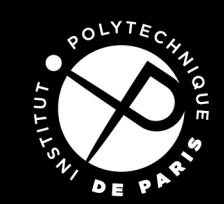
Learning Multimodal Digital Models of Disease Progression from Longitudinal Data : Methods & Algorithms for the Description, Prediction and Simulation of Alzheimer's Disease Progression

Igor Koval



INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2020IPPAX008

Thèse de doctorat

Learning Multimodal Digital Models of Disease Progression from Longitudinal Data:

Methods & Algorithms for the Description, Prediction and Simulation of Alzheimer's Disease Progression.

Thése de doctorat de l'Institut Polytechnique de Paris
préparée á Ecole Polytechnique

Ecole doctorale n°574 Ecole Doctorale de Mathématiques Hadamard (EDMH)
Spécialité de doctorat: Mathématiques Appliquées

Thèse présentée et soutenue à Paris, France, le 23 Janvier 2020, par

IGOR KOVAL

Composition du Jury :

Daniel Alexander Professeur, University College London	Rapporteur
Sach Mukherjee Professeur associé, DZNE	Rapporteur
Martin Hofmann-Apitius Professeur, Fraunhofer SCAI	Examinateur
Erwan Le Pennec Professeur, Ecole Polytechnique	Président
Stéphanie Allassnnière Professeur, Université Paris-Descartes	Co-directrice de thèse
Stanley Durrleman Directeur de Recherche, Inria	Co-directeur de thèse
Stéphane Epelbaum Neurologiste, AP-HP	Invité

Abstract

This thesis focuses on the statistical learning of digital models of neurodegenerative disease progression, especially Alzheimer's disease. It aims at reconstructing the complex and heterogeneous dynamic of evolution of the structure, the functions and the cognitive abilities of the brain, at both an average and individual level. To do so, we consider a mixed-effects model that, based on longitudinal data, namely repeated observations per subjects that present multiple modalities, in parallel recombines the individual spatiotemporal trajectories into a group-average scenario of change, and, estimates the variability of this characteristic progression which characterizes the individual trajectories. This variability results from a temporal un-alignment (in term of pace of progression and age at disease onset) along with a spatial variability that takes the form of a modification in the sequence of events that appear during the course of the disease. The 5 parts of this thesis corresponds to different aspect and features that extensively enrich the initial statistical model in order to convert it into a natural framework for the study of disease progression.

The first part of the manuscript aims at presenting the generative mixed-effects model that enables the estimation of the long-term progression of the disease and to reconstruct the individual trajectory, in the case of multivariate data. It offers a generic way to handle individual spatiotemporal trajectories that present a natural variability between patients. This variability results from a temporal un-alignment (different pace of progression and a temporal offset) along with a spatial variability that takes the form of a modification in the sequence of events that appear during the course of the disease.

The second part expands the scope of the model in order to handle data that have a spatial structure, such as images, meshes and networks. It introduces a technique to take advantage of the spatial coherence of evolution for *close* regions. It is validated on the estimation of the cortical thickness and glucose consumption evolution during the course of Alzheimer's disease.

The third part is an extensive study of the complex progression of the function (FDG-PET), the structure (cortical thickness and hippocampus meshes) and cognitive abilities (ADAS-Cog and MMSE) during the course of Alzheimer's disease. It validates the group-average multi-modal progression, evaluated by the reconstructing of individual trajectories to the noise level. The analysis of the factors modulating the evolution enables to describe the interactions between heterogeneous modalities. Furthermore, it allows to predict individual measurements up to 4 years in advance.

The fourth chapter takes advantage of the generative and mixed-effects nature of the model. It offers the possibility to first reconstruct a continuous disease timeline at the individual level and also to simulate virtual patients entirely. The former allows to impute missing values or predicting future time-points. The latter enables to simulate virtual patients that either un-bias and balance the real cohort or to augment the initial dataset in order to improve the predictive power of algorithms that requires large amount od data. It is used to reach state-of-the-art results on future stages 3 and 4 years in advance.

The fifth chapter describes the software tools that were developed along the way. They were designed to benefit to mathematical researchers that aims to develop similar models or estimation algorithms, while being sufficiently user-friendly to be used by the medical community for other diseases or even in real-life disease diagnosis and prognosis.

To conclude, this thesis introduces a general framework to grasp the complexity of the disease progression in inter-dependant heterogeneous modalities. Overall, this advanced understanding enables to characterize individual evolutions, simulate virtual cohorts and predict future disease stages for various modalities. While it focuses on the study of Alzheimer's disease solely, current works on Parkinson's disease, Huntington's disease and normal ageing highlights its capacity to generalize to other neurodegenerative diseases.

Remerciements

Il est des expériences de vie dont on ne ressort grandit que si l'on en vient à bout seul. A l'inverse, certaines acquièrent leur saveur au contact des lieux, des discussions et des personnes qui les ont vues naître et grandir. Cette thèse en a été une illustration manifeste ; elle est l'aboutissement d'une initiation intellectuelle unique et inattendue, dont le parcours a été embelli par des rencontres inestimables, et soutenu par un soutien amical et familial indéfectibles.

Mes remerciements s'adressent d'abord à Sach Mukherjee et Daniel Alexander pour avoir accepter de relire mon manuscript de thèse. Leurs rapports, traduisant l'attention qu'ils ont su porter à ce travail, ont été précieux pour en mettre en valeur certains aspects. Aussi, j'adresse toute ma gratitude à Martin Hofmann-Apitius et Erwan Le Pennec que j'ai eu la joie de compter parmi les membres de mon jury.

Ensuite, ce sont à mes directeurs de thèse, Stéphanie Allassonnière et Stanley Durrleman, que sont dédiés ces mots. Mes pensées pour eux ne sauraient décrire la joie que j'ai eu à travailler sous leur tutelle. Sans eux, ce travail n'aurait que peu des qualités, s'il en est, qui le traversent aujourd'hui : scientifiques, rédactionelles, pédagogiques, ... Leurs conseils ont fait de l'exercice que constitue une thèse, un travail passionnant et passionné.

Au delà de mes directeurs de thèse, Aramis a été un laboratoire scientifique extraordinaire. Ce lieu, ironiquement situé entre la morgue et la maternité, a été bien plus qu'un lieu de travail ; un lieu de vie, de découverte et d'épanouissement. Olivier Colliot, directeur du laboratoire, bienveillant et sincère, toujours disposé à livrer son avis éclairé, sans qui cette équipe ne serait telle qu'elle est. Emmanuelle, dont la gentillesse a apporté un brin d'humanité et de douceur au milieu de tous ces scientifiques. Ninon et Fabrizio, dont les récentes récompenses prouvent leurs grandes qualités scientifiques, mais aussi humaines. Stéphane pour qui j'ai le plus profond respect ; professionnel, d'abord, des consultations qui m'ont fait entrevoir toute l'empathie et l'altruisme du docteur, mais personnel, surtout. Enfin, une pensée pour Anne, dont le sourire nous manque toujours.

S'ajoute à ces membres, les doctorants et ingénieurs, qui, bien que de passage, auront écrit une page de l'histoire d'Aramis et de mon séjour en son sein. Mes aînés d'abord. Jean-Baptiste qui a su me transmettre ses travaux avec beaucoup de bienveillance. Hao, Catalina, Jérémy, Jorge et Alexandre R., qui en plus de nous montrer la voie scientifique, ont été des collègues et amis précieux, et pour certains, des danseurs hors pair. Mes contemporains ensuite, grâce a qui cette thèse a été un travail exaltant et collectif. Manon, rigoureuse, dédiée à son travail et ses principes, inspirante par son dévouement. Maxime, qui, en plus d'être un compagnon scientifique - et de cordée - a surtout été un complice intellectuel sur tous les sujets sociaux et sociétaux. Alexandre B., qui aura été le trublion de l'équipe, l'infatigable animateur de nos discussions et de nos pauses, à EuroPOND comme ailleurs. Juliette me fascinant par son éternel intérêt pour les convergences théoriques - plus que pratiques. Enfin, à la plus jeune génération. Celle qui me rappelle qu'en trois ans à peine, mon statut d'ancien du labo a à jamais remplacé celui de jeune doctorant. Tiziana, pour représenter haut et fort les couleurs de l'Italie et du sourire romain, inébranlable, du matin au bout de la nuit. Raphael, avec qui les sorties n'ont jamais été aussi étonnantes et sensationnelles - au propre comme au figuré. Simona, d'abord pour l'irrésistible plaisir qu'elle aura à ouvrir ces pages pour y trouver son nom. Mais par dessus tout, sa bonne humeur, surtout lorsqu'elle s'exprime par une tâche de café, de dentifrice ou de ketchup. Et pour toutes les discussions de canapé, de bar et à travers les vitres que nous avons eues et auront encore. Arnaud M. pour toutes les sensations fortes que nous avons pu partager, de jour comme de nuit.

Écrire ces lignes me rapelle que ces mots sont adressés à des amis avant tout. Cette thèse et la vie qui l'a accompagnée m'apparaissent maintenant comme des moments de

vie exceptionnels, sans précédents. Principalement liés aux personnes qui m'entourent, et dont j'essaye de tirer le meilleur enseignement, par nos discussions animés, leurs opinions et nos aventures: Paris, Québec, Florence, Barcelone, Londres, Rotterdam, ...

Le soutien de mes amis a également été précieux pendant ces années de thèse. De mes années d'études aux Ponts me reste des amis chers, et dont les réussites personnelles et professionnelles ont été inspirantes pour mes propres décisions. S'y ajoute le *groupe des prépas* ou des troubadours comme il me plaisait de les appeler. Ce quolibet fut en réalité la raison de mon attachement à eux, puisqu'ils se distinguent par leurs métiers, éloignés du mien, qui me rappelle la pluralité des avis, des opinions et des visions du monde. Enfin mes amis les plus anciens, *du lycée* comme nous avons coutumes de nous appeler, bien que la plupart de ces amitiés remontent à nos balbutiements de collégiens. Les remercier tous serait tâche déraisonnée. Mais les liens qui nous rassemblent sont d'autant plus fascinants qu'ils tiennent à nos différences, tant professionnelle, qu'idéologique, culturelle, ...

Enfin, aucune des étapes de ma vie, à l'aune de ces remerciements, n'aurait existé sans les efforts, l'assistance et les conseils déterminants que m'ont apporté ma famille. Cathy et Pascal, qui ont grandement participé à mon éducation, m'ont conseillé infailliblement. A Monique, qui n'a jamais lésiné sur les efforts. Pour me faire réciter mes leçons, dès le début. Pour inspirer et motiver mes décisions scolaires. Et pour toutes ces choses qu'elle m'a toujours affectueusement apportées. Et enfin, à mes parents, Irena et Oleg. Leurs efforts, leur invariable soutien, apportés pendant ma toute ma scolarité, sans jamais m'en faire sentir le poids, pour m'accompagner coute que coute. J'espère que la fierté qu'ils tireront de mon travail saura prouver l'amour que je leur porte. S'y ajoute une pensée particulière à ma famille en Ukraine, dont j'espère que cette thèse aura rendu fier, malgré la distance qui nous sépare.

Enfin, ce travail ne serait ce qu'il est sans celle qui accompagne mes joies et mes peines depuis tant d'années. Bien qu'aucun mot ne suffise à traduire mon absolue reconnaissance à son égard, elle a été la source, et l'objet, de mon épanouissement. Cette thèse, fruit de son soutien et de sa patience, lui est entièrement dédiée.
À mon amour, indéfectible, unique et sincère. À Louise.

Contents

Résumé en Français

Motivation

La progression des maladies neurodégénératives dépend de phénomènes biologiques complexes qui restent mal compris, d'autant qu'ils se mettent en place sur des périodes de temps longues. De ce fait, décrire un scénario typique de l'évolution de la maladie est un enjeu majeur puisqu'il permettrait de mettre en lumière les dynamiques temporelles de différents biomarqueurs comme les tests neuropsychologiques, l'imagerie médicale ou les mesures physiologiques. Cependant, la description d'un scénario *moyen* de progression est confrontée à l'expression variable de la maladie à travers les patients, variabilité qui se traduit, par exemple, par des âges de diagnostics, des vitesses d'évolutions, des séquences et intensités d'événements divers.

Au vu de ces éléments, cette thèse s'emploie à d'écrire l'évolution typique de la maladie à l'échelle de la population, ce qui nécessite une caractérisation fine de la dynamique temporelle de différentes modalités. Au dela de cette description *moyenne*, le travail entrepris tend à décrire la progression de la maladie à l'échelle individuelle, afin de (i) la comparer à l'évolution typique, (ii) prédire l'évolution future et (iii) analyser les cofacteurs à l'origine de cette variabilité, comme le sexe, les mutations génétiques ou des facteurs environnementaux. Ces analyses ne sont rendues possibles que grâce à une définition claire de la variabilité spatiotemporelle de l'évolution de la maladie.

Néanmoins, la progression des maladies neurodégénératives comprennent des spécificités qui en rendent la description plus complexe que d'autres processus temporels. D'abord, bien que décrivrant un processus similaire chez tous les patients, son expression présente des caractéristiques individuelles propres, notamment l'absence d'alignement temporel entre les patients. Par exemple, la vitesse de progression de la maladie et l'âge au diagnostic sont suceptibles de différer d'un individu à l'autre. Typiquement, deux personnes du même âge peuvent présenter des stades d'avancement différents. A l'inverse, le même stade d'avancement peut apparaître à des âges différents selon les patients. Pour ces raisons, l'âge réel n'est pas un indicateur précis d'un âge *physiologique* qui correspondrait au stade de la maladie. Et déterminer ce dernier n'est pas aisé puisqu'il présuppose de désenchevêtrer l'impact de la maladie des caractéristiques naturelles des patients : les capacités cognitives, qui varient naturellement d'un individu à l'autre, en sont une illustration concrète. Ainsi, il est nécessaire de comparer correctement les évolutions individuelles les unes aux autres, et, potentiellement, à un scenario de référence. Malheureusement, définir ce scenario normatif est un défi puisqu'il demande de reconstruire une trajectoire sur des périodes de temps longues - plus longues que n'importe quelles mesures individuelles.

Toutes ces caractéristiques sont partagées par une majorité des maladies neurodégénératives. La maladie d'Alzheimer nous livre l'exemple d'une telle dynamique temporelle, où les interactions entre la structure et les fonctions sont loin d'être parfaitement comprises, tout autant que leurs impacts sur les fonctions cognitives. Les phases précoces de la maladie sont caractérisées par des dépôts de plaques de protéines dans le cerveau, suivies par une modification de sa structure, conséquence d'une importante mort neuronale qui présente elle-même une dynamique propre. Les symptômes cliniques n'apparaissent que quelques années après, causant un dépistage de la maladie à des phases tardives où les fonctions et la structure du cerveau ont été modifiées de manière irréversible. De fait, il est devenu critique de déterminer les marqueurs précoces de la maladie.

Dans ce contexte, de nombreux modèles statistiques ont été développés pour rendre compte de l'évolution de différents biomarqueurs au cours de la progression de la maladie, à l'échelle de la population d'abord, puis des individus. L'un d'eux, introduit

dans [Schiratti et al., 2015], a permis de caractériser l'évolution moyenne de biomarqueurs scalaires au sein d'une cohorte, tout en interprétant chaque trajectoire individuelle comme la modification de l'évolution moyenne grâce à un nombre réduit de paramètres. La présente thèse étend le domaine d'application de ce modèle en introduisant des modèles de progression plus complexes, notamment pour des données d'imagerie médicale. D'autre part, la thèse présente des procédures mathématiques nécessaires à l'estimation des trajectoires individuelles, rendant possible l'imputation de données manquantes et la prédiction de variables dans le futur. Enfin, le modèle est utilisé pour générer des données longitudinales virtuelles pour lesquelles on montre qu'elles peuvent se substituer à des données réelles, et, être utilisées par des prédicteurs qui requièrent d'importants volumes de données.

Ces différentes contributions sont évaluées sur la capacité du modèle à décrire l'évolution de la maladie d'Alzheimer pour des scores cognitifs, l'épaisseur corticale, l'hypométabolisme et le maillage des hippocampes gauche et droit. A cette évolution typique s'ajoute l'analyse des cofacteurs (sexe, facteurs environnementaux, mutations génétiques) qui modifient cette évolution. Aussi, ce travail s'attache sur la capacité du modèle à prédire l'évolution future de patients atteints de troubles cognitifs précoces qui, dans certains cas, aboutissent à la maladie d'Alzheimer.

L'ensemble des modèles et algorithmes introduits dans cette thèse ont été regroupées dans le package Python `Leaspy`, permettant de conduire des analyses similaires sur d'autres cohortes, modalités et maladies neurodégénératives.

Présentation des parties

La Partie I est une introduction au modèle spatiotemporel de progression des maladies neurodégénératives. Ce modèle présente la volonté, d'abord, de décrire l'évolution moyenne d'une population, puis, d'estimer la variabilité spatiotemporelle de cette évolution dans la cohorte, et, enfin, de retracer l'histoire individuelle de la maladie à n'importe quel âge, de manière à imputer des données manquantes et prédire des valeurs futures. Le début de cette Partie s'attarde sur les notions fondamentales de géométrie riemannienne et d'estimations statistiques pour les modèles bayésiens à effets-mixtes, nécessaires à la compréhension générale du modèle. Ces notions permettent de construire un modèle spatio-temporel générique de progression des maladies neurodégénératives. S'ensuivent les instantiations de ce modèle pour la description de la progression de biomarqueurs qui présentent des profils temporels linéaires, logistiques ou exponentiels. A la suite de la description géométrique du modèle, la seconde partie pourvoie le lecteur des outils indispensables aux procédures mathématiques suivantes : la *calibration* du modèle, la *personnalisation* aux données d'un nouveau patient, et, la *simulation* de données virtuelles - ou synthétiques. La calibration permet d'estimer entièrement l'évolution typique de la maladie sur des périodes de temps longues. Elle repose sur l'algorithme *Monte Carlo Markov Chain Stochastic Approximation Expectation Maximization*, une version stochastique de l'algorithme Expectation-Maximization, où l'échantillonage des variables latentes est réalisé à l'aide d'une méthode de Monte Carlo par chaînes de Markov. La personnalisation, quant à elle, correspond à l'estimation des paramètres individuels qui décrivent l'évolution des variables d'un sujet. Cette étape rend possible l'imputation de données manquantes et la prédiction des variables dans le futur. Enfin, la simulation est un moyen de synthétiser des patients virtuels qui reproduisent les caractéristiques des patients de la cohorte réelle. Ces patients virtuels peuvent être échantillonés sur une période de temps et avec un interval entre visites arbitraires.

Tandis que la Partie I s'attarde à décrire la progression de variables scalaires, la Partie II s'intéresse à l'extension du précédent modèle pour des données qui présentent une structure

spatiale, au sens où certaines variables correspondent à l'évolution d'un même biomarqueur en des régions proches du cerveau. Cette structure se retrouve dans les images où les pixels voisins présentent a priori des caractéritiques similaires, mais également dans les atlas, les maillages, ... et toute donnée qui peut être représentée par un graphe dont chaque noeud intègre la progression au cours du temps de la valeur étudiée. Ce modèle est utilisé pour estimer l'évolution de l'épaisseur corticale au cours du temps, pour des patients qui développent la maladie d'Alzheimer. Dans un second temps, ce modèle est appliqué à l'estimation de la progression de traceurs radioactifs issus de PET scans, projetés sur la surface corticale.

La Partie III est une étude approfondie du développement de la maladie d'Alzheimer, depuis les stades précoces jusqu'aux phases avancées. Elle reprend les modèles et algorithmes introduits dans les parties précédentes afin de les appliquer à différents types de données. De l'imagerie par résonnance magnétique (IRM), on extrait le maillage des hippocampes gauche et droit, et, l'épaisseur corticale en près de 3500 régions uniformément distribués sur la surface du cerveau. Aux IRM s'ajoutent les PET scans dont le rôle est de décrire l'évolution de la consommation de glucose dans le cerveau, projeté sur 120 régions du cerveau. Enfin sont ajoutés cinq scores cognitifs dont l'évolution est une manifestation de la maladie : perte de mémoire, des capacités de concentration, puis de la praxis et du language. Dans cette étude, il est montré l'évolution conjointe de ces modalités, illustrée sur le site `www.digital-brain.org`, ainsi que des cofacteurs (sexe, facteurs génétiques et environnementaux) qui modulent cette évolution. Enfin, une estimation des trajectoires individuelles permet de montrer l'intérêt de ce modèle pour, d'une part, reconstruire les données au niveau du bruit de mesure, et d'autre, part, de prédire leurs valeurs (scores cognitifs, volume de l'hippocampe et épaisseur corticale) jusqu'à quatre ans en avance.

La Partie IV tire profit de la capacité du modèle à décrire les évolutions individuelles des sujets en évaluant la variabilité spatiotemporelle de la progression. Il est alors possible d'imputer des données manquantes et de prédire l'avancement de la maladie. D'autre part, le caractère génératif du modèle permet de simuler des patients virtuels à des stades différents de la maladie, avec un nombre de visite quelconque, et un échantillonnage temporel arbitraire. La qualité de ces patients virtuels est confirmée par l'impossibilité, pour un réseau de neurones adversarial, de distinguer des données réelles de données simulées. Enfin, ces dernières sont utilisées pour renforcer le pouvoir prédictif de réseaux de neurones récurrents, amélioraant la prédiction de certains scores cognitifs 3 et 4 ans à l'avance.

La Partie V reflète l'ensemble des développement logiciels produits au cours de la thèse. Ceux-ci incluent le package Python `Leaspy` qui permet d'utiliser les modèles et algorithmes précédents dans le cadre d'autres cohortes, modalités et maladies neurodégénératives. Ce package a pour vocation de simplifier l'analyse de données longitudinales dans la recherche médicale, mais également d'être suffisament souple et structuré pour permettre l'implémentation de nouveaux modèles de progression et d'algorithmes d'estimation. En plus de cette librairie, ce sont des outils de visualisation et d'aide à la décision qui sont développés. On citera, parmi eux, des développements sur navigateur qui permettent, à partir de données individuelles, d'établir l'évolution future du patient pour un radiologue ou un neurologue.

Introduction

Motivation

Numerous phenomenon, such as people settlements, virus spreading or climatic evolutions, are governed by temporal interplays that make their description and comprehension challenging for the scientific community. Among them, the progression of neurodegenerative diseases remains poorly understood due to complex interactions between multiple biomarkers that evolve through long periods of time. The description of a generic scenario of change is further hampered by the diversity and variability of individual evolutions in term of onset, pace, intensity and sequence of events. This consequently prevents from an accurate characterization of the individual disease progression. For these reasons, to overcome the lack of knowledge during the progression of neurodegenerative diseases, there has been a large interest over the past decades to model the disease progression, and its consequences on different modalities (e.g. cognitive assessments, medical imaging, physiological measurements), at both a population and individual level.

Arising in this historical context, the thesis aims at properly describing the typical history of long-term disease progression which inevitably implies to characterize the complex temporal dynamics of inter-dependant modalities. On top of this average description, the work aspires to personalize this representative scenario of change to individual progressions in order to compare them to the mean, to study the influence of cofactors (e.g. gender, genetic mutations, environmental factors) and to enable a proper prediction of current and future time-points. Such a personalization pushes towards the definition and estimation of the spatiotemporal variability of disease progression within a population. All these elements pave the way to a sharper and more exhaustive analysis of the consequences of disease progression on diverse modalities.

Nevertheless, describing the progression of neurodegenerative diseases faces some particular specificities compared to other temporal processes. First, even though it is, by definition, the evolution of an analogous phenomenon across patients, its expression present subject-specific patterns and characteristics. This is particularly highlighted by the temporal unalignment between individuals. For instance, the pace of progression or the age at disease onset might vary across patients. Typically, two persons sharing the same age might present a different disease stage. Said differently, the same disease stage is likely to appear at a wide range of ages. For that reason, the disease stage, as the expression of a *physiological age* along the disease development, better characterizes the disease and its progression than the observed age. However, the determination of this disease stage is restricted by the entanglement of the disease consequences with the natural characteristics of the patient. A typical example is the cognitive abilities that, despite declining during normal ageing or during the course of some diseases, are different within a population. This advocates for an unbiased comparison between individuals in order to determine properly the impact of the disease. However, the patients are observed during periods of time shorter than the long-lasting overall phenomenon - the latter never being observed fully and directly. This makes the comparison unlikely as there is no overall reference of a typical scenario of disease progression to compare to. Moreover, the very definition of a patient-wise disease stage is unclear as many evidence show that the temporal dynamics of the different biomarkers are not entirely related. This is revealed by the fact that the ordering of the (biological and clinical) events during the course of a disease, as well as their intensities, differ from one patient to another. Due to this variability, considering the existence of an overall disease stage involves that the later is characterized by potentially very different biomarker stages. Such a representation might be misleading and counterproductive as it associates into the same disease stage patients that have diverse biological

and clinical symptoms, preventing from an appropriate description of the disease. Therefore, it might be more accurate to refer to a disease status per biomarker, with complex interplays between the modalities, that finally result in a clinical stage.

All the aforementioned challenges characterize the progression of most of the neurodegenerative diseases. An example of such heterogeneous dynamics is the Alzheimer's disease where the specific role of the structure and of the functions of brain, as well as their interactions, on the cognitive decline remain unclear. First biological symptoms appear during the early, or *prodromal*, phase of the disease, such as the deposition of proteins plaques in the brain. They are essentially followed by a modification of the brain structure which is most of the time associated to a neuronal loss and consequently a diminution the brain metabolism. After a substantial amount of time (e.g. some years), they translate to clinical symptoms such as cognitive complaints and memory loss to finally end by an important dependence on relatives and medical staff. One of the obstacles to properly cure this decline is the fact that the clinical symptoms, i.e. the one that cause the medical examination and diagnosis, appear at late disease stages, when the neuronal loss is unduly important with no possible reversibility. Therefore, the importance to uncover, describe and analyze biomarkers during the course of a disease, especially those associated to early disease stage is crucial. This essentially means to properly understand both the temporal dynamic of each biomarker and their interactions. Such investigation might be undertaken at a population level by characterizing the long-term disease progression, but also at an individual level by identifying patients that will develop the disease at future stages. Some argue that such prediction is worthless as there is no potential treatment. We highlight here that this is actually a chicken-egg dilemma, the lack of treatment being the consequence of unsatisfactory disease modelings and predictions: describing the patients at risk enables to determine the critical biomarkers, common to these subjects, that result in a disease development. Moreover, one of the reason of lacking treatments is partly due to the fact that these treatment should be administered prior to the neuronal loss, at early stages of the disease i.e. in patients whose future prediction indicate a risk of disease development. Furthermore, the treatments may not be adequate for everyone but need to target subgroups of patients with similar patterns of progression. It again supports the idea that the disease should be adequately described and predicted at the different levels (e.g. structural, functional and cognitive) and to study how different cofactors modulate the biomarker progressions.

Therefore, investigating and exhibiting the biomarkers that are related to the disease progression pushes towards the development of appropriate tools that characterize the natural long-term history of the disease. Such a description is made possible only if there is a adequate correspondence between a patient observation and its physiological age along the disease timeline. Recent developments in the medical field have raised promising results in predicting current status based on various measurements. Among many examples, we can cite the detection of breast cancer metastases [Bejnordi et al., 2017], the detection of a particular form of diabetes from retinal photographs [Gulshan et al., 2016], the prediction of cardiovascular risk from the same retinal photographs [Poplin et al., 2018] or the detection of cancer cells in lungs [Zhou et al., 2002]. All these impressive performances allowed to better identify the processes at stake during the related diseases. Though, they are predominantly - if not only - achieved for tasks that present multiple common characteristics: a clear definition of the problem, well-defined labels, significant knowledge about the underlying disease, ... In a word, tasks that are well-identified and rigorously described by the medical community and whose context is clearly set.

Unfortunately, these common denominators are not present in the case of neurodegenerative diseases. In other terms, improving straightforwardly the accuracy of the disease stage prediction is an illusion when the disease is poorly understood. A typical example of such deficiency, in Alzheimer's disease, is the limited number of labels associated to a

disease evolution that is continuous: the patients are either cognitively normal (CN), presenting mild cognitive impairments (MCI) or having Alzheimer's disease (AD). The MCI stage corresponds to the premises of a dementia that can - or cannot - convert to AD: while the converters to AD are called progressive MCI, it is difficult to know whether the non converters, also called stable MCI, are intrinsically not developing AD (potentially in favor of another dementia) or because they pass away before an hypothetical conversion. Besides the lack of label granularity, the very own definition of Alzheimer's disease has evolved during the last decades to describe realities that depend on the community. For some, it corresponds to clinical symptoms. For others, this past definition was based on symptoms that might be the result of different diseases or at least different patterns of progression that cannot be tackled simultaneously. They accordingly added biological characterizations to the disease. Nowadays, there is a tendency to distinguish Alzheimer's pathology, i.e. a set of defined biological biomarkers, from Alzheimer's disease, namely clinical symptoms. This reveals the ineffectiveness to predict a label, namely a disease stage, whose definition has not been properly set nor represents a homogeneous set of characteristics. Eventually, this is worsen by the trade-off between performance and interpretability, the former being predominantly chosen by the Machine Learning community. This is helpful in cases where the problems are well-defined but their determinants are too complex to be fully controlled and established by hand (by a doctor for instance). However, problems that are poorly defined and ill-posed are not susceptible to be improved in a substantial manner. This might be a reason of the somehow constant accuracy in disease progression over the past years. This is an additional reason to believe that explanability of the methods is key for disease that lack knowledge about their causes, effects and consequences.

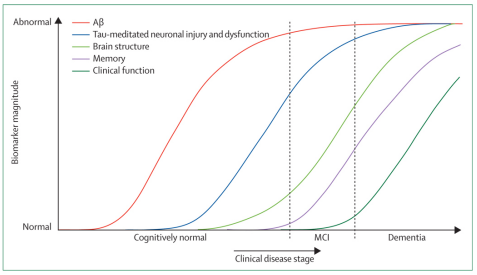
For all these reasons, there is an intensive need to model and better understand the disease progression through its repercussions on different biomarkers and modalities. This necessarily involves to properly define and estimate the variability of individual progressions within a population. To this end, these complex dynamics should be accordingly inspected, described and analysed at both population and individual levels while considering the interplay of the different biomarkers.

Disease Modeling

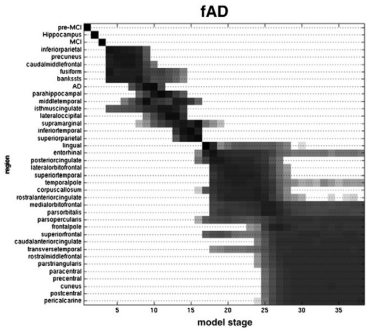
To overcome the lack of knowledge about the progression of neurodegenerative diseases, there has been a large interest in *disease modeling* over the past decades. The first models have mainly been introduced by the medical community that has synthesized years of practical knowledge into so called *hypothetical models* as they are not directly supported by data evidence but rather field work and experience. In the case of Alzheimer's disease, one of the most famous model was described in [Jack Jr et al., 2010b], shown on Fig. 1a. It was intended to present a hypothesis for the sequence of events that leads to Alzheimer's disease. They hypothesized that there exists a cascade of consequential events that starts many years before the clinical symptoms, during the prodromal phase, that is characterized by protein plaques in the brain followed by a neuronal loss. They simultaneously highlighted the multi-modal aspects of the disease progression.

While they remain good starting points, these models remain hypothetical. And even though the first mathematical frameworks to order sequence of events have been introduced almost three decades ago [Beckett, 1993], they have not received much attention by the medical community because of the lack of large cohorts that were supposed to confirm or infirm the assumptions of the hypothetical models. These databases, essentially cross-sectional at the beginning, i.e. one observation per patient, contributed to the development of *data-driven models* that produced models of disease progression based on data evidence. Among them, [Fonteijn et al., 2012b] introduced the *event-based model* to characterize the sequence of events during the progression of Alzheimer's disease and Huntington's disease. Latter improved in [Young et al., 2014, Venkatraghavan et al., 2019], it essentially orders the observations to produce a sequence of events that occur during the disease progression, as shown on Fig. 1b. This cascade does not measure the temporal evolution of each biomarker, nor the time delay between the apparition of two symptoms. Also, the variability between individuals was only defined as an uncertainty in the cascade of events, represented by blocks of biomarkers that might occurs either simultaneously or with intervertions for different patients. Studies as [Huang and Alexander, 2012] defined an explicit variability of the model within a given population. This absence of temporal characterization of the disease evolution was mainly due to the fact that age is a poor proxy of the disease stage. Among the attempts to circumvent this issue, [Iturria-Medina et al., 2016] considered the stage of the Alzheimer's disease (CN, early MCI, late MCI and AD) as a proxy of the evolution to show the role of vascular disregulation during the course of the disease.

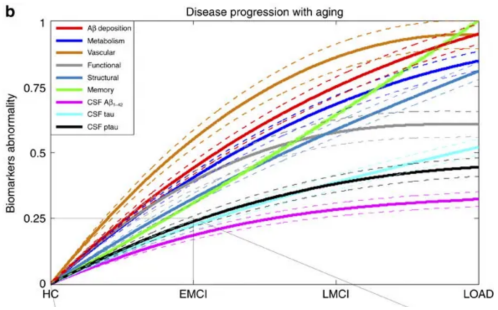
To overcome the issue of the temporal variability in the dynamic of disease progression, longitudinal databases, namely multiple visits of patients, have been gathered. Hundreds of patients were followed during many years to measure various biomarkers (e.g. cognitive assessments or medical imaging) along with abundant cofactors (e.g. gender, genetics, socio-demographic attributes, comorbidities). While undoubtedly informative, these databases bring together patients at different disease stages, with potentially different disease onset and pace of progression. To address this temporal variability, [Jedynak et al., 2012] introduced an affine time-reparametrization of the real age $t \mapsto \alpha_i t + \beta_i$ onto the physiological age, considering that each subject presents a temporal onset, or shift, β_i as well as an acceleration factor α_i. The model was later extended to spatial data (e.g. PET amyloid imaging) that were converted into a disease score that allowed to realigned the observation while showing spatial correlation in the pattern of progression [Bilgel et al., 2015, Bilgel et al., 2016]. At the same time, [Donohue et al., 2014] considered that each individual short-term measurements represent snapshots of the overall disease progression. Once reparametrized, the patients observations can retrace the long-term evolution of the different biomarkers, as shown on Fig. 1d. Similarly, [Guerrero et al., 2016] described the individual trajectories as deviation of the group-average scenario of change.



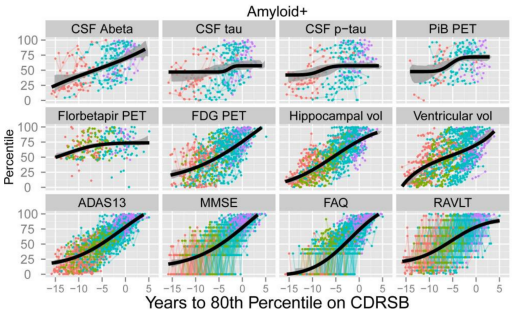
(a) Hypothetical model of the cascade of events during the disease progression. Courtesy of [Jack Jr et al., 2010b].



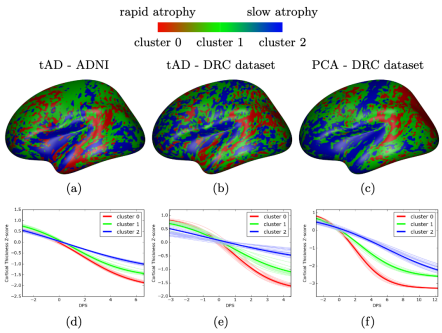
(b) Event-based model that rank the sequence of events. Courtesy of [Fonteijn et al., 2012b].



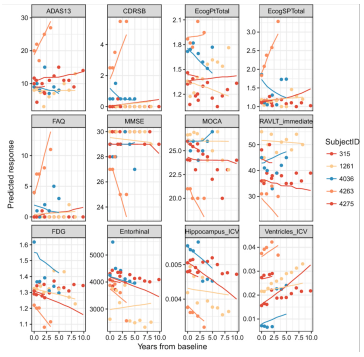
(c) Temporal realignment using the diagnosis as proxy of the disease progression. Courtesy of [Iturria-Medina et al., 2016].



(d) Temporal realignment of the individual observations to reconstruct the group-average trajectory. Courtesy of [Donohue et al., 2014].



(e) Temporal realignment of the subject measurements based on their spatial coherence. Courtesy of [Marinescu et al., 2017].



(f) Derivation of the group-average trajectory to predict individual future measurements. Courtesy of [Iddi et al., 2019].

Figure 1: Evolution of the disease modeling in the last decades, from hypothetical broad models to individual specific prediction.

Such description corresponds to mixed-effects model, where the individual derivations take the form of random effects around the fixed effects that are shared by the population.

Recently, a probabilistic setting of evolution was introduced in [Lorenzi et al., 2017]. It also allows to realign the individual measurements along the disease axis. The follow-up measurements show to be very informative in the evolution modeling. However, it lacks a natural way to represent the variability in term of spatiotemporal progression. Another recent technique was introduced in [Oxtoby et al., 2018]. The authors relate the biomarker rate of change to the biomarker value itself, exploiting differential equations to model the disease progression.

The previous model essentially examined cognitive assessments or features derived from medical imaging, such as values of the cortical thickness or of the hypometabolism in specific regions of interest, the volume of sub-cortical structures, the concentration of proteins in blood tests, ... Such studies, that go beyond univariate measures to analyse different modalities, essentially extract biomarkers prior to analyse their evolution. This is particularly true for imaged-based features. On the contrary, few studies explore the complexity of entire images. Among them, [Marinescu et al., 2017] recombine biomarkers from medical imaging measurements, taking advantage of the spatial structure of the disease progression. This allows to re-position the observations along the disease axis while aiming at clustering regions that are most likely to have a strong evolution during the disease history.

While most of the aforementioned models are able to define a group-average trajectory based on individual measurements, there are not suited to characterize individual progressions. Apart from the mixed-effects models, they do not provide a simple way to derive the long-term trajectory to individual observations. On the other hand, [Iddi et al., 2019] proposed a mixed-effects model used to predict future time-points, as shown on Fig. 1f, but whose overall explanability is made more complex due to the use of advanced Machine Learning algorithms. The authors of [Schiratti et al., 2015] introduced a generative mixed-effects model that similarly reconstructs the long-term disease progression from individual short-term measurements. Additionally, as the model considers the individual trajectories as spatiotemporal variations of the group-average one, it enables to derive estimation of the disease progression at an individual level. This is a first step to define continuous trajectories that define the evolution of the biomarkers at any time, potentially at future time-points. This description is made possible by the characterization of the overall spatiotemporal variability of evolution across subjects. The variability takes the form of random effects that, once estimated, define a probability distribution over the individual variables that modulate the typical scenario of progression. This distribution makes the model *generative* in the sense that it is possible to draw new samples. As each of them defines exactly the variations to the group-average scenario of progression, it entirely defines a new patient that reproduce the characteristics of the real patients, while being simulated and thus anonymous.

To sum up, over the last decades, models that first described the general trend of the disease progression, slowly considered a time-realignment of the individuals to get the average sequence of events. Along the way, some proposed to convert this sequence to a temporal dynamic and ordering, including more complex modalities. Finally, recent advances helped to go from a group-average trajectory to a individual description of the disease progression. These improvements were associated to the increasing complexity of the biomarkers analyzed. Unfortunately, they did not necessarily investigate the multimodal aspects of the disease, especially their separate but dependant temporal dynamics. This advocates for a general framework to study the changes of different biomarkers under the influence of the disease evolution.

The work detailed in this thesis follows this historical development, especially from the work initiated in [Schiratti, 2016]. It tackles the challenges raised by the complex temporal dynamics of each biomarkers, which also present specificities at the individual level, to :

- describe the typical scenario of disease progression, from prodromal to clinical stages,
- personalize this average trajectory to individual progressions, enabling an in-depth study of the cofactors that modulate this progression,
- characterize the individual trajectories to impute missing values and predict future stages,
- properly estimate the variability of disease progression to simulate virtual cohorts of anonymized individuals.

While [Schiratti, 2016] introduced the spatiotemporal model to essentially address the first point, especially for scalar biomarkers, this manuscript aims at further investigating the three remaining challenges for a larger family of spatiotemporal models of progression and to include multiple modalities such as the thickening of the cortical structure, the brain glucose consumption as a marker of the hypometabolism, and the decline of the cognitive abilities. It inevitably involves to analyze data that have different characteristics such as their acquisition, their dimension and resolution, their measurement noise and or inter- and intra-individual variability. Therefore, this work, while analysing and comparing heterogeneous data, is built with the intention to be adaptable to different modalities and biomarkers.

Conceptual overview

To ease the reading and understanding of the proposed model, we first start by exemplifying the problem at hand: the characterization of a temporal trajectory given some measurements. Let us describe the growth of a child given his pictures at different ages. We consider a fixed camera that shots a plain-size picture of a child at 6 and 12 months (on a white background), as shown on Fig.2. Each picture, represented by a blue dot, has N pixels, each being valued between 0 and 255, such that the feature space is $]0, 255[^N$. The picture of the same person at 9 months old also belongs to this space *but* it is easily understandable that this picture is not the mean (in the Euclidean space) of the two previous pictures: the mean, i.e. the mean of the pixels, results in the blurry version of the superposition of the first pictures, as shown on Fig. 2. On the other hand, the collection of all the pictures between 6 and 12 months defines a *curve* (in a sense to be precised) in the feature space, represented by the black curve, that corresponds to the individual trajectory. As we can theoretically define this trajectory for any individual, the set of all the resulting curves results in a subspace of the initial embedding space that we model by the blue-to-red surface on Fig. 2. This subspace, called a Riemannian manifold (see Chapter 1), allows to perform calculus between images, for instance to exhibit the picture at 18 months old by *following* the black curve on this manifold.

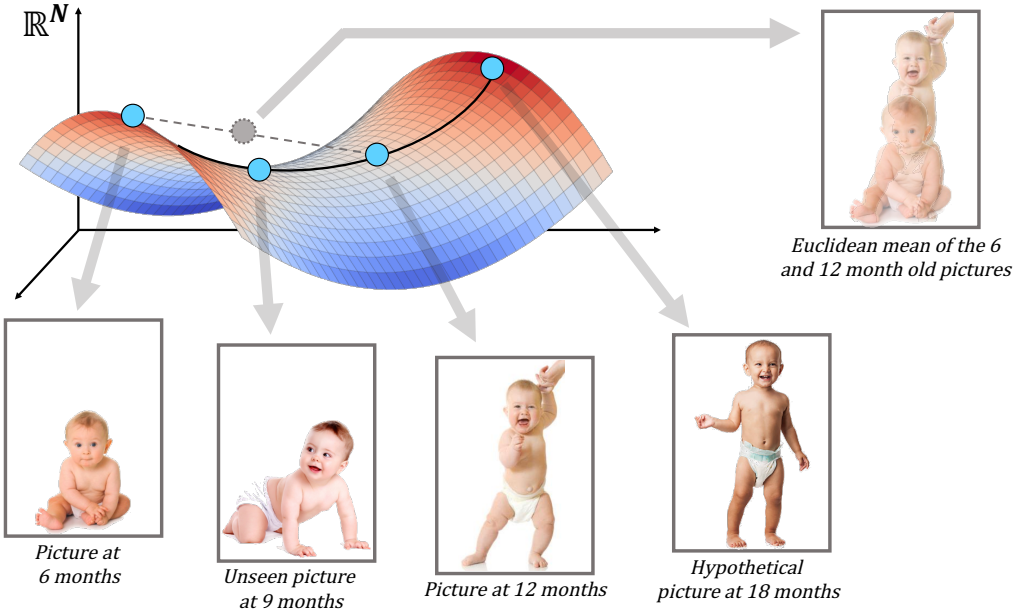


Figure 2: The pictures (blue dots in the space of measurements $\mathbb{R}^N$) belongs to a subspace, illustrated by the blue-to-red surface that corresponds to the space of possible pictures during the growing. This subspace circumvents the inability of the Euclidean mean to compute the mean picture between 6 and 12 months. It also enables to predict future pictures, for instance at 18 months, based on the trajectory (black curve) that is estimated from the first pictures.

While the mathematical formalism of this modeling is described in Chapter 1, same logic applies to the modeling of the individual spatiotemporal trajectories of disease progression. In the case of longitudinal databases, the repeated observations of the subjects are represented by the colored dots on Fig. 3. The collection of all possible observations is represented by the blue-to-red surface. As in the previous example, the individual trajectories are represented by curves on this surface. In the particular case of disease progression, while individuals are mostly observed during short-term periods, we make the hypothesis that there exists a long-term group-average spatiotemporal trajectory represented by

the black curve. The latter should be considered as the recombination of the individual snapshots, each providing consistent information about a particular disease stage. By construction, it consequently spans a longer time-window of disease progression. To account for the spatial variability, i.e. the fact that there is a distance between the group-average and individual curves, we consider that there exists a spatial shift from one to the other. On top of that, the temporal variability is modeled through a temporal reparametrization of the progression along the curves. It enables a time shift of the disease onset as well as an acceleration factor that modulates the individual speed of progression. To these spatiotemporal variability, we highlight that the subjects are observed at different stages, with different baseline ages and a different number of times. These characteristics, while being potentially difficult to handle, are actually key to provide necessary information to reconstruct the overall disease progression over a long period of time.

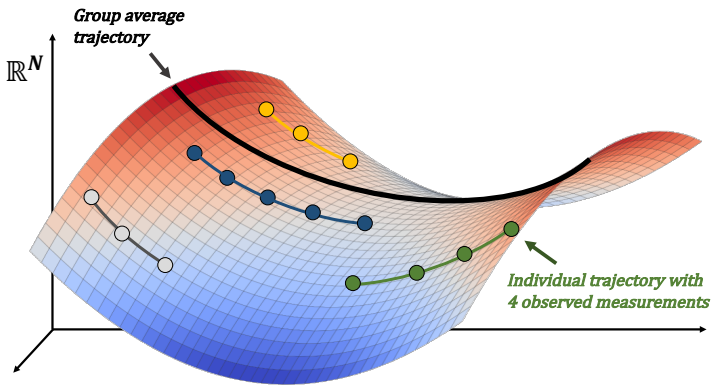


Figure 3: Given the space of possible observations represented by the blue-to-red surface, the dots are the measurements such that the colors indicate the patient they belong to. The corresponding curve is the individual trajectory that can be recombine into a long-term group-average trajectory in black. The trajectory present an important variability in term of number of measurements, distance to the mean curve, potentially the stage at the first visit.

Fig. 3 finally shows that this modeling allows to both characterize the group-average long-term trajectory while personalizing this progression to individual measurements. Beyond the possibility to achieve the two aforementioned goals, this figure illustrates that this embedding provides a framework (detailed in the next chapters) that also enables to:

- accurately compare the individual spatiotemporal progressions thanks to their relative positioning to the mean, in particular by studying the cofactors that significantly modulate the disease progression,

- reconstruct particular values along the individual trajectory to either impute missing values or predict future time-points by extrapolating the timeline,

- properly determine the space of possible measurements and trajectories to generate virtual individuals with longitudinal measurements that enables first to un-bias or balance initial cohorts and to enhance the predictive power of algorithms that require large amounts of data.

To demonstrate that this model provides an adequate framework to study the disease progression, at both an average and individual level, we concentrate our attention on the study of Alzheimer's disease, especially the synchronized evolution of the cognitive functions, the hypometabolism and the structure of the brain. This first requires to consider a generic model suited for data that present different characteristics in term of structure or dimensions, e.g. vectors of cognitive assessments, positron emission tomography (PET) and magnetic resonnance imaging (MRI). Once formulated, the model is evaluated on its capacity to characterize the long-term disease progression and to reconstruct the individual trajectories. The quality of the latter is assessed by their comparison to the intrinsic noise in the data which is to be determined and measured (e.g. test-retest data, medical imaging resolution, feature extraction). As these reconstructions are made possible thanks to the estimation of a continuous individual trajectory, it further enables to impute missing values and predict the biomarkers at future time-points, up to 4 years in advance. Finally, the ability of the model to properly estimate the variability in term of disease progression enables the simulation of virtual patients with longitudinal measurements, that once gathered into a *virtual cohort*, can be shared without violating sharing policies as anonymization.

This framework can provide promising tools to the medical community to better understand diseases and their underlying mechanisms. To this end, special attention has been given to the development of numerical tools that can be used efficiently by both the medical and research community. We dedicated the website `www.digital-brain.org` to an interactive digital model of the Alzheimer's disease progression that can be modified to exhibit individual scenarios of evolution. Furthermore, the `Leaspy` Python package has been released to enable researchers to estimate similar disease progression on other biomarkers, modalities, cohorts and diseases.

Manuscript overview

The first part is a general introduction to the generative mixed-effects model presented in this thesis. It first exposes the mathematical definitions and tools needed to define the generic model of progression. It then presents particular instantiations of the model that suit different profiles of disease progression. Finally, it describes the mathematical operations that allow to reconstruct the group-average scenario of change, to personalize it to individual measurements, to impute missing values, to predict biomarkers at future time-points and, finally, to simulate virtual patients.

The second part introduces a model that is better suited to describe the disease progression for data that present a spatial structure. It includes medical imaging and network-values measures. The model ensures a spatial coherence of the disease progression for neighbor regions. This is validated by characterizing the cortical atrophy and the brain metabolism decrease during the course of Alzheimer's disease.

The third part extensively describes and validates the possibilities offered by this disease progression framework. We consider a large scale longitudinal database of Alzheimer's disease from which we use cognitive assessments and medical imaging derived data (cortical thickness over the entire brain, deformation of the hippocampus meshes and FDG-PET data) to reconstruct the long-term disease progression. This demonstrates the reconstruction of the individual data up to the noise level and the prediction of future time-points. It comes with a finer description of the disease mechanisms during the course of the disease

The fourth part takes the most out of the generative and mixed-effects characteristics of the model: it allows to simulate patient's missing or future observations and also virtual patients. The simulated observations can either be used to impute missing values or to predict future timepoints. On the other hand, the simulation of virtual patient enables to un-bias or unbalance real cohorts for under-represented subgroups. In both cases, the resulting virtual cohort helps improving algorithm predictive power in order to reach state of the art results in the long-term prediction of cognitive impairments.

The fifth part describes the digital tools developed during the thesis. First, the Python package `Leaspy` allows to run similar analysis on new cohorts for potentially other (neurodegenerative) diseases. Then, we develop a digital model that relates for the long term progression of the disease for different modalities. Finally, we propose a clinical dashboard to monitor patients in a clinical study or in real life.

Part I

Spatiotemporal Model of Progression from Longitudinal Data

Chapter 1

Scalar Models and Extensions

This chapter first briefly introduces the key mathematical concepts of the manuscript the model is build on, i.e. the Riemannian geometry and the mixed-effects models. It cannot be considered as a detailed or exhaustive description of these mathematical notions, but rather a glimpse of the tools that are essential to the understanding of the following chapters. However readers that are eager to understand the formal mathematical background of the corresponding topics are referred to the mentioned references. In a second part, the chapter introduces the generative mixed-effects model of disease progression. It starts with the mathematical description of the model, in its generic form, thanks to the Riemannian geometry setting. It then gives multiple instantiations of the model before discussing some of its properties.

Contents

1.1 Riemannian geometry

In the introduction, we made the hypothesis that the data of interest belong to a particular subspace of the feature space, that individual trajectories are described by curves on this subspace and that the repeated observations are points on these curves. This subspace is thus central to the disease modeling as it entirely defines the space of possible measurements and consequently the individual spatiotemporal trajectories. To this end, we introduce the Riemannian geometry that is well suited to define such spaces but also to derive mathematical notions such as curves and distances on this space. Historically, it has been introduced to study differentiable topological spaces embedded in $\mathbb{R}^n$. These spaces, called manifolds, are characterized by the associated metrics that allows to generalize the notion of distances in Euclidean spaces to such manifolds. We then introduce the concept of geodesics that characterizes curves in these non Euclidean spaces. Finally, we define the concept of parallel transport which is an important tool to *shift* (in a sense to be precise) the previous curve to other regions of the manifold. This theoretical framework is extensively presented in [Do Carmo Valero, 1992].

1.1.1 Manifold

A *manifold* is a topological space for which each point presents a neighborhood that is homeomorphic to the Euclidean space. Simply said, there exist a collection of mappings (called atlas) from *regions* of this space (as defined in [Do Carmo Valero, 1992]) to linear spaces. It is possible to make calculus on each of this linear space and to derive it to the corresponding region ; that leads to a locally differentiable structure. However, if these local differentiable structures are continuous (in some sense [Do Carmo Valero, 1992]) from one local mapping to the other, then the differentiable structure is said to be globally differentiable. This defines a *differentiable manifold* or *smooth manifold*.

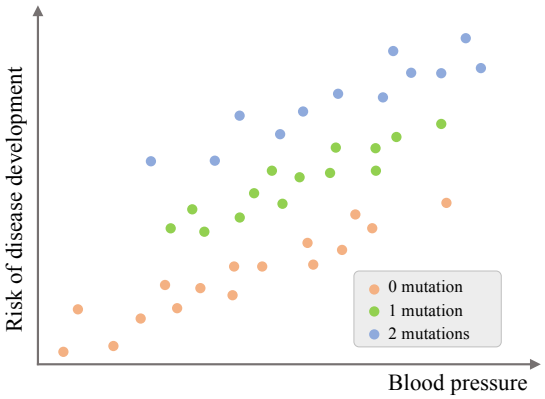
Given a smooth manifold $\mathbf{M}$ of $\mathbb{R}^n$, each point $\mathbf{p} \in \mathbf{M}$ is associated to its *tangent space* $\mathbf{T_pM}$ that is a linear approximation of the manifold in the neighborhood of $\mathbf{p}$. This tangent space contains all the possible derivations of $\mathbf{M}$ at $\mathbf{p}$, intuitively corresponding to the vectors at $\mathbf{p}$ in the direction of the derivations. These derivations are made possible at each point of the smooth manifold by the differentiable structure.

1.1.2 Metrics and Riemannian manifolds

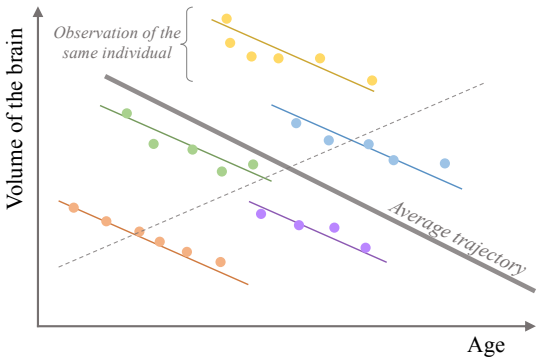
We consider a smooth manifold $\mathbf{M}$ such that each point $\mathbf{p} \in \mathbf{M}$ is associated to an inner product $g_\mathbf{p}$ on the vector field of the tangent space $\mathbf{T_pM}$, which varies smoothly from point to point. The collection $g_\mathbf{M} = (g_\mathbf{p})_{\mathbf{p}\in\mathbf{M}}$ is called a *metric* on the manifold. This generalizes the Euclidean scalar product to manifolds. Equipped with this metric, $(\mathbf{M}, g_\mathbf{M})$ is called a *Riemannian manifold*. This key concept allows to introduce, among others, the notion of distances on this differentiable structure.

1.1.3 Geodesics

The geodesics are to Riemannian geometry what straight lines are to Euclidean spaces: they correspond to curves that to some extend represent the shortest path between two points of the underlying manifold. Formally, given a smooth curve $\gamma : I \subset \mathbb{R} \to \mathbf{M}$, we say that γ is a geodesic of $\mathbf{M}$ if $\nabla_{\dot\gamma}\dot\gamma = 0$, i.e. a smooth curve with zero acceleration (∇ corresponds to the Levi-Civita connection, see [Do Carmo Valero, 1992] for technical details).



(a) Each point being the observation of a patient, the risk of disease development depends on the blood pressure (fixed effect) and the number of mutations of a given allele (random effect).



(b) Considering repeated observations per subject, a standard linear regression (grey dashed line) returns unexpected results. A random-slope random intercept model is better suited to describe the individual variability.

Figure 1.1: Examples of mixed effects models in the case of (a) independent data and (b) longitudinal data with repeated observations per subjects. They are better suited than standard tools (e.g. standard linear regressions) to combine population and individual effects.

1.1.4 Exponential mapping

We consider a point $\mathbf{p} \in \mathbf{M}$, a velocity $\mathbf{v} \in \mathbf{T_pM}$ and a geodesic γ such that $\gamma(t) = \mathbf{p}$ and $\dot{\gamma}(t) = \mathbf{v}$. It can be shown that such geodesic is unique so that we rewrite it $\gamma := \text{Exp}_{\mathbf{p},t}(\mathbf{v}) : t \mapsto \text{Exp}_{\mathbf{p},t}(\mathbf{v})(t)$. The *exponential mapping* associates the vector $\mathbf{v}$ to the point reached by this geodesic at time $t+1$. It writes $\mathbf{v} \in \mathbf{T_p} \mapsto \text{Exp}_{\mathbf{p}}(\mathbf{v}) = \text{Exp}_{\mathbf{p},t}(\mathbf{v})(t+1)$. It is essentially a *step* on the manifold from $\mathbf{p}$ in the direction of $\mathbf{v}$.

1.1.5 Parallel-transport

Given a manifold $\mathbf{M}$ and a smooth curve $\gamma : I \subset \mathbb{R} \to \mathbf{M}$, a vector field X is said to be *parallel* along γ if $\frac{DX}{dt} = 0$. Given $\mathbf{w}_0 \in \mathbf{T}_{\gamma(t_0)}\mathbf{M}$, one can show there exists a unique vector field $\mathbf{w}(t)$ parallel along γ such that $\mathbf{w}(t_0) = \mathbf{w}_0$. This corresponds to the *transport* of $\mathbf{w}_0$ along γ such that the vector field $\mathbf{w}(t)$ remains *parallel* to $\mathbf{w}_0$. This notion is crucial to compare calculus across tangent spaces along a geodesic.

