Modeling individual disease trajectory: application to the prediction of treatment effect

Pierre-Emmanuel Poulet

Modeling individual disease trajectory: application to the prediction of treatment effect

Author:

Pierre-Emmanuel POULET

Aramis lab, INRIA Paris, Institut du Cerveau, Inserm, CNRS, Sorbonne Université

Supervisor:

Stanley DURRLEMAN

Directeur de recherche INRIA

PhD Thesis for the degree of Doctor in Applied Mathematics at EDITE Paris (ED 130)

Date of defense: 2nd of October 2023

Modeling individual disease trajectory: application to the prediction of treatment effect

Author:
Pierre-Emmanuel POULET (Aramis lab, INRIA Paris, Institut du Cerveau, Inserm, CNRS, Sorbonne Université)

Supervisors:
Stanley DURRLEMAN (Directeur de recherche INRIA)

Abstract

This thesis aims at developing statistical models to describe the progression of neurodegenerative diseases from longitudinal data, i.e. repeated measurements in time for every subject. Our work builds upon the Disease Course Mapping, a spatio-temporal disease progression model. This mixed-effect model has a hierarchical structure allowing to account for the global dynamics at a population level while also accounting for individualized disease trajectories. The variability between subjects is decomposed between two effects: a temporal reparametrization to construct a common disease timeline, and a spatial effect to describe the heterogeneity in how the disease presents itself. However it suffers from several limitations in addressing the challenges of disease modelling. First this model is embedded in a Riemannian manifold, where the family of possible trajectories is restricted by the metric. Secondly it only considers continuous manifold-valued observations, while neurodegenerative cohorts are filled with discontinuous data. Thirdly it lacks a structure to identify and characterize disease subtypes. We will address all those needs.

In the first part, after introducing the Disease Course Mapping model, we propose a model that loosens the choice of possible trajectories by learning the metric of the Riemannian manifold.

In the second part we tackle discrete data and extend the model to binary and ordinal markers of disease progression. This allows us to predict symptoms and model cognitive scales. We apply it to Parkinson's disease and derive clinical interpretation from the results.

The third part is dedicated to heterogeneity. We extend the progression model to account for disease subtypes by adding a mixture layer on top of the statistical model. The mixture model is able to jointly uncover latent classes while learning the typical progression pattern of each class. We showcase the model's ability to extract distinct trajectories of disease evolution in the Alzheimer's disease case.

The fourth part focuses on the application of disease progression to treatment effect models. We define a framework to learn a therapeutic effect using only treated patients by leveraging a disease model's predictions, which act as a synthetic control arm. Application to the dopaminergic treatment in Parkinson's disease shows great promise. We then design a new model to account for disease-modifying treatment effects, i.e. long-term modifications of the disease progression trajectory.

Overall this thesis provides several tools to extend the Disease Course Mapping potential applications, easing its transfer towards having meaningful clinical impact. The methods developed here are aimed at improving our understanding of neurodegenerative diseases, by identifying subtypes, including markers of various nature and taking into account the effect of treatment over these markers.

Résumé

Cette thèse vise à développer des modèles statistiques pour décrire la progression des maladies neurodégénératives à partir de données longitudinales, c'est-à-dire des mesures répétées dans le temps pour chaque sujet. Notre travail s'appuie sur le modèle Disease Course Mapping, un modèle de progression de la maladie spatio-temporel. Ce modèle à effets mixtes a une structure hiérarchique permettant de rendre compte de la dynamique globale au niveau de la population tout en tenant compte des trajectoires individuelles de la maladie. La variabilité entre les sujets est décomposée en deux effets : une reparamétrisation temporelle pour construire une chronologie commune de la maladie, et un effet spatial pour décrire l'hétérogénéité de la manifestation de la maladie. Cependant, il présente plusieurs limitations pour relever les défis de la modélisation de telles maladies. Tout d'abord, ce modèle est défini dans une variété riemannienne, où la famille de trajectoires possibles est restreinte par la métrique. Deuxièmement, il ne considère que des observations continues à valeurs dans la variété, alors que les cohortes neurodégénératives sont remplies de données discontinues. Troisièmement, il manque une structure pour identifier et caractériser les sous-types de maladies. Nous répondrons à tous ces besoins.

Dans la première partie, après avoir présenté le modèle Disease Course Mapping, nous proposons un modèle qui assouplit le choix des trajectoires possibles en apprenant la métrique de la variété riemannienne.

Dans la deuxième partie, nous abordons les données discrètes et étendons le modèle aux marqueurs binaires et ordinaux de progression de la maladie. Cela nous permet de prédire les symptômes et de modéliser les échelles cognitives. Nous l'appliquons à la maladie de Parkinson et tirons une interprétation clinique des résultats.

La troisième partie est consacrée à l'hétérogénéité. Nous étendons le modèle de progression pour tenir compte des sous-types de maladies en ajoutant une couche de mélange au-dessus du modèle statistique. Le modèle de mélange permet d'identifier conjointement les classes latentes tout en apprenant le schéma de progression typique de chaque classe. Nous présentons la capacité du modèle à extraire des trajectoires distinctes de l'évolution de la maladie dans le cas de la maladie d'Alzheimer.

La quatrième partie se concentre sur l'application aux modèles d'effet de traitement. Nous définissons un cadre pour apprendre l'effet thérapeutique avec uniquement des patients traités en utilisant les prédictions d'un modèle de progression de la maladie, qui agissent comme un bras de contrôle synthétique. L'application au traitement dopaminergique dans la maladie de Parkinson illustre le potentiel de la méthode. Nous concevons ensuite un nouveau modèle pour tenir compte des effets de traitement modificateurs de la maladie, c'est-à-dire des modifications à long terme de la trajectoire de progression de la maladie.

Dans l'ensemble, cette thèse propose plusieurs outils pour étendre les applications potentielles du

Disease Course Mapping, facilitant son transfert vers un impact clinique. Les méthodes développées ici visent à améliorer notre compréhension des maladies neurodégénératives, en identifiant les sous-types, en incluant des marqueurs de nature variée et en tenant compte de l'effet du traitement sur ces marqueurs.

Remerciements

Tout d'abord je voudrais remercier Stanley de m'avoir encadré, d'abord pendant mon stage de fin d'études et ensuite pendant ma thèse. Tu m'as permis d'évoluer dans le monde de la recherche en stimulant mon questionnement, en me laissant explorer mes propres solutions et en me ramenant sur le droit chemin quand je partais trop loin. Ensuite je voudrais remercier les anciens du labo : Igor, Raphaël, Tiziana, Simona, Juliana, Etienne, Arnaud, Sophie, qui sont tous partis depuis (sauf Juli). C'est grâce à vous que je suis venu à l'Institut, et même si nous n'avons pas tous eu l'occasion de beaucoup nous voir à cause du Covid qui a marqué le début de ma thèse, vous avez tous participé à me faire aimer cet endroit qui est honnêtement le meilleur cadre de travail que j'ai pu voir et avoir. Ce cadre il doit beaucoup aux gens, à l'ambiance amicale et ouverte qui règne dans l'équipe. Je vous remercie donc sincèrement tous d'être là au quotidien : Némo, Juliette, Cécile, Ravi, Matthieu, Charley, Rémy, Tristan, Vito, Lisa, Camille, Elise, Arya, Sophie, Maëlys. Merci à tous les co-thésards avec qui nous avons évolué dans les méandres des publications, de l'administratif et surtout des afterworks. Merci à tous les ingénieurs qui nous rappellent au quotidien qu'un code de recherche ce n'est qu'un brouillon. Merci à Ninon de t'occuper de tous nos petits soucis informatiques.

Je remercie en particulier toute l'équipe Leaspy, avec qui les discussions sont toujours sources d'un éternel questionnement sur nos modèles. Je remercie mes collaborateurs, en particulier Bruno qui a suivi mon travail depuis deux ans. Ensuite je remercie les stagiaires que j'ai eu la chance d'encadrer, Samuel et Paul, qui sont maintenant en thèse. Vous avez apporté de la nouveauté dans mon regard sur mon travail en général et vos publications sont la preuve de la pertinence de réflexions.

Voilà j'en ai fini avec Aramis, du moins pour l'instant parce que l'aventure continue avec vous pendant encore un an.

Maintenant je me dois de remercier ma famille pour avoir soutenu mon parcours depuis le plus jeune âge. J'ai une pensée énorme pour mes grands-parents maternels qui se sont beaucoup investis dans ma réussite, et ce doctorat est une fierté de plus pour moi, une façon de leur dire merci.

Merci à tous mes amis d'école, les expats du Volley en famille élargie, dont beaucoup font une thèse et connaissent ce chemin si particulier. J'aurais cette petite fierté d'avoir été le premier du groupe à obtenir ma thèse.

Enfin merci à Nell, qui me supporte depuis maintenant plus de cinq ans. Je ne sais pas ce que j'aurais fait sans toi.

Contents

Introduction

Contents of the chapter

Motivation

In the wake of medical and sanitary progress over the last centuries, our societies have gained considerably in life expectancy. With the advent of an ever-aging population, new healthcare challenges have risen, namely cancer and neurodegenerative diseases. Focusing on the latter, this category of pathologies encompasses a large spectrum of diseases, being defined by a dysfunction in the peripheral or central nervous system. Currently neurodegenerative diseases are the second leading cause of deaths worldwide[1]. It is estimated that one in two women and one in three men will be affected by it during their life. Dementia affects 1 person in 14 after the age of 65 and 1 person in 6 after 80 years old, Alzheimer's disease (AD) being the most frequent cause of dementia with about 70% of all dementia cases. The second most common neurodegenerative disease is Parkinson's disease (PD), with more than 12 millions patients in Europe. These figures are bound to increase in the years to come, as demographic previsions forecast an increase of the 65 and older from 18% to 24% in 2060 (French population statistics[2]). However these neurodegenerative diseases raise several issues in their study.

Understanding the disease: Contrarily to infectious diseases, where the cause is easily identifiable and the pathogen mechanisms can be studied, neurodegenerative diseases remain poorly understood for the most part. There is no bacteria or virus to blame, no clear debut such as an infection, no cure nor vaccine. Their pathophysiology have been based on symptoms and clinical assessments of cognitive or motor abilities, which are the result from advanced neurodegeneration. However the process of neurodegeneration started much earlier, undetected before it is too late. For instance in AD or PD, the disease can last up to decades, and evidence shows that the biological mechanisms responsible for neuronal damage started more than twenty years prior to diagnosis[3]. This hidden evolution of neurodegeneration has been the main reason why medical research has been lagging behind in the understanding of the diseases. Diagnosed patients only account for the late part of the disease, when most of the damage has been done. The study of the natural disease history therefore is hampered by a lack of subjects with early disease development, in a state which is called prodromal. The recruitment of prodromal individuals is particularly difficult, due to another missing stone in our knowledge of neurodegenerative diseases: the causes remain mostly unknown. Risk factors have been identified[4,5], including environmental factors or genetics. However they only account for a part of the diseases cases, and they do not automatically imply that the individual at risk will develop the pathology, which is the nuance between a risk factor and a cause. We are often faced with the large majority of patients being affected spontaneously, which is referred to as the idiopathic form of the disease.

The study of those diseases have led to hypothetical models trying to shape our understanding of their progression. In Alzheimer's disease the famous amyloid cascade mechanism was proposed

[1] J. Dumurgier and C. Tzourio, "Epidemiology of neurological diseases in older adults," en, *Revue Neurologique*, vol. 176, no. 9, pp. 642–648, Nov. 2020.

[2] *Projections de population à l'horizon 2060 - Insee Première - 1320.*

[3] R. B. Postuma, A. E. Lang, J. F. Gagnon, *et al.*, "How does parkinsonism start? Prodromal parkinsonism motor changes in idiopathic REM sleep behaviour disorder," *Brain*, vol. 135, no. 6, pp. 1860–1870, Jun. 2012.

[4] G. Livingston, J. Huntley, A. Sommerlad, *et al.*, "Dementia prevention, intervention, and care: 2020 report of the Lancet Commission," English, *The Lancet*, vol. 396, no. 10248, pp. 413–446, Aug. 2020.

[5] T. Nedelec, B. Couvy-Duchesne, F. Monnet, *et al.*, "Identifying health conditions associated with Alzheimer's disease up to 15 years before diagnosis: An agnostic study of French and British health records," English, *The Lancet Digital Health*, vol. 4, no. 3, e169–e178, Mar. 2022.

in[6]. This model highlighted the role of the β-amyloid protein, whose progressive accumulation in the brain led to neuronal death, ultimately causing dementia. This hypothetical model has shaped the research in the field since then, to the point where the aggregating amyloid proteins are the primary target of drug trials. The study of this cascade over observable aspects of the disease has led to another widely known hypothetical model in AD. Introduced in[7], the authors proposed that the evolution of disease markers, such as the amyloid burden, neuronal death markers and cognitive decline, would follow a logistic progression from normal to abnormal. Similar models have also been proposed in PD, namely the Braak stages[8] and more recently the Brain-first and Body-first hypothesis[9]. The crucial part in those descriptive models is to provide an *ordering* of markers degradation, which is then used to define stages in the disease. Staging is essential as we will see when discussing the development of a treatment.

Defining the disease: This fledgling understanding of the biological processes underlying the observable pathophysiology of the disease questions even the very definition of these diseases. Until quite recently the notion of disease was related to the clinical manifestations, i.e. symptoms, and the diagnosis was assessed by a neurologist. However the patient probably *had* the disease before the symptoms' onset. Therefore recent developments push for a biological definition of the diseases, based on objective measurements of biomarkers which would allow for a diagnosis prior to the clinical one. In AD such a definition relies on the amyloid presence in the brain, or the tau burden -another protein whose misfolding induces tangles and fibrils, toxic for the neurons-, and patients are thus described under an A/T/N (for Amyloid/Tau/Neurodegeneration) classification[10].

Embrace heterogeneity: This last example of classification in AD patients echoes the wide range of possible disease manifestations. Oftentimes a neurodegenerative disease encompasses various subtypes, sometimes due to specific genetic variants. This heterogeneity in the disease spectrum also explains why it is hard to define the disease properly. Heterogeneity is unavoidable in some cases because studying smaller subgroups of patients would leave too few data to study the disease, and too few potential subjects to be recruited in a clinical trial. Therefore one always has to account for the shared part of the disease physiology as well as the differences between the subtypes. The distinction between subtypes is also not always straightforward, as there is not always a genetic variant or a separating marker. Many attempts have been made for disease subtyping in PD, for instance[11,12,13], but lines between the subgroups are blurred and not always stable as the diseases

[6] J. A. Hardy and G. A. Higgins, "Alzheimer's disease: The amyloid cascade hypothesis," eng, *Science (New York, N.Y.)*, vol. 256, no. 5054, pp. 184–185, Apr. 1992.

[7] C. R. Jack, D. S. Knopman, W. J. Jagust, *et al.*, "Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade," English, *The Lancet Neurology*, vol. 9, no. 1, pp. 119–128, Jan. 2010.

[8] H. Braak, K. D. Tredici, U. Rüb, *et al.*, "Staging of brain pathology related to sporadic Parkinson's disease," en, *Neurobiology of Aging*, vol. 24, no. 2, pp. 197–211, Mar. 2003.

[9] J. Horsager, K. B. Andersen, K. Knudsen, *et al.*, "Brain-first versus body-first Parkinson's disease: A multimodal imaging case-control study," *Brain*, vol. 143, no. 10, pp. 3077–3088, Oct. 2020.

[10] C. R. Jack, D. A. Bennett, K. Blennow, *et al.*, "A/T/N: An unbiased descriptive classification scheme for Alzheimer disease biomarkers," eng, *Neurology*, vol. 87, no. 5, pp. 539–547, Aug. 2016.

[11] R. S. Eisinger, C. W. Hess, D. Martinez-Ramirez, *et al.*, "Motor subtype changes in early Parkinson's disease," en, *Parkinsonism & Related Disorders*, vol. 43, pp. 67–72, Oct. 2017.

[12] J. Jankovic, M. McDermott, J. Carter, *et al.*, "Variable expression of Parkinson's disease: A base-line analysis of the DATATOP cohort. The Parkinson Study Group," eng, *Neurology*, vol. 40, no. 10, pp. 1529–1534, Oct. 1990.

