas N Greve, David H Salat, Randy L Buckner, Bruce Fischl and Alzheimer's Disease Neuroimaging Initiative. *Automated MRI measures identify individuals with mild cognitive impairment and Alzheimer's disease.* Brain, vol. 132, no. Pt 8, pages 2048–2057, Aug 2009. 177

[Dice 1945] L. R. Dice. *Measures of the Amount of Ecologic Association Between Species.* Ecology, vol. 26, no. 3, pages 297–302, July 1945. 112

[Donoho 2003] David L. Donoho and Carrie Grimes. *Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data.* PNAS, vol. 100, 2003. 84, 86

[Dossal 2011] Charles Dossal, Maher Kachour, Jalal Fadili, Gabriel Peyré and Christophe Chesneau. The degrees of freedom of the Lasso for general design matrix. Previously entitled "The degrees of freedom of penalized l1 minimization", August 2011. 202

[Dupuis 1998] Paul Dupuis, Ulf Grenander and Michael I. Miller. *Variational Problems on Flows of Diffeomorphisms for Image Matching*, 1998. 162

[Durrleman 2008] Stanley Durrleman, Xavier Pennec, Alain Trouvé, Paul Thompson and Nicholas Ayache. *Inferring brain variability from diffeomorphic deformations of currents: An integrative approach.* Medical Image Analysis, vol. 12, no. 5, pages 626 – 637, 2008. Special issue on the 10th international conference on medical imaging and computer assisted intervention - MICCAI 2007. 162

[Durrleman 2009] Stanley Durrleman, Xavier Pennec, Alain Trouvé and Nicholas Ayache. *Statistical models of sets of curves and surfaces based on currents.* Medical Image Analysis, vol. 13, no. 5, pages 793 – 808, 2009. Includes Special Section on the 12th International Conference on Medical Imaging and Computer Assisted Intervention. 162

[Durrleman 2010] Stanley Durrleman. *Statistical models of currents for measuring the variability of anatomical curves, surfaces and their evolution.* Thèse de sciences (phd thesis), Université de Nice-Sophia Antipolis, March 2010. 162

[Durrleman 2013] Stanley Durrleman, Stéphanie Allassonnière and Sarang Joshi. *Sparse Adaptive Parameterization of Variability in Image Ensembles.* International Journal of Computer Vision, vol. 101, pages 161–183, 2013. 162, 197

[Dyrby 2008] Tim B. Dyrby, Egill Rostrup, William F.C. Baaré, Elisabeth C.W. van Straaten, Frederik Barkhof, Hugo Vrenken, Stefan Ropele, Reinhold Schmidt, Timo Erkinjuntti, Lars-Olof Wahlund, Leonardo Pantoni, Domenico Inzitari, Olaf B. Paulson, Lars Kai Hansen and Gunhild Waldemar. *Segmentation of age-related white matter changes in a clinical multicenter study.* NeuroImage, vol. 41, no. 2, pages 335 – 345, 2008. 102

[Efron 1986] Bradley Efron. *How Biased is the Apparent Error Rate of a Prediction Rule?* Journal of the American Statistical Association, vol. 81, no. 394, pages 461–470, 1986. 202

[Ellis 2009] Kathryn A Ellis, Ashley I Bush, David Darby, Daniela De Fazio, Jonathan Foster, Peter Hudson, Nicola T Lautenschlager, Nat Lenzo, Ralph N Martins, Paul Maruff, Colin Masters, Andrew Milner, Kerryn Pike, Christopher Rowe, Greg Savage, Cassandra Szoeke, Kevin Taddei, Victor Villemagne, Michael Woodward and David Ames. *The Australian imaging, biomarkers and lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease.* Int Psychogeriatrics, vol. 21, no. 4, pages 672–87, 2009. 50, 113, 132

[Fan 2007] Yong Fan, Dinggang Shen, Ruben C Gur, Raquel E Gur and Christos Davatzikos. *COMPARE: classification of morphological patterns using adaptive regional elements.* IEEE Trans Med Imaging, vol. 26, no. 1, pages 93–105, Jan 2007. 177

[Fan 2008a] Yong Fan, Nematollah Batmanghelich, Chris M Clark, Christos Davatzikos and Alzheimer's Disease Neuroimaging Initiative. *Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline.* Neuroimage, vol. 39, no. 4, pages 1731–1743, Feb 2008. 177

[Fan 2008b] Yong Fan, Susan M Resnick, Xiaoying Wu and Christos Davatzikos. *Structural and functional biomarkers of prodromal Alzheimer's disease: a high-dimensional pattern classification study.* Neuroimage, vol. 41, no. 2, pages 277–285, Jun 2008. 177

[Fan 2009] Mingyu Fan, Hong Qiao and Bo Zhang. *Intrinsic dimension estimation of manifolds by incising balls.* Pattern Recogn., vol. 42, no. 5, pages 780–787, 2009. 86

[Farjam 2012] Reza Farjam, Hemant A. Parmar, Douglas C. Noll, Christina I. Tsien and Yue Cao. *An approach for computer-aided detection of brain metastases in post-Gd T1-W {MRI}.* Magnetic Resonance Imaging, vol. 30, no. 6, pages 824 – 836, 2012. 102