1.2 Mixed effects models

While Riemannian geometry is fundamental to characterize the space of possible measurements, it does not provide a formulation of the individual nor the population trajectory of disease progression on these Riemannian manifolds. To this end, we introduce the mixed effects models [Fisher, 1919, Fisher, 1992]. They are statistical models which combine, in the description of a phenomenon, the contribution of global effects affecting all the observations, and, the contribution of individual effects that are specific to each observation. On the one hand, the global effects are called the *fixed effects*: they are non-random quantities that best describe the whole population; one can think of the slope and the intercept as the fixed effects of a linear regression because they affect equivalently all the observations. On the other hand, the individual variability is described by random perturbations of the fixed-effects, called the *random effects*, that allow to derive the individual observations from the population-wide description. They essentially characterize the overall variability in the population. An example of such mixed-effects model is shown on Fig. 1.1a.

This modeling helps distinguishing the common dynamic from individual specific patterns. It is particularly well suited to describe observations where there is no independence between some data e.g. repeated observations of the same individual, as they enable to define effects at the individual level, shared by all the observations of the same subject [Laird and Ware, 1982, Lindstrom and Bates, 1988, Lindstrom and Bates, 1990]. Such models, that combine population and individual effects, are called *hierarchical models* where the higher levels of the model are more discriminant (in term of explained variance for instance) than the lower levels. Fig. 1.1 gives examples that highlights the importance and benefits of such models.

1.2.1 Linear mixed effects models

The first mixed-effects models to be introduced are the linear mixed-effects models (LME) [Laird and Ware, 1982, Verbeke and Lesaffre, 1996, Bates and Pinheiro, 1998]. They can be interpreted as an extension of (classic) linear regression models with an additional degree of freedom being the subject-specific derivation from the linear trend.

Given a set of n observations, such that the i-th observations presents the outcome variables $\mathbf{y}_i \in \mathbb{R}^N$ associated to the input variables $(\mathbf{x}_i, \mathbf{z}_i) \in \mathbb{R}^{N\times p} \times \mathbb{R}^{N\times q}$, the model writes:

$$\mathbf{y}_i = \mathbf{x}_i\boldsymbol{\alpha} + \mathbf{z}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i \,,$$

where $\boldsymbol{\alpha} \in \mathbb{R}^p$ corresponds to the fixed effects of the model and $\boldsymbol{\beta}_i$ to the random effects associated to observation i. Further assumptions are possible on the distribution of the $(\boldsymbol{\beta}_i)_{1\leq i\leq N}$ such as a multivariate normal distribution.

1.2.2 Non linear mixed effects models

There are multiple reasons why linear model are not sufficient to model the interaction between the input variables $(\mathbf{x}_i, \mathbf{z}_i)$ and the output variable $\mathbf{y}_i$. [Lindstrom and Bates, 1990, Pinheiro and Bates, 1995, Bates and Pinheiro, 1998], among others, have introduced non linear mixed effects models (NLMEM). Their descriptions are ad-hoc to the studied phenomenon but they can be summed up under the general writing:

$$\mathbf{y}_i = g(\mathbf{x}_i, \mathbf{z}_i) + \boldsymbol{\varepsilon}_i \,,$$

where g is a non-linear function of the input variables $(\mathbf{x}_i, \mathbf{z}_i)$ and $\mathbf{x}_i$ (resp. $\mathbf{z}_i$) corresponds to the variables associated to the fixed effects (resp. random effects).

1.2.3 Longitudinal data in the case of biological phenomenon

In principle, it is easy to interpret the dynamic of disease progression, or any biological phenomenon described with longitudinal data, as the combination of fixed effects that characterize the average temporal dynamic and of random effects that represent the individual variations to the mean [Verbeke and Molenberghs, 2009]. An example of such phenomenon is the decrease of the overall brain size, as shown on Fig. 1.1b. The natural decrease over time (the fixed effects) is modulated by the individual brain size that might vary across subjects (the random effects). We here consider a longitudinal dataset where the i-th individual is observed at times $(t_{ij})_j$ where the j-th observation at time t_{ij} is denoted $\mathbf{y}_{ij}$. To study such longitudinal data, the *random-slope, random-intercept* has been introduced to describe each individual evolution as a variation of the slope and intercept of an average trajectory, as shown on the estimation of the individual brain volume over time on Fig. 1.1b.

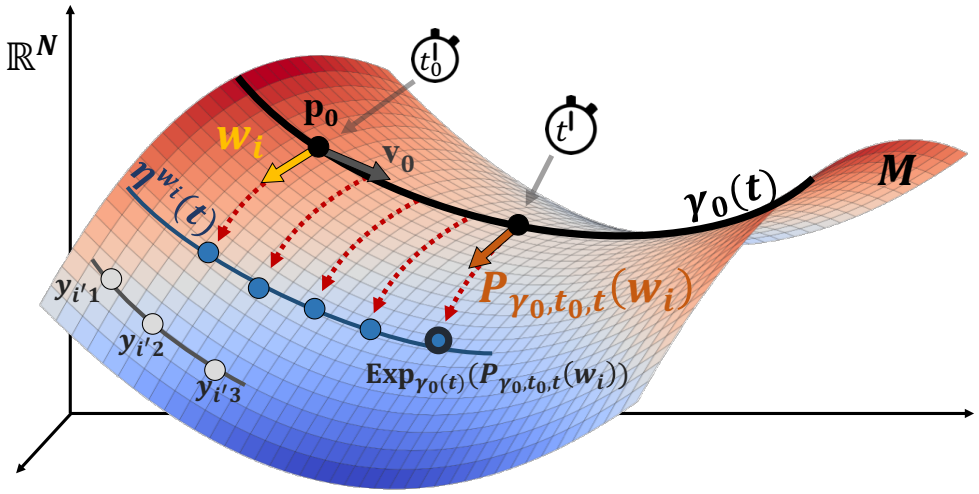


Figure 1.2: Geometric description of the model where the longitudinal observation $y_{ij} \in \mathbb{R}^N$ belongs to a Riemannian manifold $\mathbf{M}$. The group-average trajectory is characterized by the geodesic $\gamma_0(t)$. The individual trajectory $\eta^{\mathbf{w}_i}(t)$ corresponds to the exponential mapping of the vectors $P_{\gamma_0,t0,t}(\mathbf{w}_i)$, that are the parallel transport of the vector $\mathbf{w}_i \in \mathbf{T}_{p_0}\mathbf{M}$ along $\gamma_0(t)$.

This random-slope random intercept model writes :

$$\mathbf{y}_{ij} = (\alpha^1 + \beta_i^1)t_{ij} + (\alpha^2 + \beta_i^2) + \epsilon_{ij}.$$

The fixed-effects are $\alpha = (\alpha^1, \alpha^2)$ where α^1 is the model slope and α^2 is the model intercept. It defines the average trajectory $\mathbf{y} : t \mapsto \alpha^1 t + \alpha_2$. The random effects $\beta_i = (\beta_i^1, \beta_i^2)$, that are independent and identically distributed samples of a normal distribution, represent variations to the model intercept and slope. The interest of such model is shown on Fig. 1.1b where the observations of the same patients are represented by the same color. A standard linear regression that consider the observations as independent results in the dashed line that does not correspond to the dynamic of the phenomenon. On the other hand, a random-slope random-intercept model outputs an average trajectory as well as individual trajectories characterized by the variations to the mean trajectory.

This model is well suited for dynamics that are temporally aligned i.e. where any time t corresponds to the same event across patients. This is for instance accurate for dynamics with a reference time-point such as pharmacokinetics that evaluate the effect of a drug from a reference time-point which corresponds to the administration of the drug. Same logic applies for any dynamic whose starting point is known. Conversely, some phenomenon present unaligned temporal dynamics such as disease progression where the age at onset is different and unknown for each patient. In these case, it is possible to account for a reparametrization of the time, e.g. $t \mapsto t - \tau_i$ where τ_i is a temporal shift to realign the individual observations. However, this reparametrization, in the random-slope random-intercept model leads to an non indentifiable model as there are multiple sets $(\beta_i^1, \beta_i^2, \tau_i)$ that characterize the same individual trajectory.

1.3 Disease progression model

In this section, we define the mixed-effects model introduced in [Schiratti et al., 2015] that is used to characterize the long-term disease progression and the individual trajectories. To this end, we consider the repeated observations of p individuals, such that the i-th individual has been observed $k_i \in \mathbb{N}^*$ times at times $t_{i,1} < \ldots < t_{i,k_i}$. The observation at time t_{ij} is denoted $\mathbf{y}_{ij} \in \mathbb{R}^N$, where $N \in \mathbb{N}^*$. Finally, let us denote $\mathbf{y} = (t_{ij}, \mathbf{y}_{ij})_{\substack{1 \leq i \leq p \\ 1 \leq j \leq k_i}}$ the set of longitudinal observations.

1.3.1 Geometric description

Thanks to the Riemannian settings introduced in 1.1, we first consider that each observation $\mathbf{y}_{ij}$ belongs to a Riemannian manifold $\mathbf{M} \subset \mathbb{R}^N$ as shown on Fig. 1.2. We also consider that there exists a geodesic $\gamma_0 : I \subset \mathbb{R} \to \mathbb{R}^N$, reaching $\mathbf{p}_0$ at t_0 ($\gamma_0(t_0) = \mathbf{p}_0$) with velocity $\mathbf{v}_0$ ($\dot{\gamma}_0(t_0) = \mathbf{v}_0$), that represents the group-average spatiotemporal trajectory, i.e. that corresponds to the global temporal dynamic of disease progression. We consider that the individual trajectories are spatiotemporal variations of this mean trajectory in the sense that they derive from it thanks to :

- a *spatial* variation defined by $\mathbf{w}_i \in \mathbf{T}_{p_0}\mathbf{M}$, called the space-shift. It characterizes the direction in which the group-average trajectory is shifted to approximate the data $(\mathbf{y}_{ij})_{1 \leq j \leq k_i}$ of the i-th individual. As described in Chapter 1.1, it is possible to parallel transport the vector $\mathbf{w}_i$, defined at t_0, along the curve γ_0, for any time t. The resulting vector writes $P_{\gamma_0, t_0, t}(\mathbf{w}_i)$ as shown on Fig. 1.2. Accordingly, the exponential mapping of this collection of vectors, that writes $\mathrm{Exp}_{\gamma_0(t)}(P_{\gamma_0,t_0,t}(\mathbf{w}_i))$, define the individual trajectory $\eta^{\mathbf{w}_i}(t) := \mathrm{Exp}_{\gamma_0(t)}(P_{\gamma_0,t_0,t}(\mathbf{w}_i))$. This corresponds to the exponentialization of the group-average geodesic γ_0 in the direction $\mathbf{w}_i$.
- a *temporal* variability that is defined by the *acceleration factor* $\alpha_i \in \mathbb{R}$ and the *time shift* $\tau_i \in \mathbb{R}$. As there is no reason for the individual trajectory $\eta^{\mathbf{w}_i}(t)$ to progress at the same speed as $\gamma_0(t)$, we introduce a temporal reparametrization $\psi_i : t \mapsto \alpha_i(t - \tau_i - t_0) + t_0$. Therefore, the individual observation at time t_{ij} corresponds, on the disease timeline, to the age $\psi_i(t_{ij})$. The acceleration factor α_i acts on the speed of the dynamic : $\alpha_i > 1$ (resp. $\alpha_i < 1$) corresponds to faster progressors (resp. slower progressors). On the other hand, the time shift τ_i enables to shift the temporal progression of a given number of years. $\tau_i > 0$ (resp. $\tau_i < 0$) corresponds profile that present a late (resp. early) progression.

As the feature space is potentially of high dimension, we consider that $\mathbf{w}_i$ can be decomposed in an Independent Component Analysis (ICA) manner, such that $\mathbf{w}_i = As_i$ where $A \in \mathbb{R}^{N \times N_s}$ is called the mixing matrix and $s_i = (s_{ij})_{1 \leq j \leq N_s}$ are the sources. The idea is that instead of living in a high dimensional space, $\mathbf{w}_i$ can be represented is a subspace spanned by the vectors given by the columns of the mixing matrix A.

Finally, the individual measurements of the i-th individual at time t_{ij} writes

$$\mathbf{y}_{ij} = \eta^{\mathbf{w}_i}(\psi_i(t_{ij})) + \epsilon_{ij}\,, \tag{1.1}$$

where ϵ_{ij} is the residual noise not captured by the model.

As $\eta^{\mathbf{w}_i}(\psi_i(t_{ij}))$ is parametrized by $\mathbf{p}_0$, t_0 and $\mathbf{v}_0$, i.e. exactly the geodesic γ_0, the good fit between $\eta^{\mathbf{w}_i}(\psi_i(t_{ij}))$ and the real observations y_{ij} necessarily means a geodesic whose derivation is well suited to reconstruct the individual data. In other terms, this corresponds to sticking together individual data in such a way that the resulting long-term progression may be derived to describe individual spatiotemporal trajectories.

1.3.2 Statistical description

The Riemannian settings enables to describe the average and individual spatiotemporal trajectories that corresponds to the evolution of a given feature. On the one hand, the long-term disease progression corresponds to the geodesic γ_0, parametrized by $\mathbf{p}_0$, t_0 and $\mathbf{v}_0$. On the other hand, the function $t \mapsto \eta^{\mathbf{w}_i}(\psi_i(t_{ij}))$ corresponds to the geometrical description of the individual trajectory on the Riemannian manifold. It derives from γ_0 thanks to α_i, τ_i and $\mathbf{w}_i$.

The disentanglement between the population and individual trajectories, the latter deriving from the former, can be seen as a mixed-effects model: the main geodesic, i.e. $\mathbf{p}_0, t_0, \mathbf{v}_0$ corresponds to the fixed-effects while the individual parameters $\alpha_i, \tau_i, \mathbf{w}_i$ are the random-effects in the sense that they are random variations of the mean trajectory.

1.3.3 Identifiability conditions

To ensure the identifiability of the model in the presence of spatio-temporal variability, the space-shift $\mathbf{w}_i$ has to be orthogonal to the velocity $\mathbf{v}_0$. The reason of this condition is detailed in [Schiratti, 2016]. An intuitive reason to this condition is that if both vectors were not orthogonal, then the projection of $\mathbf{w}_i$ on $\gamma_0(t)$, non null, might interfere with the temporal progression which is controlled by α_i and τ_i. For instance, a larger projection on $\gamma_0(t)$ could compensate for a particular time-shift τ_i. To this end, given $g^{\mathbf{M}}$ the metric associated to the Riemannian manifold $\mathbf{M}$, we must ensure that $g^{\mathbf{M}}(\mathbf{w}_i, \mathbf{v}_0) = 0$.

As $\mathbf{w}_i = As_i$ and A_k is the k-th column of A, the orthogonality condition is ensured if $g^{\mathbf{M}}(A_k, \mathbf{v}_0) = 0 \quad \forall k \in \{1, \dots, N\}$. In practice, we use the Householder method to build an orthonormal basis $(\mathbf{v}_0, B_2, \dots, B_N)$ from which the column A_k is built as a linear combination of $(B_2, \cdots, B_N)$. This ensures the orthogonality condition between $\mathbf{w}_i$ and $\mathbf{v}_0$. The coefficients of the linear combinations for all the columns are denotes $(\beta_k)_{1 \leq k \leq N_s(N-1)}$

1.3.4 Product of 1D models

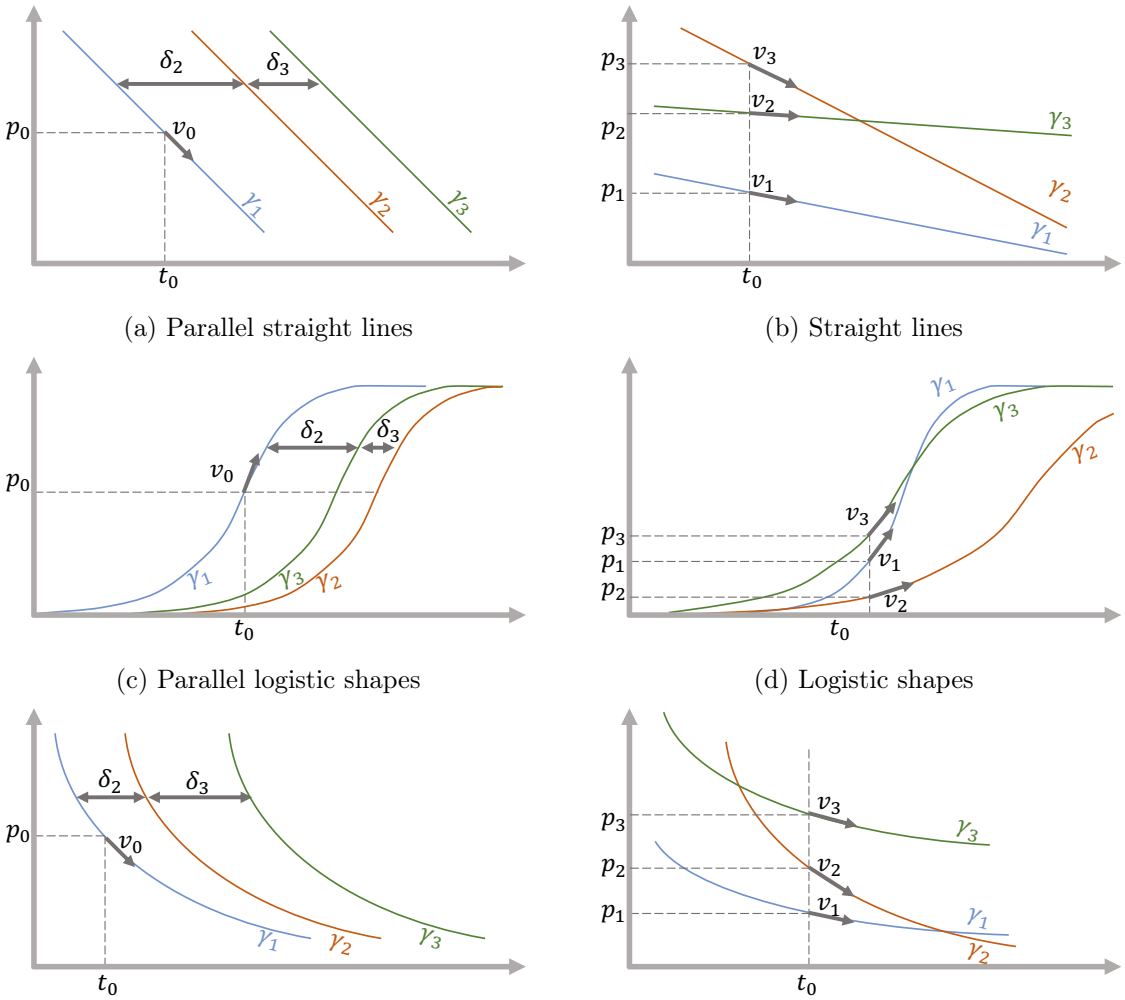
As for now, the model has been set in a very generic way as the manifold has not been described. In fact, this general writing allows to keep the same framework for multiple manifolds, and therefore multiple data types. In the following, we consider that the manifold is a product manifold of 1D manifolds such that $\gamma_0(t) = (\gamma_1(t), \dots, \gamma_n(t))$. In this case, the associated metric is a product of 1D metrics.

Given that w_{ik} the k-th coordinate of $\mathbf{w}_i$, the authors of [Schiratti et al., 2017] show that in the case of a product manifold, the k-th coordinate of $\eta^{\mathbf{w}_i}(\psi_i(t_{ij}))$ writes

$$\eta_k^{\mathbf{w}_i}(t) = \gamma_k(\frac{\mathbf{w}_{ik}}{\dot{\gamma}_k(t_0)} + \psi_i(t_{ij}))$$

1.4 Different instantiations

In this manuscript, we essentially focus on six instantiations (see Fig. 1.3) of the generic model. [Schiratti, 2016] introduced the parallel straight lines and the parallel logistic shapes. On top of them, this manuscript enriches the family of possible temporal profiles with a parallel exponential decay, and with the relaxation of the parallel constraint of the three previous models. For each of the six instantiations, we provide the metric g_p^k at point $p \in \mathbb{R}$, the equation of the corresponding one-dimensional geodesics γ_k, and the writing of the k-th coordinate of the individual trajectory $\eta_k^{\mathbf{w}_i}(\psi_i(t_{ij}))$. The proof that the given curves are geodesics is given in Appendix 1.



(a) Parallel straight lines

(b) Straight lines

(c) Parallel logistic shapes

(d) Logistic shapes

(e) Parallel exponential decays

(f) Exponential decays

Figure 1.3: Instanciations of the generic model for different type of temporal evolutions. Exemple for the progression of three variables.

1.4.1 Parallel straight lines

We consider parallel straight lines as shown on Fig. 1.3a. The k-th geodesic reaches the value p_0 with velocity v_0 at time $t_0 + \delta_k$. For identifiability purposes, we set $\delta_1 = 0$. The k-th coordinate corresponds to an horizontal translation of the first coordinate by δ_k.

In that case, the metric, the geodesic and the individual trajectory write :

$$g_p^k(u, v) = uv \tag{1.2}$$
$$\gamma_k(t) = p_0 + (t - t_0 - \delta_k)v_0 \tag{1.3}$$
$$\eta_k^{\mathbf{w}_i}(\psi_i(t_{ij})) = w_{ik} + p_0 + (\alpha_i(t_{ij} - t_0 - \tau_i) + \delta_k)v_0 \tag{1.4}$$

Note that the first coordinate of $\gamma_0(t_0) = \mathbf{p}_0$ is p_0. Similarly, the first coordinate of $\dot{\gamma}_0(t) = \mathbf{v}_0$ is v_0.

1.4.2 Straight lines

As shown on Fig. 1.3b, in the case of straight lines whose k-th coordinate, i.e. geodesic, reaches the value p_k with velocity v_k at time t_0, we have :

$$g_p^k(u, v) = uv \tag{1.5}$$
$$\gamma_k(t) = p_k + (t - t_0)v_k \tag{1.6}$$
$$\eta_k^{\mathbf{w}_i}(\psi_i(t_{ij})) = w_{ik} + p_k + \alpha_i(t_{ij} - \tau_i - t_0)v_k \tag{1.7}$$

1.4.3 Parallel logistic shapes

We here consider logistic curves whose asymptotic values are 0 at $-\infty$ and 1 at $+\infty$. The first dimension reaches the value p_0 at time t_0 with velocity v_0. The k-th dimension is delayed by a time δ_k from the first coordinate (by definition, $\delta_1 = 0$), as shown on Fig. 1.3c. For the sake of clarity, we write $E(t) = \exp\left(\frac{-v_0}{p_0(1-p_0)}(t + \delta_k - t_0)\right)$. This leads to the following :

$$g_p^k(u, v) = \frac{uv}{p^2(1-p)^2} \tag{1.8}$$
$$\gamma_k(t) = \left(1 + \left(\frac{1}{p_0} - 1\right) E(t)\right)^{-1} \tag{1.9}$$
$$\eta_k^{\mathbf{w}_i}(\psi_i(t_{ij})) = \left(1 + \left(\frac{1}{p_0} - 1\right) \exp\left(\frac{-v_0(\alpha_i(t - \tau_i - t_0) + \delta_k)}{p_0(1-p_0)} - \frac{w_{ik}(1 + gE(t_0))^2}{gE(t_0)}\right)\right)^{-1} \tag{1.10}$$

As for the parallel straight lines model, the first coordinate of $\gamma_0(t_0) = \mathbf{p}_0$ is p_0. Similarly, the first coordinate of $\dot{\gamma}_0(t) = \mathbf{v}_0$ is v_0.

1.4.4 Logistic shapes

Contrary to the parallel logistic shapes (see Fig. 1.3d), we here consider that each coordinate is independant. Therefore, the k-th geodesic reaches the value p_k at time t_0 with velocity v_k. This writes :

$$g_p^k(u,v) = \frac{uv}{p^2(1-p)^2} \tag{1.11}$$

$$\gamma_k(t) = \left(1 + \left(\frac{1}{p_k} - 1\right) \exp\left(-\frac{v_k}{p_k(1-p_k)}\,(t - t_0)\right)\right)^{-1} \tag{1.12}$$

$$\eta_k^{\mathbf{w}_i}(\psi_i(t_{ij})) = \left(1 + \left(\frac{1}{p_k} - 1\right) \exp\left(-\frac{w_{ik} + v_k\alpha_i(t_{ij} - t_0 - \tau_i)}{p_k(1-p_k)}\right)\right)^{-1} \tag{1.13}$$

1.4.5 Parallel exponential decays

We here consider geodesics that take the form of exponential decay such that the asymptotic values are $+\infty$ at $-\infty$ and 0 at $+\infty$. The first geodesic reaches p_0 with velocity v_0 at time t_0. The other geodesic are translation of the geodesic with $t \mapsto t + \delta_k$ ($\delta_1 = 0$ by definition) as shown on Fig. 1.3e. It leads to consider the following :

$$g_p^k(u,v) = \frac{uv}{p^2} \tag{1.14}$$

$$\gamma_k(t) = p_0 \exp\left(-\frac{v_0}{p_0}(t - t_0 + \delta_k)\right) \tag{1.15}$$

$$\eta_k^{\mathbf{w}_i}(\psi_i(t_{ij})) = p_0 \exp\left(-\frac{w_i k}{p_0} - \frac{v_0}{p_0}(\alpha_i(t_{ij} - t_0 - \tau_i) + \delta_k)\right) \tag{1.16}$$

$$\tag{1.17}$$

Note that the mentioned asymptotic values are correct only if $v_k > 0$. In the case that $v_k < 0$, this corresponds to a classic exponential function.

1.4.6 Exponential decays

Here, the geodesics take the form of exponential decay such that the asymptotic values are $+\infty$ at $-\infty$ and 0 at $+\infty$. They reach the value p_k at time t_0 with velocity v_k as shown on Fig. 1.3f. This writes :

$$g_p^k(u,v) = \frac{uv}{p^2} \tag{1.18}$$

$$\gamma_k(t) = p_k \exp\left(-\frac{v_k}{p_k}(t - t_0)\right) \tag{1.19}$$

$$\eta_k^{\mathbf{w}_i}(\psi_i(t_{ij})) = p_k \exp\left(-\frac{w_{ik}}{p_k} - \frac{v_k\alpha_i}{p_k}(t_{ij} - t_0 - \tau_i)\right) \tag{1.20}$$

Again, the mentioned asymptotic values are correct only if $v_k > 0$. Otherwise, it is an increasing exponential function.

1.4.7 Model variations

The list of possible instantiations of the generic model is not limited to the six mentioned models. This section intends to present some variations of the previous models. We stress that some of them are available in the Leaspy software, presented in Chapter 7.

Parametrization

For each instantiation, there are multiple ways to parametrize the model. For instance, the parallel straight lines are parametrized by a value p_0 reached with velocity v_0 at different times : $t_0, t_0 + \delta_2, \ldots, t_0 + \delta_n$. It somehow corresponds to an horizontal translation of magnitude δ_k of the first coordinate. The same straight lines could have been parametrized by values $p_1, \ldots, p_n$ reached at time t_0 with the velocity v_0. This second writing corresponds to a horizontal translation of magnitude $p_k - p_0$ of the first coordinate. Similar reparametrization are possible for the logistic shapes and the exponential decays. However, as discussed in the next Chapter, some parametrization are easier to estimate in practice. Therefore, the aforementioned models correspond to the one that have shown to be stable and robust during their estimation.

Univariate models

In the case of one-dimensional data, it is possible to use the previous models by setting $\mathbf{w}_i = (\mathbf{w}_{i1}) = 0$. For instance, the logistic parallel writes

$$\gamma(t) = \left(1 + \left(\frac{1}{p_0} - 1\right) \exp\left(-\frac{v_0}{p_0(1-p_0)}(t - t_0)\right)\right)^{-1}$$

.

All the previous and further equations holds true for univariate data.

Number of sources

The number of sources N_s defines the subspace of possible directions for the space-shift $\mathbf{w}_i$. Its selection is led by the fact that a too small value does not span $\mathbb{R}^N$ sufficiently while a too large value saturates the span of possible directions, not mentioning the associated computational cost. This is reflected in the estimation of the noise which decreases with a high number of sources until its saturation. The best value is thus the minimal value of N_s that saturates the noise.

Another option is to consider that $\mathbf{w}_i = 0$ for all the patients. This might be of interest for very noisy data where the spatial variability is a useless degree of freedom.

1.4.8 Model selection

As the model that best fit some data might not be straightforward, we here give some hints concerning its selection. The main factors to take into considerations are the following :

- **Biological process at hand.** The model should definitely depicts what happens in the data or what might be expected. For instance, in the case of cognitive assessments for which there exists minimum and maximum values, the logistic progression is a natural choice. On the other hand, the decrease of a volumetric feature might be modeled by an exponential decay or a linear decrease.

- **Number of individuals.** To reconstruct a long-term progression, the individuals should represent different part of the disease stage, along with its variability in the population.

- **Number of time points.** Similarly to the number of individuals, there should be enough observations per patient to span a sufficient disease stage. In practice, this feature is more important than the number of individuals as it tends to stabilize the disease progression.

Chapter 2

Estimation

The third chapter presents the algorithms used to estimate mixed-effect models, as well as their scope of applications and their limitations. In a second time, we exhibit how these mathematical procedures are turned into applications in the case of disease progression. It turns the spatiotemporal model of disease progression into a natural framework to handle longitudinal data.

Contents

2.1 Statistical learning

In this session, we introduce the algorithms used to estimate the parameters of statistical models. We stress the fact that the notations do not correspond to the sections above. Here, we consider that the model at hand can be written for the i-th subject as :

$$\mathbf{y}_i = f(\theta, \mathbf{z}_i) + \epsilon_i \,,$$

where

- $\mathbf{y}_i$ is the output variable (i.e. explained variable),
- $\mathbf{z}_i$ are the random variables associated to the i-th subject, also called *latent variables*,
- f corresponds to the model, i.e. the mapping from the model parameters and the random variables to the output variable,
- θ corresponds to the model parameters that might be (i) the parameters of the function f and (ii) the parameters of the latent variable distributions.

In the case of longitudinal data, y_i is a set of observations at different timepoints $y_i = (y_{ij})_j$.

In general, the objective is to find the parameters that "best describe" the observations, e.g. the parameters that maximize the likelihood

$$p(\mathbf{y}; \theta) = \int p(\mathbf{y}, \mathbf{z}; \theta) \, \mathrm{d}\mathbf{z} = \int p(\mathbf{y} \mid \mathbf{z}; \theta) p(\mathbf{z}; \theta) \, \mathrm{d}\mathbf{z}$$

2.1.1 E-M algorithm

It is sometimes difficult to maximize this likelihood directly, especially when it relies on latent variables $\mathbf{z}$ that are by definition unknown. In these cases, the Expectation-Maximization algorithm, introduced in [Dempster et al., 1977], provides an iterative algorithm to estimate either the *maximum likelihood* or the *maximum a posteriori*.

Algorithm 1: Expectation-Maximization

$\theta \leftarrow \theta_0$
$k \leftarrow 0$
while *Convergence of* θ **do**
  $k \leftarrow k + 1$
  Expectation step : Compute $Q(\theta, \theta^{(k)}) = \mathbb{E}_{p(\mathbf{z}|\mathbf{y};\theta^{(k)})}\left[\log p(\mathbf{y}, \mathbf{z}; \theta)\right]$
  Maximization step : Update $\theta^{(k+1)} = \text{argmax}_\theta \, Q(\theta, \theta^{(k)})$
end
Result: return $\theta^{(k)}$

The algorithm alternates between an Expectation step which defines a function that computes the expectation over the latent variables given a current value of the parameter θ^k at the k-th step of the algorithm, and, a Maximization step that finds parameters $\theta^{(k)}$ that maximize the previous function. The pseudo code of the algorithm is presented in Algorithm 1.

2.1.2 Stochastic Approximization Expectation Maximization

Algorithm 2: Stochastic Approximation Expectation-Maximization

$\theta \leftarrow \theta_0$
$Q_0 \leftarrow 0$
$(\epsilon_k)_{k\geq 0}$ such that $\sum_{k\geq 0} \epsilon_k = +\infty$ and $\sum_{k\geq 0} \epsilon_k^2 < +\infty$
$k \leftarrow 0$
while *Convergence of* θ **do**
  $k \leftarrow k + 1$
  Stochastic step $\mathbf{z}^{(k)} \sim p(\mathbf{z} \mid \mathbf{y}; \theta^{(k)})$
  Approximation step : Compute
  $Q_k(\theta) = Q_{k-1}(\theta) + \epsilon_k(\log p(\mathbf{y}, \mathbf{z}^{(k)}; \theta) - Q_{(k-1)})$
  Maximization step : Update $\theta^{(k+1)} = \text{argmax}_\theta \, Q_k(\theta)$
end
Result: return $\theta^{(k)}$

In the case of complex non-linear models, the integral over the latent variables during the Expectation step is intractable. On the other hand, it is possible to draw realizations $\mathbf{z}^{(k)} \sim p(\mathbf{z} \mid \mathbf{y}; \theta^{(k)})$ in order to approximate the quantity $Q(\theta \mid \theta^{(k)})$. This approximation leads to replace the Expectation step by a Stochastic and Approximization step. This algorithm, called Stochastic Approximation Expectation Maximization algorithm and introduced in [Delyon et al., 1999] is shown in Algorithm 2.

The key point of the algorithm is to require only one sample $\mathbf{z}^{(k)}$ per iteration rather than a Monte Carlo approximation of $\mathbb{E}_{p(\mathbf{z}|\mathbf{y};\theta^{(k)})}\left[\log p(\mathbf{y}, \mathbf{z}; \theta)\right]$. Besides computational cost savings, it also relies on the fact that $z^{(k)}$ is correctly sampled. This is particularly

important in the context of complex non-convex energy landscapes. Early phases of the Approximation step ($\epsilon_k \approx 1$) boils down to $Q_k(\theta) = \log p(\mathbf{y}, \mathbf{z}^{(k)}; \theta)$. This memoryless period of the algorithm (as it does not record previous values of the quantity $Q_k(\theta)$) is called the *burn-in* phase. In practice, this exploratory phase is critical for its convergence. After this phase, the parameters $\theta^{(k+1)}$ is updated given the new value of the parameters $\theta(k)$ and the previous value $\theta^{(k-1)}$.

We refer the reader to [Delyon et al., 1999] for technical details, especially the hypothesis to prove the convergence of this algorithm. Among others, it is shown that the Stochastic and Approximation steps asymptotically converge to the same set of solutions as the Expectation step of the EM algorithm.

2.1.3 Monte Carlo Markov Chain SAEM

Algorithm 3: Hasting Metropolis algorithm

Given $z^{(k)}$

begin

Choose a proposition law $q_k(\cdot \mid \mathbf{z}^{(k)})$

Draw $\mathbf{z}^c \sim q_k(\cdot \mid \mathbf{z}^{(k)})$

Update $\mathbf{z}^{(k+1)} = \mathbf{z}^c$ with probability $\tau = \min\left(\frac{p(\mathbf{z}^c|\mathbf{y};\theta^{(k)})q_k(\mathbf{z}^{(k)}|\mathbf{z}^c)}{p(\mathbf{z}^{(k)}|\mathbf{y};\theta^{(k)})q_k(\mathbf{z}^c|\mathbf{z}^{(k)})}, 1\right)$

($\mathbf{z}^{(k+1)} = \mathbf{z}^{(k)}$ otherwise)

end

Result: return $\mathbf{z}^{(k+1)}$

Algorithm 4: Monte Carlo Markov Chain Stochastic Approximation Expectation-Maximization

$\theta \leftarrow \theta_0$

$\tilde{S}_0 \leftarrow 0$

$(\epsilon_k)_{k\geq 0}$ such that $\sum_{k\geq 0} \epsilon_k = +\infty$ and $\sum_{k\geq 0} \epsilon_k^2 < +\infty$

$k \leftarrow 0$

while *Convergence of* θ **do**

$k \leftarrow k + 1$

Simulation step Given $(\mathbf{y}, \theta^{(k)})$, sample $\mathbf{z}^{(k)}$ from $\mathbf{z}^{(k-1)}$ with an Hasting Metropolis procedure

Stochastic Approximation step : $\tilde{S}_k \leftarrow \tilde{S}_k + \epsilon_k \left(S(\mathbf{y}, \mathbf{z}^{(k)}) - \tilde{S}_{k-1}\right)$

Maximization step : $\theta_k = \text{argmax}_\theta \left(-\log C(\theta) + \langle \tilde{S}_k, \Phi(\theta)\rangle\right)$

end

Result: return $\theta^{(k)}$

One additional difficulty lies in the fact that $p(\mathbf{z} \mid \mathbf{y}; \theta^{(k)})$ might be unknown. We derive from the Bayes rule that

$$p(\mathbf{z} \mid \mathbf{y}; \theta^{(k)}) = \frac{p(\mathbf{y} \mid \mathbf{z}; \theta^{(k)})p(\mathbf{z}; \theta^{(k)})}{p(\mathbf{y}; \theta^{(k)})} = \frac{p(\mathbf{y} \mid \mathbf{z}; \theta^{(k)})p(\mathbf{z}; \theta^{(k)})}{\int p(y \mid \mathbf{z}; \theta^{(k)})p(\mathbf{z}; \theta^{(k)})\,\mathrm{d}z}$$

where $p(\mathbf{y} \mid \mathbf{z}; \theta^{(k)})$ is the model itself, $p(\mathbf{z}; \theta^{(k)})$ is the known prior of the hidden variables and the denominator is a constant.

Therefore, $p(\mathbf{z} \mid \mathbf{y}; \theta^{(k)})$ is known only up to the normalizing constant. In such cases, the Hasting Metropolis algorithm presented in Algorithm 3 allows to approximate sample $\mathbf{z}^{(k+1)} \sim p(\mathbf{z}^{(k)} \mid \mathbf{y}; \theta^{(k)})$. It relies on a Markov Chain method to draw hidden variables from the probability distribution $p(\mathbf{z}^{(k)} \mid \mathbf{y}; \theta^{(k)})$.

Once we know how to sample the latent variable $\mathbf{z}^{(k)}$, this leads to Monte Carlo Markov Chain SAEM algorithm whose sampling procedure is replaced by the MCMC procedure. This algorithm is proven to converge to a critical point of the observed likelihood $p(\mathbf{y}|\theta)$ (see [Kuhn and Lavielle, 2004, Allassonnière et al., 2010, Allassonniere and Kuhn, 2015]) that is likely to be a local maximum due to the randomness of the algorithm which makes it diverge from saddle points. One of the convergence hypothesis is that the model belongs to the exponential family, which means that the log-likelihood writes :

$$\log p(\mathbf{y}, \mathbf{z}; \theta) = \langle \Phi(\theta), S(\mathbf{y}, \mathbf{z}) \rangle - \log C(\theta)$$

where $S(\mathbf{y}, \mathbf{z})$ are called sufficient statistics of the model. In such cases, the Approximation step rewrites as $Q_{k+1} = \langle \Phi(\theta), \tilde{S}_{k+1} \rangle - \log C(\theta)$ where $\tilde{S}_{k+1} = (1 - \epsilon_{k+1})\tilde{S}_k + \epsilon_{k+1} S(\mathbf{y}, \mathbf{z})$. Besides the convergence properties, this writing allows a significant computation cost reduction at each iteration of the algorithm as one only need to compute and propagate the vector $S(\mathbf{y}, \mathbf{z}^{(k)})$ instead of the entire quantity $Q(\theta, \theta^{(k)})$. Finally, the Maximization step computes $\theta^{(k+1)} = \text{argmax}_\theta \langle \Phi(\theta), \tilde{S}_{k+1} \rangle) - \log C(\theta)$. This algorithm is fully described in Algorithm 4. The reader is referred to [Kuhn and Lavielle, 2004] and [Allassonnière et al., 2010] for the proof of convergence in the case of a model belonging to the exponential family.

2.1.4 Hasting Metropolis within Gibbs sampler

Algorithm 5: Gibbs sampler

Given a set of hidden variables $\mathbf{z}^{(k)} = (\mathbf{z}_l^{(k)} 1 \leq l \leq L)$ and the parameters $\theta^{(k)}$

begin

Sample $\mathbf{z}_1^{(k+1)}$ from $q(\mathbf{z}_1 \mid \mathbf{y}, \mathbf{z}_1^{(k)}, \ldots, \mathbf{z}_L^{(k)}, \theta^{(k+1)})$

...

Sample $\mathbf{z}_l^{(k+1)}$ from $q(\mathbf{z}_l \mid \mathbf{y}, \mathbf{z}_1^{(k+1)}, \ldots, \mathbf{z}_{l-1}^{(k+1)}, \mathbf{z}_{l+1}^{(k)} \ldots, \mathbf{z}_L^{(k)}, \theta^{(k+1)})$

...

Sample $\mathbf{z}_L^{(k+1)}$ from $q(\mathbf{z}_L \mid \mathbf{y}, \mathbf{z}_1^{(k+1)}, \ldots, \mathbf{z}_{L-1}^{(k+1)}, \mathbf{z}_L^{(k)}, \theta^{(k+1)})$

end

Result: return $\theta^{(k)}$

In the previous section, we considered the hidden variable $\mathbf{z}^{(k)}$ is sampled thanks to a Markov Chain method which might be difficult in case of complex multivariate variables. Let us now consider that $\mathbf{z} = (\mathbf{z}_l)_{1 \leq l \leq L}$ and that for all $l \in \{1, \ldots, L\}$, the law $p(\mathbf{z}_l \mid \mathbf{z}_{\bar{l}}, \mathbf{y}; \theta)$ is known only up to a constant ($\mathbf{z}_{\bar{l}} = \mathbf{z} \backslash \{\mathbf{z}_l\}$), then the coordinates of $\mathbf{z}$ can be sampled one after the other. This procedure is called the Gibbs sampler and is described in Algorihm 5. If each sampling of the Gibbs sampler is done using a Hasting Metropolis procedure, we call this algorithm the Hasting Metropolis within Gibbs sampler. It leads to rewrite the MCMC step of Algorithm 4 with a coordinate by coordinate Hasting-Metropolis sampling.

2.2 Estimation of the disease progression model

In Chapter 1, thanks to the Riemannian setting, we presented a geometrical model of spatiotemporal progression as shown in Eq. 1.3.1. The model is parametrized, first,by geometrical parameters θ_{geom}, that depend on its instantiation, and, secondly, by variations of the group-average progression that take the form of individual hidden variables $\mathbf{z}_i = (\alpha_i, \tau_i, (s_{ij})_{1\leq j\leq N_s})$. For instance, $\theta_{\text{geom}} = ((p_k)_{1\leq k\leq N}, (v_k)_{1\leq k\leq N}, t_0, (\beta_k)_{1\leq k\leq (N-1)N_s}$ for the logistic curves. Given a set of longitudinal data $\mathbf{y} = (\mathbf{y}_{ij}, t_{ij})_{1\leq i\leq p, 1\leq j\leq k_i}$, Eq. 1.3.1 can be written as

$$y_{ij} = f(\theta_{\text{geom}}, \mathbf{z}_i, t_{ij}) + \epsilon_{ij} \tag{2.1}$$

Note first that we consider $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. Given that θ_{geom} corresponds the group-average trajectory and $\mathbf{z}_i$ the individual variations to the mean, Eq. 2.2 indeed characterizes a mixed-effects model as describes in Section 1.2. In that case, the hidden variables are considered as realizations of random variables which distribution also depends on parameters $\theta_{\mathbf{z}}$ such that it leads to consider the overall statistical set of parameters $\theta = (\theta_{\text{geom}}, \theta_{\mathbf{z}})$ which describes the disease progression along with its variability within the population (note that $\sigma \in \theta_{\mathbf{z}}$). The likelihood of this statistical model writes $p(\mathbf{y}; \theta) = \int p(\mathbf{y}|\mathbf{z}; \theta) p(\mathbf{z}; \theta)\, \mathrm{d}z$ where $p(\mathbf{y}|\mathbf{z}; \theta)$ is an attachment term derived from by Eq. 2.2 while $p(\mathbf{z}; \theta)$ is a regularity term that corresponds to the prior on the hidden variables.

This naturally leads to characterize the following statistical procedures :

- **calibration** : given $\mathbf{y}$, estimation of the parameters $\hat{\theta}$ that best describe the group-average spatiotemporal trajectory and its variability.
- **personalisation** : given $\hat{\theta}$, estimation of the individual hidden variables $\mathbf{z}_i^*$ that best derive the group-average trajectory to reconstruct the individual measurements.
- **reconstruction**, **imputation of missing values** and **future prediction** : given $\hat{\theta}$ and $\mathbf{z}_i^*$, estimation of $\tilde{y}_{ij} = f(\hat{\theta}_{\text{geom}}, \mathbf{z}_i^*, t_{ij})$ where the time t_{ij} has been observed $t_{ij} \in \{t_{i1}, t_{ik_i}\}$ (reconstruction), or where it is between the first and last seen visit $t_{i1} < t_{ij} < t_{ik_i}$ (missing value imputation) or where it is after the last seen visit $t_{ij} > t_{ik_i}$ (prediction).
- **simulation** : given $\hat{\theta}$ and a set of individual variables $(\mathbf{z}_i)$, drawing of a new hidden variable $\mathbf{z}_{i'}$ that entirely determine a new individual.

2.3 Calibration

The calibration procedure aims to find the value $\hat{\theta}$ in the set of possible parameters Θ that best describes the model i.e. that maximizes the likelihood $p(\mathbf{y}; \theta)$. The model of disease progression and its variability within the population $\mathbf{y}$ are fully characterized by this optimal value $\hat{\theta} \in \text{argmax}_\Theta\, p(\mathbf{y}; \theta)$. As the hidden variables $(\mathbf{z}_i)_{1\leq i\leq p}$ are unknown, the observed loglikelihood is intractable. Same for the distribution $p(\mathbf{z}|\mathbf{y}; \theta)$ which is only known up to a normalizing constant (see section 2.1.3). In such cases, the MCMC-SAEM algorithm has been proven to be a helpful algorithm to find $\hat{\theta}$ as it is proven to converge to towards a local maximum of the posterior distribution $p(\theta|\mathbf{y})$ [Kuhn and Lavielle, 2004, Allassonnière et al., 2010, Allassonnière et al., 2015]. This has to be put in regards with the existence of a maximum a posteriori of the presented model, proven in [Schiratti, 2016].

The convergence of the algorithm is proven under the assumption that the model belongs to the exponential family such that the complete log-likelihood writes

$$\log p(\mathbf{y}, \mathbf{z}; \theta) = \langle \Phi(\theta), S(\mathbf{y}, \mathbf{z}) \rangle + \log C(\theta) \,,$$

where $S(\mathbf{y}, \mathbf{z})$ are called sufficient statistics of the model as describes in 2.1.3. In our case, this writing is possible only if the fixed effects are "exponentialized" in the sense that they are considered as samples of a random variable. For instance, for a fixed-effect p, this translates into $p \sim \mathcal{N}(\overline{p}, \sigma_p^2)$ such that $\overline{p} \in \theta$ and σ_p is fixed and chosen small enough to assimilate the hidden variables $p \in \mathbf{z}$ to the model parameter $\overline{p} \in \theta$.

On top of this theoretical convergence consideration, we here also stress the identifiability properties of the model. As discussed in [Lavielle and Aarons, 2016] and [Lavielle, 2014], the model should be considered throughout its structural and practical identifiability. The former implies that there does not exist two sets of parameters θ_1 and $\theta_2 \neq \theta_1$ such that $p(\mathbf{y}; \theta_1) = p(\mathbf{y}; \theta_2)$ (the definition of the structural identifiability is in fact more complex in the case of individual and population parameters, see [Lavielle and Aarons, 2016]). On the other hand, the practical identifiability is related to the quantity and quality of the measurements that ensure a stable and robust estimation of θ. In the context of the presented model, the structural identifiability has not been fully proven but some criterion have been raised such as the orthogonality condition between $\mathbf{w}_i$ and $\mathbf{v}_0$, discussed in section 1.3.3, or such as the choice of θ_{geom} describing the temporal shape of progression. On the other hand, the practical identifiability has been benchmarked to lead to a reparametrization $\theta \leftarrow \zeta_1(\theta)$ and $\mathbf{z} \leftarrow \zeta_2(\mathbf{z})$ where ζ_1 and ζ_2 are invertible functions. Appendix I presents the reparametrization adopted for each model instantiation along with the corresponding sufficient statistics and the maximization step of the MCMC-SAEM algorithm.

2.4 Personalization : estimate individual random effects

In this section, we consider a set of longitudinal data $\mathbf{y}$ and the optimal value $\hat{\theta}$ that maximizes $p(\mathbf{y}; \theta)$. Let's $\mathbf{y}_i = (\mathbf{y}_{ij}, t_{ij})_{1 \leq j \leq k_i}$ be a set of k_i observations of the same patient, indexed by i. The personalization consists in finding the optimal value of the random effects $\mathbf{z}_i^*$ that maximizes the likelihood $p(\mathbf{y}_i, \mathbf{z}_i; \hat{\theta}) = p(\mathbf{y}_i | \mathbf{z}_i; \hat{\theta}) p(\mathbf{z}_i; \hat{\theta})$. This optimization procedure is essentially realized by a quasi Newton's method called L-BFGS [Byrd et al., 1995] or Powell's method [Powell, 1964]. Both methods are numerical schemes that find the optimal value $\mathbf{z}_i^*$ with no analytical need of the gradient of the likelihood $p(\mathbf{y}_i, \mathbf{z}_i; \hat{\theta})$ with respect to $\mathbf{z}_i$. An alternative method is to consider the estimated parameters $\hat{\theta}$ and to run K iterations of the MCMC-SAEM for the individual parameters only. It results, for each patient, in an empirical distribution of K samples of the individual parameters $(\mathbf{z}_i^k)_{1 \leq k \leq K}$ from which one can use the mode as the estimated individual parameters.

Remark: It is either possible to consider that the personalized individual belongs to the set used for calibration ($\mathbf{y}_i \in \mathbf{y}$) or not ($\mathbf{y}_i \notin \mathbf{y}$). In practice, both cases lead to similar results (see Chapter 5). In the first case, cautious reader might mention that the MCMC-SAEM procedure draw $\mathbf{z}_i$ samples. However, this value is not optimal as it first is a draw for a value $\theta^{(k)}$ that might differ from $\hat{\theta}$, and, secondly, it is a sample of a random variable, not its mode.

Remark: Maximizing $p(\mathbf{y}_i, \mathbf{z}_i; \hat{\theta}) = p(\mathbf{y}_i | \mathbf{z}_i; \hat{\theta}) p(\mathbf{z}_i; \hat{\theta})$ corresponds to the maximization of a regularity term ($p(\mathbf{z}_i; \hat{\theta})$) and an attachment term ($p(\mathbf{y}_i | \mathbf{z}_i; \hat{\theta})$) that is a sum over the number of observed time-points. This means that the more observations a patient present, the less important the regularity becomes, or equivalently, the more confidence we are in the fact that the individual measurements might different from the group-average trajectory.

2.5 Reconstruction, missing value imputation and future prediction

Let us consider $\hat{\theta}$ and the observations of a subject $\mathbf{y}_i = (\mathbf{y}_{ij}, t_{ij})_{1 \leq j \leq k_i}$ at times $(t_{i1}, \ldots, t_{ik_i})$ such that $t_{i1} < \ldots < t_{ik_i}$. The personalization procedure returns $\mathbf{z}_i^*$ that maximizes $z \mapsto p(\mathbf{y}_i, z; \hat{\theta})$, which fully specifies the individual spatiotemporal trajectory, for any t.

In the case where t is chosen among $\{t_{i1}, \ldots, t_{ik_i}\}$, the resulting $\tilde{y}_{ij} = f(\theta, \mathbf{z}_i^*, t_{ij})$ is called the *reconstruction* of the data y_{ij}. The difference $||y_{ij} - \tilde{y}_{ij}||$ is referred to as the *reconstruction error*.

In the case where t such that $t_{i1} < t < t_{ik_i}$ and $t \notin \{t_{i1}, \ldots, t_{ik_i}\}$, $\tilde{y}_{ij} = f(\theta, \mathbf{z}_i^*, t_{ij})$ corresponds to an interpolated value that might be missing as it has not been observed in the initial dataset. This corresponds to the *imputation of missing values*.

Finally, if t is chosen such that $t > t_{ik_i}$, it corresponds to the *prediction* of the measurements at future time points. The accuracy of such procedure depends on the belief that the group-average scenario describes a long-term scenario of change which can is transposed to the individual trajectory. For the same reason, predicting stages that have not been seen in the original training database is unrealistic since it relies on unknown dynamics.

2.6 Simulation

Given a longitudinal dataset $\mathbf{y} = (\mathbf{y}_{ij}, t_{ij})_{1 \leq i \leq p, 1 \leq j \leq k_i}$, the previous procedures allows us to get $\hat{\theta}$ and $(\mathbf{z}_i)_{1 \leq i \leq p}$ that defines a empirical distribution of the individual parameters from which one can eventually draw a new sample $\mathbf{z}$. Along with $\hat{\theta}$, this sample entirely defines an individual spatiotemporal trajectory. This simulation procedure aims to sample virtual subjects resulting in a simulated cohort that reproduce the characteristics of the original cohort. It enables to have

- more patients,
- more follow-up visits per patient,
- a finer temporal granularity, i.e. less time between visits.

Intuitively, one can think of the virtual patient as a "linear combination" of the real patients, e.g. if a patient has a cognitive decline of 2 points per year (given a specific neuro-psychological assessment) and another of 2.4 points per year, then a patient with a decline of 2.1 points per year is likely to exist even though he is not present in the original database. Here, the procedure allows to extend such analogy with longitudinal trajectories and potentially complex multi-dimensional observations.

The simulation procedure highly depends on the empirical distribution of the hidden variables $(\mathbf{z}_i)_{1 \leq i \leq p}$. Among others, it is possible to use a Kernel Density Estimation (KDE) or to estimate the parameters of common multivariate distributions, or a combination of both. The quality of the simulation can be assessed by comparing the original empirical distribution and the simulated hidden variables (Kullback-Leibler divergence, statistical tests as Kolgomorov Smirnov for one dimensional distributions, ...) or by comparing the distribution of the final outputs $\mathbf{y}_{\text{virtual}}$ to the original cohort $(\mathbf{y}_{ij})$. We show in Chapter 6, which extensively focuses on simulating data, that it is possible to fool a discriminator in predicting whether a longitudinal observation is real or simulated.

Remark 1: In general, for a subject in the original dataset, there is a correlation between the initial observation at t_{i1}, the number of observations k_i and the individual parameters $\mathbf{z}_i$. For instance, early progressors are included in the dataset at early ages and might be observed during longer periods of time. Therefore, if the goal is to reproduce the original

cohort identically, one need to draw the hidden variables along with these cofactors. On the other hand, it is possible to unbias the original cohort for such inclusion biases.

Remark 2: The corollary of the previous remark is that the cohort might be biased towards subjects with more visits or that are included in protocols at early stages. Another bias might happens if a certain cofactor is prevalent in the original cohort (e.g. more patients with a given genetic mutation, more educated patients, more married persons, ...). Drawing virtual cohorts might help in unbiasing the original ones by balancing the different groups.

Part II

Progression of Spatiotemporal Patterns for Spatially Structured Data

Chapter 3

Population and Individual Spatiotemporal Patterns of Progression from Longitudinal Manifold-Valued Networks

This chapter instantiates the generic mixed-effects model to data that present a spatial structure, such as images, or variation of a signal over the nodes of a mesh. The spatial structure imposes a regularity assumption on the temporal profile of "close" areas. It chapter corresponds to the article Spatiotemporal propagation of the cortical atrophy: Population and individual patterns, *Koval I., Schiratti J.-B., Routier A., Bacci M., Colliot O., Allassonniére S, and Durrleman S., in* Frontiers in Neurology, *2018. This article is itself a detailed version of the conference paper* Statistical Learning of Spatiotemporal Patterns from Longitudinal Manifold-Valued Networks, *Koval I., Schiratti J.-B., Routier A., Bacci M., Colliot O., Allassonniére S, and Durrleman S., in* International Conference on Medical Image Compyting and Computer Assisted Intervention, *2017.*

The results presented in this chapter are based on experiments that rely on `Leasp`*, a C++ code that is available at* `gitlab.com/icm-institute/aramislab/leasp`*. This code, presented in Chapter 7 is left for reproducibility purposes but is not maintained anymore as it is currently under migration to* `Leaspy`*, presented in the same Chapter.*

Contents

Abstract

Repeated failures in clinical trials for Alzheimer's Disease (AD) have raised a strong interest for the prodromal phase of the disease. A better understanding of the brain alterations during this early phase is crucial to diagnose patients sooner, to estimate an accurate disease stage and to give a reliable prognosis.

According to recent evidence, structural alterations in the brain are likely to be sensitive markers of the disease progression. Neuronal loss translates in specific spatiotemporal patterns of cortical atrophy, starting in the enthorinal cortex and spreading over other cortical regions according to specific propagation pathways.

We developed a digital model of the cortical atrophy in the left hemisphere from prodromal to diseased phases, which is built on the temporal alignment and combination of several short-term observation data to reconstruct the long-term history of the disease. The model not only provides a description of the spatiotemporal patterns of cortical atrophy at the group level, but also shows the variability of these patterns at the individual level in terms of difference in propagation pathways, speed of propagation and age at propagation onset.

Longitudinal MRI datasets of patients with mild cognitive impairments who converted to AD are used to reconstruct the cortical atrophy propagation across all disease stages. Each observation is considered as a signal spatially distributed on a network, such as the cortical mesh, each cortex location being associated to a node. We consider how the temporal profile of the signal varies across the network nodes.