[13] R. B. Postuma, A. E. Lang, J. F. Gagnon, *et al.*, "How does parkinsonism start? Prodromal parkinsonism motor changes in idiopathic REM sleep behaviour disorder," *Brain*, vol. 135, no. 6, pp. 1860–1870, Jun. 2012.

progresses. In this case, some advocate for a continuum of variability to describe the differences between individuals. Hence the need to precisely identify the types of heterogeneity: a subtype-based heterogeneity, relying on an underlying causal factor for the differences, and the natural disease heterogeneity which leads to a continuous spectrum of observed variability.

Early detection: The last challenge with neurodegenerative diseases is to identify early markers of the disease. This is crucial in the enrollment of prodromal subjects into studies, in order to increase our understanding of the early development of the disease mechanisms. This also induces the possibility of early treatment, which is necessary in order to cure such diseases. Indeed, trying to revert the loss of neurons is an impossible task. The treatment has to act to *prevent* the neuronal loss, hence the administration time has to be at an early stage. But detecting early signs of a neurodegenerative disease is challenging, since many risk factors[14] are not specific to one neurodegenerative disease: depression, constipation, or anxiety just to name a few ones. More specific markers can be detected with brain imaging or a cerebro-spinal fluid (CSF) sample, however such measurements are too costly and invasive to be performed routinely in order to detect early disease signs. More recently clinical research has striven to discover blood-based biomarkers (BBBM) of neurodegeneration[15] which would be an adequate balance between sensitivity -which is lower than with imaging markers- and cost as well as invasiveness. Some more subtle signs are also studied, for instance the lost of the olfactive sense for AD[16], or sleep troubles for PD[17].

In order to address all of these challenges, high expectations were put upon data-driven models. Longitudinal cohorts have been built to follow the slow progression of those diseases, whose duration can span dozens of years. The databases are said to be longitudinal in the sense that individuals have repeated measurements at various points in time, which we will refer to as visits. This allows for the modelling of the dynamics of the disease, which is the goal of the field of disease progression modelling. The hope is that the accumulation of enough data from numerous centers with a wide array of measurements will help us better understand the disease, and ultimately lead to the discovery of the needed biomarkers for early detection and targets for therapeutic interventions.

Measurements in the cohorts include data from various nature: biomarkers such as BBBM, CSF or imaging biomarkers; cognitive assessments; patient questionnaires; motor tests; genetic sequencing... A special mention has to go to imaging, which has enormously benefited from the last decades developments of Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET). Brain imaging has been at the root of the identification of abnormalities related to the neurodegenerative diseases. Famous large public research cohorts include the Alzheimer's Disease Neuroimaging Initiative (ADNI)[18], and Parkinson's Progression Markers Initiative (PPMI)[19].

[14] T. Nedelec, B. Couvy-Duchesne, F. Monnet, *et al.*, "Identifying health conditions associated with Alzheimer's disease up to 15 years before diagnosis: An agnostic study of French and British health records," English, *The Lancet Digital Health*, vol. 4, no. 3, e169–e178, Mar. 2022.

[15] C. E. Teunissen, I. M. W. Verberk, E. H. Thijssen, *et al.*, "Blood-based biomarkers for Alzheimer's disease: Towards clinical implementation," en, *The Lancet Neurology*, vol. 21, no. 1, pp. 66–77, Jan. 2022.

[16] C. Murphy, "Olfactory and other sensory impairments in Alzheimer disease," en, *Nature Reviews Neurology*, vol. 15, no. 1, pp. 11–24, Jan. 2019.

[17] R. B. Postuma, A. E. Lang, J. F. Gagnon, *et al.*, "How does parkinsonism start? Prodromal parkinsonism motor changes in idiopathic REM sleep behaviour disorder," *Brain*, vol. 135, no. 6, pp. 1860–1870, Jun. 2012.

[18] R. C. Petersen, P. S. Aisen, L. A. Beckett, *et al.*, "Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization," en, *Neurology*, vol. 74, no. 3, pp. 201–209, Jan. 2010.

[19] K. Marek, D. Jennings, S. Lasch, *et al.*, "The Parkinson Progression Marker Initiative (PPMI)," en, *Progress in Neurobiology*, Biological Markers for Neurodegenerative Diseases, vol. 95, no. 4, pp. 629–635, Dec. 2011.

The growth of these longitudinal cohorts has spurred the development of the disease progression modeling field. These digital models try to extract patterns of progression from the data. Their structure often allows to distinguish between two levels: global effects that describe the disease progression for the whole population, and individualized description of the disease to allow for subject-specific modelling. The goals of disease progression models are manifold.

Diagnosis and detection: One hope is that, by leveraging the vast amount of data collected from the longitudinal cohorts, the models are able to detect small signs of the disease onset ahead of the diagnosis. On the other hand, they can also be used to numerically determine thresholds for the diagnosis of the disease. Such models could be used to help the clinician with the diagnosis, and more realistically they could be implemented in the general practitioners offices to suggest when to redirect the patient to an expert neurologist consultation.

Prognosis: Disease progression models often provide individual predictions in time, thus allowing to forecast individuals in the future. This task is often used as the motivation for a disease progression model design, as predicting is more familiar to the Machine Learning community. This lead to the creation of prediction challenges for the progression of neurodegenerative diseases, such as the TADPOLE challenge[20]. Even though this competitive framework has sparked many interesting modelling attempts, performance should not be the sole aim of a model.

Interpretability: The model should provide insights into the disease typical evolution. Being able to exhibit parameters responsible for the individual specificity can be used to perform covariate analysis for instance. A common milestone is the design of what we call a disease timeline. It is well-known that age is not the proper reference axis to compare patients. Indeed some may start the disease earlier, others progress at a faster pace, or maybe both at the same time. Two patients at the same age might not be at the same disease stage. Hence the need to temporally align patients on a common disease timeline. Once this has been done, the variability left in the disease presentation can be referred to as spatial variability, as opposed to the time variability which accounts for realigning patients in time. A specific effort in our work has been put on the distinction between temporal and spatial effects.

Once a disease progression model has proven to reliably perform any of the tasks mentioned previously, the next step consists in using this model to achieve the ultimate goal of the research in neurodegenerative disease: find a cure. A first hope that is quickly dismissed is to find a potential target for a drug. This is highly unlikely, as the models are based on data that has been collected for the specific needs of the cohort. Data are thus recorded because of their already known, or hypothesized, link to the neurodegenerative process. A second potential use of disease progression models is to better identify when, how and to whom to administrate a drug. This is particularly useful in the era of personalized medicine. A final application context is the enhancement of clinical trials. Recently the US Food and Drugs Administration has approved the use of PROCOVA[21], a

[20] R. V. Marinescu, N. P. Oxtoby, A. L. Young, *et al.*, "TADPOLE Challenge: Accurate Alzheimer's disease prediction through crowdsourced forecasting of future data," eng, *PRedictive Intelligence in MEdicine. PRIME (Workshop)*, vol. 11843, pp. 1–10, Oct. 2019.

[21] D. Bertolini, K. Arnemann, D. Hall, *et al.*, "Machine Learning Enables Smaller ALS Clinical Trials (P1-13.003)," en, *Neurology*, vol. 98, no. 18 Supplement, May 2022.

method where the use of a prognostic disease progression score in an ANCOVA allows for a more powered trial, or equivalently allows for a smaller sample size.

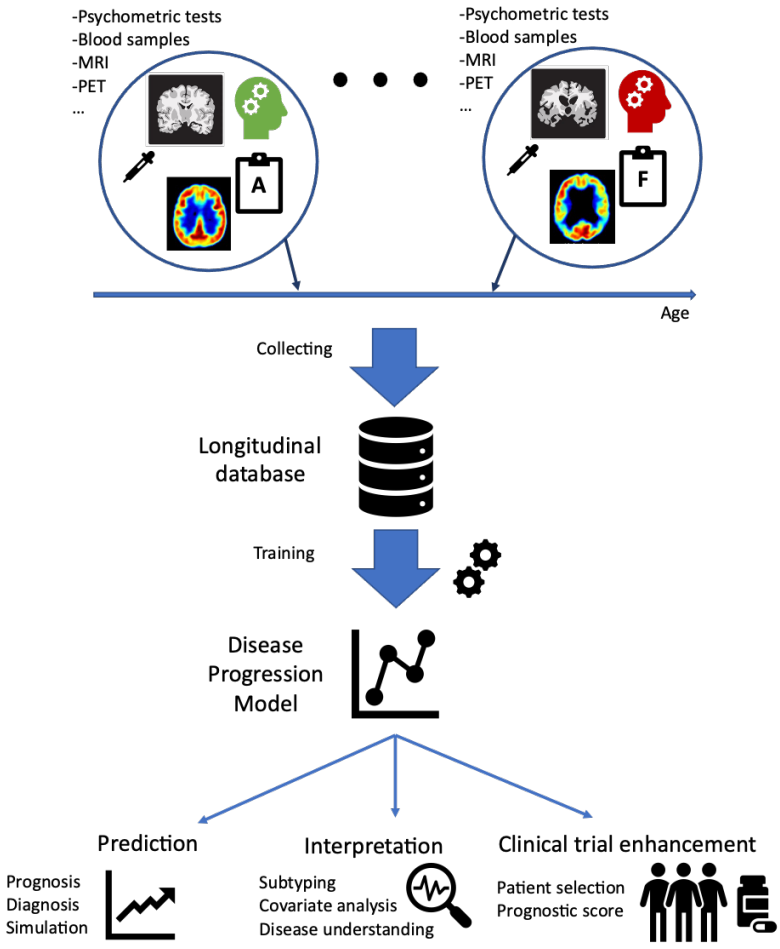

Figure 1: Schematic representation of disease progression models. Data from various sources is recorded repeatedly in time for a patient. The aggregation of multiple individual visits in a longitudinal database allows to train disease progression models. These models can be used in multiple applications.

Disease progression modelling

Unfolding from the hypothetical model from[22], disease progression models have emerged to validate the hypothesis and provide a data-driven model as evidence. A specific class of models focuses on the ordering of the biomarkers abnormality progression. These models are called Event Based Models (EBM), and date back to[23]. Their application to neurodegenerative diseases uses cross-sectional data to determine the ordering of events[24]. When markers are continuous, events are defined by setting abnormality thresholds, allowing to include any type of data in the models. The EBM have been extended to account for the heterogeneity in the SUbtyping and STAge INference (SUSTAIN) algorithm[25]. However the sequence of events is only providing disease stages, and the disease timeline is not explicitly modelled.

Many models have focused on the time alignment as the cornerstone to estimate the disease progression. An early work proposed linear mapping to extract the disease timeline from individual observations[26]. This was later improved by the introduction of time-warps[27]. The construction of the common disease timeline by mapping patients' age onto a same axis through what is called a time reparametrization considers that individuals' observations are snapshots of a global common progression pattern, as in[28]. In this work the average trajectory was estimated without any prior shape hypothesis, only using the realigned individual snapshots. This common disease timeline can be seen as a continuous version of disease staging, a disease progression score[29].

The main class of longitudinal models used in disease progression modelling has traditionally been the class of mixed-effect models[30,31]. This type of models has a hierarchical structure allowing to differentiate between population shared characteristics, or fixed effects, and individual variability, or random effects. Due to the non-linear nature of the progression of most biomarkers, the models belong to the non-linear mixed-effect models[32]. A family of template curves is chosen for the

[22] C. R. Jack, D. S. Knopman, W. J. Jagust, *et al.*, "Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade," English, *The Lancet Neurology*, vol. 9, no. 1, pp. 119–128, Jan. 2010.

[23] L. A. Beckett, "Maximum Likelihood Estimation in Mallows's Model Using Partially Ranked Data," en, in *Probability Models and Statistical Analyses for Ranking Data*, M. A. Fligner and J. S. Verducci, Eds., ser. Lecture Notes in Statistics, New York, NY: Springer, 1993, pp. 92–107.

[24] H. M. Fonteijn, M. Modat, M. J. Clarkson, *et al.*, "An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease," eng, *NeuroImage*, vol. 60, no. 3, pp. 1880–1889, Apr. 2012.

[25] A. L. Young, R. V. Marinescu, N. P. Oxtoby, *et al.*, "Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference," en, *Nature Communications*, vol. 9, no. 1, pp. 1–16, Oct. 2018.

[26] E. Yang, M. Farnum, V. Lobanov, *et al.*, "Quantifying the pathophysiological timeline of Alzheimer's disease," eng, *Journal of Alzheimer's disease: JAD*, vol. 26, no. 4, pp. 745–753, 2011.

[27] S. Durrleman, X. Pennec, A. Trouvé, *et al.*, "Toward a Comprehensive Framework for the Spatiotemporal Statistical Analysis of Longitudinal Shape Data," en, *International Journal of Computer Vision*, vol. 103, no. 1, pp. 22–59, May 2013.

[28] M. C. Donohue, H. Jacqmin-Gadda, M. Le Goff, *et al.*, "Estimating long-term multivariate progression from short-term data," *Alzheimer's & dementia : the journal of the Alzheimer's Association*, vol. 10, no. 0, S400–S410, Oct. 2014.

[29] V. Kmetzsch, E. Becker, D. Saracino, *et al.*, "Disease Progression Score Estimation From Multimodal Imaging and MicroRNA Data Using Supervised Variational Autoencoders," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 12, pp. 6024–6035, Dec. 2022.

[30] N. M. Laird and J. H. Ware, "Random-effects models for longitudinal data," eng, *Biometrics*, vol. 38, no. 4, pp. 963–974, Dec. 1982.

[31] M. J. Lindstrom and D. M. Bates, "Newton—Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data," *Journal of the American Statistical Association*, vol. 83, no. 404, pp. 1014–1022, Dec. 1988.

[32] J. Serroyen, G. Molenberghs, G. Verbeke, *et al.*, "Nonlinear Models for Longitudinal Data," *The American*

progression, for instance logistic curves, and with an affine time-reparametrization the individuals are mapped onto the average trajectory, as in[33]. Other choices of template curves are possible, as the ones proposed by[34]. In the R-package for latent classes mixed effect models (LCMM)[35], the non-linear model can be combined with the joint estimation of subtypes. Non-linear mixed effect models describing the progression of continuous markers can be coupled with event modelling, in what is called joint models[36]. An interesting approach used Gaussian Processes (GP) to account for the random effects of the mixed effect model[37,38]. The Disease Course Mapping (DCM) approach taken in[39] relies on a geometric model assuming that the observations lie on a Riemannian manifold. It is inspired from a prior geometric model[40], which was computationally intractable because the noise was defined in the manifold.

These non-linear models have also been extended to high-dimensional modalities such as imaging. Extending[41], the authors in[42] applied the model to images in ADNI. DIVE[43] is voxel-based approach for modelling the evolution of the brain in neurodegenerative diseases. The GP model was extended to shape modelling in[44]. In[45], the DCM model was applied to brain imaging, where the brain is modelled as a mesh and embedded in a shape manifold.

Modelling the dynamic progression of the biomarkers has also been performed using dynamical models. They are based on Ordinary Differential Equations (ODE)[46,47] or discretized versions of

Statistician, vol. 63, no. 4, pp. 378–388, Nov. 2009.

[33]B. M. Jedynak, A. Lang, B. Liu, *et al.*, "A Computational Neurodegenerative Disease Progression Score: Method and Results with the Alzheimer's Disease Neuroimaging Initiative Cohort," *NeuroImage*, vol. 63, no. 3, pp. 1478–1486, Nov. 2012.

[34]L. L. Raket, "Statistical Disease Progression Modeling in Alzheimer Disease," *Frontiers in Big Data*, vol. 3, 2020.

[35]C. Proust-Lima, V. Philipps, and B. Liquet, "Estimation of Extended Mixed Models Using Latent Classes and Latent Processes: The R Package lcmm," en, *Journal of Statistical Software*, vol. 78, pp. 1–56, Jun. 2017.

[36]B. He and S. Luo, "Joint modeling of multivariate longitudinal measurements and survival data with applications to Parkinson's disease," eng, *Statistical Methods in Medical Research*, vol. 25, no. 4, pp. 1346–1358, Aug. 2016.

[37]M. Lorenzi, X. Pennec, G. B. Frisoni, *et al.*, "Disentangling normal aging from Alzheimer's disease in structural magnetic resonance images," en, *Neurobiology of Aging*, Novel Imaging Biomarkers for Alzheimer's Disease and Related Disorders (NIBAD), vol. 36, S42–S52, Jan. 2015.