[Fiot 2011] Jean-Baptiste Fiot, Laurent D. Cohen, Parnesh Raniga and Jurgen Fripp. *Efficient Lesion Segmentation using Support Vector Machines.* In Computational Vision and Medical Image Processing: VipIMAGE 2011, page ISBN 9780415683951, Olhão, Portugal, September 2011. 102

[Fiot 2012] Jean-Baptiste Fiot, Laurent Risser, Laurent D. Cohen, Jurgen Fripp and François-Xavier Vialard. *Local vs global descriptors of hippocampus shape evolution for Alzheimer's longitudinal population analysis.* In 2nd International MICCAI Workshop on Spatiotemporal Image Analysis for Longitudinal and Time-Series Image Data (STIA'12), volume 7570, pages 13–24, Nice, France, October 2012. 201, 202

[Fletcher 2004] P. T. Fletcher, C. Lu, M. Pizer and S. Joshi. *Principal geodesic analysis for the study of nonlinear statistics of shape.* IEEE Transactions Medical Imaging, pages 995–1005, 2004. 165, 169, 177, 182

[Frackowiak 2004] Richard S. J. Frackowiak, John T. Ashburner, William D. Penny and Semir Zeki. Human Brain Function, Second Edition. Academic Press, 2 édition, January 2004. 159

[Fréchet 1948] Maurice Fréchet. *Les éléments aléatoires de nature quelconque dans un espace distancie.* Annales de l'institut Henri Poincaré, vol. 10, no. 4, pages 215–310, 1948. 166, 182

[Fukunaga 1971] K. Fukunaga and D. R. Olsen. *An Algorithm for Finding Intrinsic Dimensionality of Data.* IEEE Trans. Comput., vol. 20, no. 2, pages 176–183, 1971. 86

[Fukunaga 1975] K. Fukunaga and L. Hostetler. *The estimation of the gradient of a density function, with applications in pattern recognition.* Information Theory, IEEE Transactions on, vol. 21, no. 1, pages 32–40, January 1975. 93

[Gerardin 2009] Emilie Gerardin, Gaël Chételat, Marie Chupin, Rémi Cuingnet, Béatrice Desgranges, Ho-Sung Kim, Marc Niethammer, Bruno Dubois, Stéphane Lehéricy, Line Garnero, Francis Eustache, Olivier Colliot and Alzheimer's Disease Neuroimaging Initiative. *Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging.* Neuroimage, vol. 47, no. 4, pages 1476–1486, Oct 2009. 177

[Gerber 2009] S Gerber, T Tasdizen, S Joshi, R Whitaker, GZ Yang, DJ Hawkes and D Rueckert. *On the Manifold Structure of the Space of Brain Images.* In MICCAI, pages 312, 305, 2009. 91

[Gerber 2010] Samuel Gerber, Tolga Tasdizen, P. Thomas Fletcher, Sarang Joshi and Ross Whitaker. *Manifold modeling for brain population analysis.* Medical Image Analysis, vol. 14, no. 5, pages 643 – 653, 2010. 18, 91, 93, 94, 127, 140, 143

[Glaunès 2005] Joan Alexis Glaunès. *Transport par difféomorphismes de points, de mesures et de courants pour la comparaison de formes et l'anatomie numérique.* PhD thesis, Université Paris 13, 2005. 162

[Gosche 2002] K M Gosche, J A Mortimer, C D Smith, W R Markesbery and D A Snowdon. *Hippocampal volume as an index of Alzheimer neuropathology: findings from the Nun Study.* Neurology, vol. 58, no. 10, pages 1476–82, 2002. 186

[Grenander 1998] Ulf Grenander and Michael I. Miller. *Computational anatomy: an emerging discipline.* Q. Appl. Math., vol. LVI, no. 4, pages 617–694, 1998. 159

[Grundman 2004] Michael Grundman, Ronald C. Petersen, Steven H. Ferris, Ronald G. Thomas, Paul S. Aisen, David A. Bennett, Norman L. Foster, Clifford R Jack Jr, Douglas R. Galasko, Rachelle Doody, Jeffrey Kaye, Mary Sano, Richard Mohs, Serge Gauthier, Hyun T. Kim, Shelia Jin, Arlan N. Schultz, Kimberly Schafer, Ruth Mulnard, Christopher H. van Dyck, Jacobo Mintzer, Edward Y. Zamrini, Deborah Cahn-Weiner, Leon J. Thal and Alzheimer's Disease Cooperative Study . *Mild cognitive impairment can be distinguished from Alzheimer disease and normal aging for clinical trials.* Arch Neurol, vol. 61, no. 1, pages 59–66, Jan 2004. 47

[Hamm 2010] J. Hamm, D.H. Ye, R. Verma and C. Davatzikos. *GRAM: A framework for geodesic registration on anatomical manifolds.* Med Image Anal, 2010. 18, 88, 89

[Heckemann 2006] Rolf A. Heckemann, Joseph V. Hajnal, Paul Aljabar, Daniel Rueckert and Alexander Hammers. *Automatic anatomical brain MRI segmentation combining label propagation and decision fusion.* NeuroImage, vol. 33, no. 1, pages 115 – 126, 2006. 90

[Hoerl 1970] Arthur E. Hoerl and Robert W. Kennard. *Ridge Regression: Biased Estimation for Nonorthogonal Problems.* Technometrics, vol. 12, no. 1, pages 55–67, February 1970. 65, 197, 198

[Holm 2008] Darryl D Holm, Alain Trouvé, Laurent Younes, Darryl D Holm, Alain Trouvé and Laurent Younes. *The Euler Poincaré theory of metamorphosis.* Quart. Appl. Math, 2008. 164

[Jack 2010] Clifford R Jack Jr, David S. Knopman, William J. Jagust, Leslie M. Shaw, Paul S. Aisen, Michael W. Weiner, Ronald C. Petersen and John Q.