We introduce a statistical mixed-effect model to describe the evolution of the cortex alterations. To ensure a spatiotemporal smooth propagation of the alterations, we introduce a constraint on the propagation signal in the model such that neighboring nodes have similar profiles of the signal changes. Our generative model enables the reconstruction of personalized patterns of the neurodegenerative spread, providing a way to estimate disease progression stages and predict the age at which the disease will be diagnosed. The model shows that, for instance, APOE carriers have a significantly higher pace of cortical atrophy but not earlier atrophy onset.

3.1 Introduction

Neuroimaging studies have shown an alteration of the brain structure during the course of Alzheimer's Disease (AD) ([Du et al., 2001, Benzinger et al., 2013]). These lesions appears during the prodromal phase of the disease ([Amieva et al., 2008, Wilson et al., 2011, Mura et al., 2014]) whose observation have been limited due to the absence of clinical symptoms and diagnosis. The importance of the structural changes before the clinical symptoms led to hypothetical models ([Jack et al., 2010]), which have been later refined thanks to the gathering of multiple scientific evidences. These modifications took the form of a structural change of the brain in particular an important neuronal loss and an atrophy of the brain cortex ([Fan et al., 2008, Singh et al., 2006]).The study of the temporal evolution of the cerebral cortex reveals an atrophy of the grey matter ([Baron et al., 2001]). This cortical atrophy presumably relates the traces of the progression of the lesions over the brain surface. A fine-scale modeling of the atrophy propagation is likely to give a wider understanding of the disease evolution, as the structural markers seems reliable to assess the conversion to the AD stage, potentially carrying subtle indicators of the disease progression in early phases.

The spatiotemporal propagation of these alterations encloses two entangled components. On the one hand, the spatial characterization of the lesions over the brain surface at each time, and, on the other hand, a temporal dynamic of these alterations that may differ from one region to another. Characterizing the proper dynamics of these lesions relies on the possibility to reconstruct the whole time-line of AD, at both a spatial and temporal level, out of short-term observations that are not temporally aligned. Another challenging aspect consists in the variability inherent to the individual patterns of atrophy, that requires to consistently compare the subject-specific spreads of alterations. Accounting for the inter-individual variability in term of lesion propagation should allow to reconstruct individual patterns of propagation, paving the way to possible personalized model of atrophy,

that potentially informs on subject-specific age of conversion or disease stage.

Recently, large datasets have opened the opportunity to investigate data-driven models that have refined and validated these hypotheses to some extend, in particular Event-Based Models ([Fonteijn et al., 2012b, Young et al., 2014, Young et al., 2015]) that considers the propagation as a series of events, allowing to define a sequence of disease stages. They characterize the overall variability of the events ordering at a population level. However these models are not well-suited to relate for the temporal delays of the alterations at a population level, neither to determine individual cortical atrophy. Multimodal observations, including Positron Emission Tomography (PET) scans, Magnetic Resonance Imaging (MRI) and biomarkers, have been gathered within longitudinal databases, *i.e.* repeated observations of patients during significant periods of time. The underlying intention is to provide multiple individual snapshots of the disease - patients examined during short-term periods - in order to reconstruct the long-term history of the pathology ([Jedynak et al., 2012], [Donohue et al., 2014]) at a group and individual level. Moreover, it offers the possibility to describe and interpret the observed data contrary to quantiles or percentiles that require arbitrary reference distributions. A challenging aspect of AD patient comparison is the fact that, even though AD is related to age, the latter is not a good proxy of the disease stage ([Gao et al., 1998, Devanand et al., 2007, Bilgel et al., 2016]) leaving us without any easy way to align all the individual on the same time-line. In [Schiratti et al., 2015], the authors introduced mixed-effect model that consider each individual trajectory as a variation of a mean scenario of evolution, with a time-warp function that is able to realign the subjects on the same time-line ([Durrleman et al., 2013]). It allows to characterize a spatial and temporal variability of propagation in the sense that it defines a group-average trajectory of propagation with the possibility to reconstruct individual observations thanks to personalized parameters. Nevertheless, [Schiratti et al., 2015] constrain the model to parallel profiles of progression which does not hold when looking at signals that have various dynamics. Moreover, the model does not take into account the spatial correlations between the data whereas [Bilgel et al., 2016], which focus on spatiotemporal patterns of progression for images, exhibited that this led, in the case of a non-linear mixed effects model, to poor estimations of the subject-specific parameters and individual trajectories.

To account for the spatial structure of the signal, networks have been introduced ([Leuchter et al., 1992, Maguire et al., 1998]), representing the brain areas as the graph nodes. In this paper, the networks correspond to a graph representation of a signal spatially distributed, namely the cortical thickness mapped on a mesh representation of the cortex. The node values are the cortical thickness values over time on the related brain area. Extracting and projecting patients cortical thickness on the common mesh allows to compare their atrophy on the same atlas to exhibit similar patterns. As we expect the signal propagation to be spatially smooth with a similar temporal profile of change for neighbour nodes, we consider that a subset of the graph nodes act as control nodes. They define a evaluation function such that the signal at each node is an interpolation of the signal at the control nodes, enabling to smooth the high frequencies ([Broomhead and Lowe, 1988]). The proximity between nodes is defined by the distance matrix which informs on the distance between any pair. Moreover, the number of nodes of this vertex-based graph can be tuned based on the desired application, potentially the same as the resolution of the input data, *e.g.* a voxel for MRI or PET data.

The aim of this paper is to introduce a model of the cortical atrophy propagation during the long-term course of AD thanks to a graph representation of the neuroimaging data. This model is able to personalize the reconstruction of the propagation to individual longitudinal measurements, allowing to describe the stages of the disease, potentially in the future. The model is described as a general framework for any longitudinal data spatially distributed on a common graph and it is instantiated to exhibit the propagation of the cortical atrophy on the left hemisphere of the brain, across nearly 2.000 regions, thanks to

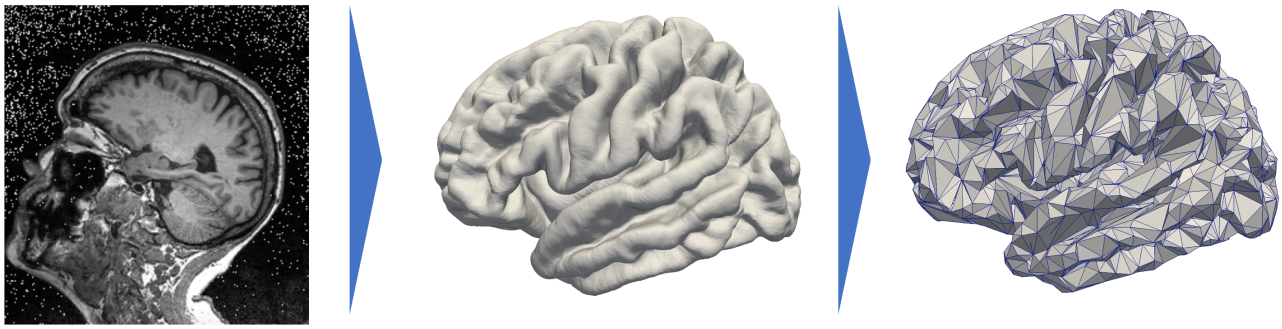


Figure 3.1: Data preprocessing that projects the cortical thickness of the raw MRI observation (left) on a mesh, namely the FSAverage atlas constituted of 163.842 nodes per hemisphere (middle) before sub-sampling it and averaging the signal onto a 1827-node graph (right).

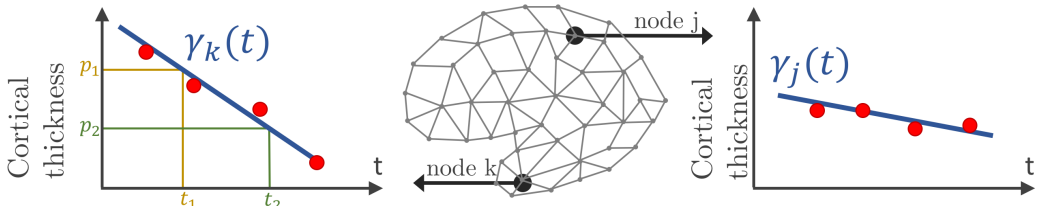
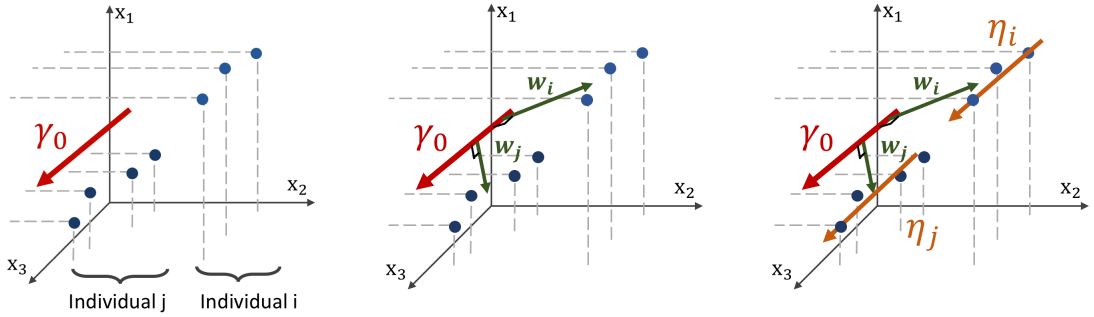


Figure 3.2: Mesh of the cortical surface where each node embeds a time-series of observations (red points). At node k, the function $\gamma_k(t)$, which can be parametrized by a velocity and two different sets (p_1, t_1) or (p_2, t_2), estimates the cortical thickness over time.

longitudinal observations of 154 Mild Cognitive Impaired (MCI) patients that were later diagnosed with AD. While exhibiting an average pattern of propagation, this mixed-effects model allows to reconstruct individual observations through time.



(a) Three dimensional space embedding individual observations (blue points) of two individuals and the mean spatiotemporal trajectory γ_0 (red curve).

(b) The spatial variations from the group-average trajectory γ_0 to the individual observations are captured in individual vectors w_i and w_j, called space-shifts.

(c) The vector w_i is parallel-transported along γ_0 (orange vectors) to define a parallel curve η_i that characterizes the individual spatiotemporal trajectory.

Figure 3.3: Geometric description of the construction of the mean and individual spatiotemporal trajectories in the space of measurement, which is the Riemannian Manifold M that embeds both the real observations and the trajectories

3.2 Materials and Methods

3.2.1 Sketch of the method

Prior to detail our method, we would like to sketch the key ideas and notations of our work to ease and guide the reading. Firstly, we consider I patients ; each patient i is observed J_i times, his jth visit being at age t_{ij}, and each observation led to an MRI scan as shown on the left hand side of Figure 3.1. Segmentation of the cortical thickness, out of the neuroimaging observations, are mapped onto a mesh, as presented on the middle part of Figure 3.1. The last step corresponds to a subsampling process that leds to a graph G of K nodes, characterized by a distance matrix D. At each node, the individual observations define a time-series describing the evolution of the signal through time.

In a second time, we assume that, at each node k of the graph G, there exists a function $t \mapsto \gamma_k(t)$ that describes a characteristic evolution of the signal at this node, as shown on Figure 3.2. The time-series of individual i at node k derives from a continuous function $\eta_{ik}(t)$, which is assumed to be a spatial and temporal variations of the representative trajectory $\gamma_k(t)$, illustrated on Figure 3.3. The temporal variation corresponds to the time realignment of individual i on the common time-line. It adjusts the individual dynamics to a mean pace of evolution, thanks to personalized parameters τ_i and α_i. τ_i stands for the individual time-shift to the mean disease onset, allowing an early ($\tau_i < 0$) or delayed ($\tau_i > 0$) age at diagnosis. The parameter α_i integrates the patient-specific possibility to have a faster ($\alpha_i > 1$) or slower ($\alpha_i < 1$) pace of atrophy compared to the mean scenario of changes. On the other side, the spatial variation corresponds to the adjustment from the mean cortical thickness to individual data. It accounts, for instance, for the difference in size or in spatial thickness distribution at the same disease stage.

We consider that the characteristic signal $\boldsymbol{\gamma}_k(t)$ at node k belongs to a family of curve, here the straight line curves, parametrized by the cortical thickness p_k and the rate of atrophy v_k. To account for the spatial structure of the signal and the large number of nodes, a subset of nodes, referred to as control nodes, is selected to control the interpolation of the cortical and atrophy values over all the nodes. The distribution of the control nodes depends on the size of the kernel bandwidth such that the kernels densities map almost uniformly the feature space.

The model introduces population parameters, that allow to define a characteristic spatiotemporal trajectory of the atrophy, and individual parameters, that not only enable to reconstruct individual trajectories but also permit the statistical study of the distribution of spatiotemporal atrophy patterns. These parameters are estimated thanks to the Monte-Carlo Markov-Chain Stochastic Approximation Expectation Maximization (MCMC-SAEM) algorithm which handle non-linear mixed-effects models, with theoretical guarantees and consistent results in practice.

3.2.2 Subjects and Data Preprocessing

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org

The Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset contains longitudinal MRI data for patients that are, at each visit, either Cognitively Normal (CN), Mild Cognitive Impaired (MCI) patients, or, AD subjects. We selected all the subjects that

presented a monotonous decline from MCI to AD, called the MCI converters, removing those that may convert from AD back to MCI or CN. Although AD patients get through an MCI phase, we could not keep CN to MCI patients as they might just as well convert to another dementia. Also, the patients that underwent from CN to MCI and then to AD are not numerous enough to give robust estimation of early stages (CN to MCI). Thus, we kept only the MCI to AD visits of such patients. All-together, the paper focuses on 154 MCI patients that represents 787 visits, each individual being examined 5 times on average, from 2 to 7 times.

Each visit led to a T1-weighted MRI acquisition, as shown on the left side of Figure 3.1. The longitudinal pipeline of FreeSurfer ([Reuter et al., 2012]) was used to extract the cortical thickness of the left hemisphere of the brain which was then projected on a common atlas, namely FSAverage ([Fischl et al., 1999]), which is a three dimensional mesh composed of 163,842 nodes for each hemisphere represented on the central part of Figure 3.1. This common fixed graph allows to compare the cortical thickness between visits or patients, node to node.

The data acquisition and inter-individual alignment led to a considerable noise, especially in terms of variability in the measures for close nodes. To smooth this noise and to reduce the computational time, we sub-sampled the initial graph into a new graph of 1827 nodes. To do so, we selected 1827 nodes uniformely distributed over the whole FSAverage graph ; the other nodes were then associated to one of the 1827 nodes thanks to a geodesic distance d on the graph ($i.e$ the length of the shortest path on the surface mesh between the nodes) using the Fast Marching Algorithm on the mesh ([Peyré et al., 2010]). Therefore it constitutes collection of nodes referred to as patches. The value of each node of the sub-sampled graph is the average value over the corresponding patch, each being constituted of approximately 89 initial nodes of the FSAverage graph. The resolution of this vertex-based approach is lower than the initial one, shown on the right hand side of Figure 3.1, but still holds the brain topology while smoothing part of the acquisition noise. In our case, each observation can be considered as a vector of size 1827 where the kth coordinate is related to the kth node of the common fixed-graph G. The latter is also described by a the distance matrix D between the 1827 nodes. It was obtained using the geodesic distance d between the 1827 nodes on the initial graph FSAverage, whose edges are weighted by a physical length. Finally, for all $i, j \in \{1, ..., 1827\}$, we set $D_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$ where $\mathbf{x}_i$ and $\mathbf{x}_j$ are two nodes of the graph.

In the following, we will present a data-driven model which allow to track the propagation of any signal spatially distributed, supposedly the cortical thickness. We consider a longitudinal dataset $\mathbf{y} = (\mathbf{y}_{i,j})_{1\leq i\leq I, 1\leq j\leq J_i}$ of I individuals, each patient i being observed J_i times during the study at ages $(t_{ij})_{1\leq j\leq J_i}$. We suppose that there exits a common fixed-graph G defined by a set $\mathcal{V} = (\mathbf{x}_1, ..., \mathbf{x}_K)$ of K nodes and a distance matrix D which accounts for the distance between the nodes. Any node $\mathbf{x}_k \in \mathbb{R}^3$ corresponds to a coordinate of a point in space. Each observation $\mathbf{y}_{ij} == (\mathbf{y}_{ij1}, ..., \mathbf{y}_{ijK}) \in \mathbb{R}^K$ corresponds to the measured signal spatially distributed over the K nodes of G, represented by a point in the multivariate space $\mathbb{R}^K$, schematically represented on Figure 3.3a for $K = 3$, as if there were only 3 vertices in the mesh. Therefore, it defines a network whose nodes are valued with the signal of interest. It follows that the collection $(\mathbf{y}_{ij})_{1\leq j\leq J_i}$ of the observations of a particular subject defines a network that embeds a time-series on each node of G, indexed by the patient age at each observation $(t_{ij})_{1\leq j\leq J_i}$.

3.2.3 Model

From short-term data to long-term history

We assume there that the repeated observations of a subject are sampled from a continuous function $t \mapsto \eta_i(t) = (\eta_{i1}(t), \ldots, \eta_{iN}(t))$, where $\eta_{ik}(t)$ describes the decrease of cortical

thickness of this ith individual at vertex k, such that

$$\forall i \in \{1, \dots, I\} \quad \forall j \in \{1, \dots, J_i\} \quad \forall k \in \{1, \dots, K\} \qquad \mathbf{y}_{ijk} = \eta_{ik}(t_{ij}) + \epsilon_{ijk}, \tag{3.1}$$

where $\epsilon_{ijk} \sim N(0, \sigma^2)$ corresponds to the model noise, whose variance is σ^2.

The function $t \mapsto \eta_{ik}(t)$ describes the evolution of the time-series at node k for the individual i. Thus the vector function $t \mapsto \eta_i(t) = (\eta_{i1}(t), \dots, \eta_{iN}(t))$ describes the continuous evolution on the graph for a particular individual *i.e.* the spatiotemporal propagation of the signal over the whole brain. It corresponds to a spatiotemporal trajectory in the space of measurements. The trajectory $t \mapsto \eta_i(t)$ is therefore able to reconstruct the existing observations $(\mathbf{y}_{ij})_{1 \leq j \leq J_i}$, defined at the related time-points $(t_{ij})_{1 \leq j \leq J_i}$, as shown on Figure 3.3b, but also generate an observation at any time t, potentially in the future.

The repeated data of each individual is a particular window in the long-term course of the disease that potentially overlaps with other patients. We aim to re-align along a common time-line these short-term sequences by carefully analyzing the spatiotemporal patterns within each short-term snapshot. Nevertheless, to do so, we also need to account for the inter-individual variability in cortical thickness measurements and trajectories of propagation across the network. The inter-individual variability prevents us from considering any individual propagation as a good representation of the disease evolution.

Consequently, we assume that there exists a mean scenario of propagation, defined by a group-average spatiotemporal trajectory $t \mapsto \boldsymbol{\gamma}_0(t)$, represented on Figure 3.3c, such that each individual trajectory $t \mapsto \eta_i(t)$ is a temporal and spatial variation of this mean scenario of changes, detailed in section 3.2.3. This typical scenario of change describes the mean pattern of spatiotemporal propagation of the signal and writes $\boldsymbol{\gamma}_0(t) = (\boldsymbol{\gamma}_1(t), \dots, \boldsymbol{\gamma}_K(t))$ where for all $k \in \{1, \dots, K\}$, $t \mapsto \boldsymbol{\gamma}_k(t)$ characterize the typical temporal evolution of the cortical thickness on the brain region related to the node k. As represented on Figure 3.2. Each node has a different temporal profile of atrophy, accounting for the variation of the cortical thickness over time.

Individual estimation

Translating the generic framework introduced by [Schiratti et al., 2015] into this case requires to exhibit individual parameters that characterize the individual spatial and temporal variations to the mean, namely the *space-shifting* and the *time reparametrization*.

Time reparametrization We introduce a time-warp function $\psi_i(t)$ that corresponds to a time reparametrization that adjust the individual dynamics on a common time-line, which here is the average spatiotemporal trajectory $\boldsymbol{\gamma}_0$. For any patient i with observations $(y_{ij})_{1 \leq j \leq J_i}$ at time-points $(t_{ij})_{1 \leq j \leq J_i}$, $\psi_i(t_{ij}) = \alpha_i(t_{ij} - t_0 - \tau_i) + t_0$ where t_0 is a common reference time of the reparametrization, α_i encodes for the individual pace of propagation and $t_0 + \tau_i$ describes subject-specific time-shift to the mean disease onset. As such, if the acceleration factor α_i is greater than 1, it corresponds to a faster pace of cortical atrophy whereas $\alpha_i < 1$ indicates a slower pace of atrophy. In the same way, the larger the value of the time-shift τ_i is, the later the disease occurs. Therefore, it leads to write $\eta_i(t) = \boldsymbol{\gamma}_0(\psi_i(t)) + \boldsymbol{\varepsilon}_{ij}$. It adjusts the pace at which the trajectory is followed for the ith individual.

Space-shifting In the space of measurements $\mathbb{R}^K$, we consider individual observations and the mean trajectory $\gamma_0(t)$ as shown on Figure 3.3a. In order to account for the spatial variability of the individual trajectories, we assume that there exists, for any individual i, a vector $w_i \in \mathbb{R}^K$ called the space-shift, that characterizes the spatial variations from $\gamma_0(t)$ to the observations as shown on Figure 3.3b. For any point on $\gamma_0(t)$, $\gamma_0(t) + w_i$

is assumed to be on the individual trajectory. Therefore, it is possible to translate all the points $(\gamma_0(t))_{t\in\mathbb{R}}$ to $(\gamma_0 + w_i)_{t\in\mathbb{R}}$ as shown on figure 3.3c. This collection defines the individual trajectory $\eta_i(t)$. This space-shift must be orthogonal to the trajectory as it ensures the identifiability of the model. In fact, if the direction w_i was not orthogonal to the trajectory, then the projection of w_i on the geodesic γ_0 would interfere with the individual time realignment induced by the dynamic parameters (α_i, τ_i).

Using mathematical tools from the Riemannian geometry beyond the scope of this study, [Schiratti et al., 2017] shows that the kth coordinate of the individual spatiotemporal trajectory writes $\eta_{ik}(t) = \boldsymbol{\gamma}_k(\frac{w_{ik}}{\dot{\boldsymbol{\gamma}}_k(t_0)} + \psi_i(t))$. As the space-shift must be estimated in $\mathbb{R}^K$, w_i is supposed to be a linear combination of few independent components, in the spirit of Independent Component Analysis (ICA)([Allassonniere et al., 2012]). It leads to consider A a $K \times N_s$ matrix of N_s independent directions, and $(s_{ij})_{1\leq i\leq I, 1\leq j\leq N_s}$ parameters to estimate. $s_i = (s_{i1}, \ldots, s_{iN_s}) \in \mathbb{R}^{N_s}$ correspond to parameters of individual i that characterize his spatial variations from the mean spatiotemporal trajectory. The orthogonality condition, mentionned in the previous paragraph, leads to consider a basis $(B_1, \ldots, B_{(K-1)N_s})$ of matrices, whose columns are orthogonal to the direction of $\boldsymbol{\gamma}_0(t)$, and parameters $(\beta_l)_{1\leq l\leq (K-1)N_s}$ such that $A = \sum_{j=1}^{(K-1)N_s} \beta_j B_j$. This procedure allows to reduce the dimension of the parameters to estimate for each w_i, from K to the chosen number of sources.

It leads to write :

$$y_{ijk} = \boldsymbol{\gamma}_k\left(\frac{w_{ik}}{\dot{\boldsymbol{\gamma}}_k(t_0)} + \alpha_i(t_{ij} - \tau_i - t_0) + t_0\right) + \boldsymbol{\varepsilon}_{ijk}\,. \tag{3.2}$$

Curve parametrization

In this paper, we consider a *straight line model* such that $\boldsymbol{\gamma}_k(t) = v_k(t - t_k) + p_k$, v_k accounting for the ratio of atrophy and p_k for the thickness value at time t_k. A linear decay in cortical atrophy is then represented by a straight line trajectory, parametrized by time, in the K-dimensional space as shown on Figure 3.3. Note that as shown on figure 3.2, it is possible to parametrize the same curve with two distinct sets (p_1, t_1) and (p_2, t_2) preventing from having an identifiable model. We decided to fix the parameter t_k among all the nodes such that for all $k \in \{1, \ldots, K\}$ $t_k = t_0$, the time reference used in section 3.2.3, without any loss of generality as $t \mapsto \boldsymbol{\gamma}_k(t)$ is defined on $\mathbb{R}$. Despite the linear form of each coordinate $t \mapsto \boldsymbol{\gamma}_k(t)$, the resulting model is non-linear as it includes among others, multiplication of individual and population parameters.

Finally, equation (3.2) becomes

$$y_{ijk} = p_k + w_{ik} + v_k\alpha_i(t_{ij} - \tau_i - t_0) + \boldsymbol{\varepsilon}_{ijk}. \tag{3.3}$$

This model therefore defines a distribution of multivariate straight line trajectories that accounts for the distribution of the individual trajectories.

Spatial smoothness

The model proposed in this paper deals with data that are spatially distributed on a graph G defined by a set of nodes $\mathcal{V} = (\mathbf{x}_1, ..., \mathbf{x}_K)$, where each node embeds a spatial coordinate in $\mathbb{R}^3$. We expect a smoothly varying profile of atrophy across nodes. The proximity between edges is given by the distance matrix D.

In order to ensure small variations of the signal, we introduce a subset $\mathcal{V}_c = (\mathbf{x}_{d_1}, ..., \mathbf{x}_{d_{N_c}}) \subset \mathcal{V}$ whose vertices are called control nodes. Instead of estimating $(p_k)_{1\leq k\leq K}$ (resp. $(v_k)_{1\leq k\leq K}$) at all the nodes, we consider only the parameters at the control nodes $(p_{d_k})_{1\leq k\leq N_c}$ (resp. $(v_{d_k})_{1\leq k\leq N_c}$). We introduce a estimation function $\mathbf{x} \mapsto p(\mathbf{x})$ (resp. $\mathbf{x} \mapsto v(\mathbf{x})$) for all $\mathbf{x} \in \mathcal{V}$

such that, at the control nodes, the function is equal to the parameters : $\forall k \in \{1, \dots, N_c\}$, $p(\mathbf{x}_{d_k}) = p^{d_k}$ (resp. $v(\mathbf{x}_{d_k}) = v^{d_k}$). At the other nodes, the function is an interpolation of the parameter value at the control nodes weighted by the distance to each of them. Therefore the control vertices controls the evaluation of the parameters among all the nodes.

We choose a Gaussian kernel K_b as interpolation splines : $\forall x, y \in \mathcal{V}, \quad K_b(x, y) = \exp\left(-\frac{d(x,y)^2}{b^2}\right)$ where d is the geodesic distance on the mesh and b is the kernel bandwidth. This interpolation allows to remove the possible high frequencies, smoothing the signal spatially. Therefore, it leads to write :

$$\forall \mathbf{x} \in \mathcal{V}, p(\mathbf{x}) = \sum_{i=1}^{N_c} \beta_p^i K_b(\mathbf{x}, \mathbf{x}_{d_i}) \quad \text{and} \quad \forall \mathbf{x} \in \mathcal{V}, v(\mathbf{x}) = \sum_{i=1}^{N_c} \beta_v^i K_b(\mathbf{x}, \mathbf{x}_{d_i}) \,. \tag{3.4}$$

The parameters $(\beta_p^i)_{1 \leq i \leq N_c}$ (resp.$(\beta_v^i)_{1 \leq i \leq N_c}$) are the solution of the linear system $p^{d_k} = \sum_{i=1}^{N_c} \beta_p^i K_b(\mathbf{x}_{d_k}, \mathbf{x}_{d_i})$ (resp. $v^{d_k} = \sum_{i=1}^{N_c} \beta_v^i K_b(\mathbf{x}_{d_k}, \mathbf{x}_{d_i})$).

Given these interpolations, Equation 3.3 writes

$$y_{ijk} = p(\mathbf{x}_k) + (As_i)_k + v(\mathbf{x}_k)\alpha_i(t_{ij} - \tau_i - t_0) + \boldsymbol{\varepsilon}_{ijk} \,. \tag{3.5}$$

Even though the distance computed for the cortical thickness corresponds to a distance on the brain cortex, it is possible to compute a connectivity distance based on the connectome, or even an appropriate combination of some of these distances. The challenging part is to put into correspondence areas defined by the connectivity matrices and other networks such as the FSAverage Atlas.

The choice of the set $\mathcal{V}_c$ of control nodes among the whole set of nodes $\mathcal{V}$ is mostly determined by the choice of the bandwidth b : their uniform distribution is such that there is an approximate distance b between them. In the case of the cortical thickness, we have chosen a bandwidth equal to 16 mm which is representative of the spatial variability of the signal.

3.2.4 Algorithm

Algorithm 6: Estimation of the general and individual cortical thickness decrease with the MCMC-SAEM algorithm.

input : Longitudinal dataset $y = (y_{i,j})_{i,j}$ of measurement maps, with the corresponding ages $(t_{i,j})_{i,j}$.
Initial parameters θ^0 and latent variables z^0.
Geometrically decreasing sequence of step-sizes ρ^k.
Sufficient statistics S^k

Initialization: set $k = 0$ and $S^0 = S(z^0)$.

repeat

  Simulation: **foreach** block of latent variables z_b **do**

    Draw a candidate $z_b^c \sim p_b(.|z_b^k)$.

    Set $z^c = (z_1^{k+1}, ..., z_{b-1}^{k+1}, z_b^c, z_{b+1}^k, ..., z_{n_b}^k)$.

    Compute the acceptance ratio $\omega = \min\left[1, \frac{q(z^c|y,\theta^k)}{q(z^k|y,\theta^k)}\right]$.

  end

  Stochastic approx.: $S^{k+1} \leftarrow S^k + \rho^k\left[S(z^{k+1}) - S^k\right]$.

  Maximization: $\theta^{k+1} \leftarrow \theta^\star(S^{k+1})$.

  Increment: set $k \leftarrow k + 1$.

until convergence

output: Estimation of θ^*.
Samples $(z^s)^s$ approximately distributed following $q(z \,|\, y, \theta^*)$.

Equation 3.5 describes a mixed-effects model, introducing population and individual parameter in this high-dimensional non-linear model. We consider that $((\alpha_i)_{1\leq i\leq I}, (\tau)_{1\leq i\leq i}, (s_{ij})_{1\leq i\leq pI, 1\leq j\leq N_s})$ are random-effects of the model, leading to write $\forall i \in \{1, \ldots, I\}\, \forall j \in \{1, \ldots, N_s\}$:

$$\alpha_i = \exp(\xi_i)\,, \xi_i \sim \mathcal{N}(0, \sigma_\xi^2)\,, \tau_i \sim \mathcal{N}(0, \sigma_\tau^2)\,, \text{ and } , s_{ij} \sim \mathit{Laplace}\left(0, \frac{1}{2}\right)\,.$$

α_i corresponds to the realization of a log-normal distribution so that it is always positive, preventing the individuals to present an increasing cortical thickness over time. Moreover, the Laplacian distribution of s_{ij} arises from theoretical considerations as we need the model to be identifiable, *i.e.* the solution of the problem to be unique. Finally, these random effects account for the statistical distribution of the individual trajectories. In the following, we consider $\mathbf{z} = ((\alpha_i)_{1\leq i\leq I}, (\tau)_{1\leq i\leq I}, (s_{ij})_{1\leq i\leq I, 1\leq j\leq N_s})$ as hidden variables.

Given Equation (3.5) and the observations $\mathbf{y}$, we would like to estimate the parameters $\theta = (t_0, (p^{d_k})_{1\leq k\leq N_c}, (v^{d_k})_{1\leq k\leq N_c}, (\beta_k)_{1\leq k\leq N_s(K-1)}, \sigma_\tau, \sigma_\xi, \sigma)$ as a Maximum Likelihood Estimate (MLE) $\theta^* = \operatorname{argmax} p(\mathbf{y}|\boldsymbol{\theta})$. The natural way to perform such estimation in mixed-effects models is the Expectation-Maximization algorithm ([Dempster et al., 1977]). Unfortunately, the E-step is intractable and it is not possible to sample according to the conditional distribution $p(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta})$. Therefore we use a stochastic version of the EM algorithm coupled with a Monte-Carlo Markov-Chain method, namely the Monte-Carlo Markov-Chain Stochastic Approximation Expectation Maximization (MCMC-SAEM) algorithm that is able to deal with non-linear equations in a high-dimensional setting. The algorithm is proven to convergence ([Allassonnière et al., 2010]) if the model belongs to the exponential family. In our case, it corresponds to consider that $p^{d_k} \sim \mathcal{N}(\overline{p}, \sigma_p^2)$, $v^{d_k} \sim \mathcal{N}(\overline{v}, \sigma_v^2)$ and $\beta_k \sim \mathcal{N}(\overline{\beta}_k, \sigma_{\overline{\beta}}^2)$

This leads to consider $\mathbf{z} = ((\xi_i)_{1\leq i\leq I}, (\tau_i)_{1\leq i\leq I}, (s_i)_{1\leq i\leq I}, (p^{d_k})_{1\leq k\leq N_c}, (v^{d_k})_{1\leq k\leq N_c}, (\beta_k)_{1\leq k\leq N_s(K-1)})$ as the extended hidden variables and $\boldsymbol{\theta} = (t_0, \overline{p}, \overline{v}, (\beta_k)_{1\leq k\leq N_s(K-1)}, \sigma_\xi, \sigma_\tau, \sigma_{\overline{p}}, \sigma_{\overline{v}}, \sigma)$ as the parameters of the model. The latter introduces sufficient statistics S of the model that are functions of the observations $\mathbf{y}$ and latent variables $\mathbf{z}$. The aim of such functions is to disentangle the maximization of the parameters θ and the simulation of the latent variables

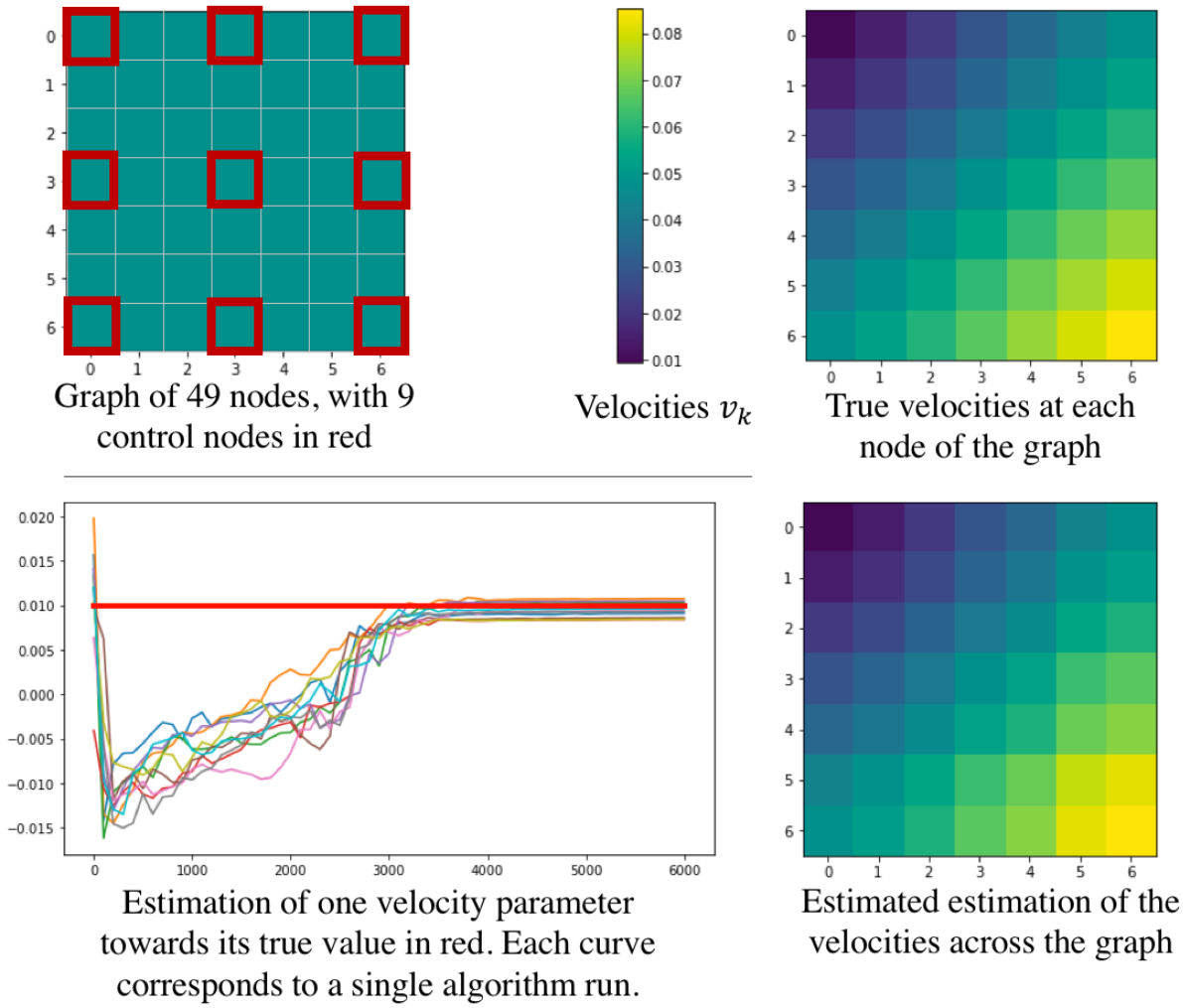


Figure 3.4: Simulation study performed to show the effectiveness of the parameter estimation procedure. The upper part describes the simulated graph (left) and the true velocities across the nodes (right). On the bottom part, a convergence example is given (left) as well as the estimated velocities estimated across the graph (right).

z.

The pseudo-code of the algorithm, reproduced in 6, shows the different steps of the optimization until convergence. For further information about the steps of the algorithm, the reader is referred to [Delyon et al., 1999, Kuhn and Lavielle, 2005, Allassonnière et al., 2010] and references therein.

3.2.5 Simulation study

Since we introduce a new approach to deal with longitudinal data spatially distributed, we performed a simulation procedure to show both the legitimacy of the model used, and the effectiveness of the estimation procedure. To this end, we define a graph represented on the top left of figure 3.4, representing a square mesh of 7 nodes per edge, thus 49 nodes in total. Among them, 9 equally distributed nodes represent the control nodes, in red on the figure. As we simulate data according to equation (3.3), we choose position and velocities across the node of the graph, as shown on the top right part of figure 3.4. Then we simulated realizations $(\xi_i, \tau_i, (s_{ij})_{1\leq j\leq N_s})_{1\leq i\leq N}$ for 350 patients, from 4 to 12 visits each (2980 visits in total) such that it represents 350 longitudinal trajectories of biomarkers

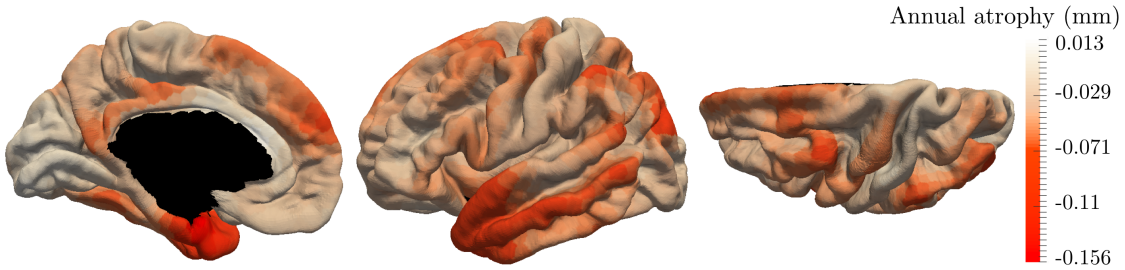


Figure 3.5: Annual rate of atrophy mapped over the brain surface used as initialization of our algorithm. Given one area, the corresponding rate of atrophy is obtained as the average regression coefficient of the linear regressions applied to each patient independently.

spatially distributed. These data were used to find the parameters used to simulate them. Thus, we have performed 10 runs of the estimation procedure. In order to account for the stochasticity of the algorithm and the motion of the Markov Chains, the results in table 3.1 are given with their standard deviation over 10 runs. As we need an initial value for the parameters, we initialized the algorithm without specific knowledge about the positions and velocities, contrary to the experience on the cortical atrophy, so it might reflect a worst-case scenario. Table 3.1 shows how well the algorithm performs on either control nodes or random nodes, as well as for the individual parameters.

The bottom part of Figure 3.4 shows some results of the estimation procedure. On the left hand side, we provide an example of the stochastic estimations of a parameter over the iterations of the algorithm - the figure shows 10 independent runs. The right hand side presents the final estimation of the velocities across the node of the graph, showing that the model is likely to reproduce the real signal. Overall, these results confirm that such procedure seems reasonable to assess the validity of the model and of the estimation procedure in order to estimate the temporal profile of longitudinal data spatially distributed, such as the cortical atrophy.

Parameter	Initial value	Final value	Real value	Error rate
p^{d_1}	2.0	$2.994\,(\pm 0.025)$	3.0	0.2%
$p^{d_{20}}$	2.0	$3.663\,(\pm 0.146)$	3.714	1.4%
$p^{d_{46}}$	2.0	$3.860\,(\pm 0.177)$	3.9	1.0%
v^{d_4}	1.0×10^{-2}	$2.84\,(\pm 0.24) \times 10^{-2}$	3×10^{-2}	5.3%
$v^{d_{21}}$	1.0×10^{-2}	$5.86\,(\pm 0.49) \times 10^{-2}$	6.25×10^{-2}	6.2%
$v^{d_{41}}$	1.0×10^{-2}	$7.83\,(\pm 0.65) \times 10^{-2}$	7.8×10^{-2}	0.4%
t_0	75	$70.9\,(\pm 2.7)$	70	1.3%
σ_τ^2	1.0×10^{-3}	$27.5\,(\pm 1.6)$	25	10%
σ_ξ^2	10^{-7}	$0.154\,(\pm 0.008)$	0.15	2.7%
σ^2	Not initialized	$1.34\,(\pm 0.03) \times 10^{-5}$	10^{-5}	34%

Table 3.1: The table shows the ability of the algorithm to estimate the real value of the model parameters.

3.3 Results

3.3.1 Initialization

We evaluated the propagation of the cortical atrophy thanks to cortical thickness values of 154 MCI converters (787 observations) distributed on a graph with 1827 nodes.

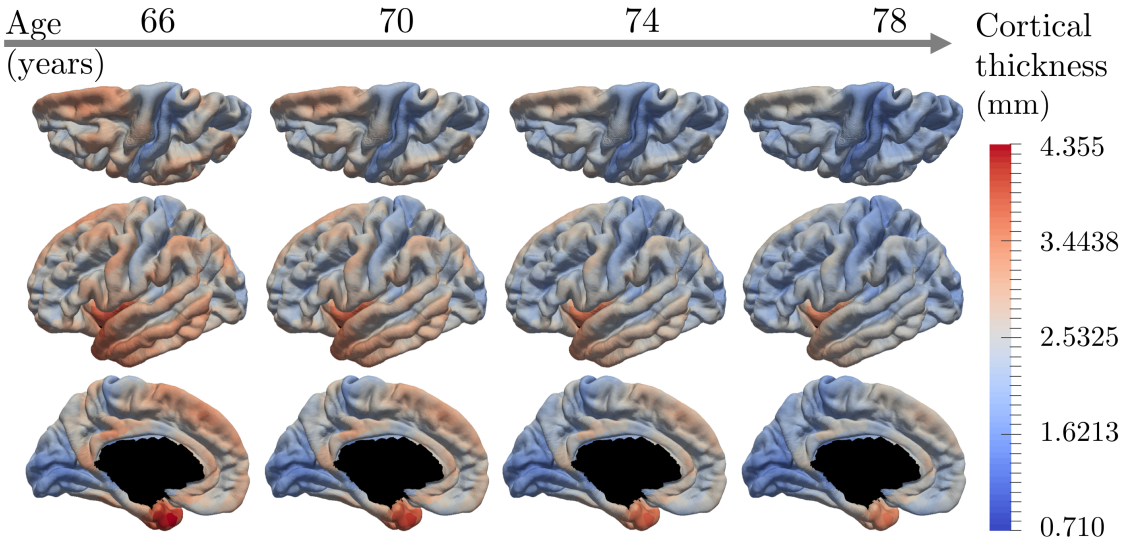


Figure 3.6: Estimated modes of evolution of the cortical thickness from 66 to 78 years old. This typical spatiotemporal pattern of atrophy propagation shows an important cortical loss in the superior frontal lobe, the temporal lobe and the hippocampus region.

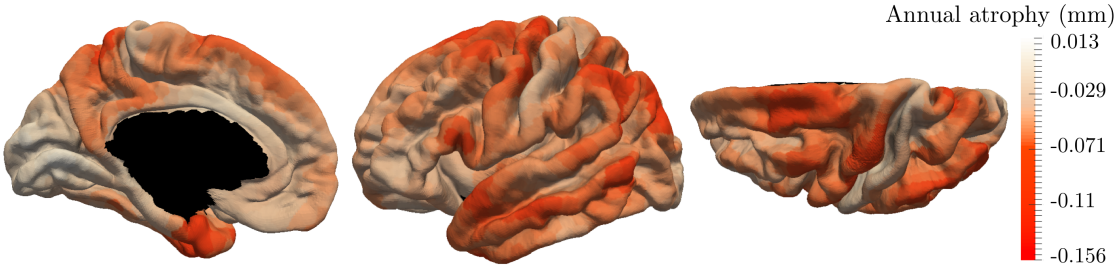
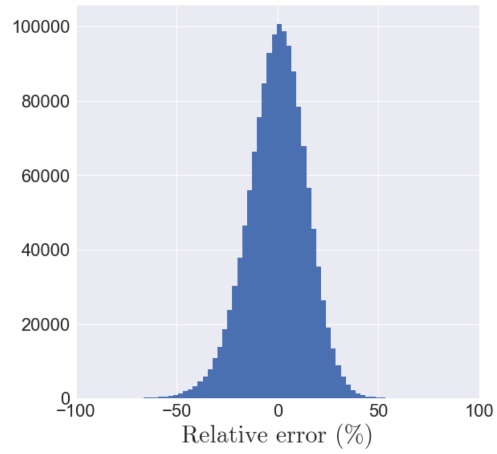


Figure 3.7: Final estimation of the annual rate of cortical loss observed during the typical pattern of atrophy propagation.

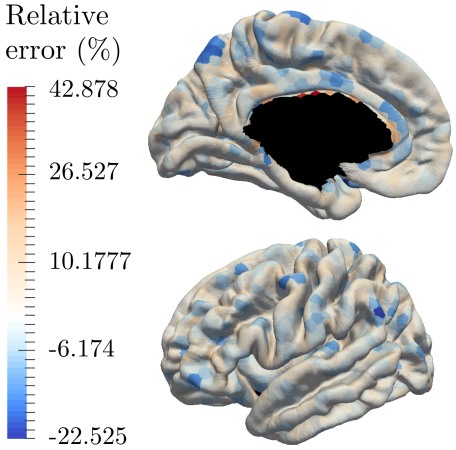
The initialization of the MCMC-SAEM algorithm requires initial values of the parameters $\boldsymbol{\theta}$ and realizations $\mathbf{z}$. We would like to draw attention on the realizations $((\alpha_i)_{1\leq i\leq I}, (\tau_i)_{1\leq i\leq I}, (s_i)_{1\leq i\leq I})$ and $((p^{d_k})_{1\leq k\leq N_c}, (v^{d_k})_{1\leq k\leq N_c}, t_0)$. The former are chosen equal to 0, leading to initial individual trajectories that are equal to the mean spatiotemporal trajectories. The pattern of atrophy is the same for everyone at the beginning. The latter variables, $((p^{d_k})_{1\leq k\leq N_c}, (v^{d_k})_{1\leq k\leq N_c}, t_0)$, are initialized based on the raw data. Besides t_0 that is chosen as the mean age of the input observations, for each control node k, we computed linear regressions on the longitudinal thickness values of each patient. Then we average the regression coefficients, each corresponding to a given subject, such that we end up with one rate of atrophy v_k per patch. Also, p_k was chosen as the average thickness on a given patch. The figure 3.5 shows the map of the initial v_k distributed over the cortical surface which looks reasonable.

The initializations of Figure 3.5 present areas with important cortical decrease over time, such as the temporal lobe and the hippocampus area. On the other hand, the primary visual cortex is less subject to a cortical atrophy. This initialization looks reasonable, however, these linear regressions are not able to reconstruct the individual observations, preventing from a characterization of personalized patterns of atrophy. It avoids describing the temporal and spatial variability of the individual propagations. Moreover, the linear regressions do not take into account the spatial coherence of the propagation as shown by the colorbar on Figure 3.5 where some areas present an important increase of the cortical thickness. It may be associated to the important noise within the data which is produced by the data acquisition, the extraction of the cortical thickness, and, the alignment on the same atlas.

Thanks to the model we introduced, we were able to reconstruct a mean (resp. individual) spatiotemporal trajectory, detailed in section 3.3.2 (resp. 3.3.3), that takes the form of the input measurements, preventing from working with percentiles or clusters that cannot be compared directly to the real observations. Due to the numerous number of hyperparameters and the stochastic behavior of the MCMC-SAEM, the algorithm was computed several times, each run of 100,000 iterations taking approximately 15 hours. The runs led to similar results. In the following, the results are presented for the run that provided the best individual reconstruction *i.e.* the smaller standard deviation σ of the noise. Its last estimation is of 0.29 mm, where 90% of the input data are between 1.5 and 4 mm.



(a) Histogram of the relative error of reconstruction of all individuals across all nodes.



(b) Average relative error of reconstruction over each patch, distributed on the graph.

Figure 3.8: The model is able to reconstruct the data at the individual level, while smoothing the signal over the brain surface, with a relative error randomly distributed.

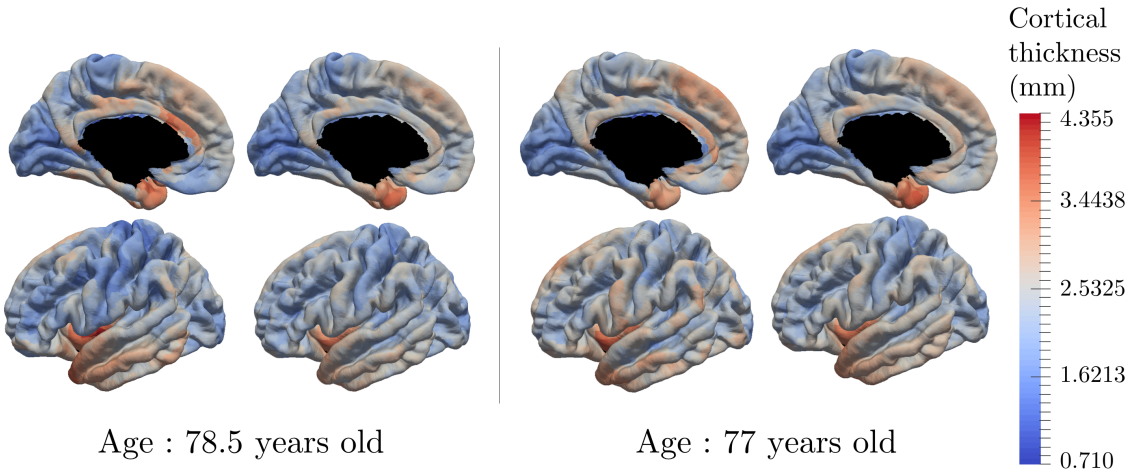


Figure 3.9: Real data and data reconstruction for subjects with a small space shift (right) and large space shift (left). The model is able to reconstruct the observed data, with a smoothing component, for subjects that present different characteristics.

3.3.2 Population level

The model exhibits a long term characteristic pattern of atrophy propagation from early MCI stage to post AD diagnosis. It corresponds to the group-average trajectory described in section 3.2.3 whose spatial ($\mathbf{w}_i$) and temporal (α_i and τ_i) variations corresponds to individual spatiotemporal trajectories. It is important to mention that this trajectory is a mean trajectory in a statistical sense, as its parameters are the mean values of the individual parameters.

Figure 3.6 shows the temporal and spatial evolution of the cortical atrophy, from 66 to 78 years old. The brain medial and lateral views shows an important atrophy on the temporal lobe and the medial temporal lobe, especially the fusiform and the parahippocampical gyrus. An important cortical decrease is also discernible on the superior frontal gyrus and at the wider region defined by the inferior parietal lobe and the angular gyrus. On the other side, the prefrontal cortex, the primary visual cortex, the calcaris sulcus and the post central gyrus are less subject to atrophy.

These results are supported by Figure 3.7 that shows the map of the annual atrophy v_k for the mean spatiotemporal trajectory, distributed over the corresponding brain areas. The areas affected by the cortical atrophy correspond to previous knowledge ([Whitwell et al., 2007, Jack et al., 1997, Scahill et al., 2002]) even tough the different measurements and methodologies lack in consensus. The patterns are still debated in order to find the best characterization of AD compared to normal aging or other neurodegenerative diseases. The proposed model may provide results for different populations on the same atlas, facilitating the comparison between diseases or with normal aging.

3.3.3 Individual reconstruction

The model is able to characterize personalized patterns of atrophy propagation thanks to a reconstruction of the individual observations. The validation is assessed thanks to the relative error of reconstruction. As mentioned previously, the input data are noisy, at both a temporal and spatial level. As for the temporal part, the 154 patients represent 281.358 temporal profiles (time-series) over the 1827 patches, from which only 6.4 % present a monotonous profile of decrease. Given all the linear regression computed for the algorithm initialization, the mean (resp. the variance) of the corresponding R-square values is equal to 0.348 (resp. 0.307). On the other side, the spatial noise corresponds to high variation of

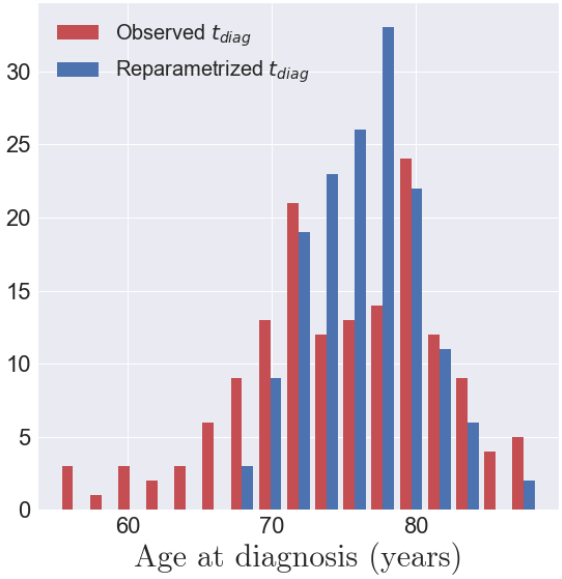


Figure 3.10: In red, the histogram of the observed age at diagnosis $t_{diag,i}$ for the 154 MCI converters. In blue, the histogram of the repatametrized age at diagnosis $\psi_i(t_{diag,i})$ once aligned on the common time-line. This shows that the age at diagnosis is mapped to a smaller range of time-points, in the model of cortical atrophy, suggesting that conversion to AD occurs at a specific stage of cortical atrophy.

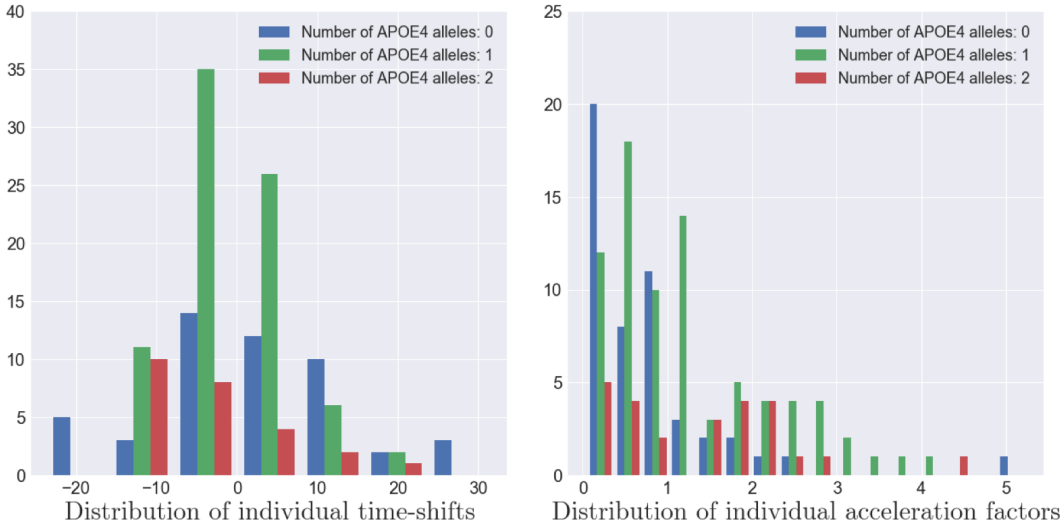


Figure 3.11: Distribution of the individual time-shifts (left) and the individual acceleration factors (right) for three types of APOE-ϵ4 population. A larger number of alleles of the APOE-ϵ4 genes is correlated to a faster pace of propagation of the Alzheimer's Disease (p-value $\simeq 0.001$) but not with an earlier atrophy onset (p-value $\simeq 0.5$).

the signal for neighbour nodes. Given this important noise, the goal of the reconstruction is not to reconstruct perfectly the data but rather to smooth the propagation over the brain and to capture individual tendencies of atrophy propagation. Thus, the 787 observations involve 1,437,849 reconstruction y_{ijk}, whose relative error of reconstruction is represented on Figure 3.8a which confirms the hypothesis that the noise is a Gaussian distribution with a zero mean (p-value = $4.24.10^{-109}$ for a t-test comparison with a theoretical distribution of mean equal to zero). As highlighted by Figure 3.8b, that represents the relative error of reconstruction over the 1827 patches, the error is mostly randomly distributed over the brain surface. It confirms that the reconstruction error does not have a spatial component as it is uniformly distributed over the brain surface. The color-bar was chosen according to the extreme values : it is important to mention that the larger error of reconstruction corresponds to areas that are close to the corpus callosum where the interpolation relies on a fewer number of control points.