[38]M. Lorenzi, M. Filippone, G. B. Frisoni, *et al.*, "Probabilistic disease progression modeling to characterize diagnostic uncertainty: Application to staging and prediction in Alzheimer's disease," en, *NeuroImage*, Mapping diseased brains, vol. 190, pp. 56–68, Apr. 2019.

[39]J.-B. Schiratti, S. Allassonnière, A. Routier, *et al.*, "A Mixed-Effects Model with Time Reparametrization for Longitudinal Univariate Manifold-Valued Data," en, in *Information Processing in Medical Imaging*, S. Ourselin, D. C. Alexander, C.-F. Westin, *et al.*, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2015, pp. 564–575.

[40]T. Fletcher, "Geodesic Regression on Riemannian Manifolds," en, Sep. 2011, p. 75.

[41]B. M. Jedynak, A. Lang, B. Liu, *et al.*, "A Computational Neurodegenerative Disease Progression Score: Method and Results with the Alzheimer's Disease Neuroimaging Initiative Cohort," *NeuroImage*, vol. 63, no. 3, pp. 1478–1486, Nov. 2012.

[42]M. Bilgel, J. L. Prince, D. F. Wong, *et al.*, "A multivariate nonlinear mixed effects model for longitudinal image analysis: Application to amyloid imaging," en, *NeuroImage*, vol. 134, pp. 658–670, Jul. 2016.

[43]R. V. Marinescu, A. Eshaghi, M. Lorenzi, *et al.*, "DIVE: A spatiotemporal progression model of brain pathology in neurodegenerative disorders," en, *NeuroImage*, vol. 192, pp. 166–177, May 2019.

[44]C. Abi Nader, N. Ayache, P. Robert, *et al.*, "Monotonic Gaussian Process for spatio-temporal disease progression modeling in brain imaging data," eng, *NeuroImage*, vol. 205, p. 116 266, Jan. 2020.

[45]I. Koval, "Learning Multimodal Digital Models of Disease Progression from Longitudinal Data : Methods & Algorithms for the Description, Prediction and Simulation of Alzheimer's Disease Progression," en, Ph.D. dissertation, Institut Polytechnique de Paris, Jan. 2020.

[46]N. P. Oxtoby, A. L. Young, D. M. Cash, *et al.*, "Data-driven models of dominantly-inherited Alzheimer's disease progression," *Brain*, vol. 141, no. 5, pp. 1529–1544, May 2018.

[47]K. Lahouel, M. Wells, V. Rielly, *et al.*, *Learning nonparametric ordinary differential equations from noisy data*,

ODE[48], where the field of the ODE is estimated from the whole cohort and is akin to the fixed effects of the mixed models while the initial conditions of an individual patients trajectory account for the random effects. Neural ODE[49] use a Deep Learning network to learn the field. Other Deep Learning approaches using Recurrent Neural Networks have also been applied for disease progression[50]. These Deep Learning methods are often less interpretable. This is quite the opposite to the multifactorial causal model proposed in[51]. This causal model allows to perform interventions, which can be leveraged to emulate therapeutic intervention.

All those statistical methods have in common the ability to describe the progression of the disease at a population level. Most of them can also compute an individualized trajectory for every patient. The main differences lie in their interpretability. Models with an explicit time-reparametrization provide a separation of temporal and spatial variability, which is an asset to us. Semi-parametric or non-parametric models, such as the dynamical models based on ODE or Deep Learning based methods, often lack on the interpretability side. Our work focuses on the geometrical approach of the DCM, which provides both the spatio-temporal disentanglement and meaningful parameters.

Parkinson's disease

Most of the work in this thesis has been developed for applications to Parkinson's disease data, so we will provide a brief overview of this disease, its hypothesized causes, its manifestation and its treatments.

PD was first described by James Parkinson in 1817. This neurodegenerative disease is the second most prevalent after AD. Its prevalence increases drastically with age, from being almost null before 50 years to close to 2% after 80 years old[52]. Men are twice more likely to develop the disease compared to women.

PD is characterized by the loss of dopaminergic neurons, starting in the substantia nigra, a region in the brainstem containing almost exclusively dopaminergic neurons. Dopamine is a neurotransmitter involved in the reward system, but is also responsible for the activation of the motor system. The decrease in dopamine levels due to neurodegeneration entails all the motor symptoms of PD that we are familiar with: bradykinesia (slowness of movement), muscle rigidity, tremors, which are the main symptoms[53], but many others are also observed including falls, slowness of speech... The loss of dopaminergic neurons spreads in time to other regions of the brain, inducing non-motor symptoms, such as apathy, hallucinations or depression. However the onset of symptom is a very late landmark of neuronal death. When PD is diagnosed, already 80% of the dopaminergic neurons in the subtantia nigra are already lost. The neurodegeneration is hypothesized to come

Feb. 2023.

[48] B. O. Taddé, H. Jacqmin-Gadda, J.-F. Dartigues, *et al.*, "Dynamic modeling of multivariate dimensions and their temporal relationships using latent processes: Application to Alzheimer's disease," en, *Biometrics*, vol. 76, no. 3, pp. 886–899, 2020.

[49] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, *et al.*, *Neural Ordinary Differential Equations*, Dec. 2019.

[50] M. Nguyen, T. He, L. An, *et al.*, "Predicting Alzheimer's disease progression using deep recurrent neural networks," en, *NeuroImage*, vol. 222, p. 117 203, Nov. 2020.

[51] Y. Iturria-Medina, F. M. Carbonell, R. C. Sotero, *et al.*, "Multifactorial causal model of brain (dis)organization and therapeutic intervention: Application to Alzheimer's disease," en, *NeuroImage*, vol. 152, pp. 60–77, May 2017.

[52] A. Elbaz, L. Carcaillon, S. Kab, *et al.*, "Epidemiology of Parkinson's disease," en, *Revue Neurologique*, Neuroepidemiology, vol. 172, no. 1, pp. 14–26, Jan. 2016.

[53] J. Jankovic, M. McDermott, J. Carter, *et al.*, "Variable expression of Parkinson's disease: A base-line analysis of the DATATOP cohort. The Parkinson Study Group," eng, *Neurology*, vol. 40, no. 10, pp. 1529–1534, Oct. 1990.

from the accumulation of the α-synuclein protein, which is currently the main target for drug development[54]. The loss of dopaminergic neurons can be observed with dopamine transporter Single Photon Emission Computed Tomography (SPECT) scan.

As mentioned in the challenges about neurodegenerative diseases, the definition of PD is complex. It is currently clinical, at the neurologist discretion, but there is no clear-cut guideline for diagnosis. The symptoms are similar to Lewy-body disease, which leads to the early diagnosis being revised sometimes. However we are still far from a biological definition of the disease. The heterogeneity of the disease plays an important role, starting from the symptom onset[55]. A first distinction can be made based on the location of the first motor signs. 60% of the patients are first impaired on one side of the body, the symptoms spreading slowly towards the rest of the body in time. Attempts at subtyping have tried to distinguish early prodromal subtypes, based on sleep behaviours[56]. One of the most common classification based on motor symptoms is the Tremor Dominant (TD) vs Postural Instability and Gait Disorder (PIGD) system proposed in[57]. However these subtypes do not encompass all of the patients, and they are not stable in time[58 ,59]. Some particular subtypes can be derived from genetic variants GBA and LRKK2, which account for about 5% of the total PD cases, the others being idiopathic.

PD has benefited from the discovery of a treatment in the early 50's, called Levodopa. This is a dopamine precursor, and when taken regularly it makes up for the lack of dopamine production in the brain. Other drugs based on dopaminergic input have since been developed such as dopamin agonists, or inhibitors of recapture including IMAOB and ICOMT. They all rely on the same mechanism, with different side-effects. Levodopa is for instance known to cause dyskinesia, i.e. spontaneous movements, and has also been related to freezing in some studies[60]. Dopamin agonists are more likely to cause Impulse Control Disorders (ICD), which causes addictive behaviours such as gambling. The treatment for PD patients nowadays is a cocktail of several dopaminergic drugs, experts are still debating what is the best treatment policy[61]. Many factors can influence the answer of a patient to the treatment, and its likelihood of developing negative side-effects.

The dopaminergic treatment is purely symptomatic however, it has not shown any improvement in the long run[62]. Its efficacy diminishes as the disease progresses towards later stages, leading to fluctuations between "ON" and "OFF" states. "ON" states correspond to the normally functioning drug state, where motor symptoms are alleviated. "OFF" states on the other hand occur when the treatment effect wears out. In later stages of PD the half-life of the dopaminergic drug becomes

[54] C. R. Fields, N. Bengoa-Vergniory, and R. Wade-Martins, "Targeting Alpha-Synuclein as a Therapy for Parkinson's Disease," *Frontiers in Molecular Neuroscience*, vol. 12, 2019.

[55] W. J. Zetusky, J. Jankovic, and F. J. Pirozzolo, "The heterogeneity of Parkinson's disease: Clinical and prognostic implications," eng, *Neurology*, vol. 35, no. 4, pp. 522–526, Apr. 1985.

[56] R. B. Postuma, A. E. Lang, J. F. Gagnon, *et al.*, "How does parkinsonism start? Prodromal parkinsonism motor changes in idiopathic REM sleep behaviour disorder," *Brain*, vol. 135, no. 6, pp. 1860–1870, Jun. 2012.

[57] J. Jankovic, M. McDermott, J. Carter, *et al.*, "Variable expression of Parkinson's disease: A base-line analysis of the DATATOP cohort. The Parkinson Study Group," eng, *Neurology*, vol. 40, no. 10, pp. 1529–1534, Oct. 1990.

[58] T. Simuni, C. Caspell-Garcia, C. Coffey, *et al.*, "How stable are Parkinson's disease subtypes in de novo patients: Analysis of the PPMI cohort?" en, *Parkinsonism & Related Disorders*, vol. 28, pp. 62–67, Jul. 2016.

[59] R. S. Eisinger, C. W. Hess, D. Martinez-Ramirez, *et al.*, "Motor subtype changes in early Parkinson's disease," en, *Parkinsonism & Related Disorders*, vol. 43, pp. 67–72, Oct. 2017.

[60] M. Gilat, N. D'Cruz, P. Ginis, *et al.*, "Freezing of gait and levodopa," English, *The Lancet Neurology*, vol. 20, no. 7, pp. 505–506, Jul. 2021.

[61] R. M. A. de Bie, C. E. Clarke, A. J. Espay, *et al.*, "Initiation of pharmacological therapy in Parkinson's disease: When, why, and how," eng, *The Lancet. Neurology*, vol. 19, no. 5, pp. 452–461, May 2020.

[62] R. Cilia, E. Cereda, A. Akpalu, *et al.*, "Natural history of motor symptoms in Parkinson's disease and the long-duration response to levodopa," *Brain*, vol. 143, no. 8, pp. 2490–2501, Aug. 2020.

really short due to the quasi-inexisting production by the brain.

Another treatment called Deep Brain Stimulation implants an electrode in the amygdala to directly stimulate the brain. This treatment has shown to be a potent alternative to dopaminergic therapy[63], but still remains quite rare due to its invasiveness. The dopaminergic treatment has long overshadowed the research for a disease-modifying drug, therefore the research field pales in comparison with its counterpart in AD.

Longitudinal cohorts used in this thesis work include PPMI[64] and a french cohort from the NS-PARK consortium. The data in PD cohorts is mostly composed of cognitive and motor assessments. Historically the Hoehn and Yahr scale[65] (from 0 to 5) was used to describe the disease in successive stages. Today the most frequently used scale for disease progression is the Movement Disorder Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS)[66], which is an aggregation of many items rating multiple aspects of the disease, including a motor examination and an assessment of the disease impact on the daily life of the patient.

Thesis goals

The work in this thesis was built upon the Disease Course Mapping model introduced in[67]. We have described some key challenges raised by the neurodegenerative diseases and we will extend the DCM model to address those challenges. The DCM model has already been built to account for time alignment and spatio-temporal decoupling. However its design is far from perfect. First it is limited to continuous data. Moreover, the DCM model does not account for the heterogeneity of disease subtypes. Finally it is not geared for applications in drug development. Our brief overview of Parkinson's disease contrasts with those limitations. Symptoms, staging and rating scales are ubiquitous, and these are non-continuous variables. Subtypes are paramount in modelling the variability of PD. Finally all the data in PD cohorts comes from patients under dopaminergic treatment, thus requiring a specific model for it. We will therefore provide the following extensions of the DCM model:

- applicability to discrete data, including binary and ordinal observations
- mixture structure to uncover subtypes and better explain disease heterogeneity
- treatment models linked to the DCM to account for treatment effect and pave the way for the development of disease-modifying drugs

[63] A. L. Benabid, "Deep brain stimulation for Parkinson's disease," en, *Current Opinion in Neurobiology*, vol. 13, no. 6, pp. 696–706, Dec. 2003.

[64] K. Marek, D. Jennings, S. Lasch, *et al.*, "The Parkinson Progression Marker Initiative (PPMI)," en, *Progress in Neurobiology*, Biological Markers for Neurodegenerative Diseases, vol. 95, no. 4, pp. 629–635, Dec. 2011.

[65] M. M. Hoehn and M. D. Yahr, "Parkinsonism: Onset, progression, and mortality," en, *Neurology*, vol. 17, no. 5, pp. 427–427, May 1967.

[66] C. G. Goetz, B. C. Tilley, S. R. Shaftman, *et al.*, "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results: MDS-UPDRS: Clinimetric Assessment," en, *Movement Disorders*, vol. 23, no. 15, pp. 2129–2170, Nov. 2008.

[67] J.-B. Schiratti, S. Allassonnière, A. Routier, *et al.*, "A Mixed-Effects Model with Time Reparametrization for Longitudinal Univariate Manifold-Valued Data," en, in *Information Processing in Medical Imaging*, S. Ourselin, D. C. Alexander, C.-F. Westin, *et al.*, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2015, pp. 564–575.

All our methodological contributions are motivated by practical needs encountered in real-life data. They are thus accompanied by applications to neurodegenerative cohorts. We used the interpretability of the model to provide clinical insights in the analysis of our experimental results.

Manuscript overview

The first chapter 1 introduces the Disease Course Mapping model and discusses its limitations. The second part of the chapter presents Geodesic Bending, an extension of the DCM model which introduces more flexibility in the model by learning the shape of the typical trajectory of the disease.

The second chapter 2 presents our extension of the DCM model to binary and ordinal data. We provide in-depth applications to Parkinson's disease cohorts, as well as a parallel with the Item Response Theory.

The third chapter 3 addresses the heterogeneity challenge of neurodegenerative diseases. Our contribution takes the form of two mixture variants for the DCM model. We showcase the differences between the two with an application to Alzheimer's disease.

The fourth chapter 4 focuses on treatment modelling. We introduce two modelling approaches. The first one consists in a generic framework using any disease progression model to estimate treatment effect from patients only, without requiring data from controls nor placebo. The second approach is called Piece-wise Geodesic model and is targeted at disease-modifying drugs, providing a framework to learn changes in disease trajectory due to the treatment.

The fifth chapter 5 summarizes all our contributions to the estimation algorithm across all of the previous extensions. It presents the MCMC-SAEM algorithm used for the DCM estimation. We then introduce our contributions in the form of specific variants of this algorithm.

Introduction

Motivation

À la suite des progrès médicaux et sanitaires réalisés au cours des derniers siècles, nos sociétés ont considérablement gagné en espérance de vie. Avec l'émergence d'une population de plus en plus âgée, de nouveaux défis en matière de santé ont surgi, notamment le cancer et les maladies neurodégénératives. En se concentrant sur ces dernières, cette catégorie de pathologies englobe un large spectre de maladies, se caractérisant par un dysfonctionnement du système nerveux périphérique ou central. Actuellement, les maladies neurodégénératives sont la deuxième cause de décès dans le monde[68]. On estime qu'une femme sur deux et un homme sur trois seront touchés au cours de leur vie. La démence affecte une personne sur 14 après l'âge de 65 ans et une personne sur 6 après 80 ans, la maladie d'Alzheimer étant la cause la plus fréquente de démence, représentant environ 70% de tous les cas de démence. La deuxième maladie neurodégénérative la plus courante est la maladie de Parkinson, qui compte plus de 12 millions de patients en Europe. Ces chiffres sont appelés à augmenter dans les années à venir, car les prévisions démographiques prévoient une augmentation de la population de 65 ans et plus de 18% à 24% en 2060 (statistiques de la population française[69]). Cependant, ces maladies neurodégénératives soulèvent plusieurs problèmes dans leur étude.