Trojanowski. *Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade.* Lancet Neurol, vol. 9, no. 1, pages 119–128, Jan 2010. 17, 33, 47, 48, 70

[Jack 2013] Clifford R Jack Jr, David S. Knopman, William J. Jagust, Ronald C. Petersen, Michael W. Weiner, Paul S. Aisen, Leslie M. Shaw, Prashanthi Vemuri, Heather J. Wiste, Stephen D. Weigand, Timothy G. Lesnick, Vernon S. Pankratz, Michael C. Donohue and John Q. Trojanowski. *Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers.* Lancet Neurol, vol. 12, no. 2, pages 207–216, Feb 2013. 33, 48, 70

[Jia 2010] Hongjun Jia, Guorong Wu, Qian Wang and Dinggang Shen. *ABSORB: Atlas Building by Self-organized Registration and Bundling.* Neuroimage, vol. 51, no. 3, pages 1057–1070, Jul 2010. 169

[Jolliffe 1986] Ian T Jolliffe. Principal component analysis, volume 487. Springer-Verlag New York, 1986. 83

[Joshi 2000] S. C. Joshi and M. I. Miller. *Landmark matching via large deformation diffeomorphisms.* IEEE Trans Image Process, vol. 9, no. 8, pages 1357–1370, 2000. 162

[Joshi 2004] S. Joshi, Brad Davis, Matthieu Jomier and Guido Gerig. *Unbiased diffeomorphic atlas construction for computational anatomy.* NeuroImage, vol. 23, 2004. 168, 169

[Karcher 1977] H. Karcher. *Riemannian center of mass and mollifier smoothing.* Communications on Pure and Applied Mathematics, vol. 30, no. 5, pages 509–541, 1977. 166, 182

[Katzman 1988] Robert Katzman, Robert Terry, Richard DeTeresa, Theodore Brown, Peter Davies, Paula Fuld, Xiong Renbing and Arthur Peck. *Clinical, pathological, and neurochemical changes in dementia: A subgroup with preserved mental status and numerous neocortical plaques.* Annals of Neurology, vol. 23, no. 2, pages 138–144, 1988. 45

[Kégl 2003] B. Kégl. *Intrinsic dimension estimation using packing numbers.* In S. Becker, S. Thrun and K. Obermayer, editeurs, Adv. Neural Inform. Processing Systems, volume 15, Cambridge, MA, 2003. MIT Press. 86

[Klöppel 2008] Stefan Klöppel, Cynthia M. Stonnington, Carlton Chu, Bogdan Draganski, Rachael I. Scahill, Jonathan D. Rohrer, Nick C. Fox, Clifford R. Jack, John Ashburner and Richard S. J. Frackowiak. *Automatic classification of MR scans in Alzheimer's disease.* Brain, vol. 131, no. 3, pages 681–689, 2008. 177

[Klöppel 2011] Stefan Klöppel, Ahmed Abdulkadir, Stathis Hadjidemetriou, Sabine Issleib, Lars Frings, Thao Nguyen Thanh, Irina Mader, Stefan J. Teipel, Michael Hüll and Olaf Ronneberger. *A comparison of different automated methods for the detection of white matter lesions in MRI data.* NeuroImage, vol. 57, no. 2, pages 416–422, 2011. 102, 122

[Kruggel 2002] F. Kruggel, M. Pélégrini-Issac and H. Benali. *Estimating the effective degrees of freedom in univariate multiple regression analysis.* Medical Image Analysis, vol. 6, no. 1, pages 63–75, March 2002. 202

[Lao 2004] Zhiqiang Lao, Dinggang Shen, Zhong Xue, Bilge Karacali, Susan M Resnick and Christos Davatzikos. *Morphological classification of brains via high-dimensional shape transformations and machine learning methods.* Neuroimage, vol. 21, no. 1, pages 46–57, Jan 2004. 177

[Lao 2008] Zhiqiang Lao, Dinggang Shen, Dengfeng Liu, Abbas F. Jawad, Elias R. Melhem, Lenore J. Launer, R. Nick Bryan and Christos Davatzikos. *Computer-Assisted Segmentation of White Matter Lesions in 3D MR Images Using Support Vector Machine.* Academic Radiology, vol. 15, no. 3, pages 300 – 313, 2008. 45, 102, 103, 122

[Lawrence 2011] Neil D. Lawrence. *Spectral dimensionality reduction via maximum entropy.* In In Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, 2011. 83

[Lee 2007] John A. Lee and Michel Verleysen. Nonlinear dimensionality reduction. Springer Publishing Company, Incorporated, 2007. 83, 86