Figure 3.9 presents the reconstruction of two different individuals who present various individual spatiotemporal trajectory, especially space shift norms that are either in the 10% bigger on the left hand side, or in the 10% smaller on the right hand side. The left part of each individual part corresponds to the input data whereas the right part is the corresponding reconstruction done by the model. It shows that the reconstruction is likely to represent the real data. The same color-bar was used as for Figure 3.6 to compare the individual data with the characteristic pattern of atrophy. Moreover, the spatiotemporal trajectory η_i of individual i is not estimated only at the observed time-points but it is a continuous function of the time, as shown on Figure 3.3c. Therefore, it is possible to reconstruct the observation at any point, potentially in the future.

One of the property of the model is to exhibit individual temporal parameters, namely the acceleration factor α_i and the time-shit τ_i, which allow to reparametrize the individual dynamics on a common time-line. As the data used here correspond to the cortical thickness, the realignment is estimated thanks to structural biomarker dynamics. On the other side, the MCI converters have an age at disease onset, t_{diag}, which corresponds to a clinical status. The latter is not straightforwardly related to the structural dynamics of the individual. In that sense, we decided to realign the age at onset t_{diag}, a clinical biomarker, on the same time-line, assessed with the structural biomarkers. The observed age at diagnosis $t_{diag,i}$ are represented by the red histogram on Figure 3.10, which is not unimodal and present an important variance. The realignment of the clinical status is represented thanks to the distribution of $(\psi_i(t_{diag,i}))_{1\leq i\leq I}$, which is centered with a reduced variance. It suggests that the clinical conversion to AD, determined with t_{diag} corresponds to a specific stage of the cortical atrophy.

As the model estimates individual spatiotemporal trajectories, it allows to describe the variability within the population. The distributions of $(\alpha_i)_{1\leq i\leq I}$, $(\tau_i)_{1\leq i\leq I}$ and $(\mathbf{w}_i)_{1\leq i\leq I}$ account for the distribution of the individual patterns of atrophy. Furthermore, the ADNI dataset provides, for each patient, multiple features, such as the number of alleles of the APOE-ϵ4 gene, the gender, the marital status, and the educational level. In the case of the APOE-ϵ4 gene, which is known as a genetic risk factor regarding AD ([Strittmatter et al., 1993, Poirier et al., 1993]), we exhibited the distribution of $(\alpha_i)_{1\leq i\leq I}$ and $(\tau_i)_{1\leq i\leq I}$ for the sub-populations defined by the number of alleles of the gene as shown on Figure 3.11. The more alleles, the more likely to have AD ([Corder et al., 1993, Strittmatter et al., 1993]).

As shown on the left hand side of Figure 3.11, the patients with two alleles (resp. one allele) present a mean time-shift of -2.98 years (resp. -0.20 years) after the mean scenario, contrary to patient without APOE-ϵ4 alleles that present an average time-shift of 1.89 years, meaning that the more alleles, the earlier the atrophy onset occurs. However, we applied Mann-Whitney two-sided statistical tests, that lead to insignificantly differences between the subpopulation. On the other side, same tests were conducted for the same

subpopulation with the mean acceleration factor whose distributions are presented on the right hand side of Figure 3.11. In this case, the group of individual with no alleles presented an average acceleration factor of 0.780, statistically different from the group of individual with one alleles (resp. two alleles) that presented an average acceleration factor of 1.415 (resp. 1.236) with a p-value equal to 0.00104 (resp.0.00511). However, this acceleration factor is not statistically different between the population with one or two alleles (p-value = 0.51518), meaning that these sub-population have similar rate of atrophy. Additional investigation on the gender, the marital status and the education level did not led to significant differences. It is important to mention that the Mann-Whitney test is sensitive to the number of samples whereas this study focuses on only 154 MCI patients that might lead to insignificant results in some cases, particularly in the case of the educational level (20 categories) or the marital status (unbalanced classes). Finally, it should mentioned that the tests conducted on the individual space shifts w_i and the related sources s_i did not lead to significant results, mainly because these parameters account for the difference in brain size, and thus thickness, between people.

3.4 Discussion

The paper presents a mixed-effects model of the atrophy propagation that is able to characterize a typical pattern of propagation, and, that reconstructs individual observations and scenarios of atrophy. The model exhibits brain areas that are the most affected by the cortical atrophy, such as the parahippocampical gyrus, the temporal lobe and the superior frontal gyrus. The lesions are less important in the primary visual cortex, the prefrontal cortex and the primary sensomotory cortex. The model allows to account for the different temporal dynamics of the alterations that can be then compared and ordered.

The proposed model offers a wide versatility of instantiation in terms of profile of temporal variations (exponential decay, sigmoid decay) and spatial variations (resolution, number of control nodes, kernel bandwidth) as it defines a generic framework for the estimation of longitudinal signals spatially distributed. It should be compared to other types of graph-related approaches, such as super-voxels ([Segovia et al., 2012]) or a vertex-cluster method ([Marinescu et al., 2017]). The latter has exhibited clusters of regression that show profiles of atrophy similar to our results. However, such models do not deal with individual characteristics neither directly with imaging data but rather with normalized values or percentiles, which restrict the interpretation. Further efforts should be concentrated on the validation and improvement of our model, possibly with more complex data and signal propagation.

The individual reconstructions also inform about subject-specific patterns of atrophy propagation, with potential personalized estimation of the cortical atrophy at future timepoints. Further investigations have to be conducted to ensure the quality of the new observations the model is able to generate, so that one can exploit the outcome that the model can predict for an individual some years after his of her last visit. This should be done with a proper validation set to determine the population parameters, and a test-set to predict the individual parameters and thus the future observations. Consistent results might provide information about the structural biomarkers related to the progression of AD, such as in [Eskildsen et al., 2015].

Another improvement of the model relies in the distance matrix computation. In this paper, the distance between the nodes is related to the distance on the brain surface, hiding potential effects of the neuronal connections. New distances might be computed based on functional connectivity or combination of different distances, in order to associate the functional and structural components of the brain that are supposed to be complementary in the disease process ([Bullmore and Sporns, 2009, Damoiseaux and Greicius, 2009, Wee et al., 2012]).

The model has the potential to exhibit the spatiotemporal propagation of any signal spatially distributed over a graph. It can be used in order to compare the patterns of propagation in distinct population e.g. normal aging or any other neurodegenerative diseases. It is also a first step to define personalized patterns that would help for a future prognosis of the patient stages.

Chapter 4

Deciphering the Progression of PET Alterations using Surface-Based Spatiotemporal Modeling

This chapter is a natural application of the model proposed in the previous chapter to the brain hypometabolism extracted from the PET-FDG scan. It has been published as the abstract Deciphering the Progression of PET Alterations using Surface-Based Spatiotemporal Modeling, *Koval I., Marcoux A., Burgos N., Allassonnière S., Colliot O. and Durrleman S., in* Organization for Human Brain Mapping, *2019.*

Contents

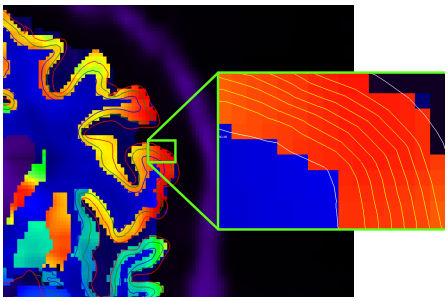
4.1 Introduction

Positron emission tomography (PET) is a central tool to study neurodegenerative diseases, allowing the measurement of hypometabolism and abnormal protein deposits (amyloid, tau). Modeling the spatiotemporal pattern of PET alterations in the cortex along the course of the illness is essential to understand disease progression and develop prognostic tools. In this study, we propose a generic method to model the spatiotemporal progression of PET alterations on the cortical surface from longitudinal images by combining two recently proposed approaches: i) a non-linear mixed-effects model for spatially distributed measurements based on Riemannian geometry [Koval et al., 2017, Schiratti et al., 2017]; ii a method for projection of PET data onto the subject's cortical surface (Marcoux et al, 2018). The model can reconstruct spatiotemporal patterns of progression at both population and individual level. We applied this approach to study the progression of hypometabolism along the course of Alzheimer's disease (AD) from the prodromal stage.

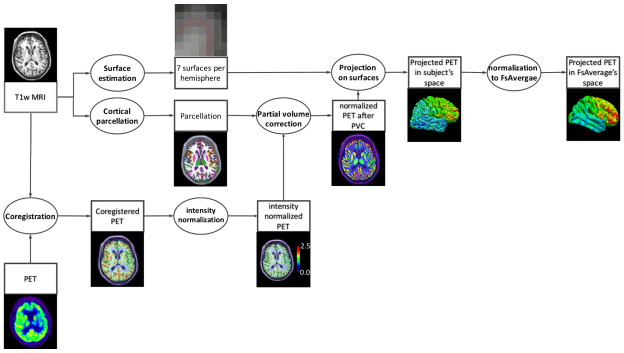
4.2 Methods

Brain metabolic activity, mainly located within the cortex, is known to be altered during the course of Alzheimer's disease (AD). Surface-based approaches, such as the one developed by [Marcoux et al., 2018], are thus well suited to analyze cortical hypometabolism derived from FDG PET images. This method, part of the open-source Clinica software [Routier et al., 2018] shown on Fig. 4.1b, includes i) co-registration of PET and T1-w MR images, ii) intensity normalization, iii) partial volume correction, iv) robust projection of the PET signal onto the subject's cortical surface shown on Fig. 4.1a, v) spatial normalization to a template.

The resulting projections once applied to repeated observations of multiples patients might inform about the spatiotemporal progression of the PET alterations. However, subjects are likely to be at different disease stages and to present different spatial patterns. [Koval et al., 2017] proposed to recombine the short-term individual observations to retrace



(a) Estimation of the FDG PET signal at 7 different surfaces representing the cortical thickness.



(b) Pipeline that includes the co-registration of PET and T1-w MR images, standard normalization and corrections and projection onto the cortical surface.

Figure 4.1: Method to project the cortical hypometabolism derived from FDG-PET images onto the cortical surface.

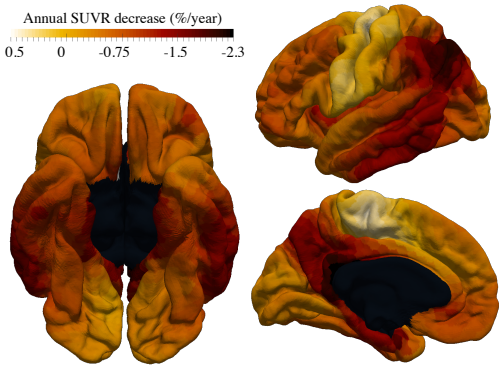


Figure 4.2: Map of the annual rate of standard uptake value ratio (SUVR) decrease.

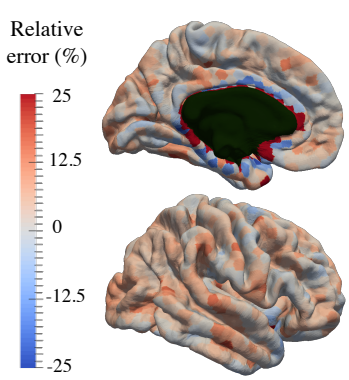
the long-term history of the disease for spatially distributed data, while accounting for the inter-subject spatiotemporal variability. This model generates an average progression profile defined at each point of the cortical surface. Furthermore, subject's trajectories are characterized by individual variations of the mean evolution, specifically an age at disease onset, a pace of progression and a spatial pattern of alteration. They enable the reconstruction of individual trajectories and the estimation of the spatiotemporal variability in the population.

We applied our approach to study progression of cortical hypometabolism along the course of AD, starting from the prodromal stage. Specifically, mild cognitively impaired patients, that progressed to AD during follow-up visits and that had at least two visits with both MRI and PET data, were selected from the ADNI database. This corresponds to 156 patients (74.0 ± 7.0 years, 89 males) with 4.4 visits on average (679 visits in total).

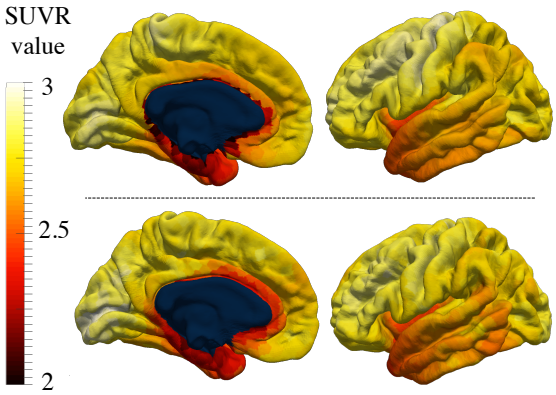
4.3 Results

Fig. 4.2 shows the metabolism decrease rate on the cortical surface for the mean profile of PET alterations. Greatest alterations are located in the precuneus, the parahippocampal gyrus, inferior and middle temporal gyri, and the inferior parietal lobule, followed by prefrontal regions. The sensory and visual cortices are spared.

The individual trajectories, considered as spatiotemporal variations of the average tra-



(a) Mean relative error between the individual observations and their reconstruction.



(b) PExample of an input data, i.e. the PET values projected onto the cortical surface on the top row, and, its reconstruction with the model on the bottom row.

Figure 4.3: The model is able to reconstruct the individual trajectories as spatiotemporal variation of the average trajectory.

jectories, allow the reconstruction of the observations. The mean relative error of reconstruction, uniformly distributed over the brain surface, is lower than 25%, except in areas close to the corpus callosum that are prone to poor preprocessing (4.3a. An example of a reconstruction is given on Fig. 4.3b.

4.4 Conclusion

We proposed a new approach to model the progression of PET alterations. Application to AD demonstrated that the method unveils relevant patterns and can adequately reconstruct the trajectory of alterations at the individual patient level. It could become a useful tool for understanding progression of neurodegenerative diseases and build new prognostic systems.

Part III

Digital Multimodal Model of the Alzheimer's Disease Progression

Chapter 5

Personalized Simulations of Alzheimer's Disease Progression with Digital Brain Models

This chapter corresponds to an article that has been submitted for revision.

Contents

Abstract

Simulating the effects of Alzheimer's disease on the brain is essential to better understand, predict and control how the disease progresses in patients. Our limited understanding of how disease mechanisms lead to the changes that are visible in brain images and clinical examination hampers the development of biophysical simulations. We develop here a statistical learning approach, where the repeated observations of several patients over time are used to synthesize personalized digital brain models. The method is built on generic geometric and statistical principles so that it can be applied to a large variety of data types such as unstructured sets of features or structured data like images or shapes.

We used it to construct a multimodal model of Alzheimer's disease progression. The model synthesizes over a period of 30 years the progression of metabolic alterations across brain regions, the deformation of the hippocampus due to atrophy, the progressive loss of grey-matter across cortical regions, together with the decline of several cognitive functions.

The model may be personalized to new subject's data by automatically adjusting age at onset, pace of progression and appearance of the model. The personalized model accurately reconstructs past and current observations and predicts future observations up to four years ahead of time for subjects at risk of developing Alzheimer's disease. We show that both reconstruction and prediction errors are of the same order as the uncertainty of the measurements. The model, therefore, acts as a digital avatar of the subject brain, which accurately simulates how brain image data and neuropsychological assessments change in time for each patient.

The personalisation of the model depends on a few and interpretable parameters, which allow a clinician to understand the specificity of the subject's progression compared to the average. We analyzed these parameters across all subjects to give a better description of the heterogeneity of the disease manifestation, highlight its main genetic and biological determinants, and give more insights into the complex interplay between the multiple effects of the disease on the brain and cognition.

5.1 Introduction

Numerical simulation has long been a central approach to understand complex systems, identify their determinants, and predict their behaviour. Recently, simulation has also proved to be key in artificial intelligence. For instance it is the ability to simulate a large number of go games that has made it possible to build a computer program that can learn to play better than a human[Silver et al., 2017]. Simulating a go game is easy because the rules are perfectly known and easy to implement. Simulating a brain developing Alzheimer's disease is more challenging because the biological mechanisms leading to the effects that are visible in brain images and clinical examinations are too imperfectly known[Khanal et al., 2016], like the reasons why these mechanisms lead to so heterogeneous effects across individuals. However, as with any complex system, simulating the disease is certainly a very promising way to better understand how it develops, identify the factors that modulate its manifestation in different individuals, and predict its progression in each patient.

We address here this simulation problem with a statistical learning approach. We design a computer program that automatically learns how Alzheimer's disease affects brain structure and function from the repeated observations of several patients in time, e.g. a longitudinal data set. It estimates a typical long-term scenario of change by normalising, re-aligning in time and combining several individual short-term data sequences. During training, the model learns how this typical scenario should be varied to reproduce the heterogeneity of progression profiles seen in the data. It does so by allowing adjustments in terms of age at onset, pace of disease progression and appearance of the model (see Fig. 5.1 and Fig. 5.2). Once trained, the model can be personalized to any new subject's data to simulate how the disease will progress at any time-point in the future, like a digital avatar of the subject's brain.

This approach may be seen as the synthesis between disease modeling and machine

learning approaches. On the one hand, numerous machine learning techniques have been proposed to predict if one patient will develop Alzheimer's disease within a given time window using clinical or imaging data[Zhou et al., 2012, Gaser et al., 2013, Moradi et al., 2015, Moore et al., 2018]. By essence, these methods predict a label, e.g. a diagnostic category, not a detailed description of the future state of the subject's brain or cognition. These black-box systems function as oracles and their lack of interpretability and explanability most probably hampers their adoption by clinicians in the daily routine. On the other hand, disease modeling approaches show how measurements continuously vary during disease progression[Fonteijn et al., 2012a, Jedynak et al., 2012, Villemagne et al., 2013, Donohue et al., 2014, Zhang et al., 2016, Guerrero et al., 2016, Khanna et al., 2018]. These modeling works have remained mostly descriptive so far, aiming to better characterize the heterogeneity of disease progression in populations. Only few of these works have been used and evaluated to predict the progression of the measurements in the future, and report error measures between predicted and true data at the individual level[Huang et al., 2016, Iddi et al., 2019]. They predict clinical assessments or some simple features extracted from the images. To the best of our knowledge, no method is general enough to allow the prediction of a full image or the shape of an anatomical structure, and to do it at any arbitrary time-point in the future.

The validation of such a simulation method is more difficult than for classification methods. In this work, we propose for the first time to compare the differences between the predicted image data and the true ones with the differences between the test and re-test image data acquired on the same day on the same patients. The differences between the test and re-test image data measure the uncertainty of the measurements due to variations in acquisition and processing. It gives an optimal value for prediction errors, since lower errors are likely to be due to an over-fit of the data. We also propose to compare our predictions with the simplest prediction method that assumes that data will not change in the future, namely the constant prediction. A recent review has shown that a third of the methods predicting diagnosis until 2 years ahead in time performs worse that assuming the diagnosis has not changed. This fact raises the need to evaluate simulations over longer time frames. We propose therefore to evaluate our simulations from 3 to 4 years in the future depending on available data.

5.2 A geometric approach of statistical learning

The proposed approach is rooted into a geometric framework, which allows an effective definition of statistical distributions of curves in high-dimensional structured spaces. It has the advantage to account for a large variety of data types including structured data such as images and shapes.

We assume that each data (from a single patient at a single visit) may be represented as a point on a multi-dimensional Riemannian manifold, a mathematical space that generalizes usual geometric operations such as addition, translation or computation of distances. Repeated observations of the same subject are then seen as noisy samples along a curve on the manifold. Furthermore, we assume that such individual curves result from random spatiotemporal transformations of a geodesic curve that is common to the population. This hierarchical structure forms therefore a mixed-effects statistical model[Schiratti et al., 2015, Schiratti et al., 2017].

Various types of data may be represented as points on a specific Riemannian manifold. In this work, we consider sets of bounded measurements such as normalized neuropsychological assessments, measurements distributed at the nodes of a fixed graph such as volumetric images or maps of cortical thickness, and shapes such as surface meshes of the hippocampus.

By an appropriate choice of the Riemannian metric, we prescribe a certain form of the

common population curve that shows how data change in time. For neuro-psychological assessments, each score is assumed to follow a logistic curve. Cortical thickness decreases at a linear rate at each vertex of the cortical surface, while ensuring that slopes and intercepts vary smoothly over the surface[Koval et al., 2017]. Image intensity at each voxel (or over a small region of interest) also decreases at a linear rate with smoothly varying parameters across neighbor voxels or regions. The shape of the hippocampus meshes is changed by the action of a smooth and invertible 3D deformation called diffeomorphism[Durrleman et al., 2014, Durrleman, 2018, Bône et al., 2018]. In all cases, this population curve is parameterized by a reference point p_0 on the manifold (e.g. a set of scores, an image or a mesh), a velocity (of the same dimension as p_0) and reference time t_0, which will be all estimated (see Methods).

Subject-specific curves derive from the population average by random spatiotemporal transformations, which is composed of a time-reparameterisation of the trajectory combined with a parallel shift of the geodesic curve on the manifold. The time-reparameterisation changes the dynamics at which the curve is followed by an individual. It is defined by a time-shift and an acceleration factor which account for individuals developing the disease earlier or later than the average and at a slower or faster pace than the average respectively. It maps the real age the subject to a physiological age on the normative time-line of the population average curve. The parallel shift changes the position of the individual curve on the manifold with respect to the population trajectory. It is defined by a direction on the tangent-space of the manifold at the reference point p_0, called "space-shift". It accounts for differences in the pattern of changes seen in the data. For neuro-psychological assessments, it accounts for different ordering and timing of alterations among the scores. For image data, it accounts for different spatiotemporal patterns of alterations across regions, vertices or voxels. For shape data, it accounts for differences in the shape of the hippocampus across subjects (see Fig. 5.1 and Fig. 5.2).

The velocity v_0, perturbed by the time-reparameterization function at the individual level, encodes the changes in data due to disease progression. The space-shifts encodes the inter-individual differences at the *same* disease stage. An orthogonality condition between the velocity v_0 and the space-shifts ensures a unique decomposition between changes due to disease progression and those due to intrinsic differences in the characteristics of the individuals[Schiratti et al., 2015, Durrleman, 2018]. It makes the model identifiable.

All in one, this procedure defines a mixed-effects statistical model. We denote $\gamma_0(t)$ the population curve where t is the physiological age on a normative time-line, $\eta^{w_i}[\gamma_0](t)$ the parallel shift of the population curve in the subject-specific direction w_i, and $\psi_i(t) = \alpha_i(t - t_0 - \tau_i) + t_0$ the time-reparameterisation function defined by the subject-specific time-shift τ_i and acceleration factor α_i. The j-th observation of the i-th subject, denoted y_{ij} acquired at age t_{ij} is then assumed to be derived from the population curve by $y_{ij} = \eta^{w_i}[\gamma_0](\psi_i(t_{ij})) + \varepsilon_{ij}$ for the ε_{ij} being a random noise (see Methods for details).

The model may be written in short as $y_{ij} = f(\theta, z_i, t_{ij}) + \varepsilon_{ij}$, for f a non-linear function that is specific to each data type, θ the vector containing the fixed-effects p_0, v_0, t_0, the variance of the random-effects and the variance of the noise, and z_i the vector of random effects: acceleration factors, time-shifts and space-shifts. We add priors on the coordinates of the vector θ in a Bayesian setting. When t is varied, the curve $f(\theta, z_i, t)$ represents the subject-specific trajectory at any time t.

We now consider three successive statistical tasks:

- **calibration:** given the longitudinal data set $\{y_{ij}, t_{ij}\}_{i=1,\dots,N, j=1,\dots,N_i}$ for a certain type of data, we find the value of parameters θ that maximises the joint likelihood $p(\{y_{ij}\}_{ij}, \theta) = p(\{y_{ij}\}_{ij}|\theta)p(\theta)$. The optimal value $\hat{\theta}$ fully specifies the model of disease progression;

- **personalisation:** for the optimal value of the parameter $\hat{\theta}$, we personalise the model

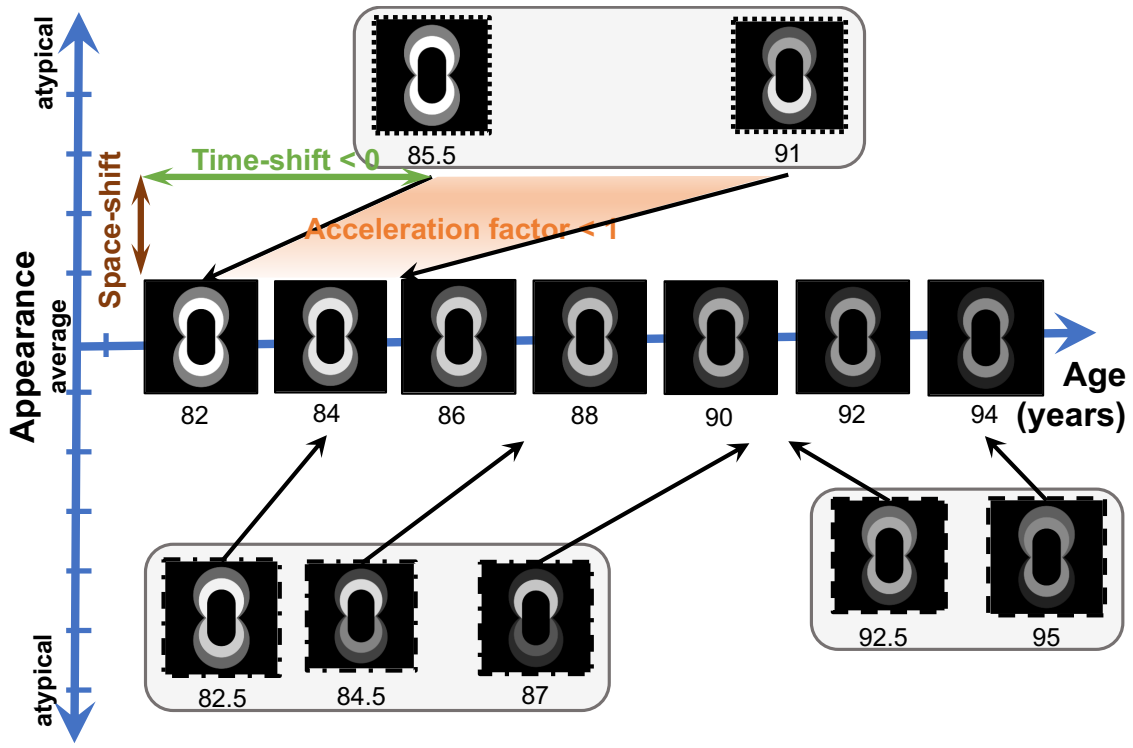


Figure 5.1: Scheme of model construction for image data. 1. A long-term scenario of change is built by normalizing and re-aligning in time several short term data sequence from different individuals (here 3 individuals are shown). 2. In turn, the model positions the progression of each individual with respect to the normative model. Time-shifts and acceleration factors encode differences in the age at onset and pace of changes, thus translating and scaling the temporal axis (x-axis). The y-axis is a schematic representation of a multi-dimensional coordinate system, where coordinates are called space-shifts. They encode here variations in model appearance. 3. The normative scenario may be personalized by translating it in the direction of the space-shift and changing its dynamics to reconstruct a personalized continuous scenario of changes and simulate the future progression of the individuals. The same concepts apply for mesh data and biomarkers. (see Fig. 5.2).

to the repeated data of a given subject (either a training subject, or a test subject in a cross-validation setting) $\{y_{test,j}, t_{test,j}\}_{j=1,\dots,N_{test}}$ by finding the optimal value of the random-effect $\hat{z}$ that maximises the conditional likelihood $p(\{y_{test,j}\}_j, z|\hat{\theta})$. The resulting $f(\hat{\theta}, \hat{z}, t_{test,j})$ is called the **reconstruction** of the data $y_{test,j}$ and its difference with the true data $y_{test,j}$ is called the reconstruction error;

- **prediction:** given a test subject with N_{test} observations, we personalize the model using only the first N_{past} ($< N_{test}$) observations to estimate $\hat{z}$, and then predict the future data after N_{past} by extrapolating the trajectory $f(\hat{\theta}, \hat{z}, t_{test,j})$, and measure the prediction error between the predicted and true (hidden) data.

We use a stochastic approximation of the Expectation-Minimisation algorithm[Allassonnière et al., 2015, Kuhn and Lavielle, 2004] for calibration, gradient-descent based method or Powell's method for personalisation (see Methods).

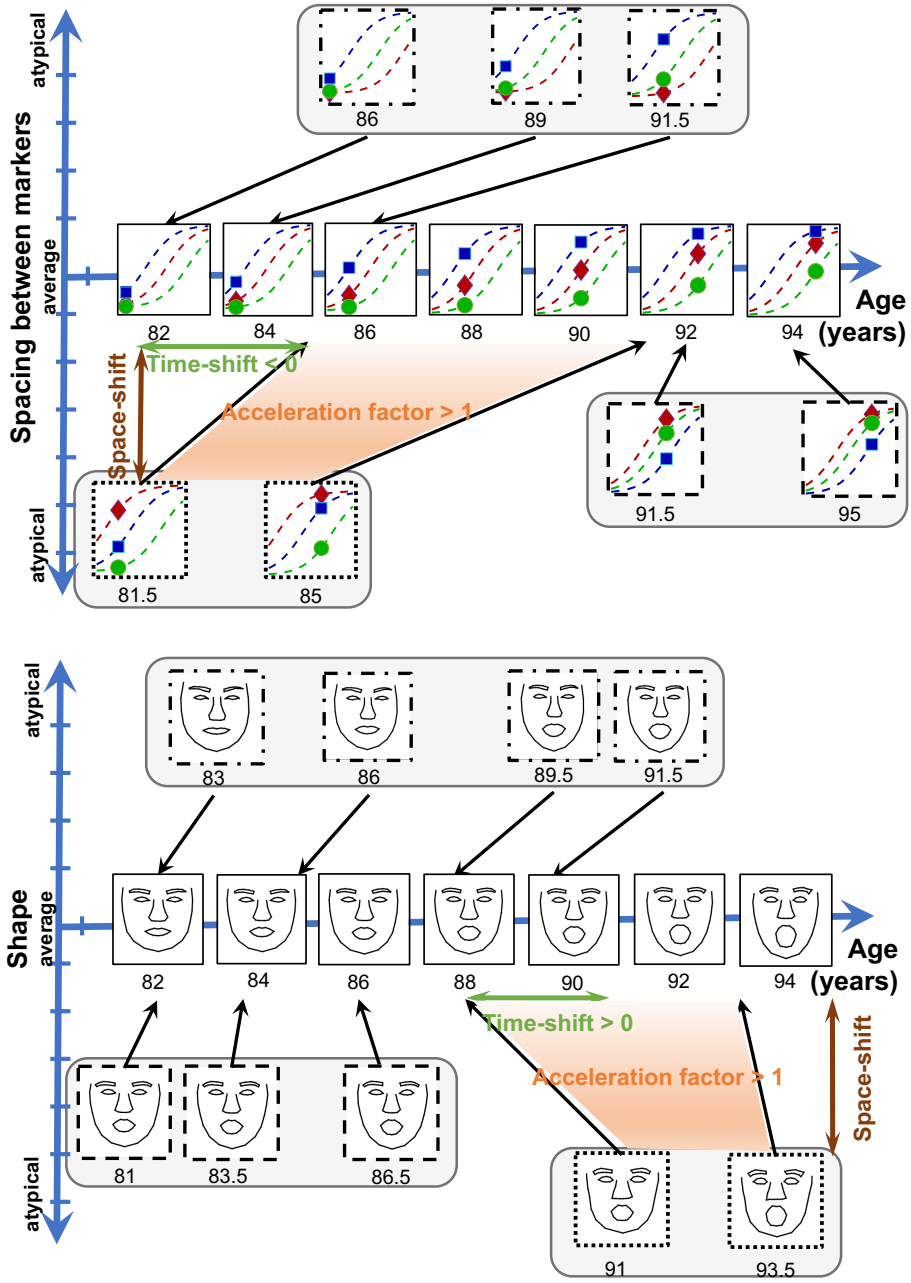


Figure 5.2: Model construction. The same method as in Fig. 5.1 may be applied for biomarkers and shape data. The only difference is that space-shifts now capture differences in the spacing between markers or the shape of the model. The use of a Riemannian framework allows to deal with all these cases with same method and very similar algorithms.

5.3 A multimodal disease progression model

We use data from the Alzheimer's Disease Neuroimaging Initiative (ADNI). In order to reproduce the natural history of the disease from the pre-clinical to the clinical stage, we selected the 322 subjects in this database who were included as cognitively normal or with mild cognitive impairments as defined in the ADNI protocol, and who had a confirmed clinical diagnosis of Alzheimer's disease at a later time-point in the study.

Whenever available, we use at each visit:

- regional measurements of standard uptake value ratio (SUVR) of fluorodeoxyglucose (FDG)-positron emission tomography (PET) to build models of hypometabolism across brain regions,
- maps of cortical thickness defined on a mesh of the cortex and extracted from T1-weighted Magnetic Resonance Images (MRI) to build models of cortical thinning,
- surface meshes of the hippocampus of both hemispheres segmented also from T1-weighted MRI to build models of hippocampal atrophy, and
- scores of the Mini-Mental State Examination[Folstein et al., 1975] (MMSE) and Alzheimer's Disease Assessment Scale - Cognitive Subscale with 13 items[Rosen et al., 1984, Mohs et al., 1997] (ADAS-Cog), the latter being divided into four sub-scores assessing memory, language, concentration and praxis, to build models of cognitive decline,

which amounts to 687 visits with PET images, 1,993 visits with MRI data and 1,235 visits with neuro-psychological assessments (See Methods and Table 5.4 for summary statistics).

For each data type, we calibrate the model parameters using all available visits of the selected subjects. The resulting models of progression are then synchronised by estimating affine time-reparameterisation maps among the normalized time-line of the different models. Finally, we use the age at diagnosis of each subject (an information that has not been used in the construction of the models) to estimate the physiological age on the normative time-line that corresponds stage at which one is diagnosed with the disease (see Methods).

Fig. 5.3 shows the synchronised models of hypometabolism, cortical thinning, hippocampal atrophy and cognitive decline at four representative time-points encompassing 16 years before diagnosis and 8 years after. It has been possible to reconstruct the disease progression over such a long period of time because we trained the model on patients data followed for much shorter periods of time but covering very different disease stages. These models may be visualised at a fine temporal resolution in the form of an interactive visualisation at the website: `www.digital-brain.org`.

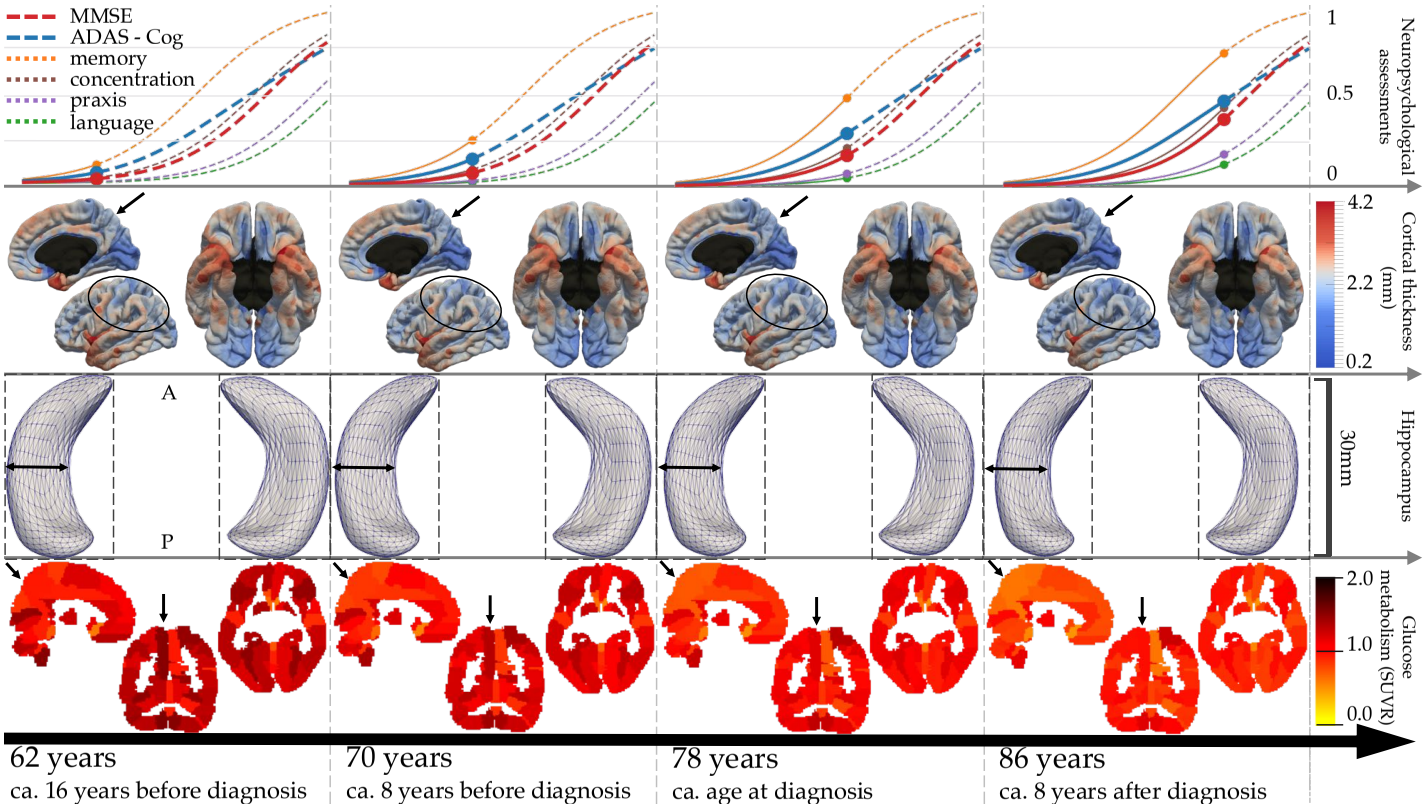


Figure 5.3: Normative models of Alzheimer's disease progression shown at 4 time-points with estimated time until/from diagnosis. Bottom to top rows show alteration of brain glucose metabolism, hippocampal atrophy, cortical thinning and onset of cognitive decline. Black arrows and ellipses indicate some areas of great changes.

Alterations shown by this digital model are in line with previous findings. For instance, the greatest alterations of glucose hypometabolism are found in the precuneus[Mosconi, 2005, Chen et al., 2010, Pagani et al., 2017], prefrontal areas[Drzezga et al., 2003] and the parahippocampal region[Mosconi et al., 2008]. Cortical atrophy also occurs in typical regions such as enthorinal cortex, hippocampal gyrus, temporal pole and fusiform gyrus[Hyman et al., 1984, Gómez-Isla et al., 1996], cortical association areas[Greene et al., 2010, Chan et al., 2001] and precuneus[Jacobs et al., 2012]. As expected, very little atrophy is shown to occur in the occipital lobe and the cingulate gyrus. More suprisingly, the model shows atrophy in the precentral gyrus and the paracentral lobule. Whether these regions are affected by cortical thinning due to Alzheimer's disease is still a debated question[Suva et al., 1999], which may be explained by the fact that the level of noise in this region is one of the largest. The model is not only confirmatory, as it integrates these heterogeneous findings into a consistent spatiotemporal view of disease progression at unprecedented temporal and spatial scales.

The model of cognitive decline shows a typical sequence of cognitive impairments starting with memory, followed by concentration 9.6 ($\pm$ 1.54) years after , praxis 9.8 ($\pm$ 1.73) years after, and finally language 3.3 ($\pm$ 2.65) years after (see Methods for the estimation of the standard deviation by cross-validation). It has been shown that Alzheimer's disease diagnosis occurs when the ADAS-Cog is comprised between 18.6 and 28.9 (i.e. between 0.21 and 0.34 on the normalised scale)[Skinner et al., 2012], which is reached between 74 and 80 years old in our normative time-line. Similarly, the diagnosis usually occurs for a MMSE score comprised between 27 and 23 (i.e. 0.1 and 0.23 on the normalised scale)[Raghavan et al., 2013], which occurs between 74 and 81 years old on our normative time-line. The age at diagnosis in the normative time-line has been estimated at 78 ($\pm$5.6) years old. The consistency of these estimates shows that the algorithm was able to correctly align the individual short term data sequences around the diagnosis time, by using solely the analysis of the spatiotemporal patterns of data changes and not the age at which the subjects were diagnosed.

5.4 Reconstruction errors and generalisation to unseen data

	Mean Error (± std)		Mean Absolute Error (± std)	
Modality (unit)	Reconstruction	Estimated measurement noise	Reconstruction	Estimated measurement noise
FDG-PET images)	$1.1 \times -10^{-4}(\pm 0.10)$	$-3.0 \times 10^{-3}(\pm 0.095)$	$7.6(\pm 6.5) \times 10^{-2}$	$6.8(\pm 9.4) \times 10^{-2}$
Cortical thickness (mm)	$5.8 \times 10^{-4}(\pm 0.44)$	$-1.1 \times 10^{-3}(\pm 0.28)$	$0.35(\pm 0.28)$	$0.19(\pm 0.20)$
Right hippocampus (mm^2)	$69.8(\pm 15.0)$	$85.2(\pm 40.1)$	$69.8(\pm 15.0)$	$85.2(\pm 40.1)$
Left hippocampus (mm^2)	$68.5(\pm 15.9)$	$83.2(\pm 36.0)$	$68.5(\pm 15.9)$	$83.2(\pm 36.0)$
Cognitive scores	$-2.2 \times 10^{-3}(\pm 0.075)$	$0(\pm 0.070)$	$5.5(\pm 5.0) \times 10^{-2}$	$5.6(\pm 4.2) \times 10^{-2}$

Table 5.1: Comparison between the statistics of the reconstruction errors and the ones of the distribution of the measurement noise. For hippocampus meshes, differences are measured by the norm of a vector, namely a positive number, so that errors and absolute errors coincide. For cognitive scores, the estimated measurements noise are computed based on the hypothesis of a centered Gaussian distribution with 7% standard deviation derived from the literature (see Methods).

We reconstruct now individual scenarios of disease progression by personalizing the model to each subject's data. The personalization finds the optimal values of the individual parameters, namely acceleration factor, time-shift and space-shift, which best fit a sequence of data of a given subject.

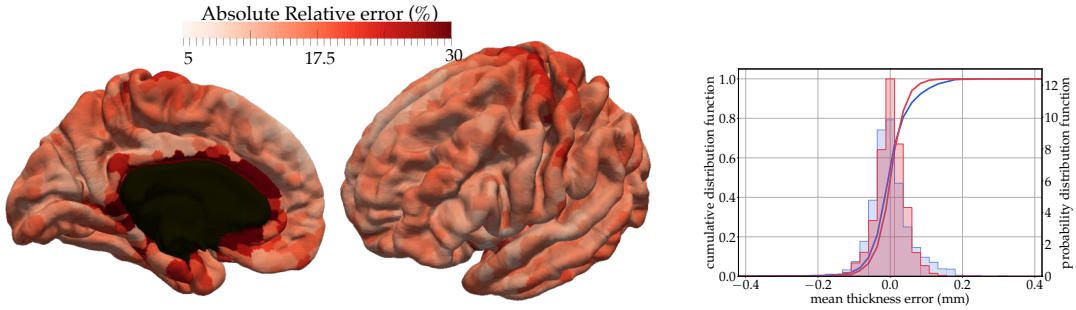
We assess the goodness-of-fit by measuring the reconstruction errors between the observed data at a given age with the data that is reconstructed by the model at the same age. We do not expect a perfect match between reconstructed data and observations as we imposed smoothness constraints in the spatial and temporal variations of the data and estimated a level of noise during model training with the aim to avoid over-fitting and allow better generalisation. Assessing the accuracy of goodness-of-fit is a difficult task, as one does not know the true level of noise of the measurements. We estimate this measurement uncertainty by measuring differences between data from test and re-test MRI sessions, PET data at baseline and follow-up for amyloid negative cognitively normals subjects and by performing a literature review of reproducibility of neuro-psychological assessments (see Methods).

Fig. 5.4 shows the superimposition of the empirical distribution of reconstruction errors with the empirical distribution of the noise for all data types. Overall, the two distributions largely overlap, and the standard error is of the same order than the measurement noise (see Table 5.1). This result shows that the model cannot be improved in the sense that smaller reconstruction errors would mean over-fitting.

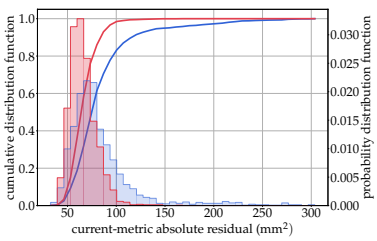
We notice that the reconstruction errors in brain regions are not evenly distributed. For PET data, the largest errors are found mostly in smaller regions. For cortical thickness, larger errors are found at the boundary of the mesh with the corpus callosum, mostly due to interpolation errors. These errors are much smaller than the best possible image resolution of 1 mm isotropic, thus making these reconstructions at sub-voxel precision.

We measure distances between hippocampus meshes using the currents distance, which is the norm of a multivariate vector of high dimension that has the unit of an area. It allows one to compare shapes with different samplings while being robust to small protrusion or topology changes[Vaillant and Glaunès, 2005]. In this case also, the distribution of reconstruction errors largely overlap with the one of the differences between test and re-test shape data. The personalisation of the model is driven by the currents distance and therefore tends to ignore the many spikes pointing outward that are often seen in the segmentations. Reconstructed meshes are smoother than observations, resulting in an under-estimation of the volume of the observations (see Fig. 5.5). It is more desirable to accurately reconstruct the shape rather than the volume, which is very sensitive to small segmentation errors. For instance, 83% of the subjects shows sequences of segmentation volume that are not monotonously decreasing, compared to only one subject for the volume of reconstructed meshes. Nevertheless, one should keep in mind that our reconstructions present a systematic bias in volume compared to the volume of the original segmentations.

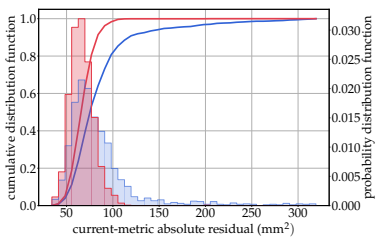
(a) FDG-PET images



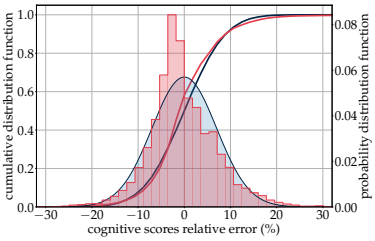
(b) Cortical Thickness maps



(c) Left hippocampus mesh



(d) Right hippocampus mesh



(e) Neuro-psychological assessments

Figure 5.4: Distributions of reconstruction errors. The empirical distribution of errors (red) is superimposed with the estimated distribution of test / re-test differences (in blue). For FDG-PET images and cortical thickness maps the absolute relative error is shown in every brain region. Mean and standard errors are given in Table 5.1.

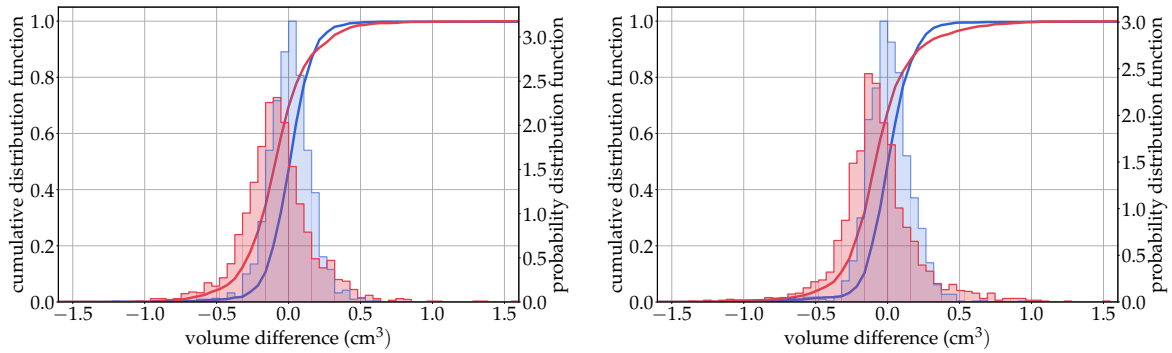
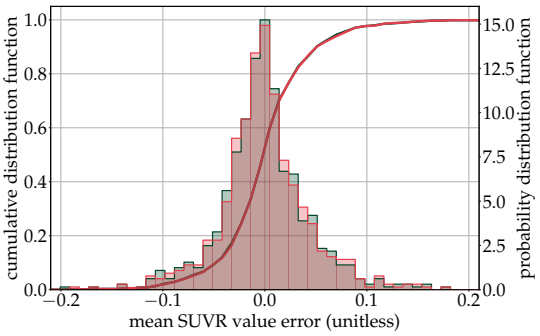
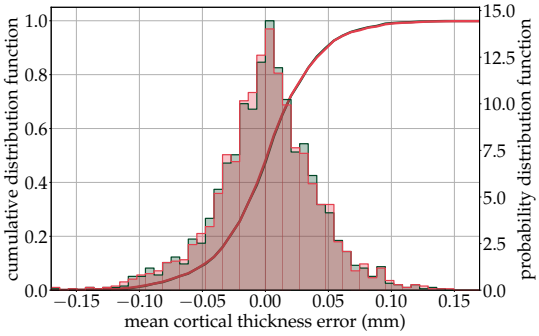


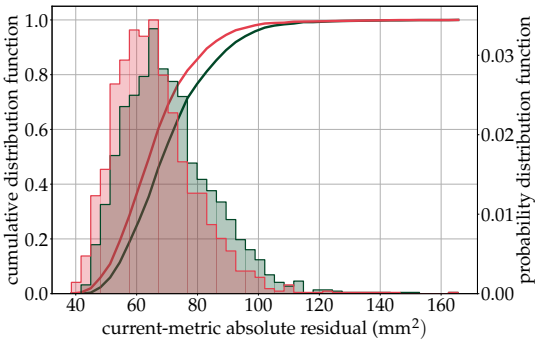
Figure 5.5: Reconstruction errors in hippocampus volume. Superimposition of the distribution of the reconstruction errors (in red) and test / re-test differences (in blue) measured as volumes for the left and right hippocampus (left and right panel respectively). Whereas the distribution of the test / re-test differences is centered (empirical mean of 0.5 mm^3 for the left hippocampus and -1.2 mm^3 for the right hippocampus), the distribution of the reconstruction errors has an empirical mean of -84.5 mm^3 for the left hippocampus and -67.3 mm^3 for the right hippocampus. The standard deviations of the distributions are: 208.6 mm^3 and 210.2 mm^3 for the test / re-test differences for left and right hippocampus respectively, to be compared to 243.2 mm^3 and 267.2 mm^3 for the reconstruction errors.



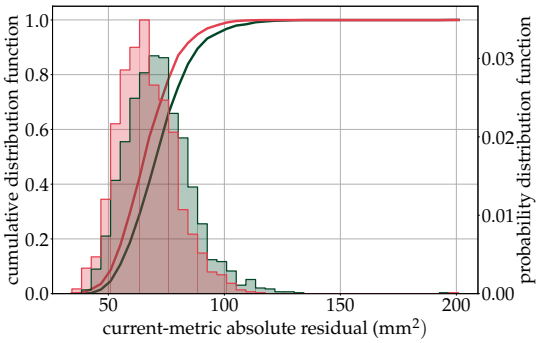
(a) FDG-PET SUVR values. The mean error is of $1.0 \times 10^{-4} \pm 0.044$ (red), and $-1.3 \times 10^{-4} \pm 0.044$ (green).



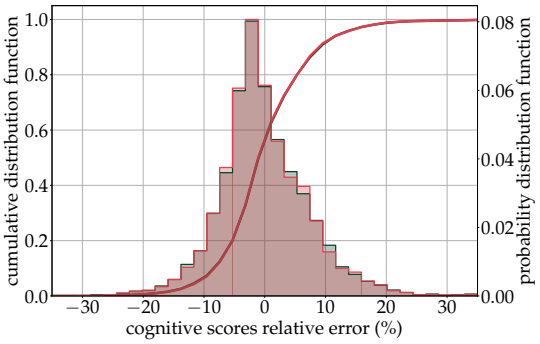
(b) Mean cortical thickness. The mean error is of $5.8 \times 10^{-4} \pm 0.040mm$ (red) and $6.1 \times 10^{-4} \pm 0.040mm$ (green).



(c) Left hippocampus. The mean error is $66.0 \pm 13.6\,mm^2$ (red), and $70.7 \pm 14.9\,mm^2$ (green).



(d) Right hippocampus. The mean error is $66.6 \pm 12.8\,mm^2$ (red), and $71.7 \pm 14.0\,mm^2$ (green).



(e) Neuro-psychological assessments. The mean error is $-0.19 \pm 7.5\,\%$ (red), and $-0.14 \pm 7.5\,\%$ (green).

Figure 5.6: Generalisation error to unseen data. The distribution of reconstruction errors when calibration and personalisation are done on the whole data set (in red, as in Fig. 5.4) is superimposed with the one estimated in the cross-validation procedure (in green).

We replicate the previous experiments in a five-fold cross validation procedure. Models are calibrated on 80% of the training data set, and personalised to the remaining 20% who were therefore not seen during model calibration. Distributions of these reconstruction errors are essentially identical with the previous ones obtained by calibrating and personalising the model on the whole data set (see Fig. 5.6). Only hippocampus shows a slightly higher generalisation errors but still below the noise level estimated with test / re-test data. The reconstruction of unseen data is therefore as good as the reconstruction of the training data, thus showing that the personalisation of the model generalises well to new individual data sequences. We also show that the discrepancy between the individual effects estimated as training or test sample is small with r^2 comprised between 0.93 and

Modality	Parameters	All data	Cross-validation
FDG-PET images	σ (no units)	0.101	0.101 ($\pm$ 0.001)
	t_0 (years)	75.5	74.9 ($\pm$ 0.9)
	σ_τ (years)	11.9	11.5 ($\pm$ 0.3)
	σ_ξ (no units)	1.30	1.28 ($\pm$ 0.03)
Cortical thickness	σ (mm)	0.442	0.442 ($\pm$ 0.001)
	t_0 (years)	82.0	82.7 ($\pm$ 0.7)
	σ_τ (years)	16.9	18.2 ($\pm$ 0.7)
	σ_ξ (no units)	0.99	1.03 ($\pm$ 0.02)
Right hippocampus	σ (mm^2)	2.49	2.60 ($\pm$ 0.03)
	t_0 (years)	76.2	75.7 ($\pm$ 0.3)
	σ_τ (years)	9.15	10.04 ($\pm$ 0.66)
	σ_ξ (no units)	0.71	0.78 ($\pm$ 0.03)
Left hippocampus	σ (mm^2)	2.67	2.74 ($\pm$ 0.04)
	t_0 (years)	76.3	76.3 ($\pm$ 0.3)
	σ_τ (years)	8.53	9.09 ($\pm$ 0.50)
	σ_ξ (no units)	0.66	0.68 ($\pm$ 0.03)
Cognitive scores	σ (no units)	0.081	0.081 ($\pm$ 0.001)
	t_0 (years)	71.5	72.4 ($\pm$ 0.8)
	σ_τ (years)	7.29	7.36 ($\pm$ 0.25)
	σ_ξ (no units)	1.07	1.11 ($\pm$ 0.11)

Table 5.2: Fixed-effects estimates using calibration on the whole data set (first column) and in a five fold cross-validation setting (second column) where mean and standard deviations of the five estimates are shown. Similarly, the delay between impairment of memory and the other cognitive functions is of 9.4 ± 1.6 yrs for concentration (9.6 yrs using all data), 19.9 ± 2.0 yrs for praxis (19.4 yrs using all data), 23.3 ± 2.6 yrs for language (22.7 yrs using all data)

0.99 (see Fig. 5.7). Furthermore, the fixed-effects parameters estimated in the five different calibration runs are consistent with the ones estimated using the whole data set as training set, thus showing the robustness of the estimation algorithm against resampling in the training set (see Table 5.2).

5.5 Personalized simulations of disease progression

Now, we evaluate the ability of the model to accurately predict the progression of the disease in the distant future. For this purpose, we select subjects and visits in the ADNI database based on criteria that can be assessed from present and past visits only, without the need to know the whole disease history of the patients as previously.

We select all the visits of all the subjects in ADNI for which the following conditions are met:

- the subject is labeled as Mild Cognitive Impairment at this visit,
- the MMSE of the subject is smaller or equal to 27 at this visit,
- the subject is amyloid positive at this visit,
- the sequence of diagnosis labels in the past visits is monotonic, meaning we exclude subjects showing reversion to control, or having AD label in the past.

These criteria aim to select subjects at risk of developing Alzheimer's disease.

From all these visits, we use the ones for which there is another visit of the same subject 3 or 4 years later in time. We personalize the previous model using the past and present visits of the subjects, extrapolate the model at 3 or 4 years, and evaluate the accuracy of the prediction by measuring the difference between the predicted and the true data (see Methods).

Note that if the test subject belongs also to the previous cohort, we used the model calibrated on the cross-validation fold that does not contain this subject. For new subjects, we use the model trained on the whole previous cohort.

Predictions of neuro-psychological assessment, for which we report the MMSE and the ADAS-Cog (as a linear combination of the 4 cognitive sub-domains predicted), are performed for 136 subjects for the prediction at 3 years, and 80 subjects for the prediction at 4 years. Prediction of the MRI data (cortical thickness maps and hippocampus shape) are performed for 72 subjects for the prediction at 3 years, and 63 subjects for the prediction at 4 years. We deem that there are not enough subjects to predict the FDG-PET data. It is worth mentioning that from the selected subjects with cognitive assessments (resp. MRI data), 36.5% and 39.1% (resp. 33.3% and 58.9%) present only one seen visit to personalize the model with at 3 and 4 years.