Comprendre les maladies: Contrairement aux maladies infectieuses, où la cause est facilement identifiable et les mécanismes pathogènes peuvent être étudiés, les maladies neurodégénératives restent largement méconnues. Il n'y a ni bactérie ni virus à blâmer, aucun début clair tel qu'une infection, aucune cure ni vaccin. Leur physiopathologie repose sur des symptômes et des évaluations cliniques des capacités cognitives ou motrices, qui résultent d'une neurodégénérescence tardive. Cependant, le processus de neurodégénérescence débute bien plus tôt, passant inaperçu jusqu'à ce qu'il soit trop tard. Par exemple, dans la maladie d'Alzheimer ou la maladie de Parkinson, la maladie peut durer des décennies, et des études montrent que les mécanismes biologiques responsables des dommages neuronaux ont commencé plus de vingt ans avant le diagnostic[70]. Cette évolution cachée de la neurodégénérescence est la principale raison pour laquelle la recherche médicale a du retard dans la compréhension de ces maladies. Les patients diagnostiqués ne représentent qu'une phase tardive de la maladie, lorsque la plupart des dommages ont déjà été causés. L'étude de l'histoire naturelle de la maladie est donc entravée par un manque de sujets présentant un stade

[68] J. Dumurgier and C. Tzourio, "Epidemiology of neurological diseases in older adults," en, *Revue Neurologique*, vol. 176, no. 9, pp. 642–648, Nov. 2020.

[69] *Projections de population à l'horizon 2060 - Insee Première - 1320.*

[70] R. B. Postuma, A. E. Lang, J. F. Gagnon, *et al.*, "How does parkinsonism start? Prodromal parkinsonism motor changes in idiopathic REM sleep behaviour disorder," *Brain*, vol. 135, no. 6, pp. 1860–1870, Jun. 2012.

précoce de la maladie, dans un état appelé prodromique. Le recrutement de personnes prodromiques est particulièrement difficile en raison d'une autre lacune dans nos connaissances sur les maladies neurodégénératives : les causes restent en grande partie inconnues. Des facteurs de risque ont été identifiés[71,72], y compris des facteurs environnementaux ou génétiques. Cependant, ils ne représentent qu'une partie des cas de la maladie, et leur présence ne signifie pas automatiquement que l'individu à risque développera la pathologie, d'où la nuance entre un facteur de risque et une cause. Nous sommes souvent confrontés à la grande majorité de patients qui sont affectés spontanément, ce que l'on appelle la forme idiopathique de la maladie.

L'étude de ces maladies a conduit à des modèles hypothétiques tentant de façonner notre compréhension de leur progression. Dans la maladie d'Alzheimer, le célèbre mécanisme de cascade amyloïde a été proposé[73]. Ce modèle mettait en évidence le rôle de la protéine β-amyloïde, dont l'accumulation progressive dans le cerveau entraînait la mort neuronale, et finalement la démence. Ce modèle hypothétique a façonné la recherche dans le domaine depuis lors, au point où les protéines amyloïdes agrégées sont la cible principale des essais thérapeutiques. L'étude de cette cascade sur les aspects observables de la maladie a également conduit à un autre modèle hypothétique bien connu dans la maladie d'Alzheimer. Introduit dans une étude[74], les auteurs ont proposé que l'évolution des marqueurs de la maladie, tels que la charge amyloïde, les marqueurs de la mort neuronale et le déclin cognitif, suivrait une progression logistique du normal à l'anormal. Des modèles similaires ont également été proposés dans la maladie de Parkinson, notamment les stades de Braak[75] et plus récemment l'hypothèse du cerveau d'abord et du corps d'abord[76]. La partie cruciale de ces modèles descriptifs consiste à fournir un *ordre* de dégradation des marqueurs, qui est ensuite utilisé pour définir les stades de la maladie. La stratification en stades est essentielle, comme nous le verrons lors de la discussion sur le développement d'un traitement.

Définir la maladie: Cette compréhension naissante des processus biologiques sous-jacents à la physiopathologie observable de la maladie remet en question même la définition de ces maladies. Jusqu'à récemment, la notion de maladie était liée aux manifestations cliniques, c'est-à-dire aux symptômes, et le diagnostic était évalué par un neurologue. Cependant, le patient avait probablement la maladie avant l'apparition des symptômes. Par conséquent, les développements récents plaident en faveur d'une définition biologique des maladies, basée sur des mesures objectives de biomarqueurs qui permettraient un diagnostic préalable au diagnostic clinique. Dans la maladie d'Alzheimer, une telle définition repose sur la présence d'amyloïde dans le cerveau, ou de la charge en tau - une autre protéine dont le mauvais repliement induit la formation de noeuds et de fibrilles toxiques pour les neurones - et les patients sont ainsi décrits selon une classification A/T/N (pour

[71] G. Livingston, J. Huntley, A. Sommerlad, *et al.*, "Dementia prevention, intervention, and care: 2020 report of the Lancet Commission," English, *The Lancet*, vol. 396, no. 10248, pp. 413–446, Aug. 2020.

[72] T. Nedelec, B. Couvy-Duchesne, F. Monnet, *et al.*, "Identifying health conditions associated with Alzheimer's disease up to 15 years before diagnosis: An agnostic study of French and British health records," English, *The Lancet Digital Health*, vol. 4, no. 3, e169–e178, Mar. 2022.

[73] J. A. Hardy and G. A. Higgins, "Alzheimer's disease: The amyloid cascade hypothesis," eng, *Science (New York, N.Y.)*, vol. 256, no. 5054, pp. 184–185, Apr. 1992.

[74] C. R. Jack, D. S. Knopman, W. J. Jagust, *et al.*, "Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade," English, *The Lancet Neurology*, vol. 9, no. 1, pp. 119–128, Jan. 2010.

[75] H. Braak, K. D. Tredici, U. Rüb, *et al.*, "Staging of brain pathology related to sporadic Parkinson's disease," en, *Neurobiology of Aging*, vol. 24, no. 2, pp. 197–211, Mar. 2003.

[76] J. Horsager, K. B. Andersen, K. Knudsen, *et al.*, "Brain-first versus body-first Parkinson's disease: A multimodal imaging case-control study," *Brain*, vol. 143, no. 10, pp. 3077–3088, Oct. 2020.

Amyloïde/Tau/Neurodégénération)[77].

Décrire l'hétérogénéité: Cet dernier exemple de classification chez les patients atteints de la maladie d'Alzheimer met en évidence la grande variété de manifestations possibles de la maladie. Souvent, une maladie neurodégénérative englobe divers sous-types, parfois en raison de variants génétiques spécifiques. Cette hétérogénéité dans le spectre des maladies explique également pourquoi il est difficile de définir correctement la maladie. Dans certains cas, l'hétérogénéité est inévitable car l'étude de sous-groupes plus restreints de patients laisserait trop peu de données pour étudier la maladie et trop peu de sujets potentiels pour être recrutés dans un essai clinique. Il est donc toujours nécessaire de tenir compte de la part commune de la physiologie de la maladie ainsi que des différences entre les sous-types. La distinction entre les sous-types n'est pas toujours claire, car il n'y a pas toujours de variant génétique ou de marqueur distinctif. De nombreuses tentatives ont été faites pour la sous-typage de la maladie de Parkinson, par exemple[78,79,80], mais les frontières entre les sous-groupes sont floues et pas toujours stables au fur et à mesure de la progression de la maladie. Dans ce cas, certains plaident en faveur d'un continuum de variabilité pour décrire les différences entre les individus. D'où la nécessité d'identifier précisément les types d'hétérogénéité : une hétérogénéité basée sur les sous-types, reposant sur un facteur causal sous-jacent pour les différences, et l'hétérogénéité naturelle de la maladie qui conduit à un spectre continu de variabilité observée.

Détection précoce: Le dernier défi des maladies neurodégénératives consiste à identifier des marqueurs précoces de la maladie. Cela est crucial pour l'inclusion de sujets prodromiques dans les études, afin d'accroître notre compréhension du développement précoce des mécanismes de la maladie. Cela ouvre également la possibilité d'un traitement précoce, qui est nécessaire pour guérir de telles maladies. En effet, tenter de rétablir la perte de neurones est une tâche impossible. Le traitement doit agir pour *prévenir* la perte neuronale, d'où la nécessité d'une administration précoce. Cependant, détecter les premiers signes d'une maladie neurodégénérative est un défi, car de nombreux facteurs de risque[81] ne sont pas spécifiques à une maladie neurodégénérative particulière : la dépression, la constipation ou l'anxiété, pour n'en nommer que quelques-uns. Des marqueurs plus spécifiques peuvent être détectés par imagerie cérébrale ou par prélèvement de liquide céphalorachidien (LCR), cependant de telles mesures sont trop coûteuses et invasives pour être réalisées systématiquement afin de détecter les premiers signes de la maladie. Plus récemment, la recherche clinique s'est efforcée de découvrir des biomarqueurs de neurodégénérescence basés sur le sang (BBBM)[82], ce qui représenterait un équilibre adéquat entre sensibilité - qui est inférieure à celle des marqueurs d'imagerie - et coût ainsi que l'aspect invasif. Certains signes plus subtils sont

[77] C. R. Jack, D. A. Bennett, K. Blennow, *et al.*, "A/T/N: An unbiased descriptive classification scheme for Alzheimer disease biomarkers," eng, *Neurology*, vol. 87, no. 5, pp. 539–547, Aug. 2016.

[78] R. S. Eisinger, C. W. Hess, D. Martinez-Ramirez, *et al.*, "Motor subtype changes in early Parkinson's disease," en, *Parkinsonism & Related Disorders*, vol. 43, pp. 67–72, Oct. 2017.

[79] J. Jankovic, M. McDermott, J. Carter, *et al.*, "Variable expression of Parkinson's disease: A base-line analysis of the DATATOP cohort. The Parkinson Study Group," eng, *Neurology*, vol. 40, no. 10, pp. 1529–1534, Oct. 1990.

[80] R. B. Postuma, A. E. Lang, J. F. Gagnon, *et al.*, "How does parkinsonism start? Prodromal parkinsonism motor changes in idiopathic REM sleep behaviour disorder," *Brain*, vol. 135, no. 6, pp. 1860–1870, Jun. 2012.

[81] T. Nedelec, B. Couvy-Duchesne, F. Monnet, *et al.*, "Identifying health conditions associated with Alzheimer's disease up to 15 years before diagnosis: An agnostic study of French and British health records," English, *The Lancet Digital Health*, vol. 4, no. 3, e169–e178, Mar. 2022.

[82] C. E. Teunissen, I. M. W. Verberk, E. H. Thijssen, *et al.*, "Blood-based biomarkers for Alzheimer's disease: Towards clinical implementation," en, *The Lancet Neurology*, vol. 21, no. 1, pp. 66–77, Jan. 2022.

également étudiés, par exemple la perte de l'odorat dans la maladie d'Alzheimer[83] ou les troubles du sommeil dans la maladie de Parkinson[84].

Pour relever tous ces défis, de grandes attentes ont été placées dans les modèles basés sur les données. Des cohortes longitudinales ont été constituées pour suivre la lente progression de ces maladies, dont la durée peut s'étendre sur des dizaines d'années. Les bases de données sont qualifiées de longitudinales dans le sens où les individus ont des mesures répétées à différents moments, que nous appellerons visites. Cela permet de modéliser la dynamique de la maladie, qui est l'objectif du domaine de la modélisation de la progression des maladies. L'espoir est que l'accumulation de suffisamment de données provenant de nombreux centres et comportant une large gamme de mesures nous aidera à mieux comprendre la maladie et à découvrir en fin de compte les biomarqueurs nécessaires pour une détection précoce et les cibles pour des interventions thérapeutiques.

Les mesures dans les cohortes comprennent des données de diverses natures : des biomarqueurs tels que BBBM, LCR ou des biomarqueurs d'imagerie ; des évaluations cognitives ; des questionnaires pour les patients ; des tests moteurs ; des séquençages génétiques... Une mention spéciale doit être faite pour l'imagerie, qui a énormément bénéficié des développements des dernières décennies en imagerie par résonance magnétique (IRM) et en tomographie par émission de positrons (TEP). L'imagerie cérébrale a été à l'origine de l'identification des anomalies liées aux maladies neurodégénératives. Des cohortes de recherche publique de renommée incluent l'initiative d'imagerie neurologique de la maladie d'Alzheimer (ADNI)[85] et l'initiative pour les marqueurs de progression de la maladie de Parkinson (PPMI)[86].

L'essor de ces cohortes longitudinales a stimulé le développement du domaine de la modélisation de la progression des maladies. Ces modèles numériques tentent d'extraire des schémas de progression à partir des données. Leur structure permet souvent de distinguer deux niveaux : des effets globaux décrivant la progression de la maladie pour l'ensemble de la population, et une description individualisée de la maladie permettant une modélisation spécifique à chaque sujet. Les objectifs des modèles de progression des maladies sont nombreux.

Diagnostic et détection: Un espoir est que, en exploitant la quantité considérable de données collectées à partir des cohortes longitudinales, les modèles soient capables de détecter de petits signes de l'apparition de la maladie avant le diagnostic. D'autre part, ils peuvent également être utilisés pour déterminer numériquement les seuils pour le diagnostic de la maladie. De tels modèles pourraient être utilisés pour aider le clinicien dans le diagnostic, et de manière plus réaliste, ils pourraient être mis en place dans les cabinets des médecins généralistes pour suggérer quand orienter le patient vers une consultation chez un neurologue expert.

Pronostic : Les modèles de progression des maladies fournissent souvent des prédictions individuelles dans le temps, ce qui permet de prévoir l'évolution des individus dans le futur. Cette tâche est souvent utilisée comme motivation pour la conception d'un modèle de progression des maladies, car la prédiction est plus familière à la communauté de l'apprentissage automatique. Cela a donné

[83] C. Murphy, "Olfactory and other sensory impairments in Alzheimer disease," en, *Nature Reviews Neurology*, vol. 15, no. 1, pp. 11–24, Jan. 2019.

[84] R. B. Postuma, A. E. Lang, J. F. Gagnon, *et al.*, "How does parkinsonism start? Prodromal parkinsonism motor changes in idiopathic REM sleep behaviour disorder," *Brain*, vol. 135, no. 6, pp. 1860–1870, Jun. 2012.

[85] R. C. Petersen, P. S. Aisen, L. A. Beckett, *et al.*, "Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization," en, *Neurology*, vol. 74, no. 3, pp. 201–209, Jan. 2010.

[86] K. Marek, D. Jennings, S. Lasch, *et al.*, "The Parkinson Progression Marker Initiative (PPMI)," en, *Progress in Neurobiology*, Biological Markers for Neurodegenerative Diseases, vol. 95, no. 4, pp. 629–635, Dec. 2011.

lieu à la création de défis de prédiction pour la progression des maladies neurodégénératives, tels que le défi TADPOLE[87]. Bien que ce cadre compétitif ait suscité de nombreuses tentatives de modélisation intéressantes, les performances ne devraient pas être le seul objectif d'un modèle.

Interprétabilité : Le modèle devrait fournir des informations sur l'évolution typique de la maladie. Être capable d'exposer les paramètres responsables de la spécificité individuelle peut être utilisé pour effectuer une analyse de covariables, par exemple. Une étape importante consiste à concevoir ce que nous appelons une chronologie de la maladie. Il est bien connu que l'âge n'est pas l'axe de référence approprié pour comparer les patients. En effet, certains peuvent commencer la maladie plus tôt, d'autres progressent plus rapidement, ou peut-être les deux en même temps. Deux patients du même âge peuvent ne pas être au même stade de la maladie. D'où la nécessité d'aligner temporellement les patients sur une chronologie commune de la maladie. Une fois cela fait, la variabilité restante dans la présentation de la maladie peut être qualifiée de variabilité spatiale, par opposition à la variabilité temporelle qui tient compte du réalignement des patients dans le temps. Un effort spécifique dans notre travail a été porté sur la distinction entre les effets temporels et spatiaux.