[Levina 2005] Elizaveta Levina and Peter J. Bickel. *Maximum Likelihood Estimation of Intrinsic Dimension.* In Lawrence K. Saul, Yair Weiss and Léon Bottou, editeurs, Advances in Neural Information Processing Systems 17, pages 777–784. MIT Press, Cambridge, MA, 2005. 86

[Lewis 2012] Adrian S. Lewis and Michael L. Overton. *Nonsmooth optimization via quasi-Newton methods.* Mathematical Programming, pages 1–29, 2012. 200

[Llad'o 2012] Xavier Llad'o, Arnau Oliver, Mariano Cabezas, Jordi Freixenet, Joan C. Vilanova, Ana Quiles, Laia Valls, Lluís Ramió-Torrentà and Àlex Rovira. *Segmentation of multiple sclerosis lesions in brain MRI: A review of automated approaches.* Information Sciences, vol. 186, no. 1, pages 164 – 185, 2012. 102

[Lorenzen 2005] Peter Lorenzen, Brad C. Davis and Sarang Joshi. *Unbiased Atlas Formation Via Large Deformations Metric Mapping.* In MICCAI 2005. 2005. 127

[Lorenzi 2011] M Lorenzi and X. Pennec. *Geodesics, Parallel Transport & One-parameter Subgroups for Diffeomorphic Image Registration.* In MFCA - Workshop MICCAI, 2011. 171, 178

[Lorenzi 2012a] Marco Lorenzi. *Deformation-based morphometry of the brain for the development of surrogate markers in Alzheimer's disease.* Ph.d. thesis, University of Nice Sophia Antipolis, December 2012. 20, 174

[Lorenzi 2012b] Marco Lorenzi and Xavier Pennec. *Geodesics, Parallel Transport & One-parameter Subgroups for Diffeomorphic Image Registration.* International Journal of Computer Vision, 2012. 164

[Lütkepohl 1996] H. Lütkepohl. Handbook of matrices. Wiley, 1996. 131

[Ma 2008] Jun Ma, Michael I. Miller, Alain Trouvé and Laurent Younes. *Bayesian template estimation in computational anatomy.* Neuroimage, vol. 42, no. 1, pages 252–261, Aug 2008. 168

[Ma 2009] J. Ma, M. I. Miller and L. Younes. *Bayesian Template Estimation for Surfaces by EM Algorithm.* IEEE PAMI, 2009. submitted. 177

[Macdonald 2000] Angus Macdonald and Delme Pritchard. *A mathematical model of Alzheimer's disease and the ApoE gene.* ASTIN Bulletin, vol. 30, pages 69–110, 2000. 148, 152

[Magnin 2009] Benoît Magnin, Lilia Mesrob, Serge Kinkingnéhun, Mélanie Pélégrini-Issac, Olivier Colliot, Marie Sarazin, Bruno Dubois, Stéphane Lehéricy and Habib Benali. *Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI.* Neuroradiology, vol. 51, no. 2, pages 73–83, Feb 2009. 177

[Manczak 2013] Maria Manczak and P Hemachandra Reddy. *Abnormal interaction of oligomeric amyloid-β with phosphorylated tau: implications to synaptic dysfunction and neuronal damage.* J Alzheimers Dis, vol. 36, no. 2, pages 285–295, Jan 2013. 153

[Mangin 2004] J-F. Mangin, D. Rivière, O. Coulon, C. Poupon, A. Cachia, Y. Cointepas, J-B. Poline, D. Le Bihan, J. Régis and D. Papadopoulos-Orfanos. *Coordinate-based versus structural approaches to brain image analysis.* Artif Intell Med, vol. 30, no. 2, pages 177–197, Feb 2004. 210

[Martinetz 1994] Thomas Martinetz and Klaus Schulten. *Topology representing networks.* Neural Netw., vol. 7, no. 3, pages 507–522, 1994. 86

[Mazziotta 2001] J. Mazziotta, A. Toga, A. Evans, P. Fox, J. Lancaster, K. Zilles, R. Woods, T. Paus, G. Simpson, B. Pike, C. Holmes, L. Collins, P. Thompson, D. MacDonald, M. Iacoboni, T. Schormann, K. Amunts, N. Palomero-Gallagher, S. Geyer, L. Parsons, K. Narr, N. Kabani, G. Le Goualher,

D. Boomsma, T. Cannon, R. Kawashima and B. Mazoyer. *A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM).* Philosophical transactions of the Royal Society of London. Series B, Biological sciences, vol. 356, no. 1412, pages 1293–1322, August 2001. 148

[Mechelli 2005] Andrea Mechelli, Cathy J. Price, Karl J. Friston and John Ashburner. *Voxel-Based Morphometry of the Human Brain: Methods and Applications.* Current Medical Imaging Reviews, vol. 1, pages 105–113, 2005. 159

[Melacci 2009] Stefano Melacci. *Manifold regularization: Laplacian SVM.* `http://www.dii.unisi.it/~melacci/lapsvmp/index.html`, September 2009. 114

[Melacci 2011] Stefano Melacci and Mikhail Belkin. *Laplacian Support Vector Machines Trained in the Primal.* Journal of Machine Learning Research, vol. 12, pages 1149–1184, March 2011. 114