We assess the prediction errors in comparison with the distribution of the noise in the measurements, using the previous empirical distributions. We also compare the prediction of our model with the "constant" prediction, where one predicts that in 3 or 4 years, the data will be the same as of today.

As shown on the box-plots on Figure 5.8, we report the absolute error for the neuro-psychological assessments, the root mean squared error for the map of cortical thickness and the current distances for the shape of the hippocampus of both hemispheres. In all cases, the errors of the prediction of image data and neuro-phychological assessements are not statistically different than the uncertainty of the measurements.

An interesting observation is that the error of the constant prediction also, though increasing with the time-to-prediction, is not statistically significant from the noise up to 4 years in time. This fact means that the effect of aging or disease progression cannot be detected with the current precision of imaging devices and reliability of neuropsychological

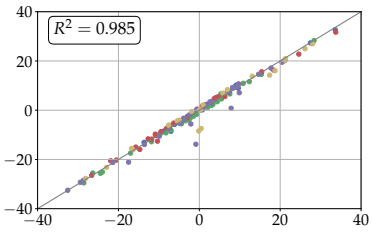
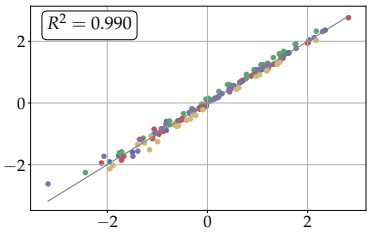
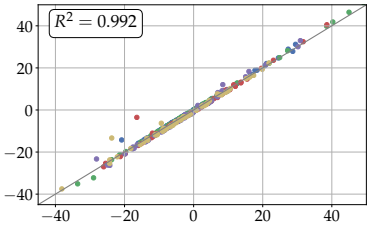
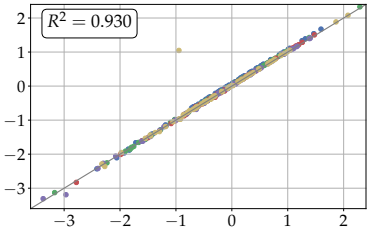
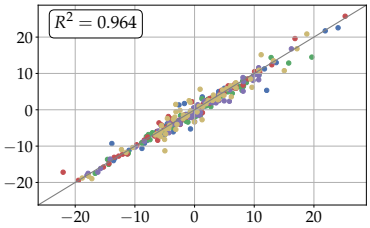
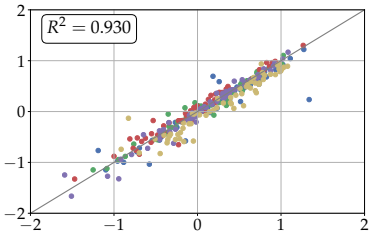
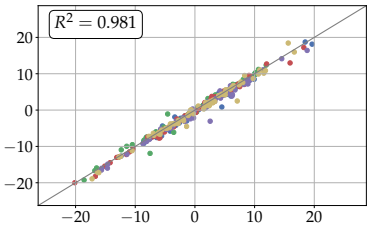
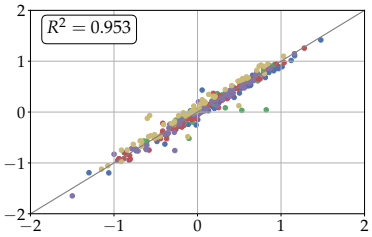
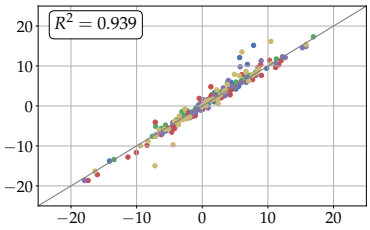
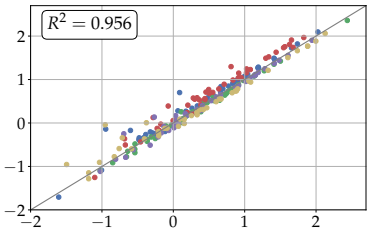
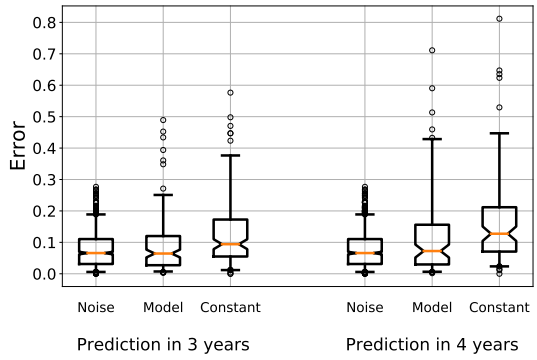
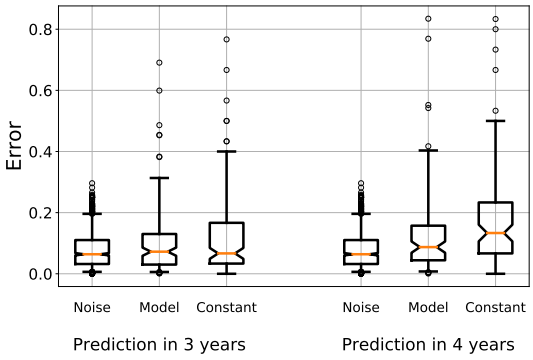
Modality	Time-shift	acceleration factor
FDG-PET SUVR		
Cortical thickness		
Right hippocampus		
Left hippocampus		
Neuro-psychological scores		

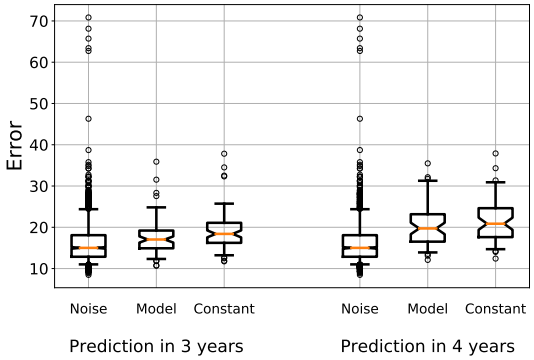
Figure 5.7: Robustness of model calibration and personalisation. Estimated time-shifts and acceleration factors when the individual belongs to the training set (x-axis) or to the test-set (y-axis). The five colors correspond to the folds the individuals belong to.



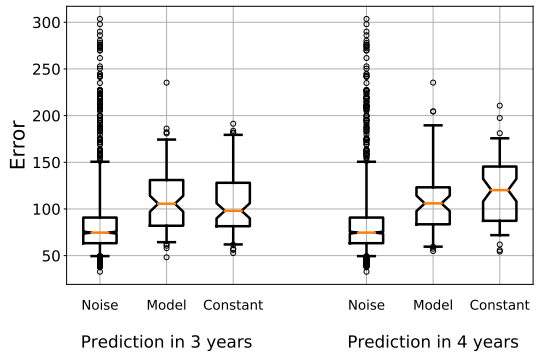
(a) ADAS-Cog prediction (absolute error)



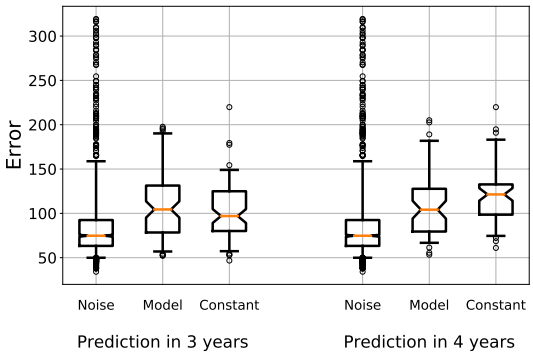
(b) MMSE prediction (absolute error)



(c) Cortical thickness prediction (L2-norm error)



(d) Left hippocampus (currents distance)



(e) Right hippocampus (currents distance)

Figure 5.8: Prediction errors of the simulated data. Box-plots show medians in orange, quartiles, and 95% confidence intervals for three image data and two cognitive assessments. Distributions of prediction errors are compared with that of the noise and the errors of the constant prediction.

assessments. It calls for longer term predictions, although the total follow-up duration in ADNI does not allow it. It also raises the question of the relevance of machine learning techniques that make predictions over periods of time that appears to be too short for the features to evolve sufficiently as compared to noise.

For MRI derived data, the noise distribution presents a very heavy tail that is due to the large heterogeneity of the image quality and its consequence in data processing. The constant prediction does not show such a heavy tail, as images of best quality 3 or 4 years apart show less variability than the test and re-test image acquired the same day. Our model shows steady performance at 3 and 4 years, above though not statistically different from noise level, whereas the constant prediction worsens as time-to-prediction increases. The prediction of cognitive performance shows a similar behavior. For the ADAS-Cog in particular, our prediction errors are closer to the noise level and shows a better contrast with the constant prediction.

Two other methods have been proposed to simulate neuro-psychological assessements, but not images or hippocampus shapes. Huang *et al.*[Huang et al., 2016] reports a Mean Absolute Error (MAE) for the MMSE of 1.81 points at 3 years and 1.66 points at 4 years. Iddi *et al.*[Iddi et al., 2019] reports an MAE of approximately 2 points at both 3 years and 4 years (see Fig.10 in this paper). We report an MAE of 3.2 points at 3 years and 4 points at 4 years. The accuracy of the test is of about 10%, so 3 points on a scale of 30.

For the ADAS-Cog, the MAE is of 3.7 pts at 3 years and 3.6 pts at 4 years in Huang *et al.*, and between 5 and 6 points at 3 years and 4 years in Iddi *et al.* (Fig. 10). We report 7.6 points at 3 years and 10.1 points at 4 years. The accuracy is also of the order of 10%, so 8.5 points on a scale of 85.

The MAE of the predicted scores greatly depends on the test cohort. It has been computed in both alternative methods on the whole ADNI data set where approximately (depending on the time to prediction) 30% are stable controls, 40% are stable MCI, and less than 15% are MCI converters. This data set would yield an accuracy of 85% if predicting a constant label. By contrast, we made predictions for subjects at risk of developing AD, defined as having a MMSE smaller than 27 and being amyloid positive, which contains only 42% of stable subjects. Our validation is therefore more stringent since it is done on a population showing greater longitudinal changes. This population better represents the characteristics of the subjects who might benefit from such simulations in the routine clinical practice.

Eventually, Iddi *et al.* used the simulated data to predict the diagnostic labels in the future. They obtained an accuracy of 80% at 2.5 years. Using a random forest classifier with our simulated data yields an accuracy of 78% at 4 years in our case.

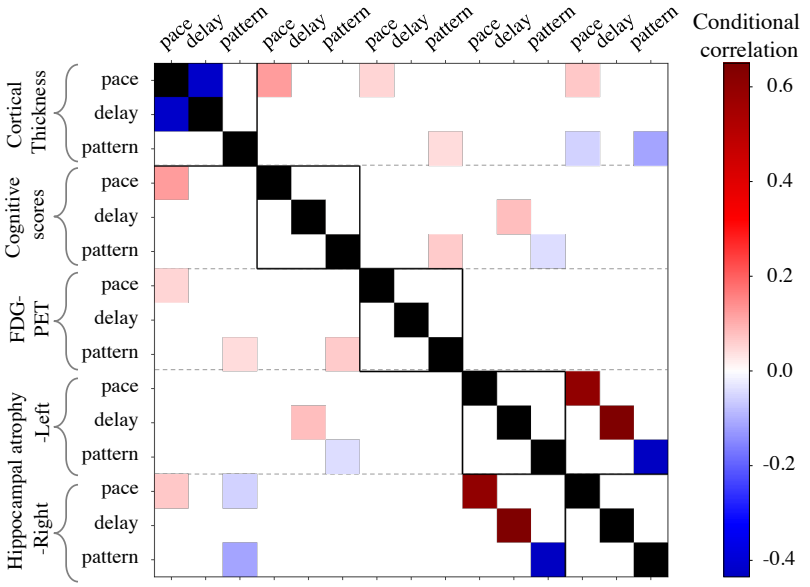


Figure 5.9: Thresholded matrix of conditional correlations from which the graph in Fig. 5.10 is built.

5.6 A holistic and dynamic view of disease progression

The interest of the model is not only to provide accurate simulations at the individual level. It allows also a systematic investigation of all aspects of disease progression and the effects that may influence it. The same types of individual parameters, all interpretable, are used for all modalities, and their relationship can therefore be studied in a quantitative manner. There are three types of parameters that quantify how much in advance or delayed the subject is, how fast or slow he is, and how different is his data at a given disease stage, as compared to a normative progression model.

It should be noted that most clinical studies does not perform temporal alignment of subject data, and that the differences that are measured between groups of subjects are likely to be partly confounded by the fact that one compares subjects at different stages of disease progression. The proposed method makes it possible to determine which observed differences are due to different progression dynamics or intrinsic differences in the subjects' evolution profile.

First, we construct a graph of a conditional correlations between all variables of all modalities (see Fig. 5.9). The statistically significant conditional correlations are represented in Figure 5.10 where three variables per modality are shown: pace of disease progression (e.g. acceleration factors), delay with respect to onset (e.g. time-shift), and pattern (e.g. space-shifts represented here as a single variable for the sake of simplicity). See Methods for details.

Interestingly, the vast majority of significant conditional correlations are found among variables of the same type across modalities, and not among different variables within the same modality. It means that the three aspects of disease progression: pace of progression, age at onset and types of progression profile are mostly independently of each other. This fact is surprising as studies reported that some early form of the disease are associated with more rapid progression, such fact being found here for the cortical thinning only.

The paces of progression of cognitive decline, hypometabolism, and hippocampus atrophy are conditionnaly independent of each other, and are all correlated with the pace of cortical thinning. The cortical thinning seems to be the main driver node, which influences

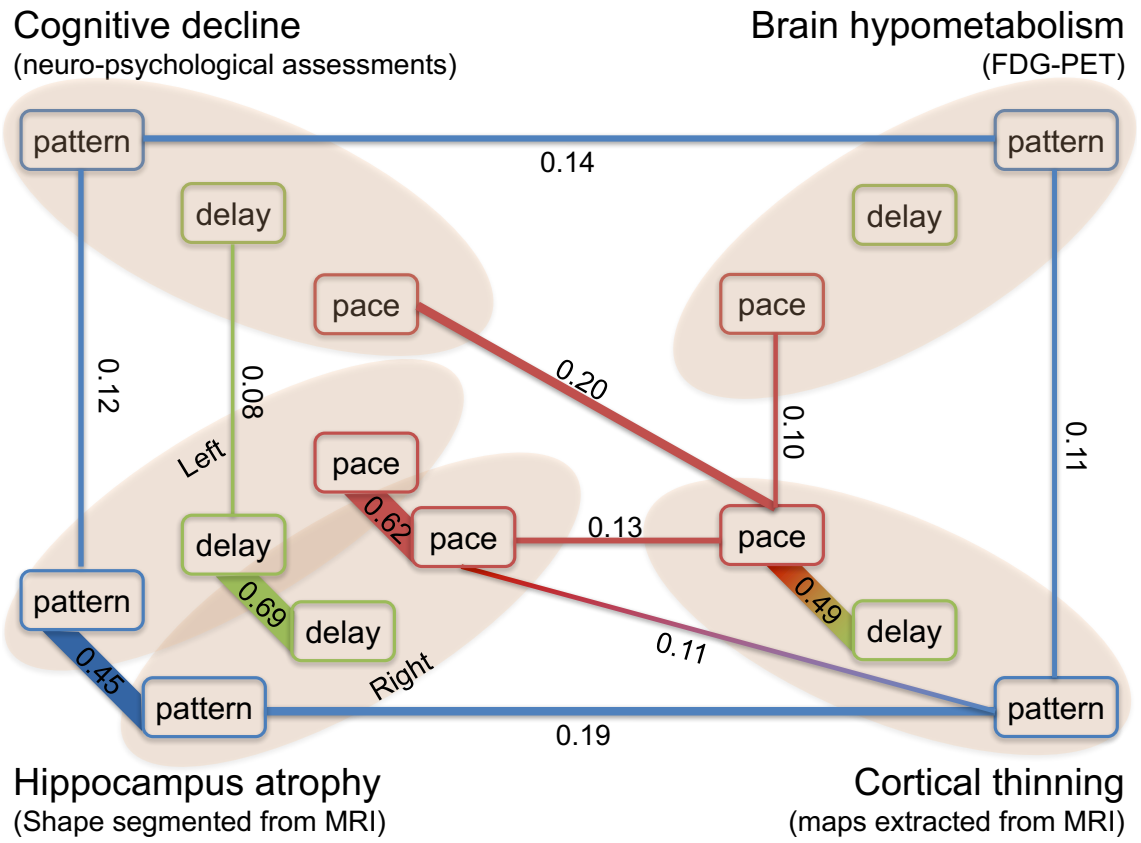


Figure 5.10: Graph of conditional correlations. An edge is shown between two parameters if there is a significant correlations between them given all other parameters. The width of the edge is proportional to the value of the conditional correlation, which is also reported on the edge. The color of the parameter denotes its type and its position the modality.

the pace of progression of the other aspect of the disease.

The age at onset of cognitive decline is only associated with the age at onset of hippocampus atrophy. The start of metabolic alterations and cortical atrophy appears to be independent.

The correlations among the pattern variables form a loop in the graph, suggesting that the profile of cognitive decline, namely the ordering and timing of alterations of difference cognitive functions, is rooted into a complex associations with the spatiotemporal patterns of hypometabolism and structural atrophy across brain regions.

This graph does not support the vision of Jack and colleagues of a cascade of events where hypometabolism induces hippocampal and cortical atrophy, which in turn induces cognitive decline. This graph shows more complex disease mechanisms with different modules inter-acting with each other. The age at onset of cognitive decline is associated with the hippocampal atrophy, and the pace of decline is associated with the cortical atrophy. The ordering and relative timing of decline of several cognitive functions depends on a more complex interplay of pattern of atrophy and hypo-metabolism.

Second, we analyze the co-factors that may influence the progression of the disease, either because they induce a delay or advance of the disease, induce a slowdown or acceleration of the disease, or determine a different profile of alterations regardless of the pace of progression. Our approach allows us to analyze in a coherent and systematic way the associations between all variables of all modalities and a given set of co-factors.

For each modality, we perform a multivariate linear regression between each individual parameters and a series of genetic, biological and environmental factors: sex, APOE-ε4

genotype, presence of amyloidosis, marital status and education level. We identify statistically significant associations using a two tailed t-test at 5% significance level corrected for multiple comparisons with the false discovery rate method (see Methods). Note that in this section, we discard subjects without assessments of amyloidosis (see Table 5.4 for corresponding number of samples).

			hypometabolism (FDG-PET)	hippocampus atrophy (MRI) left hemisphere	hippocampus atrophy (MRI) right hemisphere	cortical thinning (MRI)	cognitive decline (ADAS+MMSE)
genetic	sex female vs. male	accel. factor		$\times 1.27$ CI=[1.11, 1.45] p=2.26e-3**	$\times 1.26$ CI=[1.08, 1.45] p=6.15-3**		$\times 1.46$ CI=[1.10, 1.92] p=8.42e-3 **
		time-shift		-33.6 CI=[-55.8, -11.6] p=3.71e-3**	-29.0 CI=[-53.0, -4.91] p=2.31e-2*		-36.8 CI=[-62.0, -11.6] p=4.48e-3 **
		space-shift		± 0.55 CI=[0.28, 0.82] p=4.00e-4***	± 0.60 CI=[0.34, 0.86] p=3.89e-5****	± 0.48 CI=[0.22, 0.75] p=2.24e-3 **	
	APOE-$\varepsilon 4$ carrier vs. non-carrier	accel. factor		$\times 1.17$ CI=[1.02, 1.33] p=2.77e-2*		$\times 1.42$ CI=[1.12, 1.82] p=2.17e-2*	$\times 1.25$ CI=[1.03, 1.51] 2.17e-2 *
		time-shift		-45.0 CI=[-66.9, -23.2] p=1.57e-4***	-36.8 CI=[-60.5, -13.0] p=4.27e-3**		
		space-shift					
biological	amyloid positive vs. negative	accel. factor		$\times 1.18$ CI=[1.06, 1.32] p=8.20e-3**	$\times 1.23$ CI=[1.09, 1.39] p=4.03e-3**		
		time-shift					-21.9 CI=[-41.2,-2.5] p=2.70e-2 *
		space-shift				± 0.28 CI=[0.05, 0.50] p=2.24e-3 **	
environmental	marital married vs. non-married	accel. factor		$\times 1.25$ CI=[1.07, 1.48] p=1.08e-2*			
		time-shift		-59.5 CI=[-86.6, -32.5] p=1.06e-4***	-52.7 CI=[-82.2, -23.2] p=1.28e-3**		-32.6 CI=[1.8, 63.3] p=3.78e-2*
		space-shift					
	education nb. of years of education	accel. factor					
		time-shift		-6.04 CI=[-9.67, -2.42] p=1.95e-3**	-7.60 CI=[-11.55, -3.64] p=9.53e-4***		
		space-shift					

Table 5.3: Significant associations of individual parameters with genetic, biological and environmental factors: effect sizes, confidence intervals at 95%, and adjusted p-values. Only adjusted p-values below 5% significance level are shown. Time-shifts are in months, other quantities have no units. Directions of space-shift are not signed. The figures on the top of the column "hippocampal atrophy" reads: "hippocampus atrophy progresses faster in women than in men by a factor 1.27 and 1.26 in left and right hemispheres respectively; starts earlier in women by 33.6 and 29 months for left and right hemispheres respectively; and exhibits a different pattern of deformation for men and women in both hemispheres".

Significant associations are shown in Table 5.3. The absence of associations between cofactors and profiles of hypometabolism may be explained also by the fact that focal effects on specific brain areas may be diluted in non-specific regions of interest[Knopman et al., 2014]. Previous findings showing associations are also likely to be due to the comparison of subjects at different ages or disease stages[Knopman et al., 2014, Jack et al., 2015]. In this regard, it is interesting to notice that, except in four occasions, we found associations with parameters that modulate the dynamics of disease progression, not its trajectory. This fact suggests that previous findings showing association of these usual factors with the severity of atrophy, hypometabolism or cognitive decline are likely to be due to a non-proper temporal alignment of individual data.

Our results also show the predominant role of genetic factors to explain the heterogeneity of the manifestation of the disease. In particular, disease progression presents a strong sexual dimorphism for hippocampus atrophy and cognitive decline. The accelerated and earlier atrophy in women translates into an accelerated and even earlier cognitive decline. This dimorphism does not seem to be alleviated by compensatory mechanisms. By contrast, APOE-ε4 carriers also exhibit earlier and more pronounced alterations of their hippocampus, but this effect is, to some extend, alleviated in the onset of cognitive decline, which does not occur earlier than non-carriers, but still at a greater pace. It is as if brain plasticity is able to compensate for the advance of almost 3 years in hippocampal atrophy, but that once the compensation is made, cognitive decline still manifests itself at a faster rate than in subjects without the mutation.

The systematic investigation of association with co-factors allowed us therefore to evidence the prominent role of genetic factors to explain the heterogeneity of disease manifestation, and the presence of compensatory mechanisms in APEO-ε4 carriers.

5.7 Conclusion

We proposed a generic method to learn long-term scenarios of changes from longitudinal data sets, which temporally align and combine several short-term data sequences covering different and unknown stages of progression. The method may be applied to any data that can be represented as points in a Riemannian manifold. It includes unstructured feature vectors, images and geometric shapes. Individual parameters capture the variability in terms of age at onset, pace of progression, and shape or appearance of the model. They decompose therefore the variability due differences in the dynamics of disease progression from the inter-individual differences at the same disease stage. We used it to estimate a model of progression of Alzheimer's disease combining neuro-psychological assessments, structural magnetic resonance imaging and positron-emission tomography. It results in a holistic view of disease progression in multiple domains at an unprecedented temporal and spatial scales (see `www.digital-brain.org`).

From a biological perspective, this digital model of disease progression provides, for the first time, a comprehensive view of how structural and metabolic alterations propagate in the brain, both in space and time, and how they relate to specific sequences of decline in cognitive functions. The individual parameters allow the description and quantification of the heterogeneity of the manifestation of the disease. They allow also the systematic investigation of the co-variations among the parameters controlling the dynamics and pattern of progression for all modalities.

From a clinical perspective, the model may be personalized to new subject's data by automatically adjusting the parameters controlling for the dynamics and appearance of the model. These parameters can be interpreted by a clinician to understand the specific characteristics of each patient. We show that past data are reconstructed with an error of the same order than the uncertainty of the measurement. Likewise, the prediction of the subject's data up to four years in the future is at the same precision as the uncertainty

Table 5.4: Summary statistics of the subject subsets for each data type

	ADAS & MMSE	PET	MRI
Number of subjects	223	157	322
Number of visits	1235	690	1993
Average number of visits per subject (± std)	5.5 (±1.1)	4.4 (± 2.1)	5.8 (± 2.4)
Average age (± std)	76.2 (± 6.9)	74.0 (± 7.2)	74.0 (± 6.7)
Sex ratio (F/M in %)	39.0 / 61.0	41.8 / 58.2	41.2 / 58.8
Amyloid status (+/-/unknown in %)	65.5 / 7.2 / 27.3	77.4 / 7.3 / 15.3	73.2 / 7.1 / 19.7
APOE carriership (%)	62.8	64.2	65.2
Education (mean ± std, in years)	15.8 (± 2.8)	15.8 (± 2.7)	15.9 (± 2.8)
Marital status (married/not married in %)	81.2 / 18.8	82.3 / 17.7	80.9 / 19.1

of the measurements. The model can be used therefore as a digital avatar of the brain of each subject at risk of developing Alzheimer's disease. It accurately simulates images and cognitive performance of the subject in the future, which is key to detect subjects at risk at earlier disease stages than today, and to implement and evaluate personalized therapeutic strategies. Clinical studies need to be conducted now to assess the accuracy of the prediction in a prospective manner and to evaluate the adoption of such techniques for the recruitment of patients in trials and the implementation of early prevention strategies. As it stands, the approach might pave the way to the future advent of precision medicine in neurology.

5.8 Methods

5.8.1 Data Set

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database[1]. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

We used all available visits from ADNI, ADNI-GO and ADNI-2 data sets for all subjects who:

- have been diagnosed with Alzheimer's Disease (AD) at least at one visit;
- have been diagnosed as Mild Cognitive Impaired (MCI) subjects at least at one visit;
- did not revert to Cognitively Normal (CN) stage after being diagnosed as MCI or AD, nor revert to MCI or CN stage after being diagnosed with AD.

350 subjects satisfied the first two criteria. The third criterion excludes subjects with doubtful diagnoses: 28 subjects were then excluded, leading to a subset of 322 subjects representing a total of 2136 visits. We define 3 overlapping sub-sets by selecting different data types: ADAS-Cog & MMSE, FDG-PET images and MRI images. Table 5.4 provides summary statistics of these data sets.

For each subject, we used the following additional data: age at each visit, sex, marital status, educational level, Apolipoprotein E (ApoE) polymorphism, and presence of amyloidosis. More precisely, we define:

[1] http://adni.loni.usc.edu/

- marital status as: married versus non-married meaning widowed, divorced, or never married;
- educational level as the number of years of education;
- ApoE-$\varepsilon 4$ carriership as the presence of at least one allele $\varepsilon 4$ of the ApoE gene;
- Amyloid status as positive if one of these conditions was met at one visit at least:
  - a Standard Updake Value ratio (SUVR), normalised by the entire cerebellum, greater than 1.1 in a PET image acquired with Florbetapir (AV-45) compound [Clark et al., 2012, Landau et al., 2013];
  - an average SUVR, normalised by the cerebellum, greater than 1.47 in a PET image with a Pittsburgh compound B (PiB) [Landau et al., 2013];
  - a level of beta amyloid 1-42 (Aβ42) (measured with the Roche Elecsys assays[2]) in the cerebrospinal fluid (CSF) lower than 1098 pg/mL[Schindler et al., 2018];

  unknown if no values of CSF biomarkers and no AV45 or PiB PET images were available at any visit in the ADNI-merge file; and negative otherwise.

Not counting 7% of the population with an unknown amyloid status, 83% of the remaining held a stable positive status status across all their visits, while 9% have their visits consistently negative – the last 8% present an evolution of its status through time. The stable positive and negative individuals allows to distinguish the subjects who have developed Alzheimer's Disease in presence of amyloidosis, from those who developed the clinical signs of the disease without the significant development of amyloid plaques.

5.8.2 Pre-processing and feature extraction

We used the global MMSE score and aggregated scores from the 13 items of the ADAS-Cog. Furthermore, we pooled the 13 items into four sub-categories: memory by adding items 1, 4, 7, 8 and 9, language by adding items 2, 5, 10, 11 and 12,praxis by adding items 3 and 6, and concentration with item 13. Each value is normalised by the maximum possible value for the global score or for each category.

Regional FDG-PET SUVR were extracted using the second version of the Automated Anatomical Atlas[3] (AAL2)[Tzourio-Mazoyer et al., 2002, Rolls et al., 2015] with 120 regions covering the cortex and the main subcortical structures, using the open-source community software Clinica[4] [Routier et al., 2018]. The software performs intra-subject registration of the FDG-PET image into the space of the subject's T1-weighted MRI image using Statistical Parametric Mapping[5] (SPM) software (version 12)[Penny et al., 2011]. The PET image is then spatially normalised into MNI space using DARTEL deformation model of SPM, and its intensities normalised using the average uptake value in the pons as reference region. The SUVR map is obtained by averaging resulting intensities in each region of the atlas [Samper-González et al., 2018].

The MRI images were first processed independently with the cross-sectional pipeline of the FreeSurfer[6] software (version 5.3.0) [Fischl and Dale, 2000, Fischl et al., 2002]. The longitudinal FressSurfer pipeline is then used to create subject-specific templates from the successive data of each subject and refine image segmentations[Reuter et al., 2012]. These

[2]http://adni.loni.usc.edu/new-csf-a%CE%B21-42-t-tau-and-p-tau181-biomarkers-results-from-adni-biomarker-core-using-elecsys/

[3]http://www.gin.cnrs.fr/fr/outils/aal-aal2/

[4]http://clinica.run/doc/Pipelines/PET_Volume

[5]www.fil.ion.ucl.ac.uk/spm/

[6]https://surfer.nmr.mgh.harvard.edu

segmented images are used then to extract a cortical thickness map, and a mesh of the left and right hippocampus.

We used the cortical surface mesh projected onto the average space called FSaverage with 163,842 vertices. For dimensionality reduction purposes, we then

- inflate the FSAverage mesh to a sphere using FreeSurfer, on which 3,658 vertices (called patch-nodes) are selected to map the whole sphere uniformly,
- associate each vertex to its closest patch-node, resulting in a parcellation of the cortical mesh into 3,658 patches that are uniformly distributed over the surface, where a patch contains on average 44 vertices,
- compute the average value of the cortical thickness in each patch.

We also align the skull-stripped images with an affine 12-degrees-of-freedom transformation onto the Colin27 template brain[7], using the FSL 5.0 software[8][Woolrich et al., 2009]. Mesh representations of the geometry of the left and right hippocampus result from the following steps:

- the volumetric segmentations of the hippocampi obtained by FreeSurfer are transformed into meshes using the aseg2srf software[9],
- the resulting meshes are decimated by a 88% factor using Paraview, 5.4.1[10][Ahrens et al., 2005],
- then aligned using the previously-computed global affine transformation estimated with the FSL software,
- residual pose differences among subjects are then removed by rigidly aligning the meshes from the baseline image of each subject to the corresponding hippocampus mesh in the Colin27 atlas image, this transformation with 6 degrees of freedom being computed with the GMMReg software[11][Jian and Vemuri, 2011],
- the same transformation is eventually used to align the meshes from the follow-up images of the same subject.

5.8.3 Data representation and choice of Riemannian metrics

The statistical model may be written as:

$$y_{ij} = \eta^{w_i}\left(\gamma_0\right)\left(\psi_i(t_{ij})\right) + \varepsilon_{ij} \tag{5.1}$$

where

- $\gamma_0 : t \to \mathrm{Exp}_{p_0}\left((t - t_0)v_0\right)$ is the population average trajectory in the form of a the geodesic passing at point p_0 with velocity v_0 at time t_0 (Exp denotes the Riemannian exponential as a concise way to write geodesics),
- $\eta^{w_i}\left(\gamma_0\right) : t \to \mathrm{Exp}_{\gamma_0(t)}\left(P_{\gamma_0}^{t_0,t}(w_i)\right)$ is the exp-parallelisation of the geodesic γ_0 in the subject-specific direction w_i, called space-shift, as depicted in Fig. 5.1 ($P_{\gamma_0}^{t_0,t}(w_i)$ denotes the parallel transport of the vector w_i along the curve γ_0 from $\gamma_0(t_0)$ to $\gamma_0(t)$),

[7] http://www.bic.mni.mcgill.ca/ServicesAtlases/Colin27
[8] https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/
[9] https://brainder.org (version of July 2009)
[10] www.paraview.org
[11] https://github.com/bing-jian/gmmreg (version of July 2008)

- $\psi_i : t \rightarrow \alpha_i(t - t_0 - \tau_i) + t_0$ is a time-reparameterising function, where α_i is a subject-specific acceleration factor and τ_i a subject-specific time-shift.

For identifiability purposes, we impose the vectors w_i to be orthogonal to the velocity v_0 in the tangent-space at point p_0. Parallel transport being isometric, this property then holds at any time point. The random effects of the model are:

- an acceleration factor α_i, which accounts for the variations in pace of disease progression, and therefore distinguishes the fast from the slow progressing individuals,
- a time-shift τ_i, which accounts for the variations in age at onset, and therefore distinguishes the early from the late onset individuals,
- a space-shift w_i (a vector pointing a direction on the manifold), which accounts for the variations in the position of the individual trajectory, and therefore captures differences in patterns of disease progression (magnitude of the effects, re-ordering of events, change in the spatial pattern of alterations for instance, as detailled below).

Their prior distributions are a log-normal distribution for the acceleration factors, zero-mean Gaussian distribution for the time-shift. Space-shifts are decomposed into a series of independent components: $w_i = As_i$ where the columns of A contains a pre-defined number of vectors in the orthogonal space of v_0, called components, and s_i are random weights, called sources and distributed according to a normal distribution for non-Euclidean metrics and a Laplace distribution if the manifold is Euclidean, for identifiability purposes.

We concatenated the aggregated MMSE score and the four sub-categories of the ADAS-Cog to build a 5-dimensional feature vector, which is seen as a point in a 5-dimensional hyper-cube $[0, 1]^5$. We provide this manifold with a diagonal metric tensor which ensures that a geodesic in this hyper-cube is formed by 5 logistic curves, that are further assumed to be parallel to each others: $\gamma_{0,k}(t) = \gamma_{\text{logit}}(t+\delta_k)$ with $\gamma_{\text{logit}}(t) = \left(1 + \frac{1-p_0}{p_0} \exp\left(\frac{-v_0(t-t_0)}{p_0(1-p_0)}\right)\right)^{-1}$. A parallel shift of the population geodesics in this hyper-cube translates into a change in the temporal delay between the logistics curves of each coordinate[Schiratti et al., 2015, Schiratti et al., 2017]: $\eta_k^{w_i}(\gamma_0)(t) = \gamma_{\text{logit}}\left(t + \delta_k + \frac{w_{i,k}}{\dot{\gamma}_{\text{logit}}(t_0+\delta_k)}\right)$.

Maps of cortical thickness take the form of a vector of 3,658 coordinates corresponding to the measurements values at every patch node, seen as a point in the Euclidean space $\mathbb{R}^{3,658}$. Geodesics are straight-lines in this space, where each coordinate $k \in \{1, \ldots, 3,658\}$ is a one-dimensional straight-line of the form: $\gamma_k = p_k + v_k(t - t_0)$. The exp-parallelisation in the Euclidean space corresponds simply to a translation, so that each coordinate is transformed into[Koval et al., 2017]: $\eta_k^{w_i}(\gamma_0) = p_k + w_{i,k} + v_k(t - t_0)$. The fixed-effects p_0 and v_0 are vectors of size 3,658 whose k-th coordinate p_k and v_k are the reference intercept and slope at the k-th patch respectively. We select a sub-set of 911 control nodes $(c_i)_{1 \leq i \leq 911}$ among the patch nodes, and create a mapping which generates 3,658 values from the 911 values using a manifold-kernel smoothing interpolation. Let the k-th path node be $x_k \in \mathbb{R}^3$, corresponding to the Euclidean coordinate of the center of the path. The value $p_k = p(x_k) = \sum_{i=1}^{911} \exp\left(-\frac{d(x_k,c_i)^2}{\sigma^2}\right) \beta_i$ corresponds to the value of the parameter at the k-th node. The β_i are the 911 values at the control nodes c_i, the distance $d(x_k, c_i)$ is the geodesic distance on the cortical surface mesh between patch node x_k and control nodes c_i, and σ is a scalar parameter taken equal to $20\,mm$, which is approximately 2.5 times the average distance between neighbors control nodes (namely the three closest control nodes to a given control node). The same kernel mapping is used to generate the values $(v_k)_{1 \leq k \leq 3,658}$. By construction, the maps generated by this operation are varying smoothly over the surface mesh and are controlled by a smaller number of parameters.

Each PET measurement is characterised by a vector in $\mathbb{R}^{120}$ whose k-th coordinate correponds to the the average SUVR value on the k-th region of interest (ROI) of the

AAL2 atlas. We take the same approach as for the cortical thickness maps. The centroids of the regions in the AAL2 anatomical atlas is considered as a fully connected graph (so that the geodesic distance on the graph is the Euclidean distance between centroids), and all centroids are taken as control nodes. Spatial smoothing parameter is taken here of $\sigma = 15\,\text{voxels} = 22.5\,mm$.

For hippocampus meshes, we consider a finite-dimensional manifold of diffeomorphisms of the ambient 3D space that contains the hippocampus[Durrleman, 2018, Durrleman et al., 2014]. This manifold is parameterised by a set of momentum vectors $(m_k)_k$ attached to a set of control points $(c_k)_k$. This set of control points is seen as a dynamic system of particles which follows geodesics derived from the Hamiltonian: $H(c,m) = \sum_{k,l} \exp\left(-\frac{||c_k - c_l||^2}{\sigma^2}\right) m_k^T m_l$, where T denotes the transpose of a vector. The exponential function is a positive definite kernel defining the co-metric on this manifold as the matrix $K(c) = \left[\exp\left(-\frac{||c_i - c_j||^2}{\sigma^2}\right)\right]_{i,j}$. The deformation scale σ is an hyperparameter of this metric, and is set to $10\,mm$ in this application. For each configuration of control point $c(t)$ and momentum vector $m(t)$ at time-point t, we derive a continuous vector field $v_t(x) = \sum_k \exp\left(-\frac{||c_k(t) - x||^2}{\sigma^2}\right) m_k(t)$ for any point x. The trajectory of a set of control points and attached momenta therefore translate into a time-dependent family of vector fields. These vector fields are integrated in time from the identity map into a flow of diffeomorphisms. Diffeomorphisms along these geodesics are applied to a template shape $\mathcal{O}$ to give a smooth trajectory of shape deformation: $t \to \phi^{c,m}(t)\,(\mathcal{O})$, where we denote by $\phi^{c,m}(t)$ the diffeomorphism arising from control points c, momentum vectors m at time-point t. The set of control points and the template shape play the role of the point p_0, and momentum vectors the role of the cotangent-space vector $K(c)^{-1}v_0$.

This construction allows the exp-parallelisation of the trajectory of control points in the manifold, which translates into another trajectory of shape $\eta^{w_i}(\phi^{c,m})(t)(\mathcal{O})$. This parallel trajectory transports the deformation patterns of the baseline geodesics into a new geometry[Bône et al., 2018].

In this construction, the template shape $\mathcal{O}$ becomes a new fixed-effect of the statistical model. We use the metric on currents[Vaillant and Glaunès, 2005] to measure a distance between the deformed template and the observations, which are meshes with different topology and number of vertices. This distance appears when maximising the likelihood of the residual noise ε_{ij}[Durrleman, 2018, Gori et al., 2017]. It is homogeneous to an area, and its units is therefore in mm^2. One of its main advantage is that it smooths out small protrusions and is insensitive to small holes or topology changes in the meshes, making it robust to segmentation errors and avoiding intensive mesh pre-processing. The scale at which the metric is insensitive to these artifacts is an hyperparameter of this attachment metric[Durrleman et al., 2008, Gori et al., 2017], and is set to $5\,mm$ in this work.

5.8.4 Calibration

We use the Monte-Carlo Markov Chain Stochastic Approximation Expectation Maximisation (MCMC-SAEM) algorithm [Kuhn and Lavielle, 2004, Allassonnière et al., 2010, Allassonnière et al., 2015] to calibrate the model. It is an iterative algorithm that solves the following approximate optimisation problem at each iteration:

$$\theta_{k+1} = \text{argmax}_\theta \sum_{i=1}^{N} \int \log\left[p(\{y_{ij}\}_j, z_i; \theta)\right] p(z_i | \{y_{ij}\}_j; \theta_k) dz_i \tag{5.2}$$

At each iteration, it loops over the three following steps.

- simulation of candidate value of the random-effects z_k by running several steps of a Metropolis-Hasting method within a block Gibbs sampler with $p(z|\{y_{ij}\}_j, \theta_k)$ as

ergodic distribution. This step essentially draws a candidate from a random walk sampler and accept or reject this candidate depending notably on the value of the complete likelihood $p(\{y_{ij}\}_j, z_k, \theta_k)$, which measures how well the data generated with the candidate z_k, i.e. $f(\theta_k, z_k, \{t_{ij}\}_j)$, resembles the actual observations $\{y_{ij}\}_j$.

- stochastic approximation using a Robbins-Monro method which keeps adding the terms within the integral with decreasing gains. For distributions belonging to the curved-exponential family (which is ensured in all cases but hippocampus by assuming parameters to be drawn from a prior distribution), it amounts to keep track of a set of sufficient statistics.

- maximisation over the parameters, which is done by updating the parameters with a fixed number of gradient descent steps for hippocampus meshes, or in closed form in other cases.

The following procedures are preceded for the initialisation of the algorithm. For the hippocampus meshes, an average model was first computed by estimating an atlas [Gori et al., 2017] to initialise the template shape and the matrix A, individual geodesic regressions [Fishbaugh et al., 2014] were then estimated to initialise the velocity vector v_0. For the cortical thickness and SUVR maps, the coordinates p_k of the initial position p_0 corresponds to the mean value over all the data on the corresponding region. As for the initial velocity v_0, each coordinate v_k corresponds to the average slope of linear regressions performed on each subject independently. In the case of the cognitive scores, a random initialisation was used.

The implementation of this algorithm is available in the software Deformetrica[12] for the longitudinal shape model, and in the Leasp software[13] for the other cases.

Model synchronisation. The time-warp functions $\psi_i^{[m]}(t_{ij})$ maps the age of the i-th subject at the j-th visit, t_{ij} to a disease stage on the normative time-line for the data type m. Taking the model of cognitive decline as a reference ($m = \text{cog}$), we look for a temporal mapping $\Phi^{[m]}(t) = \lambda^{[m]} \cdot t + \mu^{[m]}$ between the normative time-line for data type m and the one of the cognitive decline so that $\Phi^{[m]} \circ \psi_i^{[m]}(t_{ij})$ is as close as possible to $\psi_i^{[\text{cog}]}(t_{ij})$ by minimising $\sum_{i=1}^{N} \sum_{j=1}^{N_i} \left| \lambda^{[m]} \cdot \psi_i^{[m]}(t_{ij}) + \mu^{[m]} - \psi_i^{[\text{cog}]}(t_{ij}) \right|^2$, which admits a closed form solution. This steps allows the synchronisation of different models of disease progression.

Estimation of age of diagnosis. The time-point $\psi_i^{[\text{cog}]}(t_i^{\text{diag}})$ maps the age at which the i-th subject was diagnosed with the disease, i.e. t_i^{diag}, to a disease stage that ideally would be the same for all subject. In practice, we used the average stage $t^{\text{diag}} = \frac{1}{N} \sum_{i=1}^{N} \psi_i^{[\text{cog}]}(t_i^{\text{diag}})$ as an estimate of the diagnosis time on the normative time-line of the model of cognitive decline. Note that this estimate is the best predictor of the age at diagnosis, as it minimises $\sum_{i=1}^{N} \left| \{\psi_i^{[m]}\}^{-1}(t^{\text{diag}}) - t_i^{\text{diag}} \right|^2$.

5.8.5 Personalisation

Once the model is calibrated on a longitudinal data set, we personalise it to the temporal sequence $\{y_{ij}, t_{ij}\}_j$ of any target subject i by finding the values of the random-effects z_i that maximises the posterior log-likelihood:

$$\log p(z_i | \{y_{ij}\}_j, \hat{\theta}) = \log p(\{y_{ij}\}_j | z_i, \hat{\theta}) + \log p(z_i | \hat{\theta}) + \text{Constant}. \tag{5.3}$$

[12] www.deformetrica.org

[13] https://gitlab.icm-institute.org/aramislab/longitudina

The first term $\log p(\{y_{ij}\}_j|z_i,\hat{\theta}) \propto -\sum_{j=1}^{N_i} \left\|y_{ij} - f(z_i,\hat{\theta},t_{ij})\right\|^2$ measures the distance between the observations and the current fit of the model to this data. The norm considered is the one appearing in the noise likelihood: sum of squared differences for neuropsychological assessments, PET images and cortical thickness maps, and the currents distance between meshes for hippocampus meshes[Vaillant and Glaunès, 2005]. The second term is a prior on the likelihood of the random-effects. This minimisation problem is solved using Powell's method for the hippocampus meshes, and the L-BFGS algorithm [Byrd et al., 1995] for all other modalities. Both algorithms were taken from the SciPy 1.1.0 library[14].

We performed model personalisation using the whole data set as a training set, or in a five fold cross-validation setting. On the one hand, we personalise the model to the training subjects using the whole data set, yielding a set of individual parameters for each subject. On the other hand, we estimate the model using 80% of the subjects and then personalise it to the remaining 20% subjects, yielding a set of individual parameters for test subjects only. After five splits, we recover a full set of individual parameters estimated in a cross-validation setting, which is compared to the first set of individual parameters. The cross-validation procedure produces five sets of fixed effects that are compared to the set of fixed effects using the whole data set as training set.

In any case, at convergence, the residual $\epsilon_{i,j} = y_{ij} - f(\hat{z}_i,\hat{\theta},t_{ij})$ for the optimal value of the random-effect $\hat{z}_i$ is called the **reconstruction error** of the j-th observation of the i-th subject. Note that in the case of the hippocampus meshes, only the absolute reconstruction error $|\epsilon_{ij}|$ can be computed, because the currents representation is a multivariate vector, of which we take the norm[Vaillant and Glaunès, 2005].

We compare the distribution of the reconstruction errors with the uncertainty in the measurements, which is estimated as follows. In the ADNI protocol[Jack Jr et al., 2008, Jack Jr et al., 2010a], most MRI sessions consist of a pair of test and re-test MRI, namely two scans performed on the same day one immediately after the other one. For 1841 out of 1993 MRI sessions, we measure therefore the differences between the MRI derived data (hippocampus meshes and cortical thickness maps) when using the test or the re-test image. These differences give an empirical distribution of the noise due to variations in image acquisition and processing.

For PET derived data, we use the baseline and follow-up scans of stable cognitively normal and amyloid negative subjects in ADNI, as a proxy to test / re-test data (125 subjects, 244 visits with a follow-up time of 18 months). For those subjects, the changes in glucose metabolism over a 18 months period is supposed to be negligible compared to all the other factors affecting the measurements such as variations in reaction to radiotracers, and methods for PET reconstruction, image correction and extraction of regional measurements.

Test / re-test studies have shown a that the MMSE, which scales from 0 to 30, is subject to a difference between two sessions, whose standard deviation ranges from 1.3 for a one-month interval[Clark et al., 1999] up to 1.82 for a 1.5 year long interval[Hensel et al., 2007], thus representing a standard deviation of 4.3% to 6%. Another study[Standish et al., 1996] measured the former ADAS-Cog that scales between 0 and 70 three times at a 2-week interval, with an agreement between raters. The inter-ratter standard deviation is of 9.64 between the first and second test, and of 6.79 between the second and third test. The intra-rater standard deviation is of 8.16 between the first and third visit. This corresponds to a standard deviation ranging from 9.7% to 13.8%. On average, we consider such neuropsychological assessments to have a zero-mean Gaussian distribution of noise with standard deviation of order 7%.

[14] https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.minimize.html

5.8.6 Prediction

Let's consider an individual such as it is possible to splits his observations $(y_{ij}, t_{ij})_{1 \leq j \leq n_i}$ into $(y_{ij}, t_{ij})_{1 \leq j \leq j^{\text{present}}}$ and $(y_{ij^{\text{futur}}}, t_{ij^{\text{futur}}})$ such that $t_{ij^{\text{futur}}} = t_{ij^{\text{present}}} + T$ where T is the time to prediction, e.g. 3 or 4 years. We consider the parameters θ previously estimated. In the case where this individual belongs to the initial cohort used to calibrate the entire model, then we consider the parameters θ estimated in the cross-validation run for which this particular patients belonged to the test set.

Given θ, we compute z_i by personalizing the model with the visits $(y_{ij}, t_{ij})_{1 \leq j \leq j^{\text{present}}}$, while fixing the pace parameter to 1 for patients that present only one past visit for the hippocampus. Then at time $t_{ij^{\text{future}}}$, it is possible to compute the prediction $\tilde{y}_{ij^{\text{future}}} = f(\theta, z_i, t_{ij^{\text{future}}}) + \epsilon_{ij}$ where ϵ_{ij} is a realization of the estimated noise distribution, and, to compare it to the real value $(y_{ij^{\text{future}}})$.

Due to the smoothness constraints in the reconstruction of cortical thickness maps, the prediction are systematically biased. We estimate the bias as $d^{\text{smooth}} = \frac{1}{j^{\text{seen}}} \sum_{j=1}^{j^{\text{seen}}} (y_{ij} - f(\theta, z_i, t_{ij}))$ and made the final prediction as $\tilde{y}_{ij^{\text{future}}} + d^{\text{smooth}}$.

5.8.7 Conditional correlation

We compute the conditional correlations among all the pairs of individual parameters (pace, delay and pattern for each modality). It is represented as a matrix whose entries are the correlations between a given pair of parameters conditionally to all the other ones. The conditional correlation matrix is seen as graph whose vertices are the individual parameters, and whose edges are weighted by the pairwise conditional correlations.

We use the GGMselect algorithm [Giraud et al., 2012] to obtain a first, very sparse, graph containing only the most basic conditional correlations. We proceeded to construct a more complex one by sequentially adding the next most potent edges: at each step of the procedure, we added a single edge to the current graph. This edge is chosen using parallel LARS [Efron et al., 2004] procedures on each node. Each LARS proposes a list of candidate graphs, among which we select the one minimizing the Kullback-Leibler divergence with regards to an estimation of the real distribution.

We stop adding edges once the Kullback-Leibler divergence between a set of unseen data and the proposed graph is at its lowest.

5.8.8 Cofactor analysis

We take the series of random-effect estimates after model calibration and personalisation on a given training data set. For each data type, we look for correlations between the values of these random-effects and a series of co-factors: sex, APOE status, marital status, level of education and amyloid status. On the one hand, the series of co-factor is regressed against the uni-dimensional temporal random effects (time-shift τ_i and acceleration factor α_i); the statistical significance of the slope coefficients is assessed by a two-sided t-test. On the other hand, for the multivariate vector of sources (s_i), we perform a 2-blocks partial least square [Abdi, 2003] method to identify correlations between a linear combination of sources and co-factors. The resulting series of p-values are corrected for multiple comparisons using the False Discovery Rate (FDR) method.

When a significant association between a linear combination of sources (i.e. a vector d in the multivariate space of sources) and a categorical co-factor has been found, we project the individual source estimates on this direction (i.e. $b_i = d^T s_i$) and compute the distance between the empirical means of each class ($\delta_{12} = \bar{b}_2 - \bar{b}_1$). We select two points in the source space at $u = \pm a\delta_{12}/2$ to represent the typical configuration of each class, where $a = 1$ (for the cortical thinning) or 3 (for the hippocampus shape) is a factor to amplify

differences for better visualisation. We then reconstruct the corresponding typical data by computing the exp-parallel curve in the direction u at a given time-point t: $\eta^{Au}(\gamma_0)(t)$.

5.8.9 Code availability

Software used for the pre-processing of the data have been listed above in footnotes. The code used for calibration, personalisation and simulation is freely available, in the Deformetrica software`www.deformetrica.org` for shape data, and in the Leasp software`https://gitlab.icm-institute.org/aramislab/longitudina` for the other cases.

Part IV

Simulation of Virtual Trajectories of Progression and Longitudinal Data Sets

Chapter 6

Simulation of Virtual Patients

This chapter extensively takes advantage of the generative and mixed-effects characteristics of the model to either simulate measurements that have been missing or that correspond to future disease stages, or, to create virtual cohorts of simulated patients that allow to significantly enhance the predictive power of algorithms, especially the prediction of the MMSE for MCI subjects up to 4 years in advance.

Contents

Abstract

Longitudinal databases, namely repeated observations per individual, provide data that are not straightforward to deal with e.g. unequal number of observations per individual, temporally unaligned measurements and varying time-steps between observations. In a medical context, this is further worsen by datasets of relatively small size that prevent from properly benefiting from the information contained in the sequences of individual observations. In this paper, we show that these challenges are related and can be tackled by a data augmentation framework that we introduce. The latter takes advantage of the information provided by each individual sequence of measurements to characterize the long-term disease progression and its spatiotemporal variability, thanks to a mixed-effects model. Based on that, we are able to derive the mean trajectory in order to reconstruct continuous individual trajectories to impute missing values but also to predict future time-points. Moreover, this generative model estimates the distribution of the random effects from which it is possible to simulate virtual patients. They can be of interest to balance or un-bias real cohorts and to simulate virtual cohorts in order to improve the predictive power of algorithms.

6.1 Introduction

To better treat diseases or prevent their apparition, there is an overwhelming need for techniques that characterize their progression. This is particularly significant for neurodegenerative diseases that are diagnosed belatedly due to the lack of knowledge about the prodromal stages and the early biomarkers. A key factor to circumvent this obstacle dwell in the fact that each patient carries information about a part of the overall disease progression. This is particularly accentuated in longitudinal databases where the patients, observed at multiple time-points, describe a short-term evolution of the disease progression. In the case of temporal dynamics that evolve during period of times that are longer than each individual snapshot, such as in neurodegenerative diseases, the patients are likely to be screened at different disease stages. Furthermore, they might present unaligned temporal dynamics (e.g. fast versus slow progressors, early versus late converters) that are not straightforward to realign and compare. On top of this temporal variability, each patient might present slight variations in term of patterns of evolution compared to a typical long-term scenario of change. An over-and-above issue is the diversity of modalities and biomarkers that favor omitted measurements in the dataset. All these concerns are in fact different aspects of the *missing value* problem : the absence of an observation at a time-point of interest. This can be a given age, a particular disease stage, but we can show that this is also tightly related to the prediction of future time-points if we consider that the missing value is in the future. Some have developed algorithms to handle missing values during the training, but fewer techniques have been developed to impute them. One of them, called the multiple imputation [Rubin, 2004] has been extended to longitudinal data [Biering et al., 2015, Young and Johnson, 2015, De Silva et al., 2017]. However it ultimately rely on data missingness assumptions that are essentially subjective or require prior knowledge.

Besides missing values for a given patient, medical longitudinal databases, comparatively to other Machine Learning (ML) fields, are often cursed by a relatively small number of individuals. This second type of unobserved data has been addressed in other fields with *data augmentation* techniques whose aim is to simulate virtual data that reproduce the characteristics of the initial database. They are essentially intended to feed algorithms with additional data in order to prevent them from being skewed by insufficient training data. Most of the literature focuses on techniques for independent and identically distributed observations which is an unrealistic hypothesis for sequence data. Some proposed methods for uni-dimensional time-series that rely on a continuous transformation of the time domain by warping, slicing or sliding the time-window [Le Guennec et al., 2016]. Such techniques do not apply to disease progression where time-invariance cannot hold true.

Other attempts have discussed data augmentation techniques in particular model designed with linear mixed effects or known laws, which is unable to properly capture complex dynamics [Ryu et al., 2011, Tang, 2015, Tang, 2019]. Alternatively, [Dalca et al., 2015] relied on general linear models to describe the effects of the neurodegenerative diseases at the voxel level, predicting their change in the next years as a regression of genetic and clinical markers . Recently, generative adversarial networks (GAN) [Goodfellow et al., 2014] have received some interest to generate virtual data but the generative process is poorly understood and more importantly, they require large datasets to be trained, which is specifically the problem we face. However, to take fully advantage of the latent spaces of GANs, some studies have introduced regularizers from the longitudinal structure of the data. They are based on the ability to encode the input images into a latent space that has a longitudinal structure which ultimately enables to decode it into MRI slices at later time-points. Recently, [Ravi et al., 2019] simulated, from a slice of the MRI at baseline, the effect of the disease and therefore the resulting MRI slice in the next few years. Additionally [Xia et al., 2019] conducted similar simulations to reproduce the effect of normal ageing on the brain alteration, particularly 2D slices of MRI scans.

A direct consequence of the limited number of samples in medical datasets is that they may be unbalanced and present some bias such as the inclusion criteria that might favor one type of progression over another. This prevents from drawing conclusions that generalize well. Due to the small size of the dataset which prevents from subsampling the most abundant sub-population, the only reliable option is to simulate new patients that mimics the characteristics of the sub-population to increase. This essentially stresses the fact that the longitudinal data augmentation technique to consider should also properly describe the interactions between the characteristics to balance (gender, genetic mutations, socio-demographic factors, ...) and the factors that modulate the disease progression. While this is an important area of interest in the representation learning community [Bengio et al., 2013], these interactions are not taken into account yet in GANs or other data augmentation techniques which restrict their use when balancing datasets. In the end, being able to simulate virtual patients that imitates initial cohort also allows to generate *virtual cohorts* whose sharing policies are less restrictive than usual anonymization requirements.