Une fois qu'un modèle de progression des maladies s'est avéré capable d'accomplir de manière fiable l'une des tâches mentionnées précédemment, l'étape suivante consiste à utiliser ce modèle pour atteindre l'objectif ultime de la recherche sur les maladies neurodégénératives : trouver un remède. Un premier espoir qui est rapidement écarté est de trouver une cible potentielle pour un médicament. Cela est très peu probable, car les modèles sont basés sur des données qui ont été collectées pour les besoins spécifiques de la cohorte. Les données sont donc enregistrées en raison de leur lien déjà connu, ou hypothétique, avec le processus neurodégénératif. Une deuxième utilisation potentielle des modèles de progression des maladies est d'identifier de manière plus précise quand, comment et à qui administrer un médicament. Cela est particulièrement utile à l'ère de la médecine personnalisée. Un dernier contexte d'application est l'amélioration des essais cliniques. Récemment, la Food and Drug Administration des États-Unis a approuvé l'utilisation de PROCOVA[88], une méthode où l'utilisation d'un score de progression de la maladie pronostique dans une analyse de covariance permet un essai plus puissant, ou de manière équivalente, permet une taille d'échantillon plus réduite.

[87] R. V. Marinescu, N. P. Oxtoby, A. L. Young, *et al.*, "TADPOLE Challenge: Accurate Alzheimer's disease prediction through crowdsourced forecasting of future data," eng, *PRedictive Intelligence in MEdicine. PRIME (Workshop)*, vol. 11843, pp. 1–10, Oct. 2019.

[88] D. Bertolini, K. Arnemann, D. Hall, *et al.*, "Machine Learning Enables Smaller ALS Clinical Trials (P1-13.003)," en, *Neurology*, vol. 98, no. 18 Supplement, May 2022.

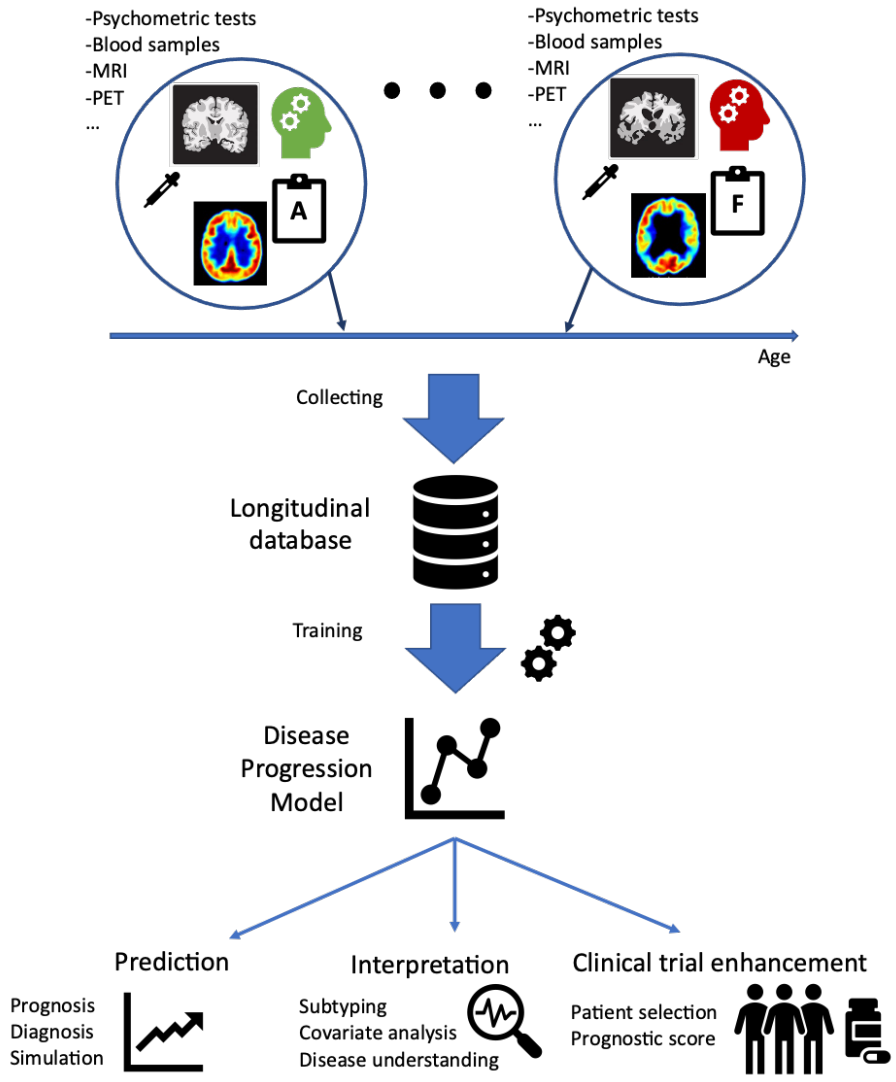

Figure 2: Représentation schématique des modèles de progression. Des données de diverses natures répétées au cours de la vie du patient sont collectées dans des bases de données longitudinales qui servent à entraîner des modèles de progression.

Modèles de progression de la maladie

Découlant du modèle hypothétique de[89], des modèles de progression des maladies ont émergé pour valider l'hypothèse et fournir un modèle basé sur les données en tant que preuve. Une classe spécifique de modèles se concentre sur l'ordre de la progression des anomalies des biomarqueurs. Ces modèles sont appelés modèles basés sur les événements (EBM), et remontent à[90]. Leur application aux maladies neurodégénératives utilise des données transversales pour déterminer l'ordre des événements[91]. Lorsque les marqueurs sont continus, les événements sont définis en fixant des seuils d'anomalie, ce qui permet d'inclure tout type de données dans les modèles. Les EBM ont été étendus pour tenir compte de l'hétérogénéité dans l'algorithme d'inférence des sous-types et des stades (SUSTAIN)[92]. Cependant, la séquence des événements ne fournit que des stades de la maladie, et la chronologie de la maladie n'est pas explicitement modélisée.

De nombreux modèles se sont concentrés sur l'alignement temporel en tant que pierre angulaire pour estimer la progression de la maladie. Un travail précoce a proposé une correspondance linéaire pour extraire la chronologie de la maladie à partir des observations individuelles[93]. Cela a ensuite été amélioré par l'introduction de déformations temporelles[94]. La construction de la chronologie commune de la maladie en faisant correspondre l'âge des patients sur un même axe, par le biais de ce qu'on appelle une reparamétrisation temporelle, considère que les observations des individus sont des aperçus d'un schéma de progression commun global, comme dans[95]. Dans cette étude, la trajectoire moyenne a été estimée sans aucune hypothèse de forme préalable, en utilisant uniquement les aperçus individuels réalignés. Cette chronologie commune de la maladie peut être considérée comme une version continue de la mise en scène de la maladie, un score de progression de la maladie[96].

La principale classe de modèles longitudinaux utilisés dans la modélisation de la progression des maladies a traditionnellement été la classe des modèles à effets mixtes[97,98]. Ce type de modèles a une

[89] C. R. Jack, D. S. Knopman, W. J. Jagust, *et al.*, "Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade," English, *The Lancet Neurology*, vol. 9, no. 1, pp. 119–128, Jan. 2010.

[90] L. A. Beckett, "Maximum Likelihood Estimation in Mallows's Model Using Partially Ranked Data," en, in *Probability Models and Statistical Analyses for Ranking Data*, M. A. Fligner and J. S. Verducci, Eds., ser. Lecture Notes in Statistics, New York, NY: Springer, 1993, pp. 92–107.

[91] H. M. Fonteijn, M. Modat, M. J. Clarkson, *et al.*, "An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease," eng, *NeuroImage*, vol. 60, no. 3, pp. 1880–1889, Apr. 2012.

[92] A. L. Young, R. V. Marinescu, N. P. Oxtoby, *et al.*, "Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference," en, *Nature Communications*, vol. 9, no. 1, pp. 1–16, Oct. 2018.

[93] E. Yang, M. Farnum, V. Lobanov, *et al.*, "Quantifying the pathophysiological timeline of Alzheimer's disease," eng, *Journal of Alzheimer's disease: JAD*, vol. 26, no. 4, pp. 745–753, 2011.

[94] S. Durrleman, X. Pennec, A. Trouvé, *et al.*, "Toward a Comprehensive Framework for the Spatiotemporal Statistical Analysis of Longitudinal Shape Data," en, *International Journal of Computer Vision*, vol. 103, no. 1, pp. 22–59, May 2013.

[95] M. C. Donohue, H. Jacqmin-Gadda, M. Le Goff, *et al.*, "Estimating long-term multivariate progression from short-term data," *Alzheimer's & dementia : the journal of the Alzheimer's Association*, vol. 10, no. 0, S400–S410, Oct. 2014.

[96] V. Kmetzsch, E. Becker, D. Saracino, *et al.*, "Disease Progression Score Estimation From Multimodal Imaging and MicroRNA Data Using Supervised Variational Autoencoders," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 12, pp. 6024–6035, Dec. 2022.

[97] N. M. Laird and J. H. Ware, "Random-effects models for longitudinal data," eng, *Biometrics*, vol. 38, no. 4, pp. 963–974, Dec. 1982.

[98] M. J. Lindstrom and D. M. Bates, "Newton—Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data," *Journal of the American Statistical Association*, vol. 83, no. 404, pp. 1014–1022, Dec.

structure hiérarchique permettant de différencier les caractéristiques partagées par la population, ou les effets fixes, de la variabilité individuelle, ou les effets aléatoires. En raison de la nature non linéaire de la progression de la plupart des biomarqueurs, les modèles sont à effets mixtes non linéaires[99]. Une famille de courbes de référence est choisie pour la progression, par exemple des courbes logistiques, et avec une reparamétrisation temporelle affine, les individus sont mis en correspondance avec la trajectoire moyenne, comme dans[100]. D'autres choix de courbes de référence sont possibles, comme celles proposées par[101]. Dans le package R pour les modèles à classes latentes à effets mixtes (LCMM)[102], le modèle non linéaire peut être combiné avec l'estimation conjointe des sous-types. Les modèles à effets mixtes non linéaires décrivant la progression des marqueurs continus peuvent être couplés à la modélisation des événements, dans ce qu'on appelle des modèles conjoints[103]. Une approche intéressante utilisant des processus gaussiens (GP) pour tenir compte des effets aléatoires du modèle à effets mixtes[104,105]. L'approche Disease Course Mapping (DCM) utilisée dans[106] repose sur un modèle géométrique supposant que les observations se trouvent sur une variété riemannienne. Elle s'inspire d'un modèle géométrique antérieur[107], qui était computationnellement inabordable car le bruit était défini dans la variété.

Ces modèles non linéaires ont également été étendus à des modalités de haute dimension telles que l'imagerie. En prolongement de[108], les auteurs de[109] ont appliqué le modèle aux images dans ADNI. DIVE[110] est une approche basée sur les voxels pour modéliser l'évolution du cerveau dans les maladies neurodégénératives. Le modèle GP a été étendu à la modélisation de formes dans[111].

1988.

[99] J. Serroyen, G. Molenberghs, G. Verbeke, *et al.*, "Nonlinear Models for Longitudinal Data," *The American Statistician*, vol. 63, no. 4, pp. 378–388, Nov. 2009.

[100] B. M. Jedynak, A. Lang, B. Liu, *et al.*, "A Computational Neurodegenerative Disease Progression Score: Method and Results with the Alzheimer's Disease Neuroimaging Initiative Cohort," *NeuroImage*, vol. 63, no. 3, pp. 1478–1486, Nov. 2012.

[101] L. L. Raket, "Statistical Disease Progression Modeling in Alzheimer Disease," *Frontiers in Big Data*, vol. 3, 2020.

[102] C. Proust-Lima, V. Philipps, and B. Liquet, "Estimation of Extended Mixed Models Using Latent Classes and Latent Processes: The R Package lcmm," en, *Journal of Statistical Software*, vol. 78, pp. 1–56, Jun. 2017.

[103] B. He and S. Luo, "Joint modeling of multivariate longitudinal measurements and survival data with applications to Parkinson's disease," eng, *Statistical Methods in Medical Research*, vol. 25, no. 4, pp. 1346–1358, Aug. 2016.

[104] M. Lorenzi, X. Pennec, G. B. Frisoni, *et al.*, "Disentangling normal aging from Alzheimer's disease in structural magnetic resonance images," en, *Neurobiology of Aging*, Novel Imaging Biomarkers for Alzheimer's Disease and Related Disorders (NIBAD), vol. 36, S42–S52, Jan. 2015.

[105] M. Lorenzi, M. Filippone, G. B. Frisoni, *et al.*, "Probabilistic disease progression modeling to characterize diagnostic uncertainty: Application to staging and prediction in Alzheimer's disease," en, *NeuroImage*, Mapping diseased brains, vol. 190, pp. 56–68, Apr. 2019.

[106] J.-B. Schiratti, S. Allassonnière, A. Routier, *et al.*, "A Mixed-Effects Model with Time Reparametrization for Longitudinal Univariate Manifold-Valued Data," en, in *Information Processing in Medical Imaging*, S. Ourselin, D. C. Alexander, C.-F. Westin, *et al.*, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2015, pp. 564–575.

[107] T. Fletcher, "Geodesic Regression on Riemannian Manifolds," en, Sep. 2011, p. 75.

[108] B. M. Jedynak, A. Lang, B. Liu, *et al.*, "A Computational Neurodegenerative Disease Progression Score: Method and Results with the Alzheimer's Disease Neuroimaging Initiative Cohort," *NeuroImage*, vol. 63, no. 3, pp. 1478–1486, Nov. 2012.

[109] M. Bilgel, J. L. Prince, D. F. Wong, *et al.*, "A multivariate nonlinear mixed effects model for longitudinal image analysis: Application to amyloid imaging," en, *NeuroImage*, vol. 134, pp. 658–670, Jul. 2016.

[110] R. V. Marinescu, A. Eshaghi, M. Lorenzi, *et al.*, "DIVE: A spatiotemporal progression model of brain pathology in neurodegenerative disorders," en, *NeuroImage*, vol. 192, pp. 166–177, May 2019.

[111] C. Abi Nader, N. Ayache, P. Robert, *et al.*, "Monotonic Gaussian Process for spatio-temporal disease progression modeling in brain imaging data," eng, *NeuroImage*, vol. 205, p. 116 266, Jan. 2020.

Dans[112], le modèle DCM a été appliqué à l'imagerie cérébrale, où le cerveau est modélisé comme un maillage intégré dans une variété de formes.

La modélisation de la progression dynamique des biomarqueurs a également été réalisée à l'aide de modèles dynamiques. Ils sont basés sur les équations différentielles ordinaires[113,114,115], où le champ des EDO est estimé à partir de l'ensemble de la cohorte et est semblable aux effets fixes des modèles mixtes, tandis que les conditions initiales de la trajectoire d'un patient individuel tiennent compte des effets aléatoires. Les EDO neurales[116] utilise un réseau d'apprentissage profond pour apprendre le champ. D'autres approches d'apprentissage profond utilisant des réseaux neuronaux récurrents ont également été appliquées à la progression des maladies[117]. Ces méthodes d'apprentissage profond sont souvent moins interprétables. C'est tout le contraire du modèle causal multifactoriel proposé dans[118]. Ce modèle causal permet de réaliser des interventions, ce qui peut être exploité pour émuler une intervention thérapeutique.

Toutes ces méthodes statistiques ont en commun la capacité de décrire la progression de la maladie au niveau de la population. La plupart d'entre elles peuvent également calculer une trajectoire individualisée pour chaque patient. Les principales différences résident dans leur interprétabilité. Les modèles avec une reparamétrisation temporelle explicite fournissent une séparation de la variabilité temporelle et spatiale, ce qui est un atout pour nous. Les modèles semi-paramétriques ou non paramétriques, tels que les modèles dynamiques basés sur les EDO ou les méthodes basées sur l'apprentissage profond, manquent souvent de côté interprétable. Notre travail se concentre sur l'approche géométrique du DCM, qui offre à la fois le désentrelacement spatio-temporel et des paramètres significatifs.

Maladie de Parkinson

La majeure partie du travail de cette thèse a été développée pour des applications aux données sur la maladie de Parkinson, nous fournirons donc un bref aperçu de cette maladie, de ses causes hypothétiques, de ses manifestations et de ses traitements.