[Michel 2011] Vincent Michel, Alexandre Gramfort, Gaël Varoquaux, Evelyn Eger and Bertrand Thirion. *Total variation regularization for fMRI-based prediction of behaviour.* IEEE Transactions on Medical Imaging, vol. 30, no. 7, pages 1328 – 1340, February 2011. 197, 198

[Miller 2004] Michael I. Miller. *Computational anatomy: Shape, growth, and atrophy comparison via diffeomorphisms.* NeuroImage, vol. 23, pages 19–33, 2004. 159

[Miller 2006] Michael I Miller, Alain Trouvé and Laurent Younes. *Geodesic Shooting for Computational Anatomy.* J Math Imaging Vis, vol. 24, no. 2, pages 209–228, Jan 2006. 179

[Modat 2010] Marc Modat, Gerard R Ridgway, Zeike A Taylor, Manja Lehmann, Josephine Barnes, David J Hawkes, Nick C Fox and Sébastien Ourselin. *Fast free-form deformation using graphics processing units.* Comput Methods Programs Biomed, vol. 98, no. 3, pages 278–284, Jun 2010. 184, 187

[Modat 2011] Marc Modat, Gerard R. Ridgway, Pankaj Daga, M. J. Cardoso, David J. Hawkes, John Ashburner and Sébastien Ourselin. *Log-Euclidean free-form deformation.* pages 79621Q–79621Q–6, 2011. 163

[Mueller 2005] Susanne G. Mueller, Michael W. Weiner, Leon J. Thal, Ronald C. Petersen, Clifford Jack, William Jagust, John Q. Trojanowski, Arthur W. Toga and Laurel Beckett. *The Alzheimer's Disease Neuroimaging Initiative.* Neuroimaging Clinics of North America, vol. 15, no. 4, pages 869 – 877, 2005. Alzheimer's Disease: 100 Years of Progress. 143, 148, 187, 200

[Niethammer 2011] Marc Niethammer, Yang Huang and François-Xavier Vialard. *Geodesic regression for image time-series.* Med Image Comput Comput Assist Interv, vol. 14, no. Pt 2, pages 655–662, 2011. 193, 210

[Ott 2008] R. Lyman Ott and Micheal T. Longnecker. An introduction to statistical methods and data analysis. Duxbury Press, 6 édition, December 2008. 112

[Ourselin 2001] S. Ourselin, A. Roche, G. Subsol, X. Pennec and N. Ayache. *Reconstructing a 3D structure from serial histological sections.* Image and Vision Computing, vol. 19, no. 1-2, pages 25 – 31, 2001. 105, 114, 148, 184, 187

[Pennec 1996] X. Pennec. *L'incertitude dans les problèmes de reconnaissance et de recalage - Application en imagerie médicale et biologie moléculaire.* PhD thesis, 1996. 165, 166

[Pennec 1999] Xavier Pennec. *Probabilities and Statistics on Riemannian Manifolds: Basic Tools for Geometric Measurements.* In A.E. Cetin, L. Akarun, A. Ertuzun, M.N. Gurcan and Y. Yardimci, editeurs, Proc. of Nonlinear Signal and Image Processing (NSIP'99), volume 1, pages 194–198, June 20-23, Antalya, Turkey, 1999. IEEE-EURASIP. 166, 182

[Pennec 2006] Xavier Pennec. *Intrinsic Statistics on Riemannian Manifolds: Basic Tools for Geometric Measurements.* J. Math. Imaging Vis., vol. 25, no. 1, pages 127–154, July 2006. 165, 166, 182

[Petersen 1999] R. C. Petersen, G. E. Smith, S. C. Waring, R. J. Ivnik, E. G. Tangalos and E. Kokmen. *Mild cognitive impairment: clinical characterization and outcome.* Arch Neurol, vol. 56, no. 3, pages 303–308, Mar 1999. 47

[Peyré 2011] Gabriel Peyré, Sébastien Bougleux and Laurent D. Cohen. *Non-local Regularization of Inverse Problems.* Inverse Problems and Imaging, vol. 5, no. 2, pages 511–530, 2011. 65

[Qiu 2008] A Qiu, L Younes, M I Miller and J G Csernansky. *Parallel transport in diffeomorphisms distinguishes the time-dependent pattern of hippocampal surface deformation due to healthy aging and the dementia of the Alzheimer's type.* NeuroImage, vol. 40, pages 68–76, 2008. 171, 177, 197

[Querbes 2009] Olivier Querbes, Florent Aubry, Jérémie Pariente, Jean-Albert Lotterie, Jean-François Démonet, Véronique Duret, Michèle Puel, Isabelle Berry, Jean-Claude Fort, Pierre Celsis and Alzheimer's Disease Neuroimaging Initiative. *Early diagnosis of Alzheimer's disease using cortical thickness: impact of cognitive reserve.* Brain, vol. 132, no. Pt 8, pages 2036–2047, Aug 2009. 177

[Raber 2004] Jacob Raber, Yadong Huang and J Wesson Ashford. *ApoE genotype accounts for the vast majority of AD risk and AD pathology.* Neurobiol Aging, vol. 25, no. 5, pages 641–650, 2004. 44