In this paper, we characterize precisely the issue of insufficient data in longitudinal setting, which can alternatively take the form of missing values for given patients, or, the simulation of an entire patient, with multiple observations. We also show that these challenges are tightly related to the prediction of future time-points but also to un-biasing and balancing initial cohorts with simulated patients. To tackle these challenges, we adapt a Bayesian mixed effects model [Schiratti et al., 2017] that captures the group-average spatiotemporal long-term trajectory of disease progression out of individual sequence of measurements. This mean trajectory is learnt in a mixed-effects setting such that the individual trajectories are variations of this long term progression. As they are continuous trajectory, in the sense that they are defined for any time t, it is possible to generate individual values at missed ages (imputation of missing values) or at later time-points (prediction of future stages). The aforementioned variability in term of disease trajectory is defined with random variables whose distribution is learnt during the estimation of the long-term progression. Therefore, this generative characteristic of the model allows to draw new reasonable variations of the mean trajectory that correspond to virtual individuals. As for real subjects, their measurements can be computed for any arbitrary age t. As the model allows to relate these variations to individual cofactors (e.g. gender, genetic mutations, marital status), it is possible to draw virtual patients with given cofactors to unbias or balance the original cohort. The resulting augmented dataset can be used to enhance standard ML algorithms. To sum up, the purpose of this framework is three-fold :

- the simulation of virtual cohorts of longitudinal patients, to unbias initial cohort or to spare sharing policies,

- the imputation of missing values and the prediction of future time-points in longitudinal studies,

- the enhancement of longitudinal data based algorithms thanks to virtual cohorts of simulated data.

We evaluated this framework with a mixed-effects model specially designed to handle cognitive assessments that are represented by logistic shapes varying from a normal to an abnormal disease stage. We performed experiments on patients with mild cognitive impairments (MCI) but also other diagnosed with Alzheimer's disease (AD). To demonstrate the the simulated data mimic the real one, we first fool a discriminator between both. Then, we show that imputing missing values and predicting future time points further ensure the ability of the framework to handle longitudinal data. Finally, we simulated virtual patients to improve state-of-the-art results on the prediction of the mini-mental state examination (MMSE) for MCI subjects up to the noise level and 4 years in advance.

6.2 Related Work

6.2.1 Missing Values Imputation

Widely used in medical communities, the multiple imputation (MI) method [Rubin, 2004] to impute missing values simulate few data guesses that have to be random variations of a given model. Its aim is to perform further analysis on all the guesses and combining the results, by averaging the parameters estimates for instance. Few methods have extended the technique to longitudinal data [Biering et al., 2015], for instance in the presence of time varying covariates [De Silva et al., 2017]. These models rely on assumptions on the missingness of the data, whether its is random (i.e. occurring independently of the data nature) or not, which can be hard to demonstrate [Young and Johnson, 2015]. Furthermore, besides the need of a proper imputation model, dealing properly with all the imputations is not straightforward and rely on diverse assumptions [Spratt et al., 2010].

6.2.2 Data Augmentation Techniques

GAN, recently introduced [Goodfellow et al., 2014], are based on a discriminative model that is able to discriminate between real and fake data, and, a generative model that samples fake data. The latter is often used to generate new realistic samples. In longitudinal settings, the generator takes the form of a recurrent neural network that outputs sequences of data. For this reason, it is non-trivial to propagate the gradient updates from the discriminator to the generator [Yu et al., 2017] as the part of the generated sequence to be updated is unclear. Said differently, it is not trivial to assess which part of the fake input data was inadequately simulated. Some attempted to overcome this challenge by embedding it in a reinforcement learning setting where the discriminator output is seen as a reward to the generator [Li et al., 2017]. However, these models ultimately rely on large databases which are typically inaccessible in the targeted medical applications. On top of that, GAN cannot properly exhibit interpretable representations in the latent space. While not in longitudinal settings, some research have discuss its disentanglement [Chen et al., 2016] but it relies on knowing which input label modulate the outcome. In our case, this corresponds to prior knowledge over the factors that modulate the disease progression. Unfortunately, this information is hidden to the observer.

6.3 Longitudinal Data Augmentation Framework

In the following, we consider a longitudinal dataset $\mathbf{y} = (y_{ij}, t_{ij})$ $(y_{ij} \in \mathbb{R}^n)$ where the measurements $(y_{ij})_{1 \leq j \leq k_i}$ of the i-th individual are observed at times $t_{i1} < \ldots < t_{ik_i}$

Recently, several generative mixed-effects model have been released to deal with longitudinal data. In [Schiratti et al., 2017], the authors introduced a technique to recombine the individual short-term data into a long-term disease progression. This work has been extended to data that present a spatial structure [Koval et al., 2018b], to deformations [Bône et al., 2018], but also to handle missing values during training [Couronne et al., 2019]. Finally, the authors of [Louis et al., 2019] designed a non-parametric version of the disease progression model. They all consider that there exists an average trajectory parametrized by θ, continuous with respect to the time t. The individual trajectory of patient i derive from this group-average scenario thanks to the random effects $\mathbf{z}_i$. For the sake of clarity, we write his observation at time t_{ij} as :

$$y_{ij} = f_\theta(t_{ij}, z_i) + \epsilon_{ij}\,, \tag{6.1}$$

where the noise is considered Gaussian $\epsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.

The mixed-effects model proposed in Eq. 6.1 relies on model parameters θ shared by the population, i.e. the fixed effects , and on individual parameters z_i, the random effects. In this paragraph, we define the three procedures we further consider to leverage the potential of this generic model :

- *Calibration*: Given a dataset $\mathbf{y}$, estimate the parameters $\hat{\theta}$ that best describe the long-term disease progression.
- *Personalization*: Given $\hat{\theta}$ and an individual y_i, estimate its individual parameters z_i. This corresponds to the variations of the mean trajectory that best fit the individual measurements at the seen ages $t_{i1}, \ldots, t_{ik_i}$.
- *Simulation*: Given θ and a set $(z_i)_i$, sample a new z'_j. The latter characterize a new variation of the mean trajectory and thus entirely describes a new subject that can be observed at any age t.

In Section 6.4, we present a particular instantiation of the function f_θ that is well-suited to describe the progression of cognitive scores. It corresponds to logistic shapes that sketch the conversion from a normal to an abnormal state. Along with this instantiation, we present the procedure that enables the three aforementioned procedures.

Given the generic model described in Eq. 6.1, we show that once arranged properly, the procedures allows to, first, generate virtual cohorts that either (i) replicate the characteristics of the initial cohort or (ii) unbias and balance real datasets, secondly, impute missing values and predict future stages, and finally, augment training sets used to train algorithms design to predict future time-points.

6.3.1 Virtual Cohort Simulation

The first application consists in simulating patients with multiple measurements. To do so, we first consider an initial longitudinal dataset $\mathbf{y}$. From it, the calibration procedure estimates the parameters $\hat{\theta}$ that describe the long-term disease progression along with its variability in term of spatiotemporal dynamics. The personalization procedures then gives, for the patient in the longitudinal dataset, their corresponding individual parameters $\mathbf{z}_i$ whose collection $(\mathbf{z}_i)_i$ defines a empirical distribution. From the latter, the simulation procedure allows to draw new samples $(\mathbf{z}'_j)_j$, each fully characterizing a virtual patient whose measurements can be computed at any age t as implied by Eq. 6.1 and shown on Fig. 6.1. Choosing these time-points depends on the usage that follows.

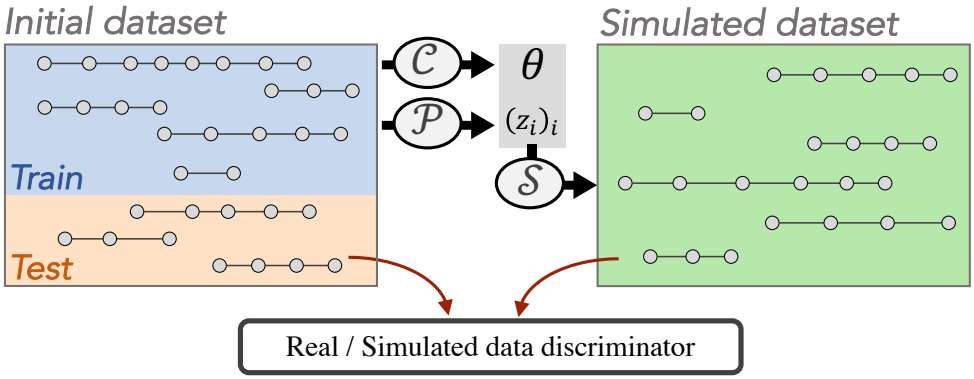


Figure 6.1: Training set of real patients used to calibrate ($\mathcal{C}$) the model, then personalize ($\mathcal{P}$) the individual parameters and finally simulate ($\mathcal{S}$) a virtual cohort. The simulated data are compared to a test set of real data.

On the one hand, it is possible to simulate virtual cohorts with patients that mimic the individual in the initial one. To do so, the age at baseline, the number of follow-up time points and the corresponding time intervals should be similar to the real patients. A direct application of such cohorts is the fact that they are constituted of simulated patients that are by definition anonymized. This can be use to share databases while overcoming some cumbersome sharing policies and anonymization processes - only if assured that the simulated data does not contain subject-specific information, which should be carefully checked based on the empirical distribution of the individual parameters. A slightly different application is to balance the initial cohort, by drawing samples $(\mathbf{z}'_j)_j$ from a subsample of the empirical distribution $(\mathbf{z}_i)_i$ which corresponds to the individual parameters of a given subpopulation (e.g. male or female, married or divorced or single, carrier of a genetic mutation or not). This allows to augment the initial dataset with virtual patients whose trajectory are similar to their subgroup.

On the other hand, the experimental settings of the database such as the inclusion criteria in medical dataset, might present some biases. It is likely to have an effect on the correlation between $\mathbf{z}_i$ and age at the first visit but also the number of follow-up measurements. To illustrate this, one might consider that for disease progression database, the recruitment process is likely to enroll earlier patients with early disease onset. Similarly, fast progressors are likely to pass away earlier so to present less visits. These biases might also affect the algorithms e.g. slow progressors may present more time-points which are ultimately more used to drive the disease progression model. Therefore, depending on the settings, the biases have to be treated thoroughly.

In a word, simulating virtual cohorts, either to replicate real one, to unbias them or to balance them for underrepresented classes, necessarily depends on the problem at hand. Careful precautions should be used in the analysis of the relation between the random effects, the cofactors and the observed time-points.

6.3.2 Missing Values Imputation and Future Time-Points Prediction

A burdensome issue to run algorithms on longitudinal data is the fact that they often present different number of time-points or that the observations are not separated equally in time. The proposed framework allows to interpolate observations at any age t.

Given a model calibrated by $\hat{\theta}$, let us consider a new subject whose observation $(y_j)_{1\leq j\leq T}$ have been seen at times $t_1 < \ldots < t_T$. The personalization procedure outputs the corresponding individual parameters $\mathbf{z}$. As Eq. 6.1 holds true for any t, it allows to impute a missing value for any age of interest $t \in [t_1, t_T]$. The procedure is shown on Fig. 6.2.

As there is no condition on the selection of t, it is legitimate to consider $f_{\hat{\theta}}(t, \mathbf{z})$ for

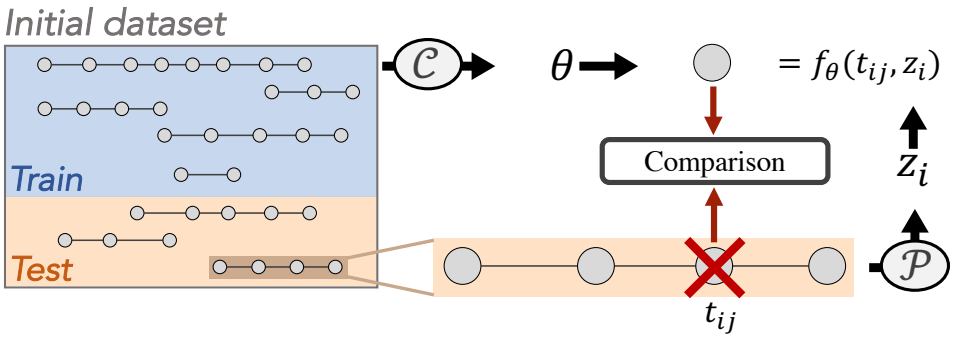


Figure 6.2: Training set used to calibrate ($\mathcal{C}$) the model before the personalization ($\mathcal{P}$) of a test patient whose individual parameters z_i are used to impute a missing value. The imputation accuracy is assessed by comparing to a hold out data.

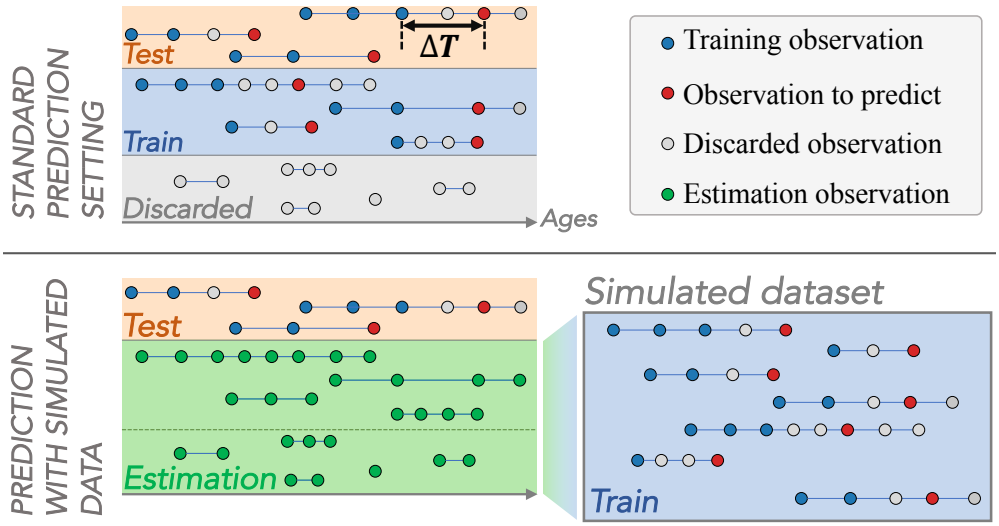


Figure 6.3: Comparison of longitudinal prediction settings in standard case (top row) or with simulated data (bottom row). The latter allows to increase the training set to avoid learning on insufficient samples.

$t > t_T$. This corresponds to the prediction of future time-points. It is made possible by the fact that $f_{\hat{\theta}}$ describes the long-term evolution of the disease. Therefore, while the personalization is computed thanks to few measurements, the resulting individual trajectory is a continuous function defined on a temporal domain that spans the whole long-term reconstruction.

6.3.3 Improved Algorithms

We here present a last application which is a natural extension of the first one : as it is possible to simulate an arbitrary number of patients with an arbitrary number of measurements, the resulting virtual cohort can be used to train algorithms that usually lacks data - which is often the case in medical datasets. To illustrate this point, we will refer to the top row of Fig. 6.3 that presents the standard setting of future prediction : past visits are used to predict values in ΔT years. We stress the fact that predicting the last observed visit might be biased, by attrition for instance. Note that the longer ΔT is, the more subjects are discarded due to insufficient temporal depth. This experimental setting is used in studies that compare the predictive power of different sets of features, and, to compare the prediction at different temporal horizons, ΔT_1 and ΔT_2 for instance. This

second comparison might be extremely inaccurate or skewed as changing ΔT inevitably change the algorithm training set. For instance, longer ΔT discard more patients and more visits for the remaining patients.

Furthermore, the small number of data in medical applications makes it difficult to know whether a given feature has saturated the algorithm with its predictive power or if new examples would have increase it. Said differently, the accuracy of an algorithm increases to a maximum value with respect to the number of training data. However, it is unsure that this value has been reached in the case of small datasets. For these reasons, to compare features and predictive power at different temporal horizons, the algorithm should relies on equivalent training and test set in order to comply to the "all other things being equal" rule.

To this end, we propose an alternative prediction setting presented on the bottom row of Fig. 6.3. Given an initial dataset, we split it into an estimation and test set. The former is used to draw virtual patients that form a simulated dataset (after a calibration and personalization procedure). This simulated dataset is used to train an algorithm whose results are reported based on the initial test set, preventing from a data leakage and constraining the metrics to be computed on unseen real subjects. Note that in this new prediction setting, the patients that were previously discarded because they presented too short time spans are now used to estimate the variability of disease progression. Therefore, the simulated dataset mimics more spatiotemporal trajectories, ultimately producing more accurate algorithms.

6.4 Longitudinal Model instantiation

The longitudinal data augmentation framework has been previously introduced in a generic manner. In this section, we give an example of such model, that can be rewritten as in Eq. 6.1. As we will study cognitive scores, we consider a model introduced in [Schiratti et al., 2017] that is particularly well suited to handle cognitive assessments as it describes long-term trajectories in the form of a logistic shape, i.e. from a normal to an abnormal state. Given this model, we also detail the three aforementionned mathematical procedures: the calibration, the personalization and the simulation.

6.4.1 Statistical Model

The longitudinal model, which considers manifold-valued data, focuses on individual observations that derive from an "average" progression in the sense that it assumes a hypothetical group-average trajectory $\gamma_0 : t \mapsto \gamma_0(t)$ in the space of measurements $\mathbb{R}^n$ that describes the long term progression of the disease. We write $\eta_i : t \mapsto \eta_i(t)$ the trajectory in $\mathbb{R}^n$ of the i-th subject. This individual trajectory derives from $\gamma_0(t)$ in two ways :

- *spatially*: the individual trajectory $\eta_i(t)$ is separated from $\gamma_0(t)$ by a distance $d_i(t)$ at time t. This distance is parametrized by time independant sources $(s_{ij})_{1\leq j\leq N_s}$ where N_s is the number of sources of an independant component analysis that helps projecting the distance $d_i(t)$ in $\mathbb{R}^n$ to a smaller subspace of possible directions.

- *temporally*: the speed of progression of $\eta_i(t)$ may be different from the one along $\gamma_0(t)$. For this reason, we consider that it progresses through time via a temporal reparametrization $t \mapsto \psi_i(t) = \alpha_i(t - \tau_i)$. The acceleration factor α_i accelerates ($\alpha_i > 1$) or decelerates ($\alpha_i < 1$) the progression, while the time-shift τ_i delays ($\tau_i > 0$) or move forward ($\tau_i < 0$) the progression compared to the group-average.

Finally, the individual trajectory $\eta_i(t)$ is a function of individual parameters $z_i = (\alpha_i, \tau_i, (s_{ij})_{1\leq j\leq N_s})$ and θ i.e. $\eta_i(t) = f_\theta(t, z_i)$. The individual parameters are hidden variables such that $\tau_i \sim \mathcal{N}(\overline{\tau}, \sigma_\tau^2)$, $\alpha_i = \exp(\xi_i)$ with $\xi_i \sim \mathcal{N}(\overline{\xi}, \sigma_\xi)$ and $s_{ij} \sim Laplace(0, 1/2)$.

The log-normal distribution corresponds to a positivity condition on the speed of propagation α_i.

6.4.2 Estimation Procedures

The model proposed in Eq. 6.1 relies on model parameters θ (shared by the population) and on individual parameters z_i. The former are the fixed effects and the latter are the random effects of this mixed-effects model. The resulting likelihood writes $p(\mathbf{y};\theta) = \int p(\mathbf{y},\mathbf{z};\theta)\mathrm{d}\mathbf{z} = \int p(\mathbf{y}|\mathbf{z};\theta)p(\mathbf{z};\theta)\mathrm{d}\mathbf{z}$ where $p(\mathbf{y}|\mathbf{z};\theta)$ corresponds to the model as $y_{ij} \sim \mathcal{N}(f_\theta(t_{ij},\mathbf{z}_i),\sigma^2)$ and $p(\mathbf{z};\theta)$ are the priors over the hidden variables. In this paragraph, we define the three actions we further consider to leverage the potential of the model.

Calibration: As the individual hidden variables $\mathbf{z}$ are unknown, the likelihood $p(\mathbf{y};\theta) = \int p(\mathbf{y}|\mathbf{z};\theta)p(\mathbf{z};\theta)\mathrm{d}\mathbf{z}$ is intractable, nor the posterior distribution $p(\mathbf{z}|\mathbf{y};\theta)$. To this end, we used the Monte-Carlo Markov Chain - Stochastic Approximation Expectation Maximization (MCMC-SAEM) algorithm, a stochastic version of the Expectation-Maximization algorithm with MCMC dynamics to sample the hidden variables z, that converge to a maximum of the likelihood under certain conditions [Allassonniere and Kuhn, 2015]. Finally, the calibration procedure leads to $\hat{\theta} \in \mathrm{argmax}_\theta p(\mathbf{y};\theta)$.

Personalization: We here consider that we have estimated $\hat{\theta}$ thanks to the calibration procedure. Given a new subject with observations $\mathbf{y}^{new} = (y_j^{new})_{1\leq j\leq T}$ at times $t_1 < \cdots < t_T$, we propose two options to estimate its hidden variables $z^{new} = (\alpha^{new}, \tau^{new}, (s_j^{new})_{1\leq j\leq N_s})$.

The first consists in drawing n samples $z_i \sim p(\cdot|\mathbf{y}^{new};\hat{\theta})$ and taking the mode of the empirical distribution. It basically corresponds to MCMC-SAEM iterations with a fixed value $\hat{\theta}$. The hidden variables z_i learnt during the estimation procedure should not be used because first it is sampled with a value $\theta^{(k)}$ relative to the k-th iteration, but, mostly, is does not not corresponds to the mode of the distribution $p(\cdot|\mathbf{y};\theta)$.

The other method is based on an optimization procedure that seeks $z^{new} \in \mathrm{argmax}_z\, p(\mathbf{y}^{new}, z;\theta) = p(\mathbf{y}^{new}|z;\theta)p(z;\theta)$, either with a quasi Newton method such as L-BFGS [Byrd et al., 1995] or Powell's conjugate direction method [Powell, 1964]. In the following experiments, we will use the Powell's method. In practice, it leads to similar results to MCMC-SAEM samples or L-BFGS optimization. This means that the individual likelihood does not present a too flat maxima.

Another practical remark is the possibility to use the same subjects to first calibrate $\hat{\theta}$ and then estimate z_i. We would expect the individual parameters to be worse if the subject is not part in the calibration procedure. In practice, these values are essentially the same if the training set is *sufficiently* large.

Simulation: We consider $\hat{\theta}$ and a set of individual parameters $(z_i)_i = (\alpha_i, \tau_i, (s_{ij})_{1\leq j\leq N_s})_i$. The latter defines a joint empirical distribution from which we propose to draw new samples $z' = (\alpha', \tau', (s'_j)_j)$. To do so, we first simulate (ξ', τ') with a kernel density estimation (KDE) of the distribution $(\xi_i, \tau_i)_i$ (we recall that $\alpha_i = \exp(\xi_i)$). Then, considering the multivariate Gaussian distribution $\mathcal{N}(\mu, \mathbf{\Sigma})$ estimated on all the random effects $(z_i)_{1\leq i\leq n}$, it is possible to draw $(s'_j)_{1\leq j\leq N_s}|\alpha', \tau' \sim \mathcal{N}(\tilde{\mu}, \tilde{\mathbf{\Sigma}})$ where $\tilde{\mu}$ and $\tilde{\mathbf{\Sigma}}$ are functions of μ and $\mathbf{\Sigma}$ [Petersen et al., 2008].

6.5 Experiments and Results

6.5.1 Data Description

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu), especially cognitive assessments of subjects that are MCI at one visit at least. It therefore excludes stable cognitively normal (CN) patients and patients with stable AD diagnosis whose do not present any significant disease progression. We rather focus on stable MCI and MCI who convert to Alzheimer's Disease that present a high progression variability. It also includes non-monotoneous profiles.

The experiments are based on the MMSE, the Alzheimer's disease assessment scale - cognitive subscale (ADAS-Cog) with 11 and 13 items, the clinical dementia rating sum of boxes (CDRSB), the Montreal cognitive assessment (MOCA) and the functional assessment questionnaire (FAQ). To be comparable, the values have been rescaled between 0 and 1 with 0 corresponding to the healthy stages and 1 to abnormal stages. Depending on the features used, the experiments rely on 721 patients (resp. 3176 visits) if all the features are considered, up to 980 patients (resp. 5659 visits) if MMSE, ADAS-11 and ADAS-13 only are considered. The code of the following experiments will be made available upon acceptance of this manuscript.

6.5.2 Virtual Cohort Validation

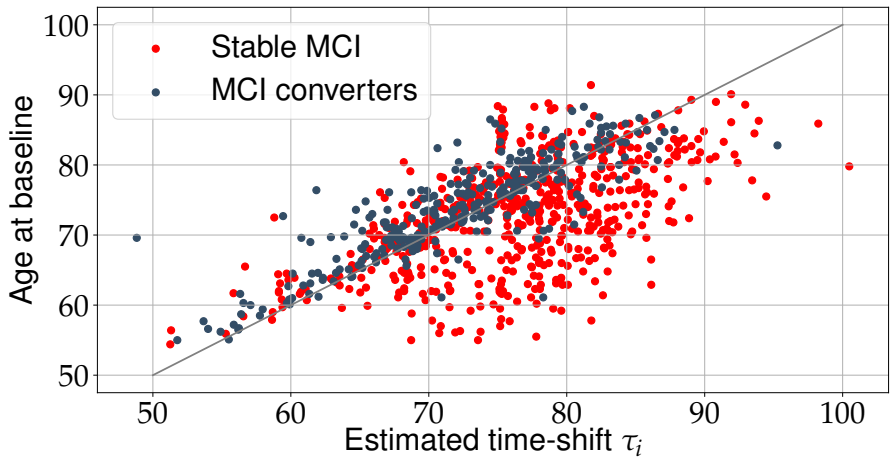


Figure 6.4: Correlation between the age of inclusion in the dataset and the estimated time-shift τ_i, colored by the type of diagnosis.

To assess the quality of the simulated data, we use a discriminative algorithm as shown in Fig. 6.1, trained to distinguish simulated from real data in the spirit of GANs. We consider a long short-term memory (LSTM) for its ability to handle longitudinal data (with 5 hidden dimension, a dropout rate of 0.5) stacked with a fully connected layer that outputs the probability of the input data to be real of simulated. The cross-entropy loss was optimized thanks to the Adam optimizer (learning rate of 10^{-3}) [Kingma and Ba, 2014].

As our goal is to fool the LSTM and get a poor accuracy, we first show that its architecture and hyperparameters are well designed for the features at hand to discriminate different tasks. The first experiment is to consider individual sequence of data and determine the diagnosis at the last seen visit (CN, MCI or AD). Based on a 10 fold cross-validation (CV), it led to an accuracy of 89.0($\pm$4.7), for which the benchmark accuracy (MCI prediction) is of 65%, corresponding to a balanced accuracy of 83.7($\pm$8.5). Considering sequence

of visits where the last one is MCI, the second experiments aims to predict the diagnosis in 3 years. The accuracy (resp. balanced accuracy) is of 85.0($\pm$3.0) (resp. 70.4 ± 9.8) on a 10 fold CV, which is comparable to state-of-the-art results [Tong et al., 2016]. These two experiments show that the LSTM is able to handle these data given its architecture.

Back to the simulation procedure, we split the initial dataset into a hidden test set of a hundred real patients, and an estimation set that is used to simulate a virtual cohorts of patients with MMSE, ADAS-Cog 11 and ADAS-Cog 13. The simulated patients are concatenated to the hidden test set and we use the aforementioned LSTM to predict whether the patients are simulated or real, in a 10 fold CV procedure. The real observations are mainly separated by one year (except for the second visit which is observed after six month), to this end, we use the same time interval to simulate the measurements of the virtual patients.

We discuss the results depending on the chosen hyper-parameters (number of simulated patients, number of visits per patient, age at baseline, time-interval, ...). Besides the fact that data simulated without noise are instantaneously spotted by the discriminator, the distribution of ages to simulate the observations with is a key feature to fool the discriminator. With a random baseline age and number of observations, the discriminator is able to distinguish the data. While surprising at first, this essentially reveals that a type of progression is associated to a scenario of seen ages, as discussed for inclusion criteria that bias the population. This is highlighted on Fig. 6.4 that shows a correlation between the age at inclusion in the dataset and the estimated time-shift τ_i: the subjects are included once they are diagnosed, which is thus related to the age at which the disease progresses. We can definitely use our framework to unbias from this correlation by sampling at any age. Similar reasoning might be applied to balance dataset for cofactors that are correlated to the individual parameters (gender, genetic mutation, ...). For instance, the distribution of the log-acceleration ξ_i is significantly different for the subjects carrier of two mutations on a gene APOE (known to be related to AD) versus the 460 non carriers (p-value = 2.110^{-9} for the Kolgomorov-Smirnov two-sided test).

Finally, if the baseline age and the number of visits are choosen in adequation to the initial cohort, the discriminator cannot separate real data from simulated one : after spliting the initial dataset in 10 folds, the resulting training set is used to simulate virtual data that are concatenated with the test set of real patient. From this concatenation, we again proceed in a 10 fold CV procedure to evaluate the LSTM on differents splits. The reported accuracy over the 100 estimations is of (51.1 ± 8.1). The consistency of these results is further demonstrated in the next experiments.

6.5.3 Missing Values Imputation

Given a dataset, we first split the data between a train and a test set. The training allows to obtain $\hat{\theta}$. We consider a patient in the test set with observations $(y_k)_{1 \leq k \leq K}$ at times $t_1 < \ldots < t_K$ from which we hide the k-th observation y_k at time t_k. The remaining observations are used to personalize the model in order to obtain the corresponding individual parameters $z = (\alpha, \tau, (s_j)_{1 \leq j \leq N_s})$. The latter are used to simulate the value $\tilde{y}_k = f_{\hat{\theta}}(t_k, z)$ and compare it to the real value y_k as shown on Fig. 6.2. We report the mean absolute error (MAE), i.e. $|\tilde{y}_k - y_k|$. We train the model on three features (MMSE, ADAS-Cog 11 and ADAS-Cog 13) in a 10 fold cross validation setting. We report the MAE of reconstruction over the three features and over the 10 runs.

As it is possible to impute different time-points, Fig. 6.5 reports the MAE for 4 different settings : imputation of the central, random, first or last visit. The orange line is the median and the upper and bottom part of the rectangles are the first and last quartile. The whisks corresponds to the 5 and 95 percentiles. For readability purposes, we removed 3 predictions (out of 2655) for the imputation of the last visit whose MAE were around

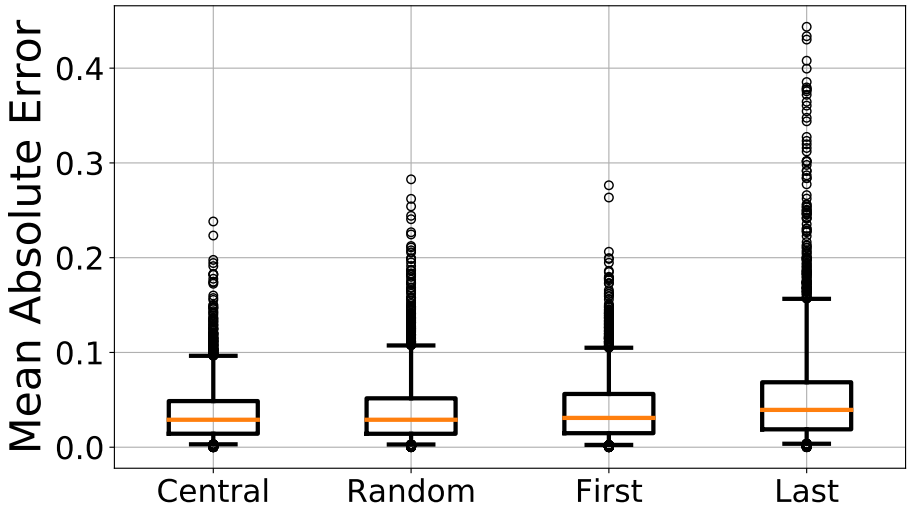


Figure 6.5: Missing value imputation for 4 different scenarios : imputation of the central, a random, the first or the last value.

0.5. It corresponds to patients whose temporal profile present a high increase in the last seen value(s) and a sudden decrease. This scenario of progression is the reason why the prediction of the last value presents more outliers, with a mean MAE of 0.0542. On the other side, the imputation of the central visit presents the best MAE (0.0357), followed by a random visit (0.0385) and the first visit (0.394). The central is unsurprisingly the easiest to interpolate as it relies on previous and future datapoints. On the other hand, to extrapolate the progression, the first visit is easier to impute than the last one as early stages shows a slow and small progression compare to advanced stages that present a higher variability of measurements - which is exactly what the scores have been designed for.

The MAE is to be compared to the natural noise in the data. [Clark et al., 1999] reports two noise values for the MMSE : a standard deviation of 1.3 and 2.8 (out of 30) for respectively CN and MCI patients. Once normalized and converted to absolute values, it corresponds to MAE errors of 0.035 and 0.074. Similarly, [Standish et al., 1996] reports value that corresponds to MAE between 0.077 and 0.11. The variability between this noise estimation comes from the experimental settings (time between two measurements, different ratters, disease stage of the subject). In any case, the error of reconstruction we present is relatively close to the noise in the data.

The reported MAE obviously depends on the number of seen visits the reconstruction has been made on. To this end, Fig. 6.6 shows the MAE when imputing last visit (right column of Fig. 6.5) for different number of seen visits. The main observation is that even one or two visits lead to a good imputation. The main effect of supplementary visits is to prevent outliers.

In fact, the imputation of the last visit can be considered as a prediction of the future. To separate predictions at different temporal horizon, we run the same experiments by ensuring a time interval of 1, 2, 3 and 4 years between the last seen visit and the imputed value. As shown on Fig. 6.7, this results in predictions that are close to the noise in the data. The prediction worsen for long temporal horizons. One reason is that the continuous logistic shape is a good approximation of short-term data but might be not well-suited for complex long-term dynamics. Another reason is that increasing the time ΔT leads to fewer examples to train/test the algorithm on and to smaller sets of calibration, of individual parameters and of testing. A final reason is that long-term extrapolation lead

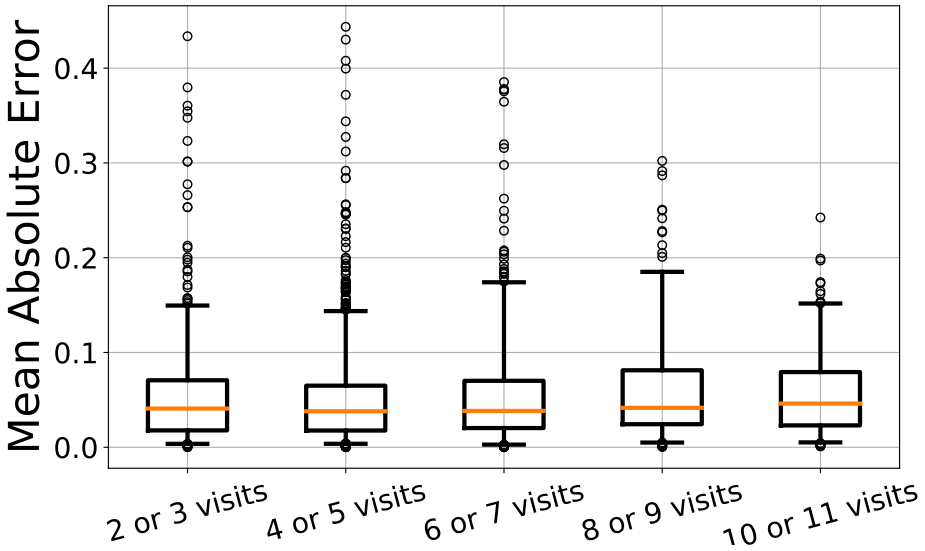


Figure 6.6: Missing value imputation of the last value depending on the number of seen visit to impute.

to consider stages that have potentially not been seen in the original database thus illusive to be predicted correctly.

6.5.4 Improved Prediction of Cognitive Scores

To improve the performance of predictive algorithms thanks to simulated data, we here focus on the prediction of the MMSE at 1, 2, 3 and 4 years, based on the MMSE, MOCA, ADAS-13, ADAS-11, CDRSB and FAQ. To predict future MMSE values, we choose a long short-term memory (LSTM) neural network, with 10 hidden dimensions, stacked with a linear layer. The mean squared error (L2-norm) loss is optimized thanks to the ADAM optimizer (learning rate of 10^{-3} and weighted decay of 10^{-5}). To prevent the model from overfitting, a subset of the train set, namely the validation set, is used to apply the early stopping criterion procedure : it stops the training if no loss improvement is detected from a given number of epoch on the validation set. To estimate the variance of the estimation procedure, the results are presented with error-bars corresponding to the mean and standard deviation of the MAE based on 10 independent runs with different test splits.

The first scenario, shown on Fig. 6.8a, corresponds to prediction of the MMSE based on three different sets of data (listed above each column) for 4 different temporal horizon (1, 2, 3 and 4 years) in a standard prediction settings presented on the top row of Fig. 6.3. The sub-scripted numbers correspond to the size of the train and test set. The results are compared first to the benchmark constant prediction, i.e. the hypothesis that there is no change of MMSE within the time interval, in dashed lines, and, on the other hand, to the noise in the data discussed previously and represented by a hatched pale orange intervals, the larger (resp. smaller) corresponding to noise of MCI (resp. CN) patients.

We conducted the same predictive experiments but instead of training the algorithms on real train data, we used the latter to calibrate the model previously introduced. We then simulated 500 virtual patients, trained the algorithm and fitted on the real test data, as described by the procedure on the bottom row of Fig. 6.3. The results are reported on Fig. 6.8. A glimpse of the hyperparameter influence is given in the Supplementarial Materials (number of patients used to calibrate the model and number of simulated patients).

Both experiments shows that it is possible to reach noise level prediction up to 2 years

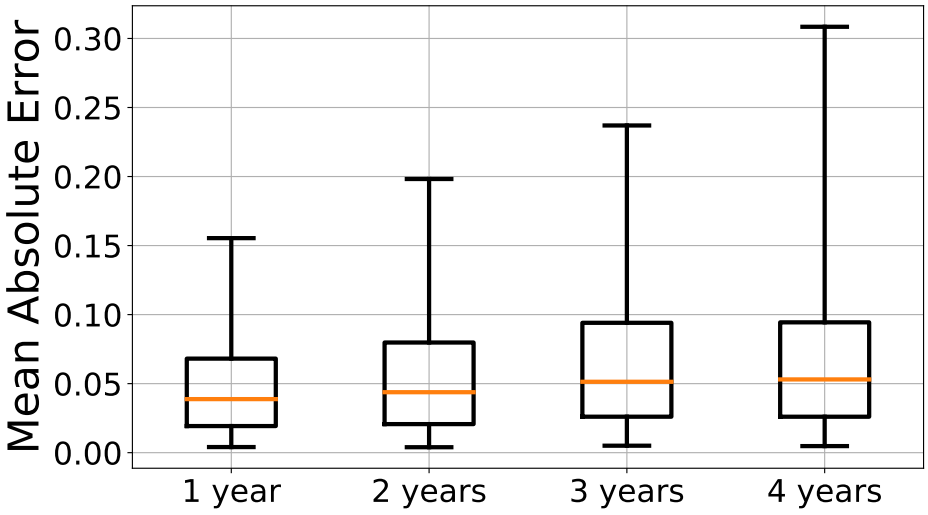


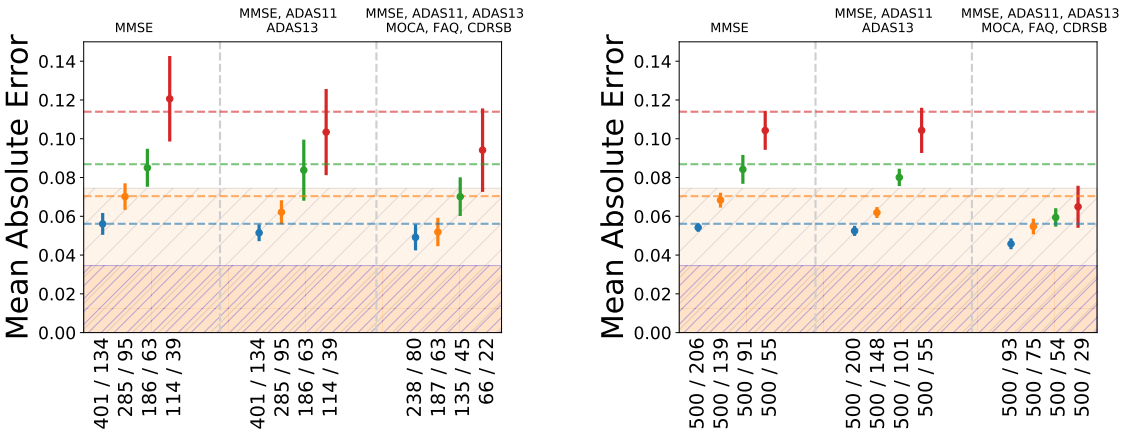
Figure 6.7: Missing value imputation of future values (prediction) at 4 temporal horizons. Outliers have been removed.

in advance. However, the nature of the conclusion made for prediction in 3 and 4 years is different depending on the prediction setting. The standard setting shows a better MAE while adding new features. However, it relies on different size of training and testing data, enabling from accurately comparing the features. This is demonstrated on Fig. 6.8b where we simulated exactly 500 virtual patients with an larger test size as less real patients are necessary to calibrate our model. While decreasing the total variance over the 10 runs, probably due to the increased test set, we improved the MAE by 20% (resp. 37%) to 0.0594 (resp. 0.0649) for prediction 3 years (resp. 4 years) in advance. This outperforms state-of-the-art results reported on Fig. 10 of [Iddi et al., 2019] that reports a MAE of 0.083 for a 4 years prediction, and, 0.0602 (resp. 0.0552) at 3 years (resp. 4 years) reported in [Huang et al., 2016]. The latter results can be challenged by the fact that the data description mentions half of the predicted visits to belong to healthy patients with a MMSE of 29.2 $\pm$ 1.1 at 4 years whereas the vast majority of our predictions concern MCI and AD stages with a high variability. Finally, the noise estimation for MCI patients can be questioned as it seems to have been over-estimated due to the fact that we almost systematically report a better MAE.

More interestingly, the predictive power of the ADAS-11, ADAS-13 and MMSE is not better than with the MMSE alone, a result that could not have been stated from the standard prediction. It essentially means that the MMSE alone is a predictor as good as the three variables but needs more patients to train the model on. In the same spirit, FAQ, MOCA and/or CDRSB provide substantial information that allow to reach noise level prediction up to 4 years in advance.

6.6 Conclusion

Longitudinal databases are promising in term of disease modeling as they convey individual measurements that derive from a long-term progression. Its counterpart lies in the heterogeneity of data it includes. To this end, we proposed a longitudinal data framework whose potential has been demonstrated by simulating virtual cohorts (that can be intentionally unbiased or unbalanced), imputing missing values potentially at future time points, and, finally improving predictive algorithms while allowing to compare them. The



(a) Standard prediction

(b) Prediction with simulated data

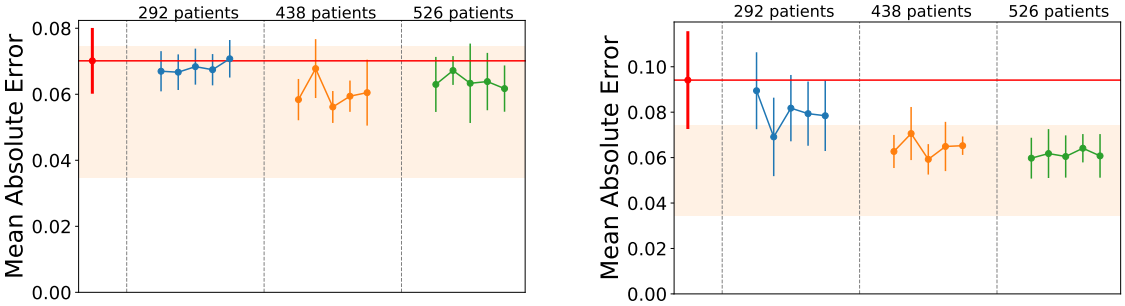
Figure 6.8: Mean and standard deviation of the prediction of the MMSE in 1 (blue), 2 (orange), 3 (green) and 4 (red) years with different sets of variables (upper part of each column) over 10 runs. The colored dashed lines corresponds to the error for the corresponding constant prediction. The hatched pale orange area corresponds to two noise estimation, resp. for MCI (top) and CN (bottom) patients. The number at the bottom presents the training and test set sizes.

latter increased state-of-the-art results of cognitive assessment prediction in the case of Alzheimer's Disease.

Further efforts will be deployed to evaluate the simulation procedure by measuring, for instance, the impact of the visits simulated, the time-interval between them or the selection of the first visit. This could benefit other studies by providing a more accurate comparison of the predictive quality of models or new biomarkers. We also wish to replicate this study on other disease progression in future work.

6.7 Supplemental materials

6.7.1 Influence of hyperparameters on the simulation



(a) MMSE prediction in 3 years

(b) MMSE prediction in 4 years

Figure 6.9: MMSE prediction based on MMSE, ADAS-11, ADAS-13, MOCA, FAQ and CRDSB. The red value on the left corresponds to the MAE without simulated data. Then, each column corresponds to a different size of the estimation set. Within each column, we simulate, from left to right, 50, 100, 250, 500 and 1000 virtual patients.

As the part of the patients used in the estimation set may vary, we tested different scenarios that lead to better results when more patients were used. On the contrary, the number of simulated patients does not seem to have a great impact on the quality of the prediction. A possible but preliminary explanation lies in the fact that even though there are not a lot of simulated patients, they already incorporate more (simulated) visits than real patients.

Part V

Software development

Chapter 7

Estimation of the Disease Progression Model

This chapter describes the software tools that have been developed to obtain the results presented in the previous chapters. The first software is a C++ package that enables the analysis of spatially structured data, used to estimate the disease progression on medical imaging data as presented in Part II. The second software is a Python package, called `Leaspy`*. It is intended to wrap up all the contribution of this manuscript to benefit the research community in providing a general framework for longitudinal data analysis, from the estimation of the long-term disease progression to the imputation of missing values and the simulation of virtual cohorts.*

Contents

This section is an adaptation of the READ ME available at `https://gitlab.com/icm-institute/aramislab/leasp`*. The repository is not maintained anymore as it is under migration to* `Leasp` *(see Section 7.2). It is left available for reproducibility purposes.*

7.1 Leasp : A C++ Software Package for the Analysis of Spatially Structured Longitudinal Data

7.1.1 Description

Leasp is a C++ 14 software package for the statistical analysis of longitudinal data, particularly medical data that come in a form of repeated observations of patients at different time-points. Considering these series of short-term data, the software aims at :

- recombining them to reconstruct the long-term spatio-temporal trajectory of evolution of the multiple signals observed
- positioning each patient observations relatively to the group-average timeline, in term of both temporal differences (time shift and acceleration factor) and spatial differences (different sequences of events, spatial pattern of progression, ...)
- quantify impact of cofactors (gender, genetic mutation, environmental factors, ...) on the evolution of the signal

The software package can be used with two different types of data :

Scalar data

The simplest type of data handled by the software are scalar data: they correspond to one (univariate) or multiple (multivariate) measurement(s) per patient observation. This includes, for instance, clinical scores, cognitive assessments, physiological measurements (e.g. blood markers, radioactive markers).

Network Inputs

As some data have a spatial coherence (e.g. cortical thickness maps, PET SUVR), it is important to integrate the spatial proximity in the long-term evolution of the signal. The important prerequisite to run this particular type of model is to provide a matrix of distance between the coordinates of the signal to the software.

7.1.2 Design

The core of the package is designed around the dialogue between the algorithm and the model. Both are virtual abstract classes that are instantiated depending on the algorithm and model used. The model encompasses the scalar and network inputs. On the other hand, the algorithms make use of different sampling algorithms.

The library provides a launch folder to directly apply the mathematical procedures such as the calibration, the personalization and the simulation. Another folder includes different utils as preprocessing functions but also a visualisation toolbox.

The google test library was used to run unit and functional tests.

7.1.3 How to use Leasp

Requirements

The C++ package depends on the following libraries :

- `tinyxml` : library for the reading and preprocessing of `xml` input files,
- `googletest` : unit and functional testing library,
- `Armadillo` : linear algebra libray, used to work with vectors and matrices.

The two first libraries can be installed by typing `git submodule init` and then `git submodule update`. Armadillo can be installed with the `brew` package manager on Mac OS.

Installation

To install the C++ software package, open a terminal and type the following commands :

1. `git clone https://gitlab.com/icm-institute/aramislab/leasp`
2. `git submodule init`
3. `git submodule update`
4. `. mkdir build \&\& cd build`
5. `. cmake --options`
6. `. make --options}`

Run the software

Leasp is a command line software. So, once added to the `\protect\T1\textdollarPATH`, it can be run with the following command (`stgs` standing for settings)

`leasp fit model_stgs.xml algorithm_stgs.xml data_stgs.xml sampler_stgs.xml`

where

- `fit` : name of the command used ; `fit` estimated the group-average trajectory for the individual measurements
- `model_settings.xml` : `xml` file that indicates which type of model should be run with additional information about hyperparameters of the model chosen (dimensions, number of sources, path to the distances matrices, ...)
- `algorithm_settings.xml` : `xml` file that defines the number of total iterations, the number of burn-in iterations, the saving settings and the display settings
- `data_settings.xml` : `xml` file that gathers the path to the data files
- `sampler_settings.xml` : `xml` file that lists the samplers used in the estimation procedure, along with their parameters

Outputs

The estimation procedure outputs 3 files that are:

- `convergence_parameters.csv` that allows to investigate the convergence of the model parameters
- `population_parameters.csv` that defines the population parameters needed to characterize the group-average long-term history of the signal
- `individual_parameters.csv` that corresponds to the temporal (time shift and acceleration factor) and spatial (space shift) individual parameters. They enables to consider the subject-specific trajectory of the signal.

7.1.4 Support

The software is hosted at `https://gitlab.com/icm-institute/aramislab/leasp`.

Examples

Several examples are to be found on the development branch (root of the document). They can be run in the terminal thanks to "sh launch_simulation.sh".

Licence

Leasp is distributed under the terms of the MIT license.

7.2 Leaspy : A Python Toolbox to Learn Spatiotemporal Patterns of Disease Progression

Abstract

This paper introduces `Leaspy`, a Python package that provides a natural framework to study short-term longitudinal data which derive from a long-term temporal phenomenon. It relies on a generative mixed-effects model that is able to reconstruct the evolution of the phenomenon at both an average and individual level. Such framework is particularly helpful to study disease progression. The package, distributed under the GNU GPLv3 licence, is hosted at `https://gitlab.com/icm-institute/aramislab/leaspy`. Documentation and examples are available at `www.leaspy.run` and in multiple `Jupyter` notebooks. Quality is ensured with unit and functional testing, `git` version control and continuous integration.

7.2.1 Introduction

When it comes to the analysis of longitudinal data, i.e. repeated observations per subject, many challenge rise from the data structure : inconsistent number of observations per patient, variable time intervals but also a potential absence of temporal alignment between the temporal trajectory. In the case where they derive from a similar long-lasting phenomenon, such as disease progression, they are likely to be observed during periods shorter that the underlying long-term process.

To this end, we introduce `Leaspy`, standing for Learning Spatiotemporal Patterns in Python, an open-source Python package conceived for the analysis of longitudinal data where the observation at each time-point takes the form of a n-dimensional vector or a $n \times m$ matrix. The mixed-effects model it relies on is able to recombine the short-term individual trajectories into a long-term progression, while dealing with two forms of variability. The temporal one, corresponding to the unalignment of the individual snapshots, is characterized by a temporal shift between subjects and an acceleration factor governing the pace of progression. On the other hand, the spatial variability relates about the geometrical shift between the individual trajectories in the space of measurements. It can take the form of a subject-wise reordering of the chronology of events or change the patterns of progression.

The methodology, that takes advantage of Riemannian geometry, has been developed in [Schiratti et al., 2017], latter extended to spatially structured data such as images or networks ([Koval et al., 2018b]) and to missing values ([Couronne et al., 2019]). Finally, an exhaustive study of Alzheimer's Disease has been conducted in [Koval et al., 2018a] to validate the approach. Additional applications, such as the imputation of missing values, the prediction of future time-points or the simulation of virtual cohorts, are introduced in Chapter 6. All these works have contributed to the development of `Leaspy`.

7.2.2 Supported Classes of Problems & Related API functions

In the following, we consider a longitudinal dataset $\mathbf{y} = (y_{ij}, t_{ij})_{1 \leq i \leq n, 1 \leq j \leq k_i}$ of n patients. The i-th patient has been observed k_i times at $t_{i1} < \cdots < t_{ik_i}$ where $y_{ij} \in \mathbb{R}^n$ (or $y_{ij} \in \mathbb{R}^{n \times m}$). The model can be simplified as :

$$y_{ij} = f(\theta_{\text{geom}}, \mathbf{z}_i, t_{ij}) + \epsilon_{ij} \tag{7.1}$$

where θ corresponds to the model parameters, $\mathbf{z}_i$ are the individual parameters are the individual variations to the group-average trajectory and ϵ_{ij} is a Gaussian noise. This statistical problem, that takes the form of a mixed-effect model, considers $\mathbf{z}_i$ as hidden

random effects parametrized by $\theta_{\mathbf{z}}$. Given $\theta = (\theta_{\text{geom}}, \theta_{\mathbf{z}})$ the set of parameters, the goal is to maximize the model likelihood that writes $p(\mathbf{y};\theta) = \int p(\mathbf{y}|\mathbf{z};\theta)p(\mathbf{z};\theta)\,\mathrm{d}\mathbf{z}$. Given this statistical writing, `Leaspy` essentially offers the four following statistical procedures :

- **Calibration** : given $\mathbf{y}$, estimation of the parameter $\hat{\theta} \in \text{argmax}_\theta\, p(\mathbf{y};\theta)$ which entirely describes the long-term group-average spatiotemporal trajectory.
- **Personalisation** : given $\hat{\theta}$, estimation of the individual hidden variables $\mathbf{z}_i^*$ that best derive the group-average trajectory to reconstruct the individual measurements. It allows to analyse how individual cofactors modulate the subject-wise disease progression which is precisely determined by the individual hidden variables.
- **Reconstruction**, **Imputation of missing values** and **Future prediction** : given $\hat{\theta}$ and $\mathbf{z}_i^*$, estimation of $\tilde{y}_{ij} = f(\hat{\theta}, \mathbf{z}_i^*, t_{ij})$ where t_{ij} is either an observed time (reconstruction), or a time between the first and last seen visit (missing value imputation) or after the last seen visit (prediction).
- **Simulation** : given $\hat{\theta}$ and a set of individual variables $(\mathbf{z}_i)_i$, drawing of a new hidden variable $\mathbf{z}_{i'}$ that entirely determine a new individual whose time-points are chosen arbitrarely. If repeated, the procedure returns a virtual cohort with potentially a larger number of patients, more follow-up time-points and finer time intervals. Apart from the data augmentation possibility, it enables to share database information while anonymizing it.

Even though the solver list is continuously increasing, we mention that the calibration is based on the MCMC-SAEM algorithm ([Delyon et al., 1999, Kuhn and Lavielle, 2004, Allassonnière et al., 2010]) which offers the Gibbs ([Geman and Geman, 1984]) and Hamiltonian Monte-Carlo ([Neal et al., 2011]) samplers within the Hasting-Metropolis algorithm ([Chib and Greenberg, 1995]). Several heuristics improve the convergence stability or speed : adaptive variance ([Atchade, 2006]) and gradient descent. The personalization includes Powell and L-BFGS methods ([Powell, 1964, Byrd et al., 1995]). The simulation rely on conditional multivariate Gaussian and kernel density estimations.

```
from leaspy import Leaspy, Data, AlgorithmSettings

data = Data.from_csv('path/to/data.csv')

leaspy = Leaspy('logistic')

calibration_settings = AlgorithmSettings('mcmc_saem', n_iter=2000)
leaspy.fit(data, algorithm_settings=calibration_settings)

personalization_settings = AlgorithmSettings('scipy_minimize')
results = leaspy.personalize(data, personalization_settings)

reconstruction = leaspy.reconstruct(results, times)

simulation_settings = AlgorithmSettings('default_simulation')
simulated_data = leaspy.simulate(results, simulation_settings)
```

Above is an example of the related API calls that were developed with a user-friendly intention. Each of the four statistical tasks (except the reconstrution) is being preceded by an `AlgorithmSettings` object that defines the parameters of the algorithm used. It is loaded with default parameters if called only by a name, to which `kwargs` are easily addable. An advanced version enables to load a parametrisable `json` file with thanks

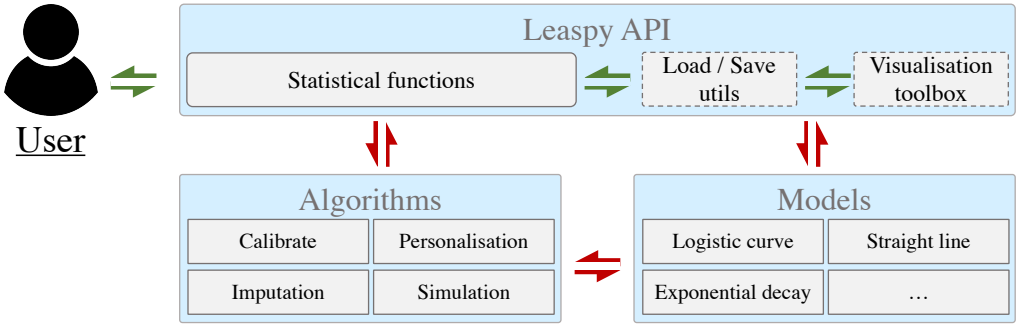


Figure 7.1: `Leaspy` architecture. Green (resp. red) arrows correspond to data structure for external (resp. internal) usage. Blue rectangles are the conceptual principles. Grey Dotted rectangles corresponds to the utils provided in the package.

to the `load(path='path/to/settings.json')` method. To check the convergence of the related stochastic algorithms, the optional `set_logs(path='path/to/logs')` method of `AlgorithmSettings` allows to save log files online in a folder created in the `path`. Besides the logs, few methods enable to efficiently save and load the model parameters, the individual hidden variables or simulated data, into standard `json` or `csv` files. The API also provides a visualization toolbox to display the results of the estimations such as the long-term progression profile or the individual trajectory reconstructions with future predictions.