La maladie de Parkinson a été décrite pour la première fois par James Parkinson en 1817. Cette maladie neurodégénérative est la deuxième plus répandue après la maladie d'Alzheimer. Sa prévalence augmente considérablement avec l'âge, passant d'une quasi-nullité avant l'âge de 50 ans à près de 2% après 80 ans[119]. Les hommes ont deux fois plus de chances de développer la maladie

[112] I. Koval, "Learning Multimodal Digital Models of Disease Progression from Longitudinal Data : Methods & Algorithms for the Description, Prediction and Simulation of Alzheimer's Disease Progression," en, Ph.D. dissertation, Institut Polytechnique de Paris, Jan. 2020.

[113] B. O. Taddé, H. Jacqmin-Gadda, J.-F. Dartigues, *et al.*, "Dynamic modeling of multivariate dimensions and their temporal relationships using latent processes: Application to Alzheimer's disease," en, *Biometrics*, vol. 76, no. 3, pp. 886–899, 2020.

[114] N. P. Oxtoby, A. L. Young, D. M. Cash, *et al.*, "Data-driven models of dominantly-inherited Alzheimer's disease progression," *Brain*, vol. 141, no. 5, pp. 1529–1544, May 2018.

[115] K. Lahouel, M. Wells, V. Rielly, *et al.*, *Learning nonparametric ordinary differential equations from noisy data*, Feb. 2023.

[116] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, *et al.*, *Neural Ordinary Differential Equations*, Dec. 2019.

[117] M. Nguyen, T. He, L. An, *et al.*, "Predicting Alzheimer's disease progression using deep recurrent neural networks," en, *NeuroImage*, vol. 222, p. 117 203, Nov. 2020.

[118] Y. Iturria-Medina, F. M. Carbonell, R. C. Sotero, *et al.*, "Multifactorial causal model of brain (dis)organization and therapeutic intervention: Application to Alzheimer's disease," en, *NeuroImage*, vol. 152, pp. 60–77, May 2017.

[119] A. Elbaz, L. Carcaillon, S. Kab, *et al.*, "Epidemiology of Parkinson's disease," en, *Revue Neurologique*, Neuroepidemiology, vol. 172, no. 1, pp. 14–26, Jan. 2016.

que les femmes.

La maladie de Parkinson est caractérisée par la perte de neurones dopaminergiques, qui commence dans la substance noire, une région du tronc cérébral contenant presque exclusivement des neurones dopaminergiques. La dopamine est un neurotransmetteur impliqué dans le système de récompense, mais elle est également responsable de l'activation du système moteur. La diminution des niveaux de dopamine due à la neurodégénérescence entraîne tous les symptômes moteurs de la maladie de Parkinson que nous connaissons : la bradykinésie (lenteur des mouvements), la rigidité musculaire, les tremblements, qui sont les principaux symptômes[120], mais de nombreux autres symptômes sont également observés, notamment les chutes, la lenteur de la parole... La perte de neurones dopaminergiques se propage dans le temps à d'autres régions du cerveau, entraînant des symptômes non moteurs, tels que l'apathie, les hallucinations ou la dépression. Cependant, l'apparition des symptômes est un indicateur très tardif de la mort neuronale. Lorsque la maladie de Parkinson est diagnostiquée, déjà 80% des neurones dopaminergiques de la substance noire ont déjà été perdus. On pense que la neurodégénérescence est due à l'accumulation de la protéine α-synucléine, qui est actuellement la principale cible du développement de médicaments[121]. La perte de neurones dopaminergiques peut être observée grâce à la scintigraphie SPECT (tomographie par émission de photons uniques) du transporteur de dopamine.

Comme mentionné dans les défis liés aux maladies neurodégénératives, la définition de la maladie de Parkinson est complexe. Elle est actuellement clinique, à la discrétion du neurologue, mais il n'y a pas de ligne directrice claire pour le diagnostic. Les symptômes sont similaires à ceux de la maladie à corps de Lewy, ce qui conduit parfois à une révision du diagnostic précoce. Cependant, nous sommes encore loin d'une définition biologique de la maladie. L'hétérogénéité de la maladie joue un rôle important, à partir du début des symptômes[122]. Une première distinction peut être faite en fonction de l'emplacement des premiers signes moteurs. 60% des patients présentent d'abord une atteinte d'un côté du corps, les symptômes se propageant lentement vers le reste du corps au fil du temps. Des tentatives de sous-typage ont été faites pour distinguer les sous-types prodromiques précoces, basés sur les comportements liés au sommeil[123]. L'une des classifications les plus courantes basées sur les symptômes moteurs est le système dominant le tremblement (TD) par rapport au trouble de l'équilibre postural et de la marche (PIGD) proposé dans[124]. Cependant, ces sous-types ne regroupent pas tous les patients, et ils ne sont pas stables dans le temps[125] ,[126]. Certains sous-types particuliers peuvent être dérivés de variants génétiques GBA et LRKK2, qui représentent environ 5% de l'ensemble des cas de la maladie de Parkinson, les autres étant idiopathiques.

La maladie de Parkinson a bénéficié de la découverte d'un traitement au début des années 1950, appelé lévodopa. Il s'agit d'un précurseur de la dopamine, et lorsqu'il est pris régulièrement, il com-

[120] J. Jankovic, M. McDermott, J. Carter, *et al.*, "Variable expression of Parkinson's disease: A base-line analysis of the DATATOP cohort. The Parkinson Study Group," eng, *Neurology*, vol. 40, no. 10, pp. 1529–1534, Oct. 1990.

[121] C. R. Fields, N. Bengoa-Vergniory, and R. Wade-Martins, "Targeting Alpha-Synuclein as a Therapy for Parkinson's Disease," *Frontiers in Molecular Neuroscience*, vol. 12, 2019.

[122] W. J. Zetusky, J. Jankovic, and F. J. Pirozzolo, "The heterogeneity of Parkinson's disease: Clinical and prognostic implications," eng, *Neurology*, vol. 35, no. 4, pp. 522–526, Apr. 1985.

[123] R. B. Postuma, A. E. Lang, J. F. Gagnon, *et al.*, "How does parkinsonism start? Prodromal parkinsonism motor changes in idiopathic REM sleep behaviour disorder," *Brain*, vol. 135, no. 6, pp. 1860–1870, Jun. 2012.

[124] J. Jankovic, M. McDermott, J. Carter, *et al.*, "Variable expression of Parkinson's disease: A base-line analysis of the DATATOP cohort. The Parkinson Study Group," eng, *Neurology*, vol. 40, no. 10, pp. 1529–1534, Oct. 1990.

[125] T. Simuni, C. Caspell-Garcia, C. Coffey, *et al.*, "How stable are Parkinson's disease subtypes in de novo patients: Analysis of the PPMI cohort?" en, *Parkinsonism & Related Disorders*, vol. 28, pp. 62–67, Jul. 2016.

[126] R. S. Eisinger, C. W. Hess, D. Martinez-Ramirez, *et al.*, "Motor subtype changes in early Parkinson's disease," en, *Parkinsonism & Related Disorders*, vol. 43, pp. 67–72, Oct. 2017.

pense le manque de production de dopamine dans le cerveau. D'autres médicaments à base d'apport dopaminergique ont depuis été développés, tels que les agonistes dopaminergiques ou les inhibiteurs de recapture, notamment IMAOB et ICOMT. Ils reposent tous sur le même mécanisme, avec des effets secondaires différents. Par exemple, on sait que la lévodopa provoque des dyskinésies, c'est-à-dire des mouvements spontanés, et a également été liée à des blocages dans certaines études[127]. Les agonistes dopaminergiques ont plus de chances de provoquer des troubles du contrôle des impulsions (TCI), qui entraînent des comportements addictifs tels que le jeu. Le traitement des patients atteints de la maladie de Parkinson consiste aujourd'hui en un cocktail de plusieurs médicaments dopaminergiques, et les experts débattent toujours de la meilleure politique de traitement[128]. De nombreux facteurs peuvent influencer la réponse d'un patient au traitement et sa probabilité de développer des effets secondaires négatifs.

Le traitement dopaminergique est purement symptomatique, il n'a montré aucune amélioration à long terme[129]. Son efficacité diminue à mesure que la maladie progresse vers des stades plus avancés, entraînant des fluctuations entre les états "ON" et "OFF". Les états "ON" correspondent à l'état de traitement fonctionnant normalement, où les symptômes moteurs sont atténués. Les états "OFF", en revanche, surviennent lorsque l'effet du traitement s'estompe. Aux stades avancés de la maladie de Parkinson, la demi-vie du médicament dopaminergique devient très courte en raison de la quasi-absence de production par le cerveau.

Un autre traitement appelé stimulation cérébrale profonde consiste à implanter une électrode dans l'amygdale pour stimuler directement le cerveau. Ce traitement s'est révélé être une alternative puissante à la thérapie dopaminergique[130], mais il reste assez rare en raison de son caractère invasif. Le traitement dopaminergique a longtemps éclipsé la recherche d'un médicament modifiant la maladie, ce qui fait que la recherche dans ce domaine est bien inférieure à celle de la maladie d'Alzheimer.

Les cohortes longitudinales utilisées dans ce travail de thèse comprennent PPMI[131] et une cohorte française du consortium NS-PARK. Les données des cohortes sur la maladie de Parkinson sont principalement composées d'évaluations cognitives et motrices. Historiquement, l'échelle de Hoehn et Yahr[132] (de 0 à 5) était utilisée pour décrire la maladie selon des stades successifs. Aujourd'hui, l'échelle la plus fréquemment utilisée pour évaluer la progression de la maladie est l'échelle unifiée de la maladie de Parkinson du Mouvement Disorder Society (MDS-UPDRS)[133], qui regroupe de nombreux éléments évaluant différents aspects de la maladie, notamment un examen moteur et une évaluation de l'impact de la maladie sur la vie quotidienne du patient.

[127] M. Gilat, N. D'Cruz, P. Ginis, *et al.*, "Freezing of gait and levodopa," English, *The Lancet Neurology*, vol. 20, no. 7, pp. 505–506, Jul. 2021.

[128] R. M. A. de Bie, C. E. Clarke, A. J. Espay, *et al.*, "Initiation of pharmacological therapy in Parkinson's disease: When, why, and how," eng, *The Lancet. Neurology*, vol. 19, no. 5, pp. 452–461, May 2020.

[129] R. Cilia, E. Cereda, A. Akpalu, *et al.*, "Natural history of motor symptoms in Parkinson's disease and the long-duration response to levodopa," *Brain*, vol. 143, no. 8, pp. 2490–2501, Aug. 2020.

[130] A. L. Benabid, "Deep brain stimulation for Parkinson's disease," en, *Current Opinion in Neurobiology*, vol. 13, no. 6, pp. 696–706, Dec. 2003.

[131] K. Marek, D. Jennings, S. Lasch, *et al.*, "The Parkinson Progression Marker Initiative (PPMI)," en, *Progress in Neurobiology*, Biological Markers for Neurodegenerative Diseases, vol. 95, no. 4, pp. 629–635, Dec. 2011.

[132] M. M. Hoehn and M. D. Yahr, "Parkinsonism: Onset, progression, and mortality," en, *Neurology*, vol. 17, no. 5, pp. 427–427, May 1967.

[133] C. G. Goetz, B. C. Tilley, S. R. Shaftman, *et al.*, "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results: MDS-UPDRS: Clinimetric Assessment," en, *Movement Disorders*, vol. 23, no. 15, pp. 2129–2170, Nov. 2008.

Buts de la thèse

Le travail de cette thèse s'est appuyé sur le modèle de Disease Course Mapping (DCM) introduit dans[134]. Nous avons décrit certains des principaux défis posés par les maladies neurodégénératives et nous allons étendre le modèle DCM pour relever ces défis. Le modèle DCM a déjà été conçu pour tenir compte de l'alignement temporel et de la séparation spatio-temporelle. Cependant, sa conception est loin d'être parfaite. Tout d'abord, il est limité aux données continues. De plus, le modèle DCM ne tient pas compte de l'hétérogénéité des sous-types de maladies. Enfin, il n'est pas adapté aux applications dans le développement de médicaments. Notre bref aperçu de la maladie de Parkinson met en évidence ces limitations. Les symptômes, la définition de stades de la maladie et les échelles de notation sont omniprésents, et toutes les variables sont non-continues. Les sous-types sont essentiels pour modéliser la variabilité de la maladie de Parkinson. Enfin, toutes les données des cohortes sur la maladie de Parkinson proviennent de patients sous traitement dopaminergique, ce qui nécessite un modèle spécifique pour cela. Nous fournirons donc les extensions suivantes du modèle DCM :

- L'applicabilité aux données discrètes, y compris les observations binaires et ordinales.
- La structure de mélange pour découvrir les sous-types et mieux expliquer l'hétérogénéité de la maladie.
- Des modèles de traitement liés au modèle DCM pour prendre en compte l'effet du traitement et ouvrir la voie au développement de médicaments modificateurs de la maladie.

Toutes nos contributions méthodologiques sont motivées par des besoins pratiques rencontrés dans des données réelles. Elles sont donc accompagnées d'applications à des cohortes neurodégénératives. Nous avons utilisé l'interprétabilité du modèle pour fournir des informations cliniques dans l'analyse de nos résultats expérimentaux.

Plan du manuscrit

Le premier chapitre introduit le modèle Disease Course Mapping (DCM) et discute de ses limitations. La deuxième partie du chapitre présente Geodesic Bending, une extension du modèle DCM qui introduit plus de flexibilité dans le modèle en apprenant la forme de la trajectoire typique de la maladie.

Le deuxième chapitre présente notre extension du modèle DCM aux données binaires et ordinales. Nous fournissons des applications approfondies aux cohortes de la maladie de Parkinson, ainsi qu'une mise en parallèle avec la théorie de la réponse à l'item.

Le troisième chapitre aborde le défi de l'hétérogénéité des maladies neurodégénératives. Notre contribution prend la forme de deux variantes de mélange pour le modèle DCM. Nous mettons en évidence les différences entre les deux avec une application à la maladie d'Alzheimer.

Le quatrième chapitre se concentre sur la modélisation du traitement. Nous présentons deux approches de modélisation. La première consiste en un cadre générique utilisant n'importe quel

[134] J.-B. Schiratti, S. Allassonnière, A. Routier, *et al.*, "A Mixed-Effects Model with Time Reparametrization for Longitudinal Univariate Manifold-Valued Data," en, in *Information Processing in Medical Imaging*, S. Ourselin, D. C. Alexander, C.-F. Westin, *et al.*, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2015, pp. 564–575.

modèle de progression de la maladie pour estimer l'effet du traitement à partir des seules données des patients, sans nécessiter de données de témoins ni de placebo. La deuxième approche s'appelle le modèle géodésique par morceaux et est destinée aux médicaments modificateurs de la maladie, fournissant un cadre pour apprendre les changements dans la trajectoire de la maladie dus au traitement.

Le cinquième chapitre résume toutes nos contributions à l'algorithme d'estimation dans toutes les extensions précédentes. Il présente l'algorithme MCMC-SAEM utilisé pour l'estimation du DCM. Nous introduisons ensuite nos contributions sous la forme de variantes spécifiques de cet algorithme.

Chapter 1

The Disease Course Mapping model: limits and first extension

This chapter will introduce disease course mapping (DCM), a particular disease modeling technique. This model was initially introduced in the original paper[1]. We refer the reader to the related thesis work that provides an in-depth presentation of the model for further details[2]. The model was later extended and successfully applied to neurodegenerative diseases by subsequent PhD students in their thesis work[3,4,5]. We will describe the structure of the model and provide an intuition of its parameters and geometric interpretation. Then we will focus on a specific variant of the model focused on logistic curves, as they will be used throughout the entirety of this thesis. All the materials about model parameters estimation is tackled in a dedicated chapter 5, as we will discuss the algorithm and all the contributions that we brought to it in the same section. This choice was made purposefully so that the focus in these first chapters is on the model itself, the intuition behind its geometry, the interpretation of the parameters and the applications.

The second part of this chapter concentrates on an extension of the DCM model called Geodesic Bending. This extension addresses one of the limitations of the DCM model, namely the requirement to select a family of template curves. We demonstrate how this limitation can be overcome by learning the shape of disease progression trajectories.

Our contributions in this chapter are the development of the Geodesic Bending model, in collaboration with Samuel Gruffaz. This lead to the paper "Learning Riemannian metric for disease progression modeling"[6].

[1] J.-B. Schiratti, S. Allassonnière, O. Colliot, *et al.*, "A Bayesian mixed-effects model to learn trajectories of changes from repeated manifold-valued observations," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 4840–4872, Jan. 2017.