[Raguet 2011] H. Raguet, J. Fadili and G. Peyré. *Generalized Forward-Backward Splitting.* Technical report, Preprint Hal-00613637, 2011. 200

[Raniga 2008] Parnesh Raniga, Pierrick Bourgeat, Jurgen Fripp, Oscar Acosta, Victor L Villemagne, Christopher Rowe, Colin L Masters, Gareth Jones, Graeme O'Keefe, Olivier Salvado and Sébastien Ourselin. *Automated 11C-PiB standardized uptake value ratio.* Acad Radiol, vol. 15, no. 11, pages 1376–1389, Nov 2008. 128

[Risser 2011a] L. Risser, D. Holm, D. Rueckert and F.-X. Vialard. *Diffeomorphic Atlas Estimation using Karcher Mean and Geodesic Shooting on Volumetric Images.* Medical Image Understanding and Analysis (MIUA'11), 2011. 166, 167, 168, 170, 177, 182, 183

[Risser 2011b] L. Risser, F.-X. Vialard, R. Wolz, M. Murgasova, D. D. Holm and D. Rueckert. *Simultaneous Multiscale Registration using Large Deformation Diffeomorphic Metric Mapping.* IEEE Trans. Med. Imaging, 2011. 163, 180

[Rosset 2003] Saharon Rosset, Ji Zhu and Trevor Hastie. *Margin Maximizing Loss Functions.* In In Advances in Neural Information Processing Systems (NIPS) 15, page 16. MIT Press, 2003. 111

[Roweis 2000] S.T. Roweis and L.K. Saul. *Nonlinear dimensionality reduction by locally linear embedding.* Science, vol. 290, no. 5500, pages 2323–2326, 2000. 84

[Rueckert 1999] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. Hill, M. O. Leach and D. J. Hawkes. *Nonrigid registration using free-form deformations: application to breast MR images.* IEEE transactions on medical imaging, vol. 18, no. 8, pages 712–721, August 1999. 105, 160, 161

[Rueckert 2006] Daniel Rueckert, Paul Aljabar, Rolf A. Heckemann, Joseph V. Hajnal and Alexander Hammers. *Diffeomorphic Registration Using B-Splines.* In Rasmus Larsen, Mads Nielsen and Jon Sporring, editeurs, MICCAI (2), volume 4191 of *Lecture Notes in Computer Science*, pages 702–709. Springer, 2006. 161

[Sabuncu 2009] Mert R Sabuncu, Serdar K Balci, Martha E Shenton and Polina Golland. *Image-driven population analysis through mixture modeling.* IEEE Trans Med Imaging, vol. 28, no. 9, pages 1473–1487, Sep 2009. 93, 127

[Sajja 2006] Balasrinivasa Sajja, Sushmita Datta, Renjie He, Meghana Mehta, Rakesh Gupta, Jerry Wolinsky and Ponnada Narayana. *Unified Approach for Multiple Sclerosis Lesion Segmentation on Brain MRI.* Annals of Biomedical Engineering, vol. 34, pages 142–151, 2006. 10.1007/s10439-005-9009-0. 102

[Salvado 2006] O. Salvado, C. Hillenbrand, S. Zhang and D.L. Wilson. *Method to Correct Intensity Inhomogeneity in MR Images for Atherosclerosis Characterization.* Medical Imaging, IEEE Transactions on, vol. 25, pages 539–552, 2006. 114

[Samaille 2012] Thomas Samaille, Ludovic Fillon, Rémi Cuingnet, Eric Jouvent, Hugues Chabriat, Didier Dormont, Olivier Colliot and Marie Chupin. *Contrast-based fully automatic segmentation of white matter hyperintensities: method and validation.* PLoS One, vol. 7, no. 11, page e48953, 2012. 102, 122

[Schölkopf 1996] Bernhard Schölkopf, Alexander Smola and Klaus-Robert Müller. *Nonlinear Component Analysis as a Kernel Eigenvalue Problem.* Technical Report 44, December 1996. 83

[Schölkopf 2001] Bernhard Schölkopf and Alexander J. Smola. Learning with kernels: Support vector machines, regularization, optimization, and beyond (adaptive computation and machine learning). MIT Press, 2001. 106, 111, 162, 186

[Scully 2010] Mark Scully, Blake Anderson, Terran Lane, Charles Gasparovic, Vince Magnotta, Wilmer Sibbitt, Carlos Roldan, Ron Kikinis and Henry Jeremy Bockholt. *An automated method for segmenting white matter lesions through multi-level morphometric feature classification with application to lupus.* Frontiers in Human Neuroscience, vol. 4, no. 0, 2010. 102

[Seghers 2004] Dieter Seghers, Emiliano D'Agostino, Frederik Maes, Dirk Vandermeulen and Paul Suetens. *Construction of a Brain Template from MR Images Using State-of-the-Art Registration and Segmentation Techniques.* In Christian Barillot, DavidR. Haynor and Pierre Hellier, editeurs, Medical Image Computing and Computer-Assisted Intervention - MICCAI 2004, volume 3216 of *Lecture Notes in Computer Science*, pages 696–703. Springer Berlin Heidelberg, 2004. 169