7.2.3 Architecture & Software Design Principles

`Leaspy` is born in an interdisciplinary research laboratory that brings together various profiles: mathematicians, computer scientists and medical practitioners and researchers. This essentially involves to provide a environment that is suited for python beginners who run medical analysis (e.g. temporal progression of new diseases, new modalities, prediction of future stages), researchers and engineers that use advanced features for research studies (e.g. imputation of missing values, simulation of virtual cohorts) and, finally, researchers that implement new start-of-the-art algorithms to improve the four statistical tasks. This is the reason of the versatility of `AlgorithmSettings` parameters calls that suit the three profiles.

`Leaspy` therefore implements as a high-level structure, as shown on Fig. 7.1, that connects different part of the project, that, given the data exchanges, are independent entities that developers can work on independently. Besides the internal data structures that enhance the computation speed (essentially `Pytorch` tensors) and the utils (visualization toolbox, and loading/saving functions), the internal project is based on models and algorithms. The models corresponds to the f function of Eq. 7.2.2 while the set of algorithms are tools to estimate θ (fit folder), estimate z_i (personnalization folder), impute $\tilde{y_{ij}}$ (imputation folder) or generate (y_{ij}, t_{ij}) (simulation folder).

7.2.4 Development

The Leaspy package has been released under the GNU GPLv3 Licence. A full presentation and documentation can be found at `www.leaspy.run` while complete tutorials are available within the example folder of the repository, hosted on a Gitlab server [1] which enables a git version control. The `example/start` tutorial is fully explained in a Medium blog post [2]. The installation has been made easy with the `conda` package manager.

[1] https://gitlab.com/icm-institute/aramislab/leaspy
[2] https://medium.com/@igoroa/analysis-of-longitudinal-data-made-easy-with-leaspy-f8d529fcb5f8

The software relies on other standard Python packages which makes it usable on Mac, Linux and Windows. The software quality is insured by continuous integration running functional and unit tests on virtual machines, one dedicated to each operating system.

Chapter 8

Enhancement of Clinical Studies with Digital Tools

This chapter presents the digital tools that have been developed to take advantage of the previous theoretical models and their related implementation in the `Leaspy` *Python package.*

Contents

8.1 Introduction

Thousands of Machine Learning applications have been developed for the medical community in the recent years, both for the the patient care (e.g. diabetic retinopathy detection [Gulshan et al., 2016], breast cancer detection [Wang et al., 2016], skins lesions [Kawahara et al., 2016], microbleed detections in the brain [Dou et al., 2016], survival prediction from imaging data in the case of brain tumors [Nie et al., 2016]) and for clinics, laboratories and clinical studies (e.g. segmentation of blood vessels in the eye [Maninis et al., 2016], cell segmentation [Ronneberger et al., 2015], landmark localization [Yang et al., 2015]). These models have shown exciting results. They occurred simultaneously to the recent open science wave : open data, free-access articles and open source code. The latter intends to enhance the medical community which in return validates and gives credit to the model and related methodology. These existing technologies require limited initial investments (compared to magnetic resonance imaging (MRI), magneto-encephalogram (MEG) or position emission tomography (PET) machines for instance) and are of limited cost to use. However, their emergence in various medical environments is still limited. One possible reason for this situation is that the theoretical methods do not always meet real-life needs, and, there access, deployment and ease of use is missing for end-users that have specific environments, frameworks and their own technical terminology. Therefore, there is an urgent need to understand the possible applications of the model and to make it usable and of value for the potential users. This will help bridging the gap between the mathematical-oriented disease modeling community with its medical counterpart.

In the previous chapters, we presented a generative mixed-effects model which is a natural framework to recombine individual longitudinal measurements in order to reconstruct the long-term disease progression. It collaterally allows to compare the patients progression and predict their future values. This model has shown promising results on

the progression of cognitive assessments and features derived from T1-weighted MRI and FDG-PET signal, during the course of Alzheimer's disease (AD). Such analysis have been made possible thanks to the Python package `Leaspy`. The code has been released in open source to enable similar studies among which the estimation of the progression for new AD biomarkers or modalities, for different AD cohorts that have particular inclusion criteria, but also for other (neurodegenerative) diseases as done in [Couronne et al., 2019] for Parkinson's disease.

To avoid the common pitfalls of not addressing real-life needs or not enabling an easy usage by end-users, we exhibit four usages that our model can meet in term of applications. The first one is the direct estimation of the long-term progression of the disease in order to exhibit the pre-symptomatic biomarkers, to understand the sequence of events and, finally, to analyze the cofactors that modulate the disease progression. The second enables the anonymized sharing of the ADNI cohort by simulating one million virtual patient with longitudinal measurements of different modalities. The third is the prediction of future time-points for patient care, e.g. in neurology offices. The fourth appears in the context of cohort monitoring in clinical studies: as the model is able to accurately predict the future natural evolution of a patient, its comparison to the same patient under medication helps measuring the drug effect. It can also exhibit subgroups of patients that present particular response to the treatment. From these use-cases, we present four proof-of-concept prototypes developed to address these goals. These digital tools are designed to support or enhance medical knowledge. They directly derive from discussion with the medical staff. They rely on outputs from `Leaspy` and are easily installable as they are used in web browsers.

8.2 Applications

8.2.1 General Requirements

The impact of a model relies on its performance but also on its adoption by the community, enabled by an easy usage by the end-users. It should be easy to install (potentially on different operating systems) and also simple to use and to interact with. From a technical point of view, the digital tool should be interfacable with `Leaspy` outputs but above all been run locally to prevent data exchanges with a server that might violate data sharing policies.

To this end, we developed web-oriented tools that can be launch on web browsers. This makes them independent from the operating system while allowing to be run locally or remotely. Furthermore, it allows to benefit from the front-end community that always enhances the visualization.

8.2.2 Long-term Disease Progression: www.digital-brain.org

For long-lasting diseases that are, first, only observed during short-term periods and secondly with an important inter-individual variability, getting a sense of the long-term progression is not easy. The Python package `Leaspy`, by recombining the individual observations, is a tool that enables to characterize the sequence of events during the course of a disease. This is the data-driven counterpart of the hypothetical model developed by [Jack Jr et al., 2010b], that enable to properly consider the spatiotemporal evolution and interplays between the biomarkers during the disease progression. It further informs about the progression variability as of its pace, its age at onset, and the specific patterns of event ordering.

To visualize the natural history of the disease on the 4 modalities studied in Chapter 5, we developed a 3D animated view available at `www.digital-brain.org`, screen-shooted on Fig. 8.1. The top row displays the decrease of both the cortical thickness over the brain

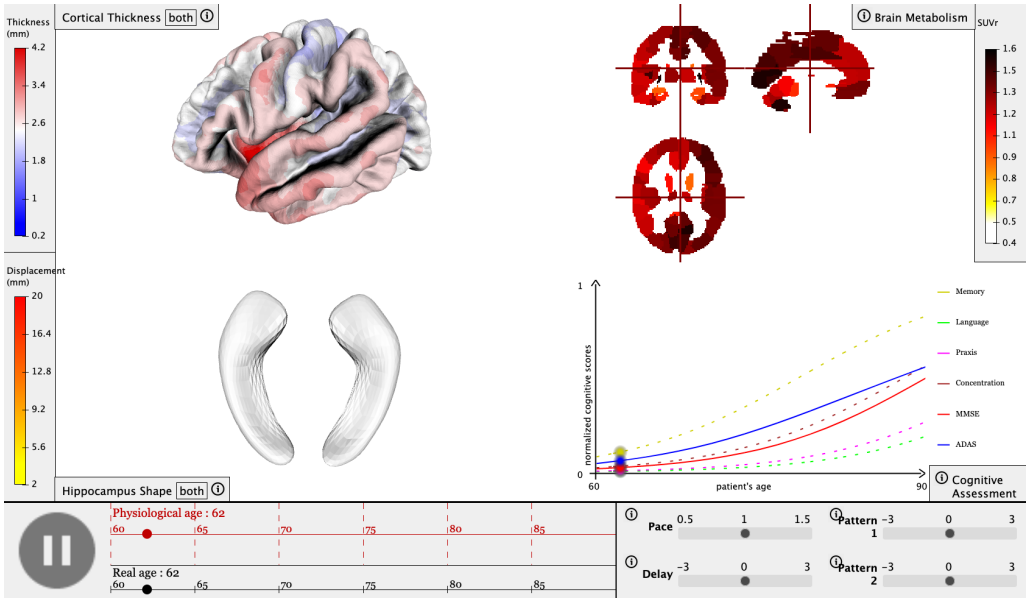


Figure 8.1: Reconstruction of the natural course of Alzheimer's disease from 60 to 90 years on 4 modalities : the cortical atrophy of the brain thickness (top left), the decrease of the brain hypometabolism derived from PET-FDG signal, the hippocampus mesh shrinkage, and, the cognitive assessments. The four triggers at the bottom right characterizes the spatiotemporal variability of disease progression. The two ages (bottom left) show the distinction between the physiological age that is related to a given stage of the disease and the observed age.

surface (left) and of the glucose consumption in the brain (right) from 60 to 90 years old. The bottom row displays the shrinkage of the hippocampi meshes (left) and the evolution of the cognitive abilities measured by 6 cognitive scores (right).

On top of these evolution, the model captures the progression variability within the cohort. It is represented by the "Pace", "Delay", "Pattern 1" and "Pattern 2" triggers at the bottom right of Fig. 8.1. The first and second one describe the speed of progression and the temporal offset (in years) to the mean progression. The *patterns* are the non-temporal parameters that are expressed by a reordering of the clinical assessments, as shown on Fig. 8.2, or the change of the spatial patterns of change for the cortical atrophy and the glucose consumption. These triggers are scaled to range between -3 and +3 standard deviation of the underlying gaussian distribution.

Given the mean trajectory, a patient is fully characterized by the set of individual parameters parameters that correspond to the four triggers. However, in real-cases, each patient presents different temporal and pattern parameters for each modality, e.g. one pace of progression for the cognitive decline and one for the cortical thinning. For the sake of clarity, we have mapped the temporal parameters to the same distribution support. On the other hand, the geometrical parameters are the first components of a principal component analysis applied to the whole space of patterns across modalities. The temporal variability is also responsible of the unalignment between the observed aged and the corresponding disease stage. To this end, we represent the real age and its physiological counterpart on the bottom left timeline.

This tool properly describes the natural history of the AD along with its variability. We believe this model to be easily personalizable to other diseases, in order to help the medical community to have a clearer view of the disease, such as the concomittent and consecutive offets, especially across modalities.

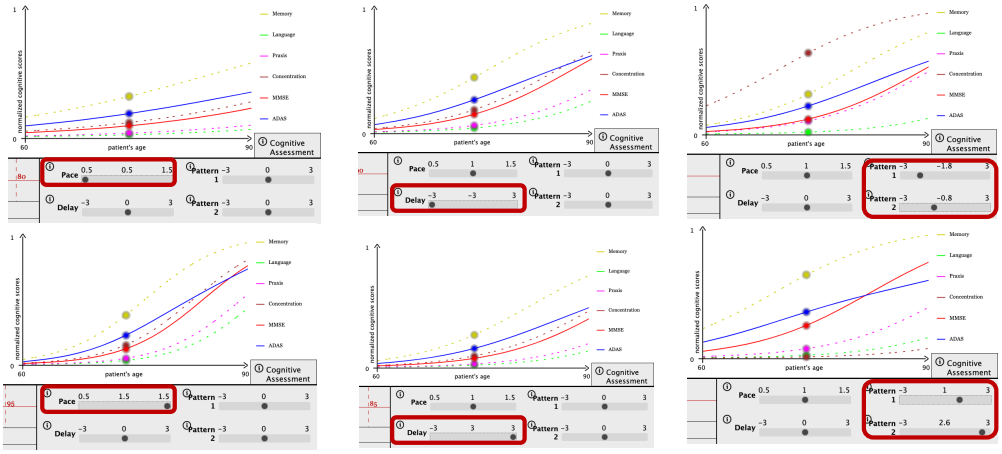


Figure 8.2: Disease progression variability illustrated on the cognitive scores. First row presents the variation of the pace of conversion, the second row displays the temporal shift in years and the third row shows examples of different sequences of cognitive alterations.

8.2.3 ADNI 1 Million

The model that estimates the long-term disease progression of the 4 aforementioned modalities and presented in `www.digital-brain.org` is generative in the sense that it estimates the distribution of the individual parameters, i.e. the values of the "Pace", "Delay", "Pattern 1" and "Pattern 2" triggers. In Chapter 6, we have shown that we are able to simulate a new set of individual parameters according to the estimated distribution. Each of this set entirely describe a virtual patient that can be observed at any arbitrary age t. We have also shown in Chapter 6 that once simulated, it is not trivial for a discriminator to distinguish real from virtual data. Furthermore, they can be used to train algorithms that require large amount of data.

The resulting virtual cohorts have interesting properties as they replicate the characteristics of the initial one, they can include an arbitrary large number of patients of time-points per patient, and, finally, as they are constituted of virtual patient, they can be shared easily without violating anonymization policies.

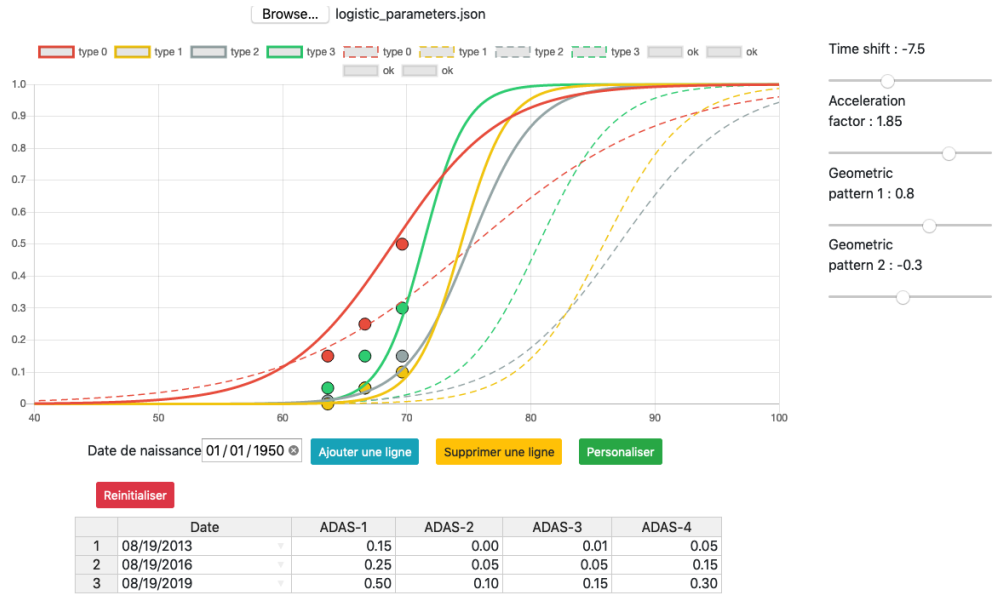
To illustrate the capacity of our model to simulate virtual cohorts at large scale, we made available 5 virtual cohort that includes :

- the cognitive assessments of 1 million subjects,
- the cortical thickness (decimated to 3658 ROI) of 100.000 subjects,
- the shape of the right and left hippocampus of 1 million subjects,
- the glucose metabolism (SUVr) projected on the AAL2 atlas of 1 million subjects.

We also provide a python 3.7 script to map the 3658 cortical thickness ROI to the FSAverage representation (+360k nodes) directly as a MGH file (to be read with Freeview). All these materials are available at `www.digital-brain.org`.

8.2.4 Patient care with future prediction

Nowadays, the patients with cognitive complaints such as memory impairments are guided towards neurologists. Unfortunately, the consultation essentially consists in a diagnosis but no possible cure, medication or at least information regarding future stages. This is due to the high variability in the evolution of the memory loss that might be the result of



	Date	ADAS-1	ADAS-2	ADAS-3	ADAS-4
1	08/19/2013	0.15	0.00	0.01	0.05
2	08/19/2016	0.25	0.05	0.05	0.15
3	08/19/2019	0.50	0.10	0.15	0.30

Figure 8.3: Browser view of the disease progression model on 4 sub-scores of the ADAS-Cog assessments. The dashed lines correspond to the mean scenario of progression. The model is personalized to individual measurements (dots) in order to show the patient disease progression.

different dementia or just being related to normal ageing. This leads to insufficient medical care of the patient and its environment (e.g. relatives, habits, place of living) that cannot be adapted to face future stages.

On the other hand, we have shown in Chapter 5 that the presented mixed-effects model can accurately predict the progression of the cognitive assessments for patient that present mild cognitive impairments 3 and 4 years in advance. This prediction, in real case scenario can inform the neurologist about the potential future disease stage of the patient, about his or her pace of progression, about the relative positioning to other individuals, and, about the particular sequence of cognitive impairments. This paves the way to a better patient care, e.g. by allowing the patient to adapt and anticipate future stages. To allow such prediction, we developed a toolbox that takes the form of a website as it is displayed in browsers, even though it runs locally to respect restrictions in data sharing policies. Fig. 8.3 shows the disease progression model fitted on 4 subscores of the ADAS-Cog scale. The individual measurements are then used to personalize the progression to the patient progression. This visualisation is both helpful to compare the individual sequence of events to the average scenario of change and to predict future stages.

Chapter 6 has shown the quality of the prediction on the whole cohort. While accurate for most of the patients, there still are outliers for which the prediction is inaccurate. Further works, besides improving the quality of the prediction, should focus on providing a metric on the confidence of the prediction. This results is as important as the prediction itself to properly inform the medical staff and the patient.

8.2.5 Dashboard for clinical studies

The characterization of the individual trajectories can be extended to multiple patients. This is particularly interesting to monitor cohorts in clinical studies. We developed a dashboard (here in the case of the Alzheimer's Disease Neuroimaging Initiative (ADNI) database used in previous chapters) that runs on browsers as shown on Fig. 8.4. For each patient, it displays the demographic information and the temporal parameters (acceleration

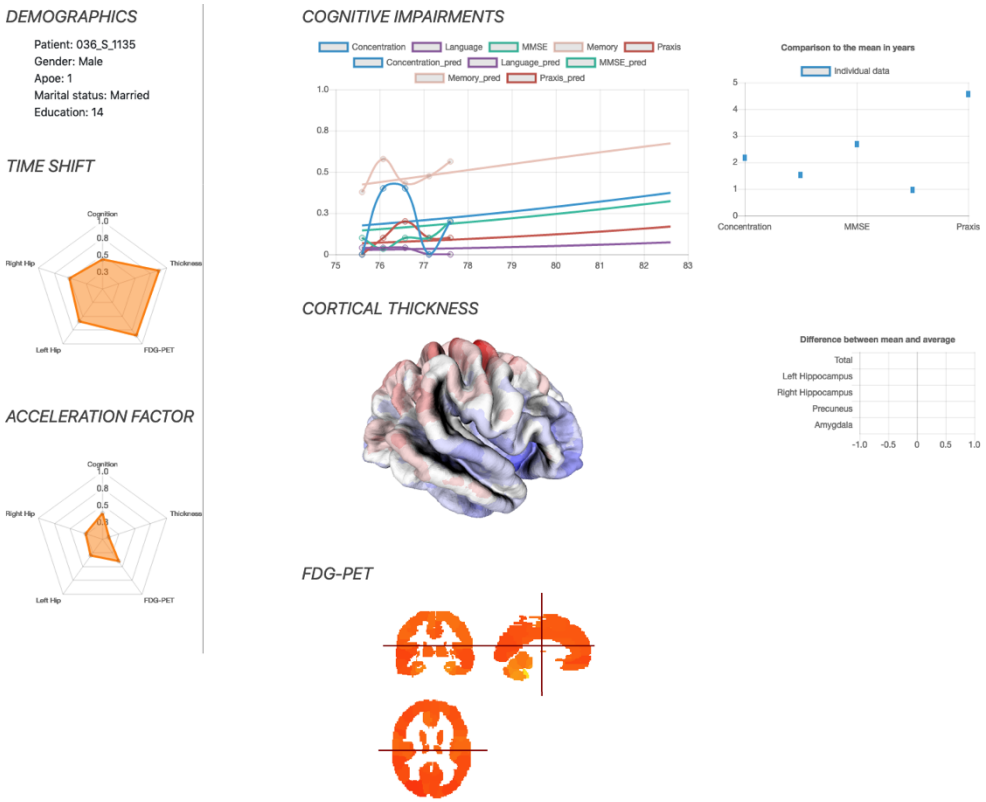


Figure 8.4: Example of a dashboard to monitor individual patients in clinical studies. It shows the individual demographic information and individual parameters (left) as well as the reconstructed data and their prediction at different stages.

factor and time-shift) on the left column. The right column groups the reconstruction of the different modalities : the cognitive impairments over time, a 3D interactive view of the cortical thickness and the three slices of the PET-FDG. This dashboard is a proof-of-concept but has the potential to be personalizable to any ad-hoc study.

Another characteristic of the clinical studies that test drug responses is that they split the cohort in subgroups (e.g. test and placebo) to measure the drug effect. This relies on the hypothesis that the disease progression is comparable between groups. Furthermore, this group-wise analysis prevents from properly measuring the relation between a drug effect and a particular cofactor, unless it has sufficient subgroups in the test and placebo groups with a given cofactor.

Contrary to classic methodologies in clinical studies, our model is able to characterize the individual spatiotemporal trajectory with an accurate prediction up to 4 years. The latter, that characterize the natural evolution, can be compared to the evolution in presence of medication, resulting in a drug effect measurement at the individual level. This indeed allows to measure the drug effect between placebo and test subgroups but it enables to exhibit the characteristics and cofactors of the individuals that are responsive to the drug, in case the overall drug effect is not sufficient. It opens the way to clustering the patients targeted by each drug, allowing a finer classification of the disease in term of targeted cofactors. Furthermore, these cofactors and characteristics might be potentially detected years before the disease appear, so that the drug might be tested at earlier disease stages, prior to the incurable neuronal loss.

8.3 Conclusion

The generative mixed effects model is a framework that can be adapted to different use cases. This toolbox allows a better understanding of the long-term disease progression, as well as a prediction of future stages at an individual level. The latter can be used to monitor cohorts in clinical studies and to report the drug effect at the subject level, in order to exhibit the cofactors that are most likely to respond to the drug.

The related digital tools, directly interfacable with the outputs of `Leaspy` can be used easily as they rely on web browsers. The latter can be run locally to be conform to the data sharing policies.

Conclusion and perspectives

Conclusion

This thesis naturally emerges in the context of *disease modeling*. It presents a generative mixed-effects model that takes the best out of individual longitudinal observations in order to estimate the natural course of a disease, at both an average and individual level. As the consequences of the disease might be very heterogeneous across modalities, such description has been made available for different type of data. While exhibiting and validating the possibilities offered by the model, we described the Alzheimer's disease progression to a whole new level. This has been made possible thanks to numerical tools that have been released to benefit the entire medical community.

Natural framework to study multimodal longitudinal data

The thesis presents a generative mixed-effects model that is particularly well suited to deal with longitudinal data. It especially recombines individual short-term measurements into a long-term scenario of change. The estimation of this average trajectory relies on the model ability to accurately reconstruct the individual progressions. Its corollary enables to impute missing values but also to predict future values up to 4 years in advance.

The model, initially designed to handle vector data such as cognitive assessments, was extended to deal with high dimensional data that present a spatial structure such as images or meshes. In such context, the model is constrained to enforce similar temporal profile of progression for *close* regions. This model allowed to analyse the cortical atrophy extracted from T1-weighted MRI, and, the brain hypometabolism extracted from PET-FDG.

Furthermore, the generative property of the model is particularly appealing as it permits to simulate an arbitrary number of virtual patients with more time-points and less time between measurements. Similarly to standard data augmentation techniques, if simulated correctly, these patients can be used to unbias and balance initial cohorts, but also to enhance other algorithms by providing additional longitudinal data.

Better understanding of the Alzheimer's disease progression

To validate the various applications provided by the model along with its ability to deal with different data type, we characterized the long-term disease progression of Alzheimer's disease. Out of the patients from the Alzheimer's Disease Neuroimaging Initiative dataset that converted from mild cognitive impairments (MCI) to Alzheimer's disease (AD), we extracted the cognitive scores, PET-FDG images and T1-weighted MRI features (cortical thickness and hippocampi meshes). We reconstructed the evolution of these feature during the course of the disease, from early stages to post-conversion ages.

Simultaneously, the model was able to reconstruct the individual profiles of evolution up to the noise level. By studying the individual parameters that enable this reconstruction, we were able to clearly reveal the cofactors that modulate the disease progression. Finally, their study provides an in-depth comprehension of the interactions between the modalities.

In the meantime, we selected the larger group of MCI patients (potentially with no conversion to AD) to characterize their evolution, to impute missing values and to predict future time-points. These patients were used to simulate virtual longitudinal cohorts, thereafter used to train recurrent neural networks. It led to state-of-the-art results to predict cognitive assessments up to 4 years in advance.

Succession of tools to improve medical research

In parallel to the model development and its validation on the Alzheimer's disease progression, we rationalized the code development and integrated it into the `Leaspy` Python package. It provides an easy API to reproduce the previous applications to new biomarkers or cohorts that track other diseases. The code was designed to fit the needs of both Python beginners that focus on new medical findings, but also for researchers that aim to develop and integrate new models and algorithms.

This release was followed by the development of digital tools directly intended to enhance the medical community. First, `www.digital-brain.org` is a summary of the evolution of the four aforementioned modalities during the course of Alzheimer's disease. It exhibits this progression over 30 years along with its variability. Then, we developed a tool that allows to predict future time-points based on patient visits and that can be of high interest in neurology offices. It describes the predicted temporal progression while comparing it to an average profile. Finally, we conceived a dashboard that might be used in clinical study to monitor a large cohort but also to measure the drug effect at the individual level.

Limitation & Perspectives

Overcome the monotonic and parametric progression constraints

One of the main limitation of the model lies in the parametric temporal profile of progression, that is also related to the monotonic assumptions. While the current model instantiations have shown promising results for the data at hand, they are not likely to describe complicated spatiotemporal profiles. This has to be put in perspective to the quality and quantity of data : in the future, the longitudinal measurements will provide more follow-up time-points with probably less noise. This will improve the *signal to noise* ratio for these longitudinal data such that more complex dynamics might be estimated. Capturing them accurately would require to potentially learn more complex Riemannian metrics - or to change the Riemannian setting.

From estimation heuristics to proven convergence

In previous chapters, we introduced few heuristics to improve the estimation stability and robustness. Among them, we can cite a tempered profile (simulated annealing) for the variances of the laws that control the MCMC-SAEM algorithm, and also a gradient descent or variational inference of the model parameters. These heuristics were used as initialization of the MCMC-SAEM, which, without the heuristics, is proven to converge to a local maximum. Further studies are worth to investigate the convergence properties of the MCMC-SAEM under these additional techniques that accelerate the convergence in practice.

Distance in the model designed for spatially structured data

The model introduced in Part II, that estimates the temporal progression of spatially-structured data, relies on a predefined estimation of the distance between the regions. This distance has been pre-computed thanks to a geodesic distance on the cortical surface. However, studies have shown that the interaction between brain areas is not only proportional to a geometrical distance but that there exists areas strongly connected despite their apparent distance. Some have hypothesized that the cortical atrophy propagates through different regions via the fiber bundles that connect regions of the brain. Therefore, the pre-computed distance might take into account the number of fiber bundles connecting different areas. Another interesting approach that can be investigated is to iteratively

estimate the disease progression (as previously) along with the estimation of the distances between regions. The outcoming distance matrix might relate about how *close*, in term of disease progression, different regions are.

Measuring the quality of a prediction & the drug effect

In chapter 8, we presented two digital tools that rely on an accurate prediction of future time-points. While we have shown that the prediction is of the noise level for a large majority of predictions, there still exists outliers for whom the prediction is inaccurate. As this is a critical information, especially when it is delivered to a patient and its relatives or when it is used for further medical actions, assessing the quality of the predictions is critical to see them integrated to support-decision systems. Further attention should be paid to the detection of this outliers, for instance with Machine Learning techniques that can identify the patient that are likely to have an inaccurate prediction. Said differently, the underlying paradigm consists in giving accurate predictions to a smaller group of patients rather than to a larger group where some prediction are unreliable.

Valorization

Scientific publications

Articles in journal

- [Koval et al., 2018b] I. Koval, J.-B. Schiratti, A. Routier, M. Bacci, O. Colliot, S. Allassonnière and S. Durrleman. Spatiotemporal propagation of the cortical atrophy: Population and individual patterns. In : *Frontiers in Neurology, 2019.* Volume 9. pp. 235.

Peer-reviewed conference papers

- [Louis et al., 2019] M. Louis, R. Couronné, I. Koval, B. Charlier and S. Durrleman. Riemannian Geometry Learning for Disease Progression Modelling. In : *International Conference on Information Processing in Medical Imaging - IPMI 2019.* pp. 543-553.

- [Koval et al., 2017] I. Koval, J.-B. Schiratti, A. Routier, M. Bacci, O. Colliot, S. Allassonnière and S. Durrleman. Statistical Learning of Spatiotemporal Patterns from Longitudinal Manifold-Valued Networks. In : *International Conference on Medical Image Computing and Computer-Assisted Intervention - MICCAI 2017.* pp. 451-459.

Abstract

- [Koval et al., 2019] I. Koval, A. Marcoux, N. Burgos, S. Allassonniére, O. Colliot and S. Durrleman. Deciphering the Progression of PET Alterations using Surface-Based Spatiotemporal Modeling. In : *Organization of Humain Brain Mapping - OHBM 2019. This paper corresponds to Chapter 4.*

- [Ansart et al., 2018] M. Ansart, I. Koval, A. Bertrand, D. Dormont and S. Durrleman. Design of a Decision Support System for Predicting the Progression of Alzheimer's Disease. In : *Alzheimer's & Dementia : The Journal of Alzheimer's Association.* Volume 7. page 433.

Preprints

- I. Koval, S. Allassonnière and S. Durrleman. Longitudinal Data Augmentation Framework for the Study of Disease Progression. *This preprint corresponds to Chapter 6.*

- I Koval, R. Couronné, S. Allassonnière and S. Durrleman. Leaspy : a Python Package Toolbox to Learn Spatiotemporal Patterns of Disease Progression. *This preprint is included in Chapter 7.*

- I. Koval, A. Bône, M. Louis, S. Bottani, A. Marcoux, J. Samper-Gonzalez, N. Burgos, B. Charlier, A. Bertrand, S. Epelbaum, O. Colliot, S. Allassonnière and S. Durrleman. Simulating Alzheimer's Disease Progression with Personalised Digital Brain Models. *This preprint corresponds to the Chapter 5.*

Talks

- Numerical Model of Neurodegenerative Disease peogression : Description, Personalization, Prediction. In : *«Numerical tools to improve the biological diagnosis »woskhop, French Radiology Society.* October 2019. Paris, France.
- Numerical Model of Neurodegenerative Disease Progression, especially Alzheimer's Disease. In : *Inserm - Quebec workshop on ageing.* October 2019. Paris, France.
- Numerical Model of Neurodegenerative Disease Progression : Description, Personalization, Prediction, Simulation. In : *Les mathématiques de l'imagerie. Agence pour les Mathématiques en Interaction avec l'Entreprise et la Société..* March 2019. Henri Poincaré Institute. Paris, France.
- Construct a Digital Model of Alzheimer's Disease with Leaspy and Deformetrica. In : *«Unified Modelling Framework »workshop, European Progression Of Neurodegenerative Diseases initiative - EuroPOND.* February 2019. Milan, Italy.
- Learning Digital Models of Alzheimer's Disease Progression. In *«Mathematical Methods for Spatiotemporal Imaging »workshop, SIAM Conference on Medical Imaging.* June 2018. Bologna, Italy.
- Statistical Learning of Spatiotemporal Patterns from Longitudinal Manifold-Valued Networks. In : *Bayes Comp.* March 2018. Barcelona, Spain.
- Numerical Models of Brain Disease Progression. In : *Microsoft & ICM days.* March 2018. Station F. Paris, France.
- Digital models of Brain Aging. Spatiotemporal Evolution of Biomarkers. In : *Epidemiology team meeting, Erasmus Medical Center.* January 2018. Rotterdam, The Netherlands.
- Network Propagation Model. In : *European Progression Of Neurodegenerative Diseases initiative workshop - EuroPOND.* February 2017. London, United-Kingdom.

Software & Website

- `Leasp` C++ 14 software [1].
- `Leaspy` Python package [2].
- `www.digital-brain.org` website.

Miscelleanous

- Blog post on Medium «Analysis of longitudinal data made easy with Leaspy »[3] . August 2019.
- Participation to the brain-related atlas «Le Grand Atlas du Cerveau ». Edited by *Glénat, Le Monde, ICM.* December 2018.

[1] `https://gitlab.com/icm-institute/aramislab/leasp`
[2] `https://gitlab.com/icm-institute/aramislab/leaspy/`
[3] `https://medium.com/@igoroa/analysis-of-longitudinal-data-made-easy-with-leaspy-f8d529fcb5f8`

Acknowledgements

This thesis has been financed by the European progression of neurological disease initiative (EuroPOND), which has received fundings from the European Union's Horizon 2020 research and innovation programme.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.;Cogstate;Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies;Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Bibliography

[Abdi, 2003] Abdi, H. (2003). Partial least square regression (pls regression). *Encyclopedia for research methods for the social sciences*, 6(4):792–795.

[Ahrens et al., 2005] Ahrens, J., Geveci, B., and Law, C. (2005). Paraview: An end-user tool for large data visualization. *The visualization handbook*, 717.

[Allassonnière et al., 2015] Allassonnière, S., Durrleman, S., and Kuhn, E. (2015). Bayesian mixed effect atlas estimation with a diffeomorphic deformation model. *SIAM Journal on Imaging Science*, 8:1367–1395.

[Allassonniere and Kuhn, 2015] Allassonniere, S. and Kuhn, E. (2015). Convergent stochastic expectation maximization algorithm with efficient sampling in high dimension. application to deformable template model estimation. *Computational Statistics & Data Analysis*, 91:4–19.

[Allassonnière et al., 2010] Allassonnière, S., Kuhn, E., and Trouvé, A. (2010). Construction of bayesian deformable models via a stochastic approximation algorithm: a convergence study. *Bernoulli*, 16(3):641–678.

[Allassonniere et al., 2012] Allassonniere, S., Younes, L., et al. (2012). A stochastic algorithm for probabilistic independent component analysis. *The Annals of Applied Statistics*, 6(1):125–160.

[Amieva et al., 2008] Amieva, H., Le Goff, M., Millet, X., Orgogozo, J. M., Pérès, K., Barberger-Gateau, P., Jacqmin-Gadda, H., and Dartigues, J. F. (2008). Prodromal alzheimer's disease: Successive emergence of the clinical symptoms. *Annals of Neurology*, 64(5):492–498.

[Ansart et al., 2018] Ansart, M., Koval, I., Bertrand, A., Dormont, D., and Durrleman, S. (2018). Design of a decision support system for predicting the progression of alzheimer's disease. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 14(7):P433.

[Atchade, 2006] Atchade, Y. F. (2006). An adaptive version for the metropolis adjusted langevin algorithm with a truncated drift. *Methodology and Computing in applied Probability*, 8(2):235–254.

[Baron et al., 2001] Baron, J., Chételat, G., Desgranges, B., Perchey, G., Landeau, B., de la Sayette, V., and Eustache, F. (2001). In vivo mapping of gray matter loss with voxel-based morphometry in mild alzheimer's disease. *NeuroImage*, 14(2):298 – 309.

[Bates and Pinheiro, 1998] Bates, D. M. and Pinheiro, J. C. (1998). Linear and nonlinear mixed-effects models.

[Beckett, 1993] Beckett, L. A. (1993). Maximum likelihood estimation in mallows's model using partially ranked data. In *Probability models and statistical analyses for ranking data*, pages 92–107. Springer.

[Bejnordi et al., 2017] Bejnordi, B. E., Veta, M., Van Diest, P. J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J. A., Hermsen, M., Manson, Q. F., Balkenhol, M., et al. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210.

[Bengio et al., 2013] Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

[Benzinger et al., 2013] Benzinger, T. L., Blazey, T., Jack, C. R., Koeppe, R. A., Su, Y., Xiong, C., Raichle, M. E., Snyder, A. Z., Ances, B. M., Bateman, R. J., et al. (2013). Regional variability of imaging biomarkers in autosomal dominant alzheimer's disease. *Proceedings of the National Academy of Sciences*, 110(47):E4502–E4509.

[Biering et al., 2015] Biering, K., Hjollund, N. H., and Frydenberg, M. (2015). Using multiple imputation to deal with missing data and attrition in longitudinal studies with repeated measures of patient-reported outcomes. *Clinical epidemiology*, 7:91.

[Bilgel et al., 2015] Bilgel, M., Jedynak, B., Wong, D. F., Resnick, S. M., and Prince, J. L. (2015). Temporal trajectory and progression score estimation from voxelwise longitudinal imaging measures: Application to amyloid imaging. In *International Conference on Information Processing in Medical Imaging*, pages 424–436. Springer.

[Bilgel et al., 2016] Bilgel, M., Prince, J. L., Wong, D. F., Resnick, S. M., and Jedynak, B. M. (2016). A multivariate nonlinear mixed effects model for longitudinal image analysis: Application to amyloid imaging. *NeuroImage*, 134:658–670.

[Bône et al., 2018] Bône, A., Colliot, O., and Durrleman, S. (2018). Learning distributions of shape trajectories from longitudinal datasets: a hierarchical model on a manifold of diffeomorphisms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9271–9280.

[Broomhead and Lowe, 1988] Broomhead, D. S. and Lowe, D. (1988). Radial basis functions, multi-variable functional interpolation and adaptive networks. Technical report, Royal Signals and Radar Establishment Malvern (United Kingdom).

[Bullmore and Sporns, 2009] Bullmore, E. and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198.

[Byrd et al., 1995] Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208.

[Chan et al., 2001] Chan, D., Fox, N. C., Scahill, R. I., Crum, W. R., Whitwell, J. L., Leschziner, G., Rossor, A. M., Stevens, J. M., Cipolotti, L., and Rossor, M. N. (2001). Patterns of temporal lobe atrophy in semantic dementia and alzheimer's disease. *Annals of neurology*, 49(4):433–442.

[Chen et al., 2010] Chen, K., Langbaum, J. B., Fleisher, A. S., Ayutyanont, N., Reschke, C., Lee, W., Liu, X., Bandy, D., Alexander, G. E., Thompson, P. M., et al. (2010). Twelve-month metabolic declines in probable alzheimer's disease and amnestic mild cognitive impairment assessed using an empirically pre-defined statistical region-of-interest: findings from the alzheimer's disease neuroimaging initiative. *Neuroimage*, 51(2):654–664.

[Chen et al., 2016] Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. (2016). Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180.

[Chib and Greenberg, 1995] Chib, S. and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The american statistician*, 49(4):327–335.

[Clark et al., 2012] Clark, C. M., Pontecorvo, M. J., Beach, T. G., Bedell, B. J., Coleman, R. E., Doraiswamy, P. M., Fleisher, A. S., Reiman, E. M., Sabbagh, M. N., Sadowsky, C. H., et al. (2012). Cerebral pet with florbetapir compared with neuropathology at autopsy for detection of neuritic amyloid-β plaques: a prospective cohort study. *The Lancet Neurology*, 11(8):669–678.

[Clark et al., 1999] Clark, C. M., Sheppard, L., Fillenbaum, G. G., Galasko, D., Morris, J. C., Koss, E., Mohs, R., and Heyman, A. (1999). Variability in annual mini-mental state examination score in patients with probable alzheimer disease: a clinical perspective of data from the consortium to establish a registry for alzheimer's disease. *Archives of neurology*, 56(7):857–862.

[Corder et al., 1993] Corder, E., Saunders, A., Strittmatter, W., Schmechel, D., Gaskell, P., Small, G., Roses, A., Haines, J., and Pericak-Vance, M. (1993). Gene dose of apolipoprotein e type 4 allele and the risk of alzheimer's disease in late onset families. *Science*, 261(5123):921–923.

[Couronne et al., 2019] Couronne, R., Vidailhet, M., Corvol, J.-C., Lehéricy, S., and Durrleman, S. (2019). Learning disease progression models with longitudinal data and missing values. In *ISBI 2019-International Symposium on Biomedical Imaging*.

[Dalca et al., 2015] Dalca, A. V., Sridharan, R., Sabuncu, M. R., and Golland, P. (2015). Predictive modeling of anatomy with genetic and clinical data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 519–526. Springer.

[Damoiseaux and Greicius, 2009] Damoiseaux, J. S. and Greicius, M. D. (2009). Greater than the sum of its parts: a review of studies combining structural connectivity and resting-state functional connectivity. *Brain Structure and Function*, 213(6):525–533.

[De Silva et al., 2017] De Silva, A. P., Moreno-Betancur, M., De Livera, A. M., Lee, K. J., and Simpson, J. A. (2017). A comparison of multiple imputation methods for handling missing values in longitudinal data in the presence of a time-varying covariate with a non-linear association with time: a simulation study. *BMC medical research methodology*, 17(1):114.

[Delyon et al., 1999] Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of a stochastic approximation version of the expectation-maximization algorithm. *Annals of Statistics*, pages 94–128.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

[Devanand et al., 2007] Devanand, D. P., Pradhaban, G., Liu, X., Khandji, A., De Santi, S., Segal, S., Rusinek, H., Pelton, G. H., Honig, L. S., Mayeux, R., Stern, Y., Tabert, M. H., and de Leon, M. J. (2007). Hippocampal and entorhinal atrophy in mild cognitive impairment: Prediction of alzheimer disease. *Neurology*, 68(11):828–836.

[Do Carmo Valero, 1992] Do Carmo Valero, M. P. (1992). *Riemannian geometry*. Birkhäuser.

[Donohue et al., 2014] Donohue, M., Jacqmin-Gadda, H., Goff, M. L., Thomas, R., Raman, R., Gams, A., Beckett, L., Jack, C., Weiner, M., Dartigues, J.-F., Aisen, P., and the ADNI (2014). Estimating long-term multivariate progression from short-term data. *Alzheimer's & Dementia*, 10(5):400–410.

[Dou et al., 2016] Dou, Q., Chen, H., Yu, L., Zhao, L., Qin, J., Wang, D., Mok, V. C., Shi, L., and Heng, P.-A. (2016). Automatic detection of cerebral microbleeds from mr images via 3d convolutional neural networks. *IEEE transactions on medical imaging*, 35(5):1182–1195.

[Drzezga et al., 2003] Drzezga, A., Lautenschlager, N., Siebner, H., Riemenschneider, M., Willoch, F., Minoshima, S., Schwaiger, M., and Kurz, A. (2003). Cerebral metabolic changes accompanying conversion of mild cognitive impairment into alzheimer's disease: a pet follow-up study. *European journal of nuclear medicine and molecular imaging*, 30(8):1104–1113.

[Du et al., 2001] Du, A. T., Schuff, N., Amend, D., Laakso, M. P., Hsu, Y. Y., Jagust, W. J., Yaffe, K., Kramer, J. H., Reed, B., Norman, D., Chui, H. C., and Weiner, M. W. (2001). Magnetic resonance imaging of the entorhinal cortex and hippocampus in mild cognitive impairment and alzheimer's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, 71(4):441–447.

[Durrleman, 2018] Durrleman, S. (2018). Geometrical approaches in statistical learning for the construction of digital models of the human brain. Habilitation à diriger des recherches, Pierre and Marie Curie University, Paris.

[Durrleman et al., 2013] Durrleman, S., Pennec, X., Trouvé, A., Braga, J., Gerig, G., and Ayache, N. (2013). Toward a comprehensive framework for the spatiotemporal statistical analysis of longitudinal shape data. *International Journal of Computer Vision*, 103(1):22–59.

[Durrleman et al., 2008] Durrleman, S., Pennec, X., Trouvé, A., Thompson, P., and Ayache, N. (2008). Inferring brain variability from diffeomorphic deformations of currents: an integrative approach. *Medical image analysis*, 12(5):626–637.

[Durrleman et al., 2014] Durrleman, S., Prastawa, M., Charon, N., Korenberg, J. R., Joshi, S., Gerig, G., and Trouvé, A. (2014). Morphometry of anatomical shape complexes with dense deformations and sparse parameters. *NeuroImage*.

[Efron et al., 2004] Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of statistics*, 32(2):407–499.

[Eskildsen et al., 2015] Eskildsen, S. F., Coupé, P., Fonov, V. S., Pruessner, J. C., Collins, D. L., Initiative, A. D. N., et al. (2015). Structural imaging biomarkers of alzheimer's disease: predicting disease progression. *Neurobiology of aging*, 36:S23–S31.

[Fan et al., 2008] Fan, Y., Resnick, M., Wu, X., and Davatzikos, C. (2008). Structural and functional biomarkers of prodromal alzheimer's disease: a high dimensional pattern classification study. *NeuroImage*, 41(2):277–285.

[Fischl and Dale, 2000] Fischl, B. and Dale, A. M. (2000). Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences*, 97(20):11050–11055.

[Fischl et al., 2002] Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Van Der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., et al. (2002).

Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3):341–355.

[Fischl et al., 1999] Fischl, B., Sereno, M., Tootell, R., and Dale, A. (1999). High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human Brain Mapp*, 8:272–284.

[Fishbaugh et al., 2014] Fishbaugh, J., Prastawa, M., Gerig, G., and Durrleman, S. (2014). Geodesic regression of image and shape data for improved modeling of 4D trajectories. In *ISBI 2014 - 11th International Symposium on Biomedical Imaging*, pages 385 – 388.

[Fisher, 1919] Fisher, R. A. (1919). Xv.—the correlation between relatives on the supposition of mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 52(2):399–433.

[Fisher, 1992] Fisher, R. A. (1992). Statistical methods for research workers. In *Breakthroughs in statistics*, pages 66–70. Springer.

[Folstein et al., 1975] Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). Mini-mental state: a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, 12(3):189–198.

[Fonteijn et al., 2012a] Fonteijn, H. M., Modat, M., Clarkson, M. J., Barnes, J., Lehmann, M., Hobbs, N. Z., Scahill, R. I., Tabrizi, S. J., Ourselin, S., Fox, N. C., and Alexander, D. C. (2012a). An event-based model for disease progression and its application in familial, alzheimer's disease and huntington's disease. *NeuroImage*, 60(3):1880–1889.

[Fonteijn et al., 2012b] Fonteijn, H. M., Modat, M., Clarkson, M. J., Barnes, J., Lehmann, M., Hobbs, N. Z., Scahill, R. I., Tabrizi, S. J., Ourselin, S., Fox, N. C., et al. (2012b). An event-based model for disease progression and its application in familial alzheimer's disease and huntington's disease. *NeuroImage*, 60(3):1880–1889.

[Gao et al., 1998] Gao, S., Hendrie, H., Hall, K., and Hui, S. (1998). The relationships between age, sex, and the incidence of dementia and alzheimer disease: A meta-analysis. *Archives of General Psychiatry*, 55(9):809–815.

[Gaser et al., 2013] Gaser, C., K, K. F., Kloppel, S., Koutsouleris, N., and Sauer, H. (2013). BrainAGE in Mild Cognitive Impaired Patients: Predicting the Conversion to Alzheimer's Disease. *PLoS ONE*, 8(6).

[Geman and Geman, 1984] Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, PAMI-6(6):721–741.

[Giraud et al., 2012] Giraud, C., Huet, S., and Verzelen, N. (2012). Graph selection with ggmselect. *Statistical applications in genetics and molecular biology*, 11(3).

[Gómez-Isla et al., 1996] Gómez-Isla, T., Price, J. L., McKeel Jr, D. W., Morris, J. C., Growdon, J. H., and Hyman, B. T. (1996). Profound loss of layer ii entorhinal cortex neurons occurs in very mild alzheimer's disease. *Journal of Neuroscience*, 16(14):4491–4500.

[Goodfellow et al., 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

[Gori et al., 2017] Gori, P., Colliot, O., Marrakchi-Kacem, L., Worbe, Y., Poupon, C., Hartmann, A., Ayache, N., and Durrleman, S. (2017). A Bayesian Framework for Joint Morphometry of Surface and Curve meshes in Multi-Object Complexes. *Medical Image Analysis*, 35:458–474.

[Greene et al., 2010] Greene, S. J., Killiany, R. J., Initiative, A. D. N., et al. (2010). Subregions of the inferior parietal lobule are affected in the progression to alzheimer's disease. *Neurobiology of aging*, 31(8):1304–1311.

[Guerrero et al., 2016] Guerrero, R., Schmidt-Richberg, A., Ledig, C., Tong, T., Wolz, R., Rueckert, D., and the ADNI (2016). Instantiated mixed effects modeling of alzheimer's disease markers. *Neuroimage*, 142(142):113–125.

[Gulshan et al., 2016] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410.

[Hensel et al., 2007] Hensel, A., Angermeyer, M. C., and Riedel-Heller, S. G. (2007). Measuring cognitive change in older adults: reliable change indices for the mmse. *Journal of Neurology, Neurosurgery & Psychiatry*.

[Huang and Alexander, 2012] Huang, J. and Alexander, D. (2012). Probabilistic event cascades for alzheimer's disease. In *Advances in neural information processing systems*, pages 3095–3103.

[Huang et al., 2016] Huang, L., Jin, Y., Gao, Y., Thung, K.-H., Shen, D., Initiative, A. D. N., et al. (2016). Longitudinal clinical score prediction in alzheimer's disease with soft-split sparse regression based random forest. *Neurobiology of aging*, 46:180–191.

[Hyman et al., 1984] Hyman, B. T., Van Hoesen, G. W., Damasio, A. R., and Barnes, C. L. (1984). Alzheimer's disease: cell-specific pathology isolates the hippocampal formation. *Science*, 225(4667):1168–1170.

[Iddi et al., 2019] Iddi, S., Li, D., Aisen, P. S., Rafii, M. S., Thompson, W. K., Donohue, M. C., Initiative, A. D. N., et al. (2019). Predicting the course of alzheimer's progression. *Brain informatics*, 6(1):6.

[Iturria-Medina et al., 2016] Iturria-Medina, Y., Sotero, R. C., Toussaint, P. J., Mateos-Pérez, J. M., Evans, A. C., Weiner, M. W., Aisen, P., Petersen, R., Jack, C. R., Jagust, W., et al. (2016). Early role of vascular dysregulation on late-onset alzheimer's disease based on multifactorial data-driven analysis. *Nature communications*, 7:11934.

[Jack et al., 2010] Jack, C. R., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., Petersen, R. C., and Trojanowski, J. Q. (2010). Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade. *The Lancet Neurology*, 9(1):119–128.

[Jack et al., 1997] Jack, C. R., Petersen, R. C., Xu, Y. C., Waring, S. C., O'brien, P. C., Tangalos, E. G., Smith, G. E., Ivnik, R. J., and Kokmen, E. (1997). Medial temporal atrophy on mri in normal aging and very mild alzheimer's disease. *Neurology*, 49(3):786–794.

[Jack et al., 2015] Jack, C. R., Wiste, H. J., Weigand, S. D., Knopman, D. S., Vemuri, P., Mielke, M. M., Lowe, V., Senjem, M. L., Gunter, J. L., Machulda, M. M., et al. (2015). Age, sex, and apoe $\varepsilon 4$ effects on memory, brain structure, and β-amyloid across the adult life span. *JAMA neurology*, 72(5):511–519.

[Jack Jr et al., 2010a] Jack Jr, C. R., Bernstein, M. A., Borowski, B. J., Gunter, J. L., Fox, N. C., Thompson, P. M., Schuff, N., Krueger, G., Killiany, R. J., DeCarli, C. S., et al. (2010a). Update on the magnetic resonance imaging core of the alzheimer's disease neuroimaging initiative. *Alzheimer's & Dementia*, 6(3):212–220.

[Jack Jr et al., 2008] Jack Jr, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., L. Whitwell, J., Ward, C., et al. (2008). The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691.

[Jack Jr et al., 2010b] Jack Jr, C. R., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., Petersen, R. C., and Trojanowski, J. Q. (2010b). Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade. *The Lancet Neurology*, 9(1):119–128.

[Jacobs et al., 2012] Jacobs, H. I., Van Boxtel, M. P., Jolles, J., Verhey, F. R., and Uylings, H. B. (2012). Parietal cortex matters in alzheimer's disease: an overview of structural, functional and metabolic findings. *Neuroscience & Biobehavioral Reviews*, 36(1):297–309.

[Jedynak et al., 2012] Jedynak, B. M., Lang, A., Liu, B., Katz, E., Zhang, Y., Wyman, B. T., Raunig, D., Jedynak, C. P., Caffo, B., Prince, J. L., et al. (2012). A computational neurodegenerative disease progression score: method and results with the alzheimer's disease neuroimaging initiative cohort. *Neuroimage*, 63(3):1478–1486.

[Jian and Vemuri, 2011] Jian, B. and Vemuri, B. C. (2011). Robust point set registration using gaussian mixture models. *IEEE transactions on pattern analysis and machine intelligence*, 33(8):1633–1645.

[Kawahara et al., 2016] Kawahara, J., BenTaieb, A., and Hamarneh, G. (2016). Deep features to classify skin lesions. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 1397–1400. IEEE.

[Khanal et al., 2016] Khanal, B., Lorenzi, M., Ayache, N., and Pennec, X. (2016). A biophysical model of brain deformation to simulate and analyze longitudinal mris of patients with alzheimer's disease. *NeuroImage*, 134:35 – 52.

[Khanna et al., 2018] Khanna, S., Domingo-Fernàndez, D., Iyappan, A., Emon, M. A., Hofmann-Apitius, M., and Frohlich, H. (2018). Using multi-scale genetic, neuroimaging and clinical data for predicting alzheimer's disease and reconstruction of relevant biological mechanisms. *Scientific Reports*, 8(1).

[Kingma and Ba, 2014] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[Knopman et al., 2014] Knopman, D. S., Jack Jr, C. R., Wiste, H. J., Lundt, E. S., Weigand, S. D., Vemuri, P., Lowe, V. J., Kantarci, K., Gunter, J. L., Senjem, M. L., et al. (2014). 18f-fluorodeoxyglucose positron emission tomography, aging, and apolipoprotein e genotype in cognitively normal persons. *Neurobiology of aging*, 35(9):2096–2106.

[Koval et al., 2018a] Koval, I., Bône, A., Louis, M., Bottani, S., Marcoux, A., Samper-Gonzalez, J., Burgos, N., Charlier, B., Bertrand, A., Epelbaum, S., et al. (2018a). Simulating alzheimer's disease progression with personalised digital brain models.

[Koval et al., 2019] Koval, I., Marcoux, A., Burgos, N., Allassonnière, S., Colliot, O., and Durrleman, S. (2019). Deciphering the progression of pet alterations using surface-based spatiotemporal modeling.

[Koval et al., 2018b] Koval, I., Schiratti, J.-B., Routier, A., Bacci, M., Colliot, O., Allassonnière, S., and Durrleman, S. (2018b). Spatiotemporal propagation of the cortical atrophy: Population and individual patterns. *Frontiers in Neurology*, 9:235.

[Koval et al., 2017] Koval, I., Schiratti, J.-B., Routier, A., Bacci, M., Colliot, O., Allassonnière, S., Durrleman, S., Initiative, A. D. N., et al. (2017). Statistical learning of spatiotemporal patterns from longitudinal manifold-valued networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 451–459. Springer.

[Kuhn and Lavielle, 2004] Kuhn, E. and Lavielle, M. (2004). Coupling a stochastic approximation version of em with an mcmc procedure. *ESAIM: Probability and Statistics*, 8:115–131.

[Kuhn and Lavielle, 2005] Kuhn, E. and Lavielle, M. (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics & Data Analysis*, 49(4):1020–1038.

[Laird and Ware, 1982] Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, pages 963–974.

[Landau et al., 2013] Landau, S. M., Breault, C., Joshi, A. D., Pontecorvo, M., Mathis, C. A., Jagust, W. J., Mintun, M. A., et al. (2013). Amyloid-β imaging with pittsburgh compound b and florbetapir: comparing radiotracers and quantification methods. *Journal of nuclear medicine: official publication, Society of Nuclear Medicine*, 54(1):70.

[Lavielle, 2014] Lavielle, M. (2014). *Mixed effects models for the population approach: models, tasks, methods and tools*. CRC press.

[Lavielle and Aarons, 2016] Lavielle, M. and Aarons, L. (2016). What do we mean by identifiability in mixed effects models? *Journal of pharmacokinetics and pharmacodynamics*, 43(1):111–122.

[Le Guennec et al., 2016] Le Guennec, A., Malinowski, S., and Tavenard, R. (2016). Data augmentation for time series classification using convolutional neural networks. In *ECML/PKDD workshop on advanced analytics and learning on temporal data*.

[Leuchter et al., 1992] Leuchter, A. F., Newton, T. F., Cook, I. A., Walter, D. O., Rosenberg-Thompson, S., and Lachenbruch, P. A. (1992). Changes in brain functional connectivity in alzheimer-type and multi-infarct dementia. *Brain*, 115(5):1543–1561.

[Li et al., 2017] Li, J., Monroe, W., Shi, T., Jean, S., Ritter, A., and Jurafsky, D. (2017). Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*.

[Lindstrom and Bates, 1988] Lindstrom, M. J. and Bates, D. M. (1988). Newton—raphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404):1014–1022.