[2] J.-B. Schiratti, "Methods and algorithms to learn spatio-temporal changes from longitudinal manifold-valued observations," en, Ph.D. dissertation, Université Paris Saclay (COmUE), Jan. 2017.

[3] I. Koval, "Learning Multimodal Digital Models of Disease Progression from Longitudinal Data : Methods & Algorithms for the Description, Prediction and Simulation of Alzheimer's Disease Progression," en, Ph.D. dissertation, Institut Polytechnique de Paris, Jan. 2020.

[4] A. Bône, "Learning adapted coordinate systems for the statistical analysis of anatomical shapes. Applications to Alzheimer's disease progression modeling," en, Ph.D. dissertation, Sorbonne Université, Jan. 2020.

[5] R. Couronné, "Progression models for Parkinson's Disease," en, Ph.D. dissertation, Sorbonne Université, Sep. 2021.

[6] S. Gruffaz, P.-E. Poulet, E. Maheux, *et al.*, "Learning Riemannian metric for disease progression modeling," in

Contents of the chapter

Advances in Neural Information Processing Systems, M. Ranzato, A. Beygelzimer, Y. Dauphin, *et al.*, Eds., vol. 34, Curran Associates, Inc., 2021, pp. 23 780–23 792.

1.1 The Disease Course Mapping model

1.1.1 Notations and mixed-effect modelling

Notations

We provide here the introduction of our standard notations for a longitudinal dataset $\mathcal{D}$ and the model trajectories that will be used throughout this work. Our longitudinal dataset is composed of several measurements $(y_{ijk})_{1\leq i\leq N, 1\leq j\leq N_i, 1\leq k\leq d}$ where d features were observed on the patient i at N_i visits. We often aggregate these features as the coordinates of a single vector in a multidimensional space $\boldsymbol{y}_{ij}$. t_{ij} is the age of the patient i at the jth visit. Some observations can be missing, as the probabilistic framework of the model will allows us to ignore those. The number of visits N_i per patient is variable. Usually observations y_{ijk} are continuous values. We will use the notation η specifically for the model's predicted trajectory in a continuous space. With a longitudinal dataset, η will always be a function of time t such that the observations follow a distribution p depending on this model $\boldsymbol{y}_{ij} \sim p(\eta(t_{ij};\theta))$, where θ generally refer to model parameters. p will be called noise model. We call trajectory the curve : $t \mapsto \eta(t;\theta)$. The model can use additional covariates $\boldsymbol{X}_{ij}$ for individual i, which may depend on the time of the visit t_{ij}.

Let us introduce a simple model to use these notations.

Linear mixed effect model

Trying to model the observations as a random gaussian noise on top of a simple function of time $y_{ijk} = \eta(t_{ij}) + \epsilon_{ij}$ with $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ is too simple since the independence assumption of the noise model between all observations does not hold. Indeed the successive observations of a same individual i are correlated, and this structure in the observations should be taken into account. This is where mixed-effect models come into play.

Mixed-effect modelling[7] is the gold standard method to deal with longitudinal data. In this framework, the model η has two types of parameters $\theta = (\boldsymbol{z}_{pop}, (\boldsymbol{z}_i)_{i\in[1,N]})$. The first set of parameters $\boldsymbol{z}_{pop}$ represent common effects across the population while $\boldsymbol{z}_i$ will relate to model parameters that are specific to individual i. $\boldsymbol{z}_{pop}$ are called fixed effects, and we will equally refer to them as *population parameters*. $\boldsymbol{z}_i$ are called random effects, or alternatively *individual parameters*. Taking the previous example as a starting point, we would have the model $y_{ijk} = \eta(t_{ij}; \boldsymbol{z}_{pop}, \boldsymbol{z}_i) + \epsilon_{ij}$ with $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ and η parametrized by both fixed and random effects.

From a conceptual point of view, this type of modelling allows to distinguish between two levels. The population level is defined only by fixed effects, and the longitudinal trajectory where only fixed effects are kept, i.e. $\boldsymbol{z}_i = 0$, should be representative of the average trend over all the individuals. We often call $\gamma_0 : t \mapsto \eta(t; \boldsymbol{z}_{pop}, 0)$ the average trajectory. For neurodegenerative diseases, the average trajectory γ_0 should reflect the knowledge of the typical disease progression. These fixed effects therefore should be more important in terms of total variance explained compared to the random effects. The individual parameters on the other hand are here to adapt the average trajectory to each individual trajectory. These random effects act as slight perturbations around the average trajectory and allow for a better description at an individual level. Again, in the context of neurodegenerative diseases, this is crucial for personalized medicine by adapting the treatment to an individual specific disease progression.

[7] R. A. Fisher, "XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance.," en, *Earth and Environmental Science Transactions of The Royal Society of Edinburgh*, vol. 52, no. 2, pp. 399–433, Jan. 1919.

A linear mixed-effect model[8] is a simple longitudinal mixed-effect model which will be our reference to get the intuition on how the disease course mapping model works. We assume that observations are continuous in the first place, $\boldsymbol{y}_{ij} \in \mathbb{R}^d$, and that the noise is Gaussian. Let us write the model linear trajectory and its noise model as follows:

$$\begin{aligned} \eta(t_{ij}; \boldsymbol{X}_{ij}, \boldsymbol{z}_{pop}, \boldsymbol{z}_i) &= \boldsymbol{C}\boldsymbol{X}_{ij} + (\boldsymbol{v} + \boldsymbol{v}_i)t_{ij} + (\boldsymbol{\beta} + \boldsymbol{\beta}_i) && (1.1.1) \\ \boldsymbol{y}_{ij} &\sim \mathcal{N}(\eta(t_{ij}; \boldsymbol{z}_{pop}, \boldsymbol{z}_i), \Sigma) && (1.1.2) \end{aligned}$$

with the fixed effects (population parameters) $\boldsymbol{z}_{pop} = (\boldsymbol{v}, \boldsymbol{\beta}, \boldsymbol{C})$, the random effects (individual parameters) $\boldsymbol{z}_i = (\boldsymbol{v}_i, \boldsymbol{\beta}_i)$ and the noise covariance matrix Σ which is often assumed to be diagonal. The random effects are assumed to follow a centered distribution, typically Gaussian $\boldsymbol{z}_i \sim \mathcal{N}(0, \boldsymbol{\Omega})$ where $\boldsymbol{\Omega}$ is the covariance matrix of the random effects. $\boldsymbol{\Omega}$ can be fixed, or learned with sometimes a specified structure (generally diagonal) in which case $\boldsymbol{\Omega}$ is part of the model parameters θ.

Having in mind the challenges presented in the introduction about spatiotemporal disentangling, we introduce a new formulation of the model above. We keep the same dimension for the parameters, but we separate time-related effects from spatial effects. This means that the random intercept $\boldsymbol{\beta}_i$ is replaced by two new terms $-\tau_i \boldsymbol{v}$ and $\mathbf{w}_i \in Span(\boldsymbol{v})^\perp$. We also decompose the random slope in a multiplicative component α_i on $\boldsymbol{v}$ and a random slope change $\boldsymbol{\zeta}_i$ in $Span(\boldsymbol{v})^\perp$. The new formula is:

$$\eta(t_{ij}; \boldsymbol{X}_{ij}, \boldsymbol{z}_{pop}, \boldsymbol{z}_i) = \boldsymbol{C}\boldsymbol{X}_{ij} + \boldsymbol{\beta} + (\alpha_i t_{ij} - \tau_i)(\boldsymbol{v} + \boldsymbol{\zeta}_i) + \boldsymbol{w}_i \quad (1.1.3)$$

We thus have three terms that are related to identifiable effects:

- the covariate effect $\boldsymbol{X}_{ij}^T \boldsymbol{C}$ reflects how the trajectory η will be influenced by the covariates. Note that the effect for a given covariate X_{ijk} can also be decomposed into a temporal and a spatial component, by decomposing the k-th column of the matrix $\boldsymbol{C}$ on $Span(\boldsymbol{v})$ and $Span(\boldsymbol{v})^\perp$
- the temporal effect $\alpha_i t_{ij} - \tau_i$ such that the random temporal effect is an affine time-warping where we can identify the acceleration factor α_i and the time intercept τ_i. These two parameters do not affect the trajectory in $\mathbb{R}^d$, they only modify the speed of progression and the starting time
- the spatial effects $\boldsymbol{\zeta}_i$ and $\boldsymbol{w}_i$ which corresponds to changes of the trajectory in $\mathbb{R}^d$, where $\boldsymbol{w}_i$ is a random shift of the curve parallel to the average trajectory defined by the fixed effects and $\boldsymbol{\zeta}_i$ is a random shift of direction of the trajectory. These correspond to the intercept and slope random effects which are not related to the axis of time

For the rest of this chapter we will focus on the last two terms, assuming that we have no covariates in the model. Modelling with covariates is the purpose of another ongoing work by a colleague[9]. We will therefore simplify the equations hereafter by positing that $\boldsymbol{X}_{ij} = 0$. We introduce the time-warping function $\psi_i(t) = \alpha_i t - \tau_i$ and the spatial curve function $\gamma_i(t) = \boldsymbol{\beta} + t(\boldsymbol{v} + \boldsymbol{\zeta}_i) + \boldsymbol{w}_i$ such that the model becomes $\eta(t; \boldsymbol{z}_{pop}, \boldsymbol{z}_i) = \gamma_i(\psi_i(t))$. Notice that γ_0 corresponds to a spatial curve with $\boldsymbol{z}_i = 0$.

Figure 1.1 illustrates these two effects.

[8] N. M. Laird and J. H. Ware, "Random-effects models for longitudinal data," eng, *Biometrics*, vol. 38, no. 4, pp. 963–974, Dec. 1982.

[9] N. Fournier and S. Durrleman, "Covariate-Aware Longitudinal Modelling for Neurodegenerative Diseases," en,

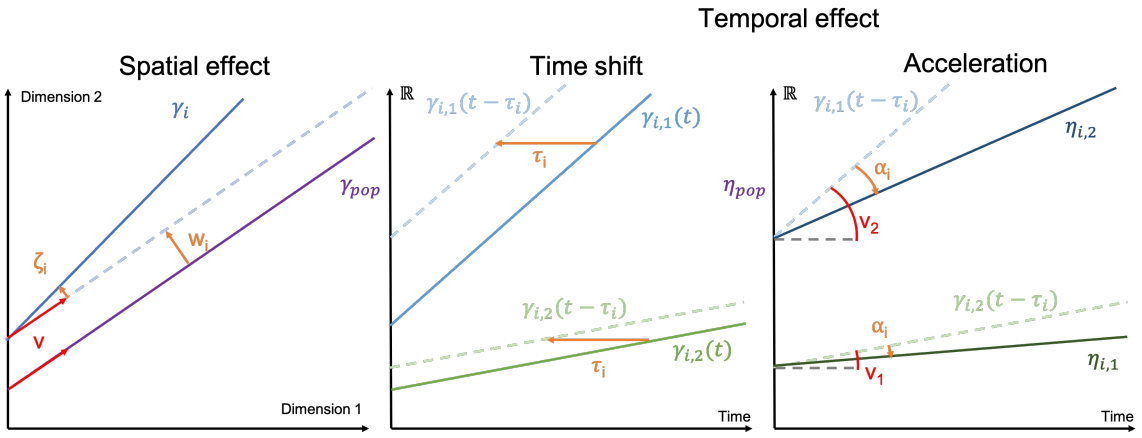

Figure 1.1: Spatial and temporal effects on the trajectory. Left plot: spatial effect, γ_{pop} is the mean trajectory defined by population parameters while γ_i is the individualized trajectory of individual i. Middle and right plot: temporal effect, showing successively the effect of time-shifting with τ_i and acceleration α_i on γ_i to obtain η_i.

As mentioned previously, trajectories for neurodegenerative markers tend to be non-linear. Therefore we need to introduce a non-linearity in the previous model.

Non-linear mixed effect model

One option would be to simply apply a non-linear deformation to the linear model above. This is the approach of generalized linear models (GLM)[10] and one that is often taken in non-linear mixed-effect modelling. For instance the LCMM library uses the linear mixed-effect model as a latent variable which is then passed through a non-linear link function such as a spline or a logit. If we denote Λ the previous linear model, the model trajectory would become $\eta(t;\theta_f,\boldsymbol{z}_{pop},\boldsymbol{z}_i) = f(\Lambda(t;\boldsymbol{z}_{pop},\boldsymbol{z}_i);\theta_f)$ where the link function f is parametrized by θ_f.

This approach is perfectly sound from a theoretical and a practical point of view. However the use of a non-linear function on top of random effects which are normally distributed in the Euclidean latent space might lead to an awkward distribution in the observation space. Especially when small random effects on the slope in the linear case can lead to large changes in the long run, which can be further amplified by the non-linear deformation, leaving the distribution of random-effects $\boldsymbol{\zeta}_i$ very sensitive to the choice of link function and prone to cause optimization issues. This remark motivated the geometric model developed by[11] which lead to the disease course mapping model.

[10] P. McCullagh, *Generalized Linear Models*, en. Routledge, Oct. 2018.

[11] J.-B. Schiratti, S. Allassonnière, A. Routier, *et al.*, "A Mixed-Effects Model with Time Reparametrization for Longitudinal Univariate Manifold-Valued Data," en, in *Information Processing in Medical Imaging*, S. Ourselin, D. C. Alexander, C.-F. Westin, *et al.*, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2015, pp. 564–575.

1.1.2 Generic disease course mapping framework

The disease course mapping model (DCM) is a non-linear mixed-effect model based on a geometric model which we describe below. The geometric model is based on Riemannian geometry, and the spatial variability is defined directly in the observation space instead of a latent space. This allows for a better control of the random effects distributions while also being more directly interpretable.

We will not delve into the concrete mathematical notions about the geometric model in this section. This has already been done in previous theses[12,13,14,15]. We prefer to provide an intuition about the model. Some introductory notions of Riemannian geometry are provided in the appendix 15. For more details about Riemannian geometry, we refer the reader to the excellent book[16]. All mathematical formulas for the DCM model used in this thesis are also found in the appendix 15, which provides a concrete summary of the notions presented in this chapter.

Model definition and intuition

The observations $\boldsymbol{y}_{ij}$ are assumed to be points belonging to a Riemannian manifold $(\mathcal{M}, g)$. A Riemannian manifold is a manifold $\mathcal{M}$ equipped with a metric g. The model considers that repeated observations in time for a subject i follow a trajectory γ_i on the manifold. The average trajectory γ_0 is assumed to be a *geodesic* in the manifold, which is the equivalent of a straight line in the Euclidean setting. Indeed a geodesic γ is the only solution to the no-acceleration differential equation $\nabla_{\dot\gamma}\dot\gamma = 0$ with specified initial conditions for $\gamma(t_0)$ and $\dot\gamma(t_0)$ (where ∇ is the Levi-Civita connection, which is equivalent to the notion of derivation in a Riemannian manifold).

The DCM model builds individual trajectories γ_i around the average trajectory γ_0. The set of all individual trajectories form a bundle of trajectories in the manifold. The random effects control the divergence of the trajectories in the bundle. In order to avoid trajectories diverging too much from each other in the space, we add a crucial hypothesis: the parallel trajectories. The DCM model stipulates that the random effects on the slopes ζ_i are negligible and should be removed. Therefore the spatial variability boils down to the shift of the curve γ_0 parallel to itself on the manifold. The notion of parallels in the DCM has been defined thanks to the Exp-parallelization 10, which extends parallels to Riemannian manifolds. The set of individual trajectories γ_i is therefore a bundle of parallel trajectories in a Riemannian sense. However note that the Exp-parallelization does not conserve the dynamic properties of the trajectory γ_0. Even though γ_0 is a geodesic, there is no guarantee that parallel trajectories γ_i are also geodesics. Figure 1.2 schematizes the process of Exp-parallelization.

This modelling implies that the choice of the Riemannian manifold and metric defines the possible trajectories. One simple choice is to use $\mathbb{R}^d$ equipped with the euclidean metric, where all the trajectories will be simple parallel straight lines. In Figure 1.3, one can see another choice of manifold and metric, leading to trajectories being logistic curves.

[12] J.-B. Schiratti, "Methods and algorithms to learn spatio-temporal changes from longitudinal manifold-valued observations," en, Ph.D. dissertation, Université Paris Saclay (COmUE), Jan. 2017.