[Singh 2010] Nikhil Singh, P. Fletcher, J. Preston, Linh Ha, Richard King, J. Marron, Michael Wiener and Sarang Joshi. *Multivariate Statistical Analysis of Deformation Momenta Relating Anatomical Shape to Neuropsychological Measures.* In Tianzi Jiang, Nassir Navab, Josien Pluim and Max Viergever, editeurs, Medical Image Computing and Computer-Assisted Intervention - MICCAI 2010, volume 6363 of *Lecture Notes in Computer Science*, pages 529–537. Springer Berlin / Heidelberg, 2010. 177

[Sommer 2011] Stefan Sommer, Mads Nielsen, François Lauze and Xavier Pennec. *A multi-scale kernel bundle for LDDMM: towards sparse deformation description across space and scales.* Inf Process Med Imaging, vol. 22, pages 624–635, 2011. 163

[Sommer 2013] Stefan Sommer, François Lauze, Mads Nielsen and Xavier Pennec. *Sparse Multi-Scale Diffeomorphic Registration: the Kernel Bundle Framework.* Journal of Mathematical Imaging and Vision, vol. 46, no. 3, pages 292–308, 2013. In press. 162, 163

[Stern 2012] Yaakov Stern. *Cognitive reserve in ageing and Alzheimer's disease.* Lancet Neurol, vol. 11, no. 11, pages 1006–1012, Nov 2012. 45

[Styner 2008] M. Styner, J. Lee, B. Chin, M.S. Chin, O. Commowick, H. Tran, S. Markovic-Plese, V. Jewells and S.K. Warfield. *3D Segmentation in the Clinic: A Grand Challenge II: MS Lesion Segmentation.* In MIDAS Journal, Special Issue on 2008 MICCAI Workshop - MS Lesion Segmentation, pages 1–5, sep 2008. 102

[Tenenbaum 2000] J. B. Tenenbaum, V. Silva and J. C. Langford. *A Global Geometric Framework for Nonlinear Dimensionality Reduction.* Science, vol. 290, no. 5500, pages 2319–2323, 2000. 84, 88

[Thompson 1917] D.A.W. Thompson. On growth and form. C.U.P., 1917. 157, 158, 159

[Thompson 2000] Paul Thompson, Dr. Paul Thompson, Michael S. Mega, Arthur W. Toga, Paul Thompson Phd, Paul Thompson Phd, Roger P. Woods, Roger P. Woods Md, Michael S. Mega Md Phd, Michael S. Mega Md Phd, Arthur W. Toga Phd and Arthur W. Toga Phd. *Mathematical/Computational Challenges in Creating Deformable and Probabilistic Atlases of the Human Brain.* Human Brain Mapping, vol. 9, pages 81–92, 2000. 165

[Tibshirani 1994] Robert Tibshirani. *Regression Shrinkage and Selection Via the Lasso.* Journal of the Royal Statistical Society, Series B, vol. 58, pages 267–288, 1994. 65, 197, 198

[Tikhonov 1943] A.N. Tikhonov. *On the stability of inverse problems.* Doklady Akademii nauk SSSR, vol. 39, no. 5, pages 195–198, 1943. 36, 65

[Tikhonov 1963] Andrey Nikolayevich Tikhonov. *Solution of Incorrectly Formulated Problems and the Regularization Method.* Soviet Math. Dokl., vol. 5, page 1035/1038, 1963. 36, 65

[Tikhonov 1977] A.N. Tikhonov and V.Y. Arsenin. Solutions of ill-posed problems. W.H. Winston, Washington, D.C., 1977. 36, 65

[Trouvé 2005] Alain Trouvé and Laurent Younes. *Metamorphoses Through Lie Group Action.* Foundations of Computational Mathematics, vol. 5, no. 2, pages 173–198, 2005. 164

[Trouvé 2012] Alain Trouvé and Francçois-Xavier Vialard. *Shape splines and stochastic shape evolutions: A second order point of view.* Quarterly of Applied Mathematics, 2012. 193, 210

[Tukey 1962] John W. Tukey. *The Future of Data Analysis.* Annals of Mathematical Statistics, vol. 33, no. 1, pages 1–67, March 1962. 34, 71

[Vaillant 2004] M. Vaillant, M. I. Miller, L. Younes and A. Trouvé. *Statistics on diffeomorphisms via tangent space representations.* Neuroimage, vol. 23 Suppl 1, pages S161–S169, 2004. 177

[Vaillant 2005] Marc Vaillant and Joan Glaunès. *Surface matching via currents.* In Proceedings of Information Processing in Medical Imaging (IPMI 2005), number 3565 in Lecture Notes in Computer Science, pages 381–392, 2005. 162

[Vaiter 2013] S. Vaiter, C. Deledalle, G. Peyré, J. Fadili and C. Dossal. *The degrees of freedom of the Group Lasso for a General Design.* In Proc. SPARS'13, 2013. 202

[van der Maaten 2007] L.J.P. van der Maaten, E.O. Postma and H.J. van den Herik. *Dimensionality reduction: A comparative review.* vol. 10, no. February, pages 1–35, 2007. 83, 127

[Van Leemput 2009] Koen Van Leemput. *Encoding probabilistic brain atlases using Bayesian inference.* IEEE Trans Med Imaging, vol. 28, no. 6, pages 822–837, Jun 2009. 90