[Lindstrom and Bates, 1990] Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, pages 673–687.

[Lorenzi et al., 2017] Lorenzi, M., Filippone, M., Frisoni, G. B., Alexander, D. C., Ourselin, S., Initiative, A. D. N., et al. (2017). Probabilistic disease progression modeling to characterize diagnostic uncertainty: application to staging and prediction in alzheimer's disease. *NeuroImage*.

[Louis et al., 2019] Louis, M., Couronne, R., Koval, I., Charlier, B., and Durrleman, S. (2019). Riemannian geometry learning for disease progression modelling. In *International Conference on Information Processing in Medical Imaging*, pages 542–553. Springer.

[Maguire et al., 1998] Maguire, E. A., Burgess, N., Donnett, J. G., Frackowiak, R. S. J., Frith, C. D., and O'Keefe, J. (1998). Knowing where and getting there: A human navigation network. *Science*, 280(5365):921–924.

[Maninis et al., 2016] Maninis, K.-K., Pont-Tuset, J., Arbeláez, P., and Van Gool, L. (2016). Deep retinal image understanding. In *International conference on medical image computing and computer-assisted intervention*, pages 140–148. Springer.

[Marcoux et al., 2018] Marcoux, A., Burgos, N., Bertrand, A., Teichmann, M., Routier, A., Wen, J., Samper-González, J., Bottani, S., Durrleman, S., Habert, M.-O., et al. (2018). An automated pipeline for the analysis of pet data on the cortical surface. *Frontiers in neuroinformatics*, 12:94.

[Marinescu et al., 2017] Marinescu, R. V., Eshaghi, A., Lorenzi, M., Young, A. L., Oxtoby, N. P., Garbarino, S., Shakespeare, T. J., Crutch, S. J., Alexander, D. C., Initiative, A. D. N., et al. (2017). A vertex clustering model for disease progression: Application to cortical thickness images. In *International Conference on Information Processing in Medical Imaging*, pages 134–145. Springer.

[Mohs et al., 1997] Mohs, R. C., Knopman, D., Petersen, R. C., Ferris, S. H., Ernesto, C., Grundman, M., Sano, M., Bieliauskas, L., Geldmacher, D., Clark, C., et al. (1997). Development of cognitive instruments for use in clinical trials of antidementia drugs: additions to the alzheimer's disease assessment scale that broaden its scope. *Alzheimer disease and associated disorders*.

[Moore et al., 2018] Moore, P., Lyons, T., and Gallacher, J. (2018). Random forest prediction of Alzheimer's disease using pairwise selection from time series data. *arXiv:1808.03273 [q-bio, stat]*. arXiv: 1808.03273.

[Moradi et al., 2015] Moradi, E., Pepe, A., Gaser, C., Huttunen, H., and Tohka, J. (2015). Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *NeuroImage*, 104:398–412.

[Mosconi, 2005] Mosconi, L. (2005). Brain glucose metabolism in the early and specific diagnosis of alzheimer's disease. *European journal of nuclear medicine and molecular imaging*, 32(4):486–510.

[Mosconi et al., 2008] Mosconi, L., De Santi, S., Li, J., Tsui, W. H., Li, Y., Boppana, M., Laska, E., Rusinek, H., and de Leon, M. J. (2008). Hippocampal hypometabolism predicts cognitive decline from normal aging. *Neurobiology of aging*, 29(5):676–692.

[Mura et al., 2014] Mura, T., Proust-Lima, C., Jacqmin-Gadda, H., Akbaraly, T. N., Touchon, J., Dubois, B., and Berr, C. (2014). Measuring cognitive change in subjects with prodromal alzheimer's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, 85(4):363–370.

[Neal et al., 2011] Neal, R. M. et al. (2011). Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2.

[Nie et al., 2016] Nie, D., Zhang, H., Adeli, E., Liu, L., and Shen, D. (2016). 3d deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 212–220. Springer.

[Oxtoby et al., 2018] Oxtoby, N. P., Young, A. L., Cash, D. M., Benzinger, T. L., Fagan, A. M., Morris, J. C., Bateman, R. J., Fox, N. C., Schott, J. M., and Alexander, D. C. (2018). Data-driven models of dominantly-inherited alzheimer's disease progression. *Brain*, 141(5):1529–1544.

[Pagani et al., 2017] Pagani, M., Nobili, F., Morbelli, S., Arnaldi, D., Giuliani, A., Öberg, J., Girtler, N., Brugnolo, A., Picco, A., Bauckneht, M., et al. (2017). Early identification of mci converting to ad: a fdg pet study. *European Journal of Nuclear Medicine and Molecular Imaging*, 44(12):2042–2052.

[Penny et al., 2011] Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J., and Nichols, T. E. (2011). *Statistical parametric mapping: the analysis of functional brain images*. Elsevier.

[Petersen et al., 2008] Petersen, K. B., Pedersen, M. S., et al. (2008). The matrix cookbook. *Technical University of Denmark*, 7(15):40–41.

[Peyré et al., 2010] Peyré, G., Péchaud, M., Keriven, R., Cohen, L. D., et al. (2010). Geodesic methods in computer vision and graphics. *Foundations and Trends® in Computer Graphics and Vision*, 5(3–4):197–397.

[Pinheiro and Bates, 1995] Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of computational and Graphical Statistics*, 4(1):12–35.

[Poirier et al., 1993] Poirier, J., Bertrand, P., Kogan, S., Gauthier, S., Davignon, J., and Bouthillier, D. (1993). Apolipoprotein e polymorphism and alzheimer's disease. *The Lancet*, 342(8873):697–699.

[Poplin et al., 2018] Poplin, R., Varadarajan, A. V., Blumer, K., Liu, Y., McConnell, M. V., Corrado, G. S., Peng, L., and Webster, D. R. (2018). Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, 2(3):158.

[Powell, 1964] Powell, M. J. (1964). An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The computer journal*, 7(2):155–162.

[Raghavan et al., 2013] Raghavan, N., Samtani, M. N., Farnum, M., Yang, E., Novak, G., Grundman, M., Narayan, V., DiBernardo, A., Initiative, A. D. N., et al. (2013). The adas-cog revisited: novel composite scales based on adas-cog to improve efficiency in mci and early ad trials. *Alzheimer's & Dementia*, 9(1):S21–S31.

[Ravi et al., 2019] Ravi, D., Alexander, D. C., Oxtoby, N. P., Initiative, A. D. N., et al. (2019). Degenerative adversarial neuroimage nets: Generating images that mimic disease progression. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 164–172. Springer.

[Reuter et al., 2012] Reuter, M., Schmandsky, N., Rosas, H., and Fischl, B. (2012). Within-subject template estimation for unbiased longitudinal image analysis. *Neuroimage*, 61(4):1402–1418.

[Rolls et al., 2015] Rolls, E. T., Joliot, M., and Tzourio-Mazoyer, N. (2015). Implementation of a new parcellation of the orbitofrontal cortex in the automated anatomical labeling atlas. *Neuroimage*, 122:1–5.

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

[Rosen et al., 1984] Rosen, W. G., Mohs, R. C., and Davis, K. L. (1984). A new rating scale for alzheimer's disease. *The American journal of psychiatry*.

[Routier et al., 2018] Routier, A., Guillon, J., Burgos, N., Samper-Gonzalez, J., Wen, J., Fontanella, S., Bottani, S., Jacquemont, T., Marcoux, A., Gori, P., et al. (2018). Clinica: an open source software platform for reproducible clinical neuroscience studies. In *Annual meeting of the Organization for Human Brain Mapping-OHBM 2018*.

[Rubin, 2004] Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons.

[Ryu et al., 2011] Ryu, D., Li, E., and Mallick, B. K. (2011). Bayesian nonparametric regression analysis of data with random effects covariates from longitudinal measurements. *Biometrics*, 67(2):454–466.

[Samper-González et al., 2018] Samper-González, J., Burgos, N., Bottani, S., Fontanella, S., Lu, P., Marcoux, A., Routier, A., Guillon, J., Bacci, M., Wen, J., et al. (2018). Reproducible evaluation of classification methods in alzheimer's disease: framework and application to mri and pet data. *bioRxiv*, page 274324.

[Scahill et al., 2002] Scahill, R. I., Schott, J. M., Stevens, J. M., Rossor, M. N., and Fox, N. C. (2002). Mapping the evolution of regional atrophy in alzheimer's disease: Unbiased analysis of fluid-registered serial mri. *Proceedings of the National Academy of Sciences*, 99(7):4703–4707.

[Schindler et al., 2018] Schindler, S. E., Gray, J. D., Gordon, B. A., Xiong, C., Batrla-Utermann, R., Quan, M., Wahl, S., Benzinger, T. L., Holtzman, D. M., Morris, J. C., et al. (2018). Cerebrospinal fluid biomarkers measured by elecsys assays compared to amyloid imaging. *Alzheimer's & Dementia*.

[Schiratti, 2016] Schiratti, J.-B. (2016). *Models and algorithms to learn spatiotemporal changes from longitudinal manifold-valued observations*. PhD thesis, EDMH.

[Schiratti et al., 2015] Schiratti, J.-B., Allassonnière, S., Colliot, O., and Durrleman, S. (2015). Learning spatiotemporal trajectories from manifold-valued longitudinal data. In *Advances in Neural Information Processing Systems*, pages 2404–2412.

[Schiratti et al., 2017] Schiratti, J.-B., Allassonniere, S., Colliot, O., and Durrleman, S. (2017). A bayesian mixed-effects model to learn trajectories of changes from repeated manifold-valued observations. *Journal of Machine Learning Research*, 18(133):1–33.

[Segovia et al., 2012] Segovia, F., Górriz, J., Ramírez, J., Salas-Gonzalez, D., Álvarez, I., López, M., Chaves, R., Initiative, A. D. N., et al. (2012). A comparative study of feature extraction methods for the diagnosis of alzheimer's disease using the adni database. *Neurocomputing*, 75(1):64–71.

[Silver et al., 2017] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chena, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., and Hassabis, D. (2017). Mastering the game of go without human knowledge. *Nature*, 550:354–359.

[Singh et al., 2006] Singh, V., Chertkow, H., Lerch, J. P., Evans, A. C., Dorr, A. E., and Kabani, N. J. (2006). Spatial patterns of cortical thinning in mild cognitive impairment and alzheimer's disease. *Brain*, 129(11):2885.

[Skinner et al., 2012] Skinner, J., Carvalho, J. O., Potter, G. G., Thames, A., Zelinski, E., Crane, P. K., Gibbons, L. E., Initiative, A. D. N., et al. (2012). The alzheimer's disease assessment scale-cognitive-plus (adas-cog-plus): an expansion of the adas-cog to improve responsiveness in mci. *Brain imaging and behavior*, 6(4):489–501.

[Spratt et al., 2010] Spratt, M., Carpenter, J., Sterne, J. A., Carlin, J. B., Heron, J., Henderson, J., and Tilling, K. (2010). Strategies for multiple imputation in longitudinal studies. *American journal of epidemiology*, 172(4):478–487.

[Standish et al., 1996] Standish, T. I., Molloy, D. W., Bédard, M., Layne, E. C., Murray, E. A., and Strang, D. (1996). Improved reliability of the standardized alzheimer's disease assessment scale (sadas) compared with the alzheimer's disease assessment scale (adas). *Journal of the American Geriatrics Society*, 44(6):712–716.

[Strittmatter et al., 1993] Strittmatter, W. J., Saunders, A. M., Schmechel, D., Pericak-Vance, M., Enghild, J., Salvesen, G. S., and Roses, A. D. (1993). Apolipoprotein e: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial alzheimer disease. *Proceedings of the National Academy of Sciences*, 90(5):1977–1981.

[Suva et al., 1999] Suva, D., Favre, I., Kraftsik, R., Esteban, M., Lobrinus, A., and Miklossy, J. (1999). Primary motor cortex involvement in alzheimer disease. *Journal of neuropathology and experimental neurology*, 58(11):1125–1134.

[Tang, 2015] Tang, Y. (2015). An efficient monotone data augmentation algorithm for bayesian analysis of incomplete longitudinal data. *Statistics & Probability Letters*, 104:146–152.

[Tang, 2019] Tang, Y. (2019). A monotone data augmentation algorithm for longitudinal data analysis via multivariate skew-t, skew-normal or t distributions. *Statistical methods in medical research*, page 0962280219865579.

[Tong et al., 2016] Tong, T., Gao, Q., Guerrero, R., Ledig, C., Chen, L., Rueckert, D., Initiative, A. D. N., et al. (2016). A novel grading biomarker for the prediction of conversion from mild cognitive impairment to alzheimer's disease. *IEEE Transactions on Biomedical Engineering*, 64(1):155–165.

[Tzourio-Mazoyer et al., 2002] Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. (2002). Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage*, 15(1):273–289.

[Vaillant and Glaunès, 2005] Vaillant, M. and Glaunès, J. (2005). Surface matching via currents. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 381–392. Springer.

[Venkatraghavan et al., 2019] Venkatraghavan, V., Bron, E. E., Niessen, W. J., Klein, S., Initiative, A. D. N., et al. (2019). Disease progression timeline estimation for alzheimer's disease using discriminative event based modeling. *NeuroImage*, 186:518–532.

[Verbeke and Lesaffre, 1996] Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, 91(433):217–221.

[Verbeke and Molenberghs, 2009] Verbeke, G. and Molenberghs, G. (2009). *Linear mixed models for longitudinal data*. Springer Science & Business Media.

[Villemagne et al., 2013] Villemagne, V. L., Burnham, S., Bourgeat, P., Brown, B., Ellis, K. A., Salvado, O., Szoeke, C., Macaulay, S. L., Martins, R., Maruff, P., Ames, D., Rowe, C. C., and Masters, C. L. (2013). Amyloid β deposition, neurodegeneration, and cognitive decline in sporadic alzheimer's disease: a prospective cohort study. *The Lancet Neurology*, 12(4):357 – 367.

[Wang et al., 2016] Wang, D., Khosla, A., Gargeya, R., Irshad, H., and Beck, A. H. (2016). Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*.

[Wee et al., 2012] Wee, C.-Y., Yap, P.-T., Zhang, D., Denny, K., Browndyke, J. N., Potter, G. G., Welsh-Bohmer, K. A., Wang, L., and Shen, D. (2012). Identification of mci individuals using structural and functional connectivity networks. *Neuroimage*, 59(3):2045–2056.

[Whitwell et al., 2007] Whitwell, J. L., Przybelski, S. A., Weigand, S. D., Knopman, D. S., Boeve, B. F., Petersen, R. C., and Jack, Jr, C. R. (2007). 3d maps from multiple mri illustrate changing atrophy patterns as subjects progress from mild cognitive impairment to alzheimer's disease. *Brain*, 130(7):1777–1786.

[Wilson et al., 2011] Wilson, R., Leurgans, S., Boyle, P., and Bennett, D. (2011). Cognitive decline in prodromal alzheimer disease and mild cognitive impairment. *Archives of Neurology*, 68(3):351–356.

[Woolrich et al., 2009] Woolrich, M. W., Jbabdi, S., Patenaude, B., Chappell, M., Makni, S., Behrens, T., Beckmann, C., Jenkinson, M., and Smith, S. M. (2009). Bayesian analysis of neuroimaging data in fsl. *Neuroimage*, 45(1):S173–S186.

[Xia et al., 2019] Xia, T., Chartsias, A., Tsaftaris, S. A., Initiative, A. D. N., et al. (2019). Consistent brain ageing synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 750–758. Springer.

[Yang et al., 2015] Yang, D., Zhang, S., Yan, Z., Tan, C., Li, K., and Metaxas, D. (2015). Automated anatomical landmark detection ondistal femur surface using convolutional neural network. In *2015 IEEE 12th international symposium on biomedical imaging (ISBI)*, pages 17–21. IEEE.

[Young et al., 2014] Young, A. L., Oxtoby, N. P., Daga, P., Cash, D. M., Fox, N. C., Ourselin, S., Schott, J. M., and Alexander, D. C. (2014). A data-driven model of biomarker changes in sporadic alzheimer's disease. *Brain*, 137(9):2564–2577.

[Young et al., 2015] Young, A. L., Oxtoby, N. P., Huang, J., Marinescu, R. V., Daga, P., Cash, D., Fox, N., Ourselin, S., Schott, J. M., Alexander, D. C., and the ADNI (2015). Multiple orderings of events in disease progression. In *Information Processing in Medical Imaging*, pages 711–722. Springer.

[Young and Johnson, 2015] Young, R. and Johnson, D. R. (2015). Handling missing values in longitudinal panel data with multiple imputation. *Journal of Marriage and Family*, 77(1):277–294.

[Yu et al., 2017] Yu, L., Zhang, W., Wang, J., and Yu, Y. (2017). Seqgan: Sequence generative adversarial nets with policy gradient. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[Zhang et al., 2016] Zhang, X., Mormino, E. C., Sun, N., Sperling, R. A., Sabuncu, M. R., Yeo, B. T. T., and (2016). Bayesian model reveals latent atrophy factors with dissociable cognitive trajectories in alzheimer's disease. *Proceedings of the National Academy of Sciences*, 113(42):E6535–E6544.

[Zhou et al., 2012] Zhou, J., Liu, J., Narayan, V. A., and Ye, J. (2012). Modeling Disease Progression via Fused Sparse Group Lasso. In *Proceedings of International Conference on Knowledge Discovery & Data Mining*, volume 2012, pages 1095–1103.

[Zhou et al., 2002] Zhou, Z.-H., Jiang, Y., Yang, Y.-B., and Chen, S.-F. (2002). Lung cancer cell identification based on artificial neural network ensembles. *Artificial Intelligence in Medicine*, 24(1):25–36.

Appendix 1

In the following, we consider a longitudinal dataset $\mathbf{y} = (t_{ij}, y_{ij})_{\substack{1\leq i\leq p \\ 1\leq j\leq k_i}}$, a N-dimensional Riemannian manifold $\mathbb{M}$ equiped with the metric g, a smooth curve $\gamma : I \subset \mathbb{R} \to \mathbb{M}$ (that is shown to be a geodesic), statistical parameters θ and random effects $\mathbf{z}$.

8.4 Preambule

In Chapter 1, we have introduced a generic mixed-effects model along with 4 different instantiations, namely the *parallel logistic shapes*, the *logistic shapes*, the *exponential decays* and the *straight lines*. In this appendix, we will exhibit for each of this instantiation :

- a short proof that the curve presented in 1 is a geodesic on the manifold induced by the corresponding metric,
- the reparametrization $\theta \leftarrow \zeta_1(\theta)$ and $\mathbf{z} \leftarrow \zeta_2(\mathbf{z})$ (see Chapter 2) of the parameters θ and the random variables $\mathbf{z}$,
- the likelihood associated to the instantiation,
- the sufficient statistics of the MCMC-SAEM algorithm (see Chapter 2),
- the parameter updates that derive from the Maximization step of the MCMC-SAEM algorithm.

8.4.1 Geodesic hypothesis

One of the central hypothesis of the generic model is to describe the group-average spatiotemporal trajectory as a geodesic on $\mathbb{M}$. We will first show that the smooth curve γ introduced in 1 is a geodesic. Given a system $\mathbf{x} = (x_1, ..., x_n)$ a system of coordinates around $\gamma(t)$, it is possible to decompose the curve $\gamma = (\gamma_1, \ldots, \gamma_n)$. We can show that γ is a geodesic if and only if it satisfies the following system of differential equations

$$\forall k \in \{1, \ldots, n\} \frac{d^2\gamma_k}{dt^2} + \sum_{1\leq i,j\leq n} \Gamma^k_{i,j}(\gamma(t)) \frac{d\gamma_i}{dt}\frac{d\gamma_j}{dt} = 0$$

where the Christoffel symboles $\Gamma^k_{i,j}$ are defined by

$$\Gamma^k_{i,j} = \frac{1}{2}\sum_{l=1}^{n} g^{k,l}\left(\frac{\partial g_{j,l}}{\partial x_i} + \frac{\partial g_{i,l}}{\partial x_j} - \frac{\partial g_{i,j}}{\partial x_l}\right)$$

Given two Riemannian manifolds $(\mathbb{M}_1, g^{\mathbb{M}_1})$ and $(\mathbb{M}_2, g^{\mathbb{M}_2})$, we first recall that the manifold $\mathbb{M} = \mathbb{M}_1 \times \mathbb{M}_2$ equipped with the product metric is a Riemannian manifold. Given γ_1 (resp. γ_2) a geodesic on $\mathbb{M}_1$ (resp. $\mathbb{M}_2$), the geodesics of $\mathbb{M}$ are of the form $t \to (\gamma_1(t), \gamma_2(t))$.

In the following, we consider that the N-dimensional Riemannian manifold $\mathbb{M}$ is a product of N one-dimensional manifolds $\mathbb{M}_1 \times \ldots \times \mathbb{M}_N$, the manifold $\mathbb{M}_i$ being equipped with the metric g^i. We remind that for $p \in \mathbb{M}_i$ and $(u, v) \in \mathbf{T}_p\mathbb{M}$ the tangent space of $\mathbb{M}$ at p, $g^i_p(u, v) = uf_i(p)v$ where $f_i : \mathbb{M} \to]0; +\infty[$. To prove that the curve $\gamma = (\gamma_1, \ldots, \gamma_N)$ is a geodesic of $\mathbb{M}$, one can show that $\forall i \in \{1, \ldots, N\}$, γ_i is a geodesic of $\mathbb{M}_i$.

Finally, given a metric g^i, γ_i is a geodesic of $\mathbb{M}_i$ if and only if

$$\ddot{\gamma}_i(t) + \frac{1}{2}\frac{f'(\gamma_i(t))}{f(\gamma_i(t))}(\dot{\gamma}_i(t))^2 = 0$$

as the metric is characterized only by the Christoffel symbol $\Gamma^1_{1,1}(p) = \frac{1}{2}\frac{f'(p)}{f(p)}$

8.4.2 Reparametrization and Likelihood

To ensure a better practical identifiability e.g. a better stability and robustness of the parameters θ estimated, the model instantiated in Chapter 1 that depends θ and $\mathbf{z}$ are reperametrized such that in fact we estimate $\theta \leftarrow \zeta_1(\theta)$ and we sample random variables $\mathbf{z} \leftarrow \zeta_2(\mathbf{z})$.

Then, the general writing of the model leads to write the likelihood as :

$$p(\mathbf{y};\theta) = \int p(\mathbf{y},\mathbf{z};\theta)\,\mathrm{d}z = \int p(\mathbf{y}|\mathbf{z};\theta)p(\mathbf{z};\theta)\,\mathrm{d}z \tag{8.1}$$

As we have shown that the random variables contain individual effects $\mathbf{z}_i$ but also population effects $\mathbf{z}_{\text{pop}}$ that arises from the *exponentialization* of the model (see Chapter 1), the regularization term $p(\mathbf{z};\theta)$ writes

$$p(\mathbf{z};\theta) = \sum_{\text{pop}} p(\mathbf{z}_{\text{pop}};\theta) + \sum_{i=1}^{p} p(\mathbf{z}_i;\theta) \tag{8.2}$$

where $p(\mathbf{z}_{\text{pop}};\theta)$ and $p(\mathbf{z}_i;\theta)$ are the priors of the random variables. On the other hand, the attachment term $p(\mathbf{y}|\mathbf{z};\theta)$ writes as a sum over the patients and the visits :

$$p(\mathbf{y}|\mathbf{z};\theta) = \sum_{i=1}^{p}\sum_{j=1}^{k_i} p(y_{ij}|\mathbf{z};\theta) \tag{8.3}$$

where $p(y_{ij}|\mathbf{z}_i;\theta) \sim \mathcal{N}(y|f(\theta_{\text{geom}},\mathbf{z}_i,t_{ij}),\sigma^2)$ (from Eq. 2.2 and $\epsilon_{ij} \sim \mathcal{N}(0,\sigma^2)$).

8.4.3 Sufficient Statistics and Parameter Updates

As an hypothesis to prove the convergence of the MCMC-SAEM algorithm, we remind that the log-likelihood writes $\log p(\mathbf{y},\mathbf{z};\theta) = \langle\Phi(\theta), S(\mathbf{y},\mathbf{z})\rangle - \log C(\theta)$ where $S(\mathbf{y},\mathbf{z})$ are the sufficient statistics. After the Approximation step that writes $\tilde{S}^{k+1} = \tilde{S}^k + \epsilon_k(S(\mathbf{y},\mathbf{z}) - \tilde{S}^k)$ at iteration k, the maximization step is computing $\theta^{k+1} = \text{argmax}_\theta\langle\Phi(\theta),\tilde{S}\rangle + \log C(\theta)$.

To this end, we write for each model the log-likelihood with the associated sufficient statistics. From it, we derive the update of the parameters in the Maximization step.

8.5 Parallel logistic shapes

8.5.1 Geodesic hypothesis

The proof of the parallel logistic curve being a geodesic is provided in [Schiratti et al., 2017]

8.5.2 Reparametrization and Log-likelihood

First, as we want $p_0 \in]0,1[$, we write $p_0 = \frac{1}{1+\exp(\tilde{g})}$ (i.e. $\tilde{g} = \ln(\frac{1}{p_0} - 1)$) with $\tilde{g} \in \mathbb{R}$. For the sake of clarity, we write $g = \exp(\tilde{g})$. Also, we consider that for each $k \in \{2, \dots, N\}$, $\tilde{\delta}_k = \frac{v_0 \delta_k}{p_0(1-p_0)}$. Finally, for all $i \in \{1, \dots, p\}$ $\tilde{\alpha}_i = \frac{\alpha_i v_0}{p_0(1-p_0)}$ and $\tilde{\tau}_i = t_0 + \tau_i$.

This leads to the rewrite the individual spatiotemporal trajectory in Eq. 1.10 as :

$$\eta_k^{\mathbf{w}_i}(\psi_i(t_{ij})) = \left(1 + g\exp\left(-w_{ik}\frac{\left(g\exp(-\tilde{\delta}_k)+1\right)^2}{g\exp(-\tilde{\delta}_k)} - \tilde{\delta}_k - \tilde{\alpha}_i(t_{ij} - \tilde{\tau}_i)\right)\right)^{-1} + \epsilon_{ijk} \quad (8.4)$$

We consider the following laws :

- $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$
- $\tilde{g} \sim \mathcal{N}(\overline{g}, \sigma_g^2)$
- $\tilde{\delta}_k \sim \mathcal{N}(\overline{\delta}_k, \sigma_\delta^2) \quad \forall k \in \{2, \dots, N\}$
- $\beta_k \sim \mathcal{N}(\overline{\beta}_k, \sigma_\beta^2) \quad \forall k \in \{1, \dots, (N-1)N_s\}$
- $\tau_i \sim \mathcal{N}(\overline{\tau}, \sigma_\tau^2) \quad \forall i \in \{1, \dots, p\}$
- $\tilde{\alpha}_i = \exp(\xi_i) \quad$ where $\quad \xi_i \sim \mathcal{N}(\overline{\xi}, \sigma_\xi^2) \quad \forall i \in \{1, \dots, p\}$
- $s_{ij} \sim \mathcal{N}(0,1) \quad \forall i \in \{1, \dots, p\} \quad \forall j \in \{1, \dots, N_s\}$

which leads to $\theta = (\sigma, \overline{g}, (\overline{\delta}_k)_{1\leq k\leq n}, (\overline{\beta}_k)_{1\leq k\leq (n-1)N_s}, \overline{\tau}, \sigma_\tau, \overline{\xi}, \sigma_\xi)$ and

$\mathbf{z} = (\tilde{g}, (\tilde{\delta}_k)_{2\leq k\leq n}, (\beta_k)_{1\leq k\leq (n-1)N_s}), (\tau_i)_{1\leq i\leq p}, (\xi_i)_{1\leq i\leq p}, (s_{ij})_{1\leq i\leq p, 1\leq j\leq k_i})$. $(\sigma_g, \sigma_\delta, \sigma_\beta)$ are fixed.

The log-likelihood then writes

$$
\begin{aligned}
\log p(\mathbf{y},\mathbf{z};\theta) = &- NK\ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{p}\sum_{j=1}^{k_i}\|y_{ij} - \eta^{\mathbf{w}_i}_{\gamma_0,\mathbf{p}_0,t_0}(\psi_i(t_{ij}))\|^2 \\
&- \ln(\sigma_g\sqrt{2\pi}) - \frac{1}{2\sigma_g^2}(\tilde{g}-\overline{g})^2 \\
&- (N-1)\ln(\sigma_\delta\sqrt{2\pi}) - \frac{1}{2\sigma_\delta^2}\sum_{k=2}^{N}(\tilde{\delta}_k - \overline{\delta}_k)^2 \\
&- (N-1)N_s\ln(\sigma_\beta\sqrt{2\pi}) - \frac{1}{2\sigma_g^2}\sum_{k=1}^{(N-1)N_s}(\beta_k - \overline{\beta}_k)^2 \\
&- p\log(\sigma_\xi\sqrt{2\pi}) - \frac{1}{2\sigma_\xi^2}\sum_{i=1}^{p}(\xi_i - \overline{\xi})^2 \\
&- p\log(\sigma_\tau\sqrt{2\pi}) - \frac{1}{2\sigma_\tau^2}\sum_{i=1}^{p}(\tau_i - \overline{\tau})^2 \\
&- pN_s\log(2\sqrt{2\pi}) - \frac{1}{2\sigma_s^2}\sum_{i=0}^{p}\sum_{j=0}^{N_s}s_{ij}^2
\end{aligned}
$$

8.5.3 Sufficient Statistics and Parameter Updates

The decomposition of the log-likelihood into the sufficient statistics is as follow :

$$
\begin{aligned}
\log p(\mathbf{y},\mathbf{z};\theta) = &- \langle \underbrace{[\|y_{ij}\|^2]_{ij}}_{S_1(\mathbf{y},\mathbf{z})}, \underbrace{-2[y_{ij}^T\eta_\gamma^{\mathbf{w}_i}]_{ij}}_{S_2(\mathbf{y},\mathbf{z})} + \underbrace{[\|\eta_\gamma^{\mathbf{w}_i}\|^2]_{ij}}_{S_3(\mathbf{y},\mathbf{z})}, \frac{1}{2\sigma^2}\mathbf{1}_{\sum k_i}\rangle - NK\ln(\sigma\sqrt{2\pi}) \\
&+ \langle \underbrace{\tilde{g}^2}_{S_4(\mathbf{y},\mathbf{z})}, -\frac{1}{2\sigma_g^2}\rangle + \langle \underbrace{\tilde{g}}_{S_5(\mathbf{y},\mathbf{z})}, \frac{1}{\sigma_g^2}\overline{g}\rangle - \frac{1}{2\sigma_g^2}\overline{g}^2 - \ln(\sigma_g\sqrt{2\pi}) \\
&+ \langle \underbrace{[\tilde{\delta}_k^2]_k}_{S_6(\mathbf{y},\mathbf{z})}, -\frac{1}{2\sigma_\delta^2}\mathbf{1}_{N-1}\rangle + \langle \underbrace{[\tilde{\delta}_k]_k}_{S_7(\mathbf{y},\mathbf{z})}, \frac{1}{\sigma_\delta^2}[\overline{\delta}_k]_k\rangle - \sum_{k=2}^{N}\frac{1}{2\sigma_\delta^2}\overline{\delta}_k^2 - (N-1)\ln(\sigma_\delta\sqrt{2\pi}) \\
&+ \langle \underbrace{[\beta_k^2]_k}_{S_8(\mathbf{y},\mathbf{z})}, -\frac{1}{2\sigma_\beta^2}\mathbf{1}_{(N-1)N_s}\rangle + \langle \underbrace{[\beta_k]_k}_{S_9(\mathbf{y},\mathbf{z})}, \frac{1}{\sigma_\beta^2}[\overline{\beta}_k]_k\rangle - \sum_{k=1}^{(N-1)N_s}\frac{1}{2\sigma_\beta^2}\overline{\beta}_k^2 - (N-1)N_s\ln(\sigma_\beta\sqrt{2\pi}) \\
&+ \langle \underbrace{[\xi_i^2]_i}_{S_{10}(\mathbf{y},\mathbf{z})}, -\frac{1}{2\sigma_\xi^2}\mathbf{1}_p\rangle + \langle \underbrace{[\xi_i]_i}_{S_{11}(\mathbf{y},\mathbf{z})}, \frac{1}{\sigma_\xi^2}\overline{\xi}\mathbf{1}_p\rangle - \frac{1}{2\sigma_\xi^2}p\overline{\xi}^2 - p\log(\sigma_\xi\sqrt{2\pi}) \\
&+ \langle \underbrace{[\tau_i^2]_i}_{S_{12}(\mathbf{y},\mathbf{z})}, -\frac{1}{2\sigma_\tau^2}\mathbf{1}_p\rangle + \langle \underbrace{[\tau_i]_i}_{S_{13}(\mathbf{y},\mathbf{z})}, \frac{1}{\sigma_\tau^2}\overline{\tau}\mathbf{1}_p\rangle - \frac{1}{2\sigma_\tau^2}p\overline{\tau}^2 - p\log(\sigma_\tau\sqrt{2\pi}) \\
&+ \langle \underbrace{[\tilde{s_{il}}^2]_{il}}_{S_{14}(\mathbf{y},\mathbf{z})}, -\frac{1}{2\sigma_s^2}\mathbf{1}_{pN_s}\rangle + \sum_{k=1}^{N_s}\langle \underbrace{[\tilde{s_{ik}}]_i}_{S_{15}(\mathbf{y},\mathbf{z})}, \frac{1}{\sigma_s^2}[\overline{s_k}]_k\rangle - p\sum_{k=1}^{N_s}\frac{1}{2\sigma_s^2}\overline{s_k} - pN_s\log(\sqrt{2\pi})
\end{aligned}
$$

At iteration k, the optimization procedure at the maximization step gives us the following updates:

$$
\begin{aligned}
(\sigma^2)^{k+1} &\leftarrow \frac{1}{NK}[\tilde{S}_1^{(k+1)} - 2\tilde{S}_2^{(k+1)} + \tilde{S}_3^{(k+1)}]^T \mathbf{1}_K \\
(\overline{g})^{k+1} &\leftarrow \tilde{S}_5^{(k+1)} \\
(\overline{\delta}_j)^{k+1} &\leftarrow \tilde{S}_7^{(k+1)} \\
(\overline{\beta}_j)^{k+1} &\leftarrow \tilde{S}_9^{(k+1)} \\
(\overline{\xi})^{k+1} &\leftarrow \frac{1}{p}\tilde{S}_{11}^{(k+1)} \\
(\sigma_\xi^2)^{k+1} &\leftarrow \frac{1}{p}[\tilde{S}_{10}^{(k+1)} - 2\overline{\xi}\tilde{S}_{11}^{(k+1)}]^T \mathbf{1}_p + \overline{\xi}^2 \\
(\overline{\tau})^{k+1} &\leftarrow \frac{1}{p}\tilde{S}_{13}^{(k+1)} \\
(\sigma_\tau^2)^{k+1} &\leftarrow \frac{1}{p}[\tilde{S}_{12}^{(k+1)} - 2\overline{\tau}\tilde{S}_{13}^{(k+1)}]^T \mathbf{1}_p + \overline{\tau}^2
\end{aligned}
$$

8.6 Logistic shapes

8.6.1 Geodesic hypothesis

The proof given in [Schiratti et al., 2017] for the parallel logistic holds true for the non-parallel instantiation of the logistic curves.

8.6.2 Reparametrization and Log-likelihood

In order to keep $p_k \in]0,1[$ $(\forall k \in \{1,\dots,N\})$, we consider $p_k = \frac{1}{1+g_k}$ where $g_k = \exp(\tilde{g}_k)$ with $\tilde{g}_k \in \mathbb{R}$. Furthermore, let's have $\tilde{\tau}_i = t_0 + \tau_i$ for all $i \in \{1,\dots,p\}$.

The equation 1.13 rewrites

$$\eta_k^{\mathbf{w}_i}(\psi_i(t_{ij})) = \left(1 + g_k \exp\left(-\frac{(1+g_k)^2}{g_k^2}\left(w_{ik} + v_k \alpha_i (t_{ij} - \tilde{\tau}_i)\right)\right)\right)^2 \tag{8.5}$$

To ensure identifiability conditions, we stress the fact that we cannot have normal laws on α_i and v_k where both mean and scales are learnt.

We have the following laws

- $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$
- $v_k \sim \mathcal{N}(\overline{v}_k, \sigma_v^2) \quad \forall k\{1,\dots,N\}$
- $\tilde{g}_k \sim \mathcal{N}(\overline{g}_k, \sigma_g^2) \quad \forall k\{1,\dots,N\}$
- $\beta_k \sim \mathcal{N}(\overline{\beta}_k, \sigma_\beta^2) \quad \forall k \in \{1,\dots,(N-1)N_s\}$
- $\tau_i \sim \mathcal{N}(\overline{\tau}, \sigma_\tau^2) \quad \forall i \in \{1,\dots,p\}$
- $\alpha_i = \exp(\xi_i) \quad$ where $\xi_i \sim \mathcal{N}(0, \sigma_\xi^2) \quad \forall i \in \{1,\dots,p\}$
- $s_{ij} \sim \mathcal{N}(0,1) \quad \forall i \in \{1,\dots,p\} \quad \forall j \in \{1,\dots,N_s\}$

We thus have $\theta = (\sigma, (\overline{g}_k)_{1\le k\le N}, (\overline{v}_k)_{1\le k\le N}, (\overline{\beta}_k)_{1\le k\le (N-1)N_s}, \overline{\tau}, \sigma_\tau, \sigma_\xi)$
and $\mathbf{z} = ((g_k)_{1\le k\le N}, (\tilde{v}_k)_{1\le k\le N}, (\beta_k)_{1\le k\le (N-1)N_s}, (\tau_i)_{1\le i\le p}, (\xi_i)_{1\le i\le p}, (s_{ij})_{1\le i\le p, 1\le j\le N_s},)$.
$(\sigma_g, \sigma_v, \sigma_\beta)$ are fixed
The log likelihood writes

$$
\begin{aligned}
\log p(\mathbf{y},\mathbf{z};\theta) = &- NK\ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{p}\sum_{j=1}^{k_i}\|y_{ij} - \eta^{\mathbf{w}_i}_{\gamma_0,\mathbf{p}_0,t_0}(\psi_i(t_{ij}))\|^2 \\
&- N\ln(\sigma_g\sqrt{2\pi}) - \frac{1}{2\sigma_g^2}\sum_{k=1}^{N}(\tilde{g}_k - \overline{g}_k)^2 \\
&- N\ln(\sigma_v\sqrt{2\pi}) - \frac{1}{2\sigma_v^2}\sum_{k=1}^{N}(\tilde{v}_k - \overline{v}_k)^2 \\
&- (N-1)N_s\ln(\sigma_\beta\sqrt{2\pi}) - \frac{1}{2\sigma_g^2}\sum_{k=1}^{(N-1)N_s}(\beta_k - \overline{\beta}_k)^2 \\
&- p\log(\sigma_\xi\sqrt{2\pi}) - \frac{1}{2\sigma_\xi^2}\sum_{i=1}^{p}\xi_i^2 \\
&- p\log(\sigma_\tau\sqrt{2\pi}) - \frac{1}{2\sigma_\tau^2}\sum_{i=1}^{p}(\tau_i - \overline{\tau})^2 \\
&- pN_s\log(2\sqrt{2\pi}) - \frac{1}{2\sigma_s^2}\sum_{i=0}^{p}\sum_{j=0}^{N_s}s_{ij}^2
\end{aligned}
$$

8.6.3 Sufficient Statistics and Parameters Update

The sufficient statistics are

$$
\begin{aligned}
\log p(\mathbf{y},\mathbf{y};\theta) = &- \langle \underbrace{[\|y_{ij}\|^2]_{ij}}_{S_1(\mathbf{y},\mathbf{z})}\underbrace{-2[y_{ij}^T\eta_\gamma^{\mathbf{w}_i}]_{ij}}_{S_2(\mathbf{y},\mathbf{z})} + \underbrace{[\|\eta_\gamma^{\mathbf{w}_i}\|^2]_{ij}}_{S_3(\mathbf{y},\mathbf{z})}, \frac{1}{2\sigma^2}\mathbf{1}_{\sum k_i}\rangle - NK\ln(\sigma\sqrt{2\pi}) \\
&+ \langle \underbrace{[\tilde{v}_k^2]_k}_{S_4(\mathbf{y},\mathbf{z})}, -\frac{1}{2\sigma_v^2}\mathbf{1}_N\rangle + \langle \underbrace{[\tilde{v}_k]_k}_{S_5(\mathbf{y},\mathbf{z})}, \frac{1}{\sigma_v^2}[\overline{v}_k]_k\rangle - \sum_{k=1}^{N}\frac{1}{2\sigma_v^2}\overline{v}_k^2 - N\ln(\sigma_v\sqrt{2\pi}) \\
&+ \langle \underbrace{[\tilde{g}_k^2]_k}_{S_6(\mathbf{y},\mathbf{z})}, -\frac{1}{2\sigma_g^2}\mathbf{1}_N\rangle + \langle \underbrace{[\tilde{g}_k]_k}_{S_7(\mathbf{y},\mathbf{z})}, \frac{1}{\sigma_g^2}[\overline{g}_k]_k\rangle - \sum_{k=1}^{N}\frac{1}{2\sigma_g^2}\overline{g}_k^2 - N\ln(\sigma_g\sqrt{2\pi}) \\
&+ \langle \underbrace{[\beta_k^2]_k}_{S_8(\mathbf{y},\mathbf{z})}, -\frac{1}{2\sigma_\beta^2}\mathbf{1}_{(N-1)N_s}\rangle + \langle \underbrace{[\beta_k]_k}_{S_9(\mathbf{y},\mathbf{z})}, \frac{1}{\sigma_\beta^2}[\overline{\beta}_k]_k\rangle - \sum_{k=1}^{(N-1)N_s}\frac{1}{2\sigma_\beta^2}\overline{\beta}_k^2 - (N-1)N_s\ln(\sigma_\beta\sqrt{2\pi}) \\
&+ \langle \underbrace{[\xi_i^2]_i}_{S_{10}(\mathbf{y},\mathbf{z})}, -\frac{1}{2\sigma_\xi^2}\mathbf{1}_p\rangle - p\log(\sigma_\xi\sqrt{2\pi}) \\
&+ \langle \underbrace{[\tau_i^2]_i}_{S_{11}(\mathbf{y},\mathbf{z})}, -\frac{1}{2\sigma_\tau^2}\mathbf{1}_p\rangle + \langle \underbrace{[\tau_i]_i}_{S_{12}(\mathbf{y},\mathbf{z})}, \frac{1}{\sigma_\tau^2}\overline{\tau}\mathbf{1}_p\rangle - \frac{1}{2\sigma_\tau^2}p\overline{\tau}^2 - p\log(\sigma_\tau\sqrt{2\pi}) \\
&+ \langle \underbrace{[\tilde{s_{il}}^2]_{il}}_{S_{13}(\mathbf{y},\mathbf{z})}, -\frac{1}{2\sigma_s^2}\mathbf{1}_{pN_s}\rangle + \sum_{k=1}^{N_s}\langle \underbrace{[\tilde{s_{ik}}]_i}_{S_{14}(\mathbf{y},\mathbf{z})}, \frac{1}{\sigma_s^2}[\overline{s_k}]_k\rangle - p\sum_{k=1}^{N_s}\frac{1}{2\sigma_s^2}\overline{s_k} - pN_s\log(\sqrt{2\pi})
\end{aligned}
$$

Which leads to the following updates

$$
\begin{aligned}
(\sigma^2)^{k+1} &\leftarrow \frac{1}{NK}[\tilde{S}_1^{(k+1)} - 2\tilde{S}_2^{(k+1)} + \tilde{S}_3^{(k+1)}]^T \mathbf{1}_K \\
(\overline{v}_j)^{k+1} &\leftarrow \tilde{S}_5^{(k+1)} \\
(\overline{g}_j)^{k+1} &\leftarrow \tilde{S}_7^{(k+1)} \\
(\overline{\beta}_j)^{k+1} &\leftarrow \tilde{S}_9^{(k+1)} \\
(\sigma_\xi^2)^{k+1} &\leftarrow \frac{1}{p}[\tilde{S}_{10}^{(k+1)}]^T \mathbf{1}_p \\
(\overline{\tau})^{k+1} &\leftarrow \frac{1}{p}\tilde{S}_{12}^{(k+1)} \\
(\sigma_\tau^2)^{k+1} &\leftarrow \frac{1}{p}[\tilde{S}_{11}^{(k+1)} - 2\overline{\tau}\tilde{S}_{12}^{(k+1)}]^T \mathbf{1}_p + \overline{\tau}^2
\end{aligned}
$$

8.7 Exponential decays

8.7.1 Geodesic hypothesis

We consider the one dimensional metric $g_p(u,v) = \frac{uv}{p^2}$ (i.e. $f(p) = \frac{1}{p^2}$ and $f'(p) = -\frac{2}{p^3}$) and $\gamma(t) = p\exp(\frac{v}{p}(t-t_0))$ ($\gamma(t_0) = p$ and $\dot{\gamma}(t_0) = v$). This leads to

$$\dot{\gamma}(t) = v\exp(\frac{v}{p}(t-t_0)) = \frac{v}{p}\gamma(t)$$

and

$$\ddot{\gamma}(t) = \frac{v^2}{p}\exp(\frac{v}{p}(t-t_0)) = \frac{v^2}{p^2}\gamma(t)$$

The differential equation is satisfied as

$$\begin{aligned} E &= \ddot{\gamma}(t) + \frac{1}{2}\frac{f'(\gamma(t))}{f(\gamma(t))}(\dot{\gamma}(t))^2 \\ &= \frac{v^2}{p^2}\gamma(t) + \frac{1}{2}\Big(-\frac{2}{(\gamma(t))^3}\Big)\Big(\gamma(t)\Big)^2(\frac{v^2}{p^2}(\gamma(t))^2) \\ &= 0 \end{aligned}$$

8.7.2 Reparametrization and Log-likelihood

We here consider $\tilde{v}_k = \frac{v_k}{p_k}$ and $\tilde{\tau}_i = t_0 + \tau_i$. The equation 1.20 rewrites

$$\eta_k^{\mathbf{w}_i}(\psi_i(t_{ij})) = p_k \exp\left(\frac{w_{ik}}{p_k} + \alpha_i\tilde{v}_k(t_{ij} - \tilde{\tau}_i)\right) + \epsilon_{ijk} \tag{8.6}$$

To ensure identifiability conditions, we stress the fact that we cannot have normal laws on α_i and v_k where both mean and scales are learnt.

We have the following laws

- $\epsilon_{ijk} \sim \mathcal{N}(0,\sigma^2)$
- $p_k \sim \mathcal{N}(\overline{p}_k, \sigma_p^2) \quad \forall k\{1,\dots,N\}$
- $\tilde{v}_k \sim \mathcal{N}(\overline{v}_k, \sigma_p^2) \quad \forall k\{1,\dots,N\}$
- $\beta_k \sim \mathcal{N}(\overline{\beta}_k, \sigma_\beta^2) \quad \forall k \in \{1,\dots,(N-1)N_s\}$
- $\tau_i \sim \mathcal{N}(\overline{\tau}, \sigma_\tau^2) \quad \forall i \in \{1,\dots,p\}$
- $\alpha_i = \exp(\xi_i)$ where $\xi_i \sim \mathcal{N}(0,\sigma_\xi^2) \quad \forall i \in \{1,\dots,p\}$
- $s_{ij} \sim \mathcal{N}(0,1) \quad \forall i \in \{1,\dots,p\} \quad \forall j \in \{1,\dots,N_s\}$

We note that it is possible to parametrize $p = \exp(p)$ and or $v = \exp(v)$ to have positive of negative values depending on the biological process at hand.

We thus have $\theta = (\sigma, (\overline{p}_k)_{1\le k\le N}, (\overline{v}_k)_{1\le k\le N}, (\beta_k)_{1\le k\le (N-1)N_s}, \overline{\tau}, \sigma_\tau, \sigma_\xi)$
and $\mathbf{z} = ((p_k)_{1\le k\le N}, (\tilde{v}_k)_{1\le k\le N}, (\overline{\beta}_k)_{1\le k\le (n-1)N_s}), (\tau_i)_{1\le i\le p}, (\xi_i)_{1\le i\le p}, (s_{ij})_{1\le i\le p, 1\le j\le k_i})$.
$(\sigma_p, \sigma_v, \sigma_\beta)$ are fixed.
The log likelihood writes

$$
\begin{aligned}
\log p(\mathbf{y},\mathbf{z};\theta) = &- NK\ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{p}\sum_{j=1}^{k_i}\|y_{ij} - \eta_{\gamma_0,\mathbf{p}_0,t_0}^{\mathbf{w}_i}(\psi_i(t_{ij}))\|^2 \\
&- N\ln(\sigma_v\sqrt{2\pi}) - \frac{1}{2\sigma_v^2}\sum_{k=1}^{N}(\tilde{v}_k - \overline{v}_k)^2 \\
&- N\ln(\sigma_p\sqrt{2\pi}) - \frac{1}{2\sigma_p^2}\sum_{k=1}^{N}(p_k - \overline{p}_k)^2 \\
&- (N-1)N_s\ln(\sigma_\beta\sqrt{2\pi}) - \frac{1}{2\sigma_g^2}\sum_{k=1}^{(N-1)N_s}(\beta_k - \overline{\beta}_k)^2 \\
&- p\log(\sigma_\xi\sqrt{2\pi}) - \frac{1}{2\sigma_\xi^2}\sum_{i=1}^{p}\xi_i^2 \\
&- p\log(\sigma_\tau\sqrt{2\pi}) - \frac{1}{2\sigma_\tau^2}\sum_{i=1}^{p}(\tau_i - \overline{\tau})^2 \\
&- pN_s\log(2\sqrt{2\pi}) - \frac{1}{2\sigma_s^2}\sum_{i=0}^{p}\sum_{j=0}^{N_s}s_{ij}^2
\end{aligned}
$$

8.7.3 Sufficient Statistics and Parameters Update

$$
\begin{aligned}
\log q(\mathbf{y},\mathbf{z};\theta) = &- \langle \underbrace{[\|y_{ij}\|^2]_{ij}}_{S_1(\mathbf{y},\mathbf{z})}\underbrace{-2[y_{ij}^T\eta_\gamma^{\mathbf{w}_i}]_{ij}}_{S_2(\mathbf{y},\mathbf{z})} + \underbrace{[\|\eta_\gamma^{\mathbf{w}_i}\|^2]_{ij}}_{S_3(\mathbf{y},\mathbf{z})}, \frac{1}{2\sigma^2}\mathbf{1}_{\sum k_i}\rangle - NK\ln(\sigma\sqrt{2\pi}) \\
&+ \langle \underbrace{[\tilde{v}_k^2]_k}_{S_4(\mathbf{y},\mathbf{z})}, -\frac{1}{2\sigma_v^2}\mathbf{1}_N\rangle + \langle \underbrace{[\tilde{v}_k]_k}_{S_5(\mathbf{y},\mathbf{z})}, \frac{1}{\sigma_v^2}[\overline{v}_k]_k\rangle - \sum_{k=1}^{N}\frac{1}{2\sigma_v^2}\overline{v}_k^2 - N\ln(\sigma_v\sqrt{2\pi}) \\
&+ \langle \underbrace{[p_k^2]_k}_{S_6(\mathbf{y},\mathbf{z})}, -\frac{1}{2\sigma_p^2}\mathbf{1}_N\rangle + \langle \underbrace{[p_k]_k}_{S_7(\mathbf{y},\mathbf{z})}, \frac{1}{\sigma_p^2}[\overline{p}_k]_k\rangle - \sum_{k=1}^{N}\frac{1}{2\sigma_p^2}\overline{p}_k^2 - N\ln(\sigma_p\sqrt{2\pi}) \\
&+ \langle \underbrace{[\beta_k^2]_k}_{S_8(\mathbf{y},\mathbf{z})}, -\frac{1}{2\sigma_\beta^2}\mathbf{1}_{(N-1)N_s}\rangle + \langle \underbrace{[\beta_k]_k}_{S_9(\mathbf{y},\mathbf{z})}, \frac{1}{\sigma_\beta^2}[\overline{\beta}_k]_k\rangle - \sum_{k=1}^{(N-1)N_s}\frac{1}{2\sigma_\beta^2}\overline{\beta}_k^2 - (N-1)N_s\ln(\sigma_\beta\sqrt{2\pi}) \\
&+ \langle \underbrace{[\xi_i^2]_i}_{S_{10}(\mathbf{y},\mathbf{z})}, -\frac{1}{2\sigma_\xi^2}\mathbf{1}_p\rangle - p\log(\sigma_\xi\sqrt{2\pi}) \\
&+ \langle \underbrace{[\tau_i^2]_i}_{S_{11}(\mathbf{y},\mathbf{z})}, -\frac{1}{2\sigma_\tau^2}\mathbf{1}_p\rangle + \langle \underbrace{[\tau_i]_i}_{S_{12}(\mathbf{y},\mathbf{z})}, \frac{1}{\sigma_\tau^2}\overline{\tau}\mathbf{1}_p\rangle - \frac{1}{2\sigma_\tau^2}p\overline{\tau}^2 - p\log(\sigma_\tau\sqrt{2\pi}) \\
&+ \langle \underbrace{[\tilde{s_{il}}^2]_{il}}_{S_{13}(\mathbf{y},\mathbf{z})}, -\frac{1}{2\sigma_s^2}\mathbf{1}_{pN_s}\rangle + \sum_{k=1}^{N_s}\langle \underbrace{[\tilde{s_{ik}}]_i}_{S_{14}(\mathbf{y},\mathbf{z})}, \frac{1}{\sigma_s^2}[\overline{s_k}]_k\rangle - p\sum_{k=1}^{N_s}\frac{1}{2\sigma_s^2}\overline{s_k} - pN_s\log(\sqrt{2\pi})
\end{aligned}
$$

Which leads to the following updates

$$
\begin{aligned}
(\sigma^2)^{k+1} &\leftarrow \frac{1}{NK}[\tilde{S}_1^{(k+1)} - 2\tilde{S}_2^{(k+1)} + \tilde{S}_3^{(k+1)}]^T \mathbf{1}_K \\
(\overline{v}_j)^{k+1} &\leftarrow \tilde{S}_5^{(k+1)} \\
(\overline{p}_j)^{k+1} &\leftarrow \tilde{S}_7^{(k+1)} \\
(\overline{\beta}_j)^{k+1} &\leftarrow \tilde{S}_9^{(k+1)} \\
(\sigma_\xi^2)^{k+1} &\leftarrow \frac{1}{p}[\tilde{S}_{10}^{(k+1)}]^T \mathbf{1}_p \\
(\overline{\tau})^{k+1} &\leftarrow \frac{1}{p}\tilde{S}_{12}^{(k+1)} \\
(\sigma_\tau^2)^{k+1} &\leftarrow \frac{1}{p}[\tilde{S}_{11}^{(k+1)} - 2\overline{\tau}\tilde{S}_{12}^{(k+1)}]^T \mathbf{1}_p + \overline{\tau}^2
\end{aligned}
$$

Titre: Apprentissage de Modèles Multimodaux Numériques de la progression des Maladies à partir de Données Longitudinales: Méthodes & Algorithmes pour la Description, la Prédiction et la Simulation de la Progression de la Maladie d'Alzheimer.

Mots clés: Apprentissage statistique - Trajectoires spatiotemporelles - Données longitudinales - Maladies neurodégénératives

Résumé: La thése s'intéresse à l'apprentissage statistique de modèles digitaux de progression des maladies neurodégénératives, en particulier la maladie d'Alzheimer. Ces modéles ont pour but de reconstruire la dynamique complexe et hétérogène de l'évolution de la structure, des fonctions et des facultés cognitives du cerveau, à un niveau moyenne mais également à l'échelle individuelle. Pour répondre à cet objectif, la thèse considére un modéle génératif à effets mixtes qui, à partir de données longitudinales, c'est à dire des observations répétées pour chaque patient, et éventuellement multimodales, recombine les trajectoires spatiotemporelles individuelles en un scénario moyen de progression de la maladie, estimant conjointement la variabilité de cette progression caractéristique. Cette variabilité est le resultat du non alignement temporel (en terme de vitesse de progression et âge de début de la maladie) et d'une variabilité spatiale qui prend la forme d'une modification de la séquence d'événements qui interviennent durant l'apparition et la progression de la maladie. Les différentes parties de la thèse forme une suite logique, depuis la problématique médicale, en passant par la description du modèle statistique associé, l'application de celui-ci pour la description de l'évolution de la maladie d'Alzheimer, et, enfin, le développement d'outils numériques à destination du corps médical pour tirer pleinement parti des méthodes présentées.

Title: Learning Multimodal Digital Models of Disease Progression from Longitudinal Data: Methods & Algorithms for the Description, Prediction and Simulation of Alzheimer's Disease Progression

Keywords: Statistical Learning - Machine Learning - Spatiotemporal trajectories - Longitudinal data - Neurodegenerative diseases

Abstract: This thesis focuses on the statistical learning of digital models of neurodegenerative disease progression, especially Alzheimer's disease. It aims at reconstructing the complex and heterogeneous dynamic of evolution of the structure, the functions and the cognitive abilities of the brain, at both an average and individual level. To do so, we consider a mixed-effects model that, based on longitudinal data, namely repeated observations per subjects that present multiple modalities, in parallel recombines the individual spatiotemporal trajectories into a group-average scenario of change, and, estimates the variability of this characteristic progression which characterizes the individual trajectories. This variability results from a temporal un-alignment (in term of pace of progression and age at disease onset) along with a spatial variability that takes the form of a modification in the sequence of events that appear during the course of the disease. The different parts of the thesis are ordered in a coherent sequence: from the medical problematic, followed by the statistical model introduced to tackle the aforementioned challenge and its application to the description of the course of Alzheimer's disease, and, finally, numerical tools developed to make the previous model available to the medical community.

Institut Polytechnique de Paris
91120 Palaiseau, France