[Vapnik 1995] Vladimir N. Vapnik. The nature of statistical learning theory. Springer-Verlag New York, Inc., New York, NY, USA, 1995. 56

[Vemuri 2008] Prashanthi Vemuri, Jeffrey L. Gunter, Matthew L. Senjem, Jennifer L. Whitwell, Kejal Kantarci, David S. Knopman, Bradley F. Boeve, Ronald C. Petersen and Clifford R. Jack Jr. *Alzheimer's disease diagnosis in individual subjects using structural MR images: Validation studies.* NeuroImage, vol. 39, no. 3, pages 1186 – 1197, 2008. 177

[Vercauteren 2008] Tom Vercauteren, Xavier Pennec, Aymeric Perchant and Nicholas Ayache. *Symmetric log-domain diffeomorphic Registration: a demons-based approach.* Med Image Comput Comput Assist Interv, vol. 11, no. Pt 1, pages 754–761, 2008. 163

[Vialard 2012] François-Xavier Vialard, Laurent Risser, Daniel Rueckert and Colin J. Cotter. *Diffeomorphic 3D Image Registration via Geodesic Shooting Using an Efficient Adjoint Calculation.* Int. J. Comput. Vision, vol. 97, no. 2, pages 229–241, April 2012. 167, 177, 180, 187

[Wang 2007] Lei Wang, Faisal Beg, Tilak Ratnanather, Can Ceritoglu, Laurent Younes, John C Morris, John G Csernansky and Michael I Miller. *Large deformation diffeomorphism and momentum based hippocampal shape discrimination in dementia of the Alzheimer type.* IEEE Trans Med Imaging, vol. 26, no. 4, pages 462–470, Apr 2007. 177

[Wei 2008] Jia Wei, Hong Peng, Yi-Shen Lin, Zhi-Mao Huang and Jia-Bing Wang. *Adaptive neighborhood selection for manifold learning.* volume 1, pages 380 –384, july 2008. 85

[Wittman 2005] Todd Wittman. *MANIfold Learning Matlab Demo.* `http://www.math.ucla.edu/~wittman/mani`, 2005. Version 2.5. 132

[Wolz 2009] Robin Wolz, Paul Aljabar, Joseph V. Hajnal, Alexander Hammers, Daniel Rueckert and the Alzheimer's Disease Neuroimaging Initiative. *LEAP: learning embeddings for atlas propagation.* NeuroImage, 2009. 90, 127

[Wolz 2011] Robin Wolz, Paul Aljabar, Joseph V. Hajnal, Jyrki Lötjönen and Daniel Rueckert. *Manifold learning combining imaging with non-imaging information.* In ISBI, pages 1637–1640, 2011. 23, 143, 144, 148, 149

[Younes 2007] L. Younes. *Jacobi fields in groups of diffeomorphisms and applications.* Quart. Appl. Math., vol. 65, pages 113–134, 2007. 171, 177

[Younes 2008] Laurent Younes, Anqi Qiu, Raimond L. Winslow and Michael I. Miller. *Transport of Relational Structures in Groups of Diffeomorphisms.* J Math Imaging Vis, vol. 32, no. 1, pages 41–56, Sep 2008. 171

[Younes 2009] Laurent Younes, Felipe Arrate and Michael I. Miller. *Evolutions equations in computational anatomy.* Neuroimage, vol. 45, no. 1 Suppl, pages S40–S50, Mar 2009. 171

[Zacharaki 2008] Evangelia I. Zacharaki, Stathis Kanterakis, R. Nick Bryan and Christos Davatzikos. *Measuring Brain Lesion Progression with a Supervised Tissue Classification System.* In Proceedings of MICCAI 2008, pages 620–627, 2008. 102

[Zhan 2009] Yubin Zhan, Jianping Yin, Xinwang Liu and Guomin Zhang. *Adaptive Neighborhood Select Based on Local Linearity for Nonlinear Dimensionality Reduction.* In Zhihua Cai, Zhenhua Li, Zhuo Kang and Yong Liu, editeurs, ISICA, volume 5821 of *Lecture Notes in Computer Science*, pages 337–348. Springer, 2009. 85

[Zhang 2003] Zhenyue Zhang and Hongyuan Zha. *Nonlinear Dimension Reduction via Local Tangent Space Alignment.* In Intelligent Data Engineering and Auto. Learning. 2003. 85

[Zhou 2004] Dengyong Zhou, Jason Weston, Arthur Gretton, Olivier Bousquet and Bernhard Schölkopf. *Ranking on Data Manifolds.* In Sebastian Thrun, Lawrence Saul and Bernhard Schölkopf, editeurs, Advances in Neural Information Processing Systems 16. MIT Press, Cambridge, MA, 2004. 85

[Zou 2005] Hui Zou and Trevor Hastie. *Regularization and variable selection via the Elastic Net*. Journal of the Royal Statistical Society, Series B, vol. 67, pages 301–320, 2005. 65, 197, 198

[Zou 2007] Hui Zou, Trevor Hastie and Robert Tibshirani. *On the "degrees of freedom" of the lasso*. The Annals of Statistics, vol. 35, no. 5, pages 2173–2192, October 2007. 202

Index