[13] A. Bône, "Learning adapted coordinate systems for the statistical analysis of anatomical shapes. Applications to Alzheimer's disease progression modeling," en, Ph.D. dissertation, Sorbonne Université, Jan. 2020.

[14] I. Koval, "Learning Multimodal Digital Models of Disease Progression from Longitudinal Data : Methods & Algorithms for the Description, Prediction and Simulation of Alzheimer's Disease Progression," en, Ph.D. dissertation, Institut Polytechnique de Paris, Jan. 2020.

[15] R. Couronné, "Progression models for Parkinson's Disease," en, Ph.D. dissertation, Sorbonne Université, Sep. 2021.

[16] M. Do Carmo, *Riemannian Geometry*, en.

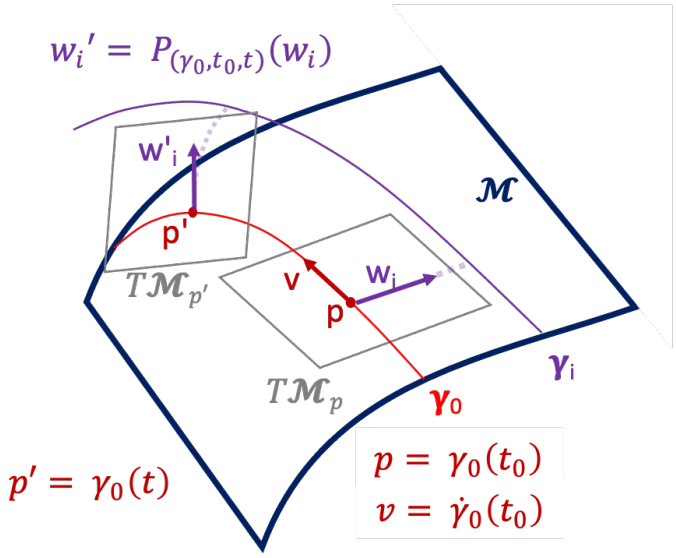

Figure 1.2: Scheme of the Exp-parallelization on a Riemannian manifold. γ_0 is the reference geodesic and γ_i is the trajectory constructed parallelly to γ_0 with the space-shift $\mathbf{w}_i$. $\mathbf{w}_i$ belongs to the tangent plane at $\mathbf{p}$, and we obtain the first point of γ_i by following the geodesic starting from $\mathbf{p}$ with derivative $\mathbf{w}_i$ (this operation is called shooting). The other points are obtained by shifting $\mathbf{w}_i$ along the γ_0 curve using parallel transport. More specifically this is done by shooting from $\gamma_0(t)$ with $P_{(\gamma_0,t_0,t)}(\mathbf{w}_i)$ which is the vector $\mathbf{w}_i$ transported from t_0 to t along the γ_0 curve.

In order to parametrize the distribution of those trajectories together, the model has a mixed-effects hierarchical structure that separates the population fixed parameters and the individual random parameters:

- **Population parameters**: the individual trajectories form a bundle centered around an "average" trajectory. This central trajectory is taken as a geodesic γ_0 in the manifold $\mathcal{M}$. This geodesic is parametrized by its initial conditions at time t_0: the initial point on the manifold $\gamma_0(t_0) = \mathbf{p}$ (thus $\boldsymbol{p} \in \mathbb{R}^d$) and the initial speed $\dot{\gamma}_0(t_0) = \mathbf{v}$ (of dimension d as well).
- **Individual parameters**: here we separate temporal and spatial parameters
  - **Spatial parameters**: each individual trajectory γ_i stems from the central geodesic γ_0 with a small shift on the manifold $\mathbf{w}_i$ called a space-shift, or inter-marker spacing. The individual follows a trajectory parallel to γ_0 in the sense of the Exp-parallelisation. The space-shift $\mathbf{w}_i$ is a vector in the tangent space at $\mathbf{p}$ (thus of dimension d) defined so that shooting with the Riemannian exponential from $\mathbf{p}$ at speed w_i yields a point of γ_i, cf Figure 1.2.
  - **Temporal parameters**: The individual temporality of the disease can be modelled with a time reparametrization $\psi_i(t)$ so that the observation $\mathbf{y}_{ij}$ is $\gamma_i(\psi_i(t_{ij}))$. The chosen time-warp is $\psi_i(t) = \alpha_i(t - t_0 - \tau_i) + t_0$, where τ_i is called the time-shift and models an early or late onset, and $\alpha_i = e^{\xi_i}$ is called the acceleration factor.

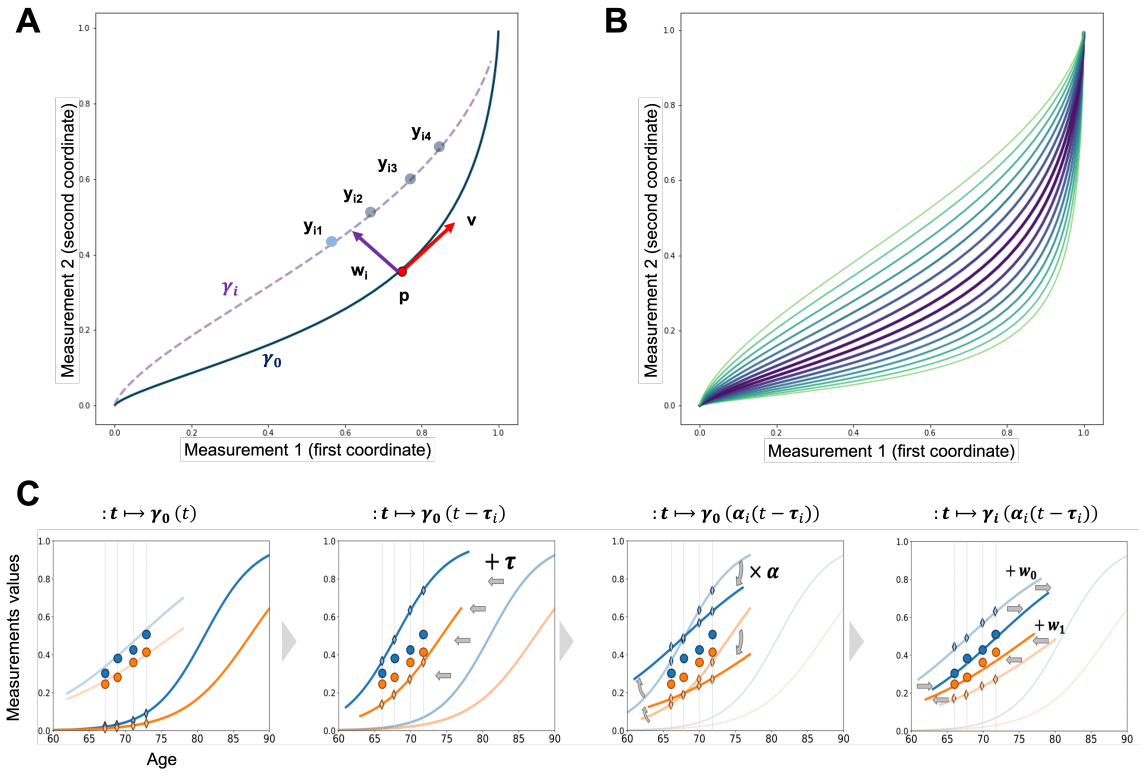

Figure 1.3: **A.** Generic model scheme. The average population trajectory γ_0 is a geodesic in the manifold. An individual trajectory γ_i is defined as an exp-parallellisation of γ_0 by a space-shift $\mathbf{w}_i$. $\mathbf{p}$ and $\mathbf{v}$ are shown on γ_0, with a space-shift $\mathbf{w}_i$ pointing to the resulting trajectory γ_i upon which lie the measurements $(\mathbf{y}_{ij})_j$. **B.** Logistic curves model. A bundle of trajectories corresponding to the logistic curves model in the 2-dimensional manifold is shown. **C.** Personalization process in the logistic curves model. The plots show the successive influence of the individual parameters on the average trajectory (left) until the trajectory is personalized and matches the measurements (right). Note that the x-axis introduces the time and the two dimensions of the measurements are shown as two different colors.

Figure 1.3 provides a visual understanding of the geometric model.

Compared to Figure 1.1, we see in that the spatial variability does not include the random effect on the slope $\boldsymbol{\zeta}_i$ in sub-figure A. Therefore we only have "parallel" trajectories, and the bundle resulting from it is shown in the sub-figure B. Sub-figure C illustrates the temporal effect, which is similar to the one observed in the linear case. The small difference is that the reference of the time axis is centered on t_0 rather than the arbitrary 0 in Figure 1.1.

For the spatial parameters, the space-shifts $(\mathbf{w}_i)_i$ have the same dimension as the observations and are a vector in the tangent space at $\mathbf{p}$, as does $\mathbf{v}$. In order to have identifiability with the time delay τ_i which operates as a shift along the curve, the space-shifts are required to be orthogonal to $\mathbf{v}$, which we enforce by restraining $\mathbf{w}_i$ to $Span(\mathbf{v})^{\perp}$, where the orthogonality is defined with the

inner product of the Riemannian manifold, which is defined for two vectors $\mathbf{x}, \mathbf{y} \in \mathcal{TM}_{\boldsymbol{p}}$ in the tangent space of a given point $\boldsymbol{p}$ depending on the metric at this point $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_p = g_p(u, v)$.

For better interpretability the model reduces the dimension of the space-shift parameter. It is estimated with an independent component analysis (ICA decomposition)[17] with N_s independent sources $(\mathbf{s}_i)_{1 \leq i \leq N_s}$. The formulation unfolds as $\mathbf{w}_i = \mathbf{A}\mathbf{s}_i$ where the columns of $\mathbf{A}$ are orthogonal to $\mathbf{v}$. The columns of the matrix $\mathbf{A}$ are defined as $A_l = \sum_{k=1}^{d-1} \beta_{lk} B_k$ a linear combination of an orthonormal basis $(B_k)_{1 \leq k \leq d-1}$ of the orthogonal hyperplane to $Span(\mathbf{v})$. This orthonormal basis is computed using a Householder method. More details on the formulas are provided in the appendix .0.45.

Note that the model is not structurally identifiable: there is indeed an infinity of triplets $(t_0, \mathbf{p}, \mathbf{v})$ defining the same geodesic γ_0. At t_0 fixed, the parameters $(\mathbf{p}, \mathbf{v})$ are unique, and so are the individual time parameters (τ_i, ξ_i). The space related parameters $\boldsymbol{A}$ and $\mathbf{s}_i$ are unique up to a permutation of the columns of the matrix $\mathbf{A}$, parametrized by the $(\beta_{lk})_{lk}$. Identifiability comes with the statistical model.

The logistic curves model

The disease course mapping model is a generic framework, and we can apply it to different types of observations by choosing the adequate manifold. For instance one can model symmetric definite positive matrices[18] or meshes[19]. For every choice, one has to compute the metric, the geodesic formulas and the Exp-parallelization formulas. However in most Riemannian manifolds, the Exp-parallelization does not yield an explicit formula, thus requiring heavy computations for each individual trajectory γ_i. This limits greatly the possible choices of manifolds. The most useful ones are the real-valued intervals manifolds with the metric resulting from the push-forward of a function f. In this case the resulting geodesics are the result of a straight line to which we applied the function f. For instance we can take the push-forward of the logit, or the exponential function, and the resulting geodesics would be logistics and exponential curves respectively.

Notice how we end up coming back to the simpler formulation of a non-linear mixed-effect model initially proposed 1.1.1, which consisted in applying a non-linearity to a linear mixed-effect model. The result of us using this fancy geometric framework is that the parametrization of the slope $\mathbf{v}$ and the space-shifts $\mathbf{w}_i$ will be adapted to the geometry of the data, with the distribution of the space-shifts having desirable properties in the manifold.

We will focus on the manifold whose geodesics are logistic curves. The choice of logistic curves also echoes a well-known hypothesis about the progression of biomarkers in Alzheimer's disease[20], but the use of logistic curves is common in most neurodegenerative diseases. This is very convenient as we will see, since the logistic is also fundamental for discrete data, which we will tackle in the next chapter. All starts with the one-dimensional manifold where the metric is the push-forward of the logit. There the Riemanninan manifold is $(0, 1)$ equipped with the metric $g_p(u, v) = uG(p)v$ where

[17] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," en, *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, Jun. 2000.

[18] J.-B. Schiratti, S. Allassonnière, O. Colliot, *et al.*, "A Bayesian mixed-effects model to learn trajectories of changes from repeated manifold-valued observations," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 4840–4872, Jan. 2017.

[19] I. Koval, "Learning Multimodal Digital Models of Disease Progression from Longitudinal Data : Methods & Algorithms for the Description, Prediction and Simulation of Alzheimer's Disease Progression," en, Ph.D. dissertation, Institut Polytechnique de Paris, Jan. 2020.

[20] C. R. Jack, D. S. Knopman, W. J. Jagust, *et al.*, "Update on hypothetical model of Alzheimer's disease biomarkers," *Lancet neurology*, vol. 12, no. 2, pp. 207–216, Feb. 2013.

$G(p) = \frac{1}{p^2(1-p)^2}$. For a d-dimensional dataset, we use the product manifold of this 1-dimensional manifold.

The specific formulation of the curve of one dimension k for one individual i at time t_{ij} in the model is the following:

$$\eta_k(\psi_i(t_{ij}); \mathbf{w}_i) = \left(1 + (\frac{1}{p_k} - 1) \exp\left(-\frac{v_k(e^{\xi_i}(t_{ij} - t_0 - \tau_i) + t_0) + w_{ik}}{p_k(1-p_k)}\right)\right)^{-1} \tag{1.1.4}$$

where (p_k, v_k, t_0) are the fixed-effects parameters for the average trajectory on item k defined by $\gamma_k(t_0) = p_k$ and $\dot{\gamma}_k(t_0) = v_k$. (τ_i, ξ_i, w_{ik}) constitute the random effects, ξ_i being the log-acceleration and τ_i the time-shift both modulating the individual disease timeline while w_{ik} is the space-shift parameter described in the previous section.

In Figure 1.3 we can observe the typical trajectories on the logistic curves manifold in 2 dimensions, as well as the trajectories as a function of time.

1.1.3 Bayesian model

The DCM model is a generative model. Therefore there is a probabilistic model on top of the geometrical model. One thing to note is that the DCM model is not structurally identifiable as there can be several triplets $(t_0, \mathbf{p}, \mathbf{v})$ describing the same set of trajectories. Identifiability comes with the probability distributions over the parameters. The authors in[21] chose the DCM model to be Bayesian, in order to have posterior distributions over the parameters rather than point estimates. We provide next the list of priors. Some new parameters are introduced to have priors over individual parameters.

Priors

We first need to introduce new notations to have priors respecting the definition domains of the variables: $p_k = \frac{1}{1+g_k}$, $g_k = e^{\tilde{g_k}}$ and $v_k = e^{\tilde{v_k}}$. This will result in a log-normal prior for g_k and v_k, while the other latent parameters will have a normal prior. Here is the complete prior list:

[Population parameters (fixed effects)]

$$\tilde{g_k} \sim \mathcal{N}(\bar{g_k}, \sigma_g^2) \tag{1.1.5}$$
$$\tilde{v_k} \sim \mathcal{N}(\bar{v_k}, \sigma_v^2) \tag{1.1.6}$$
$$\beta_{ml} \sim \mathcal{N}(\bar{\beta_{ml}}, \sigma_\beta^2) \tag{1.1.7}$$

[Individual parameters (random effects)]

$$\tau_i \sim \mathcal{N}(0, \sigma_\tau^2) \tag{1.1.8}$$
$$\xi_i \sim \mathcal{N}(0, \sigma_\xi^2) \tag{1.1.9}$$
$$s_{il} \sim \mathcal{N}(0, 1) \tag{1.1.10}$$

[21] J.-B. Schiratti, S. Allassonnière, A. Routier, *et al.*, "A Mixed-Effects Model with Time Reparametrization for Longitudinal Univariate Manifold-Valued Data," en, in *Information Processing in Medical Imaging*, S. Ourselin, D. C. Alexander, C.-F. Westin, *et al.*, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2015, pp. 564–575.

[Noise model]

$$y_{ijk} \sim \mathcal{N}(\eta_k(\psi_i(t_{ij}); \mathbf{w}_i), \sigma_k) \tag{1.1.11}$$

with many new parameters added. We provide below a summary of those parameters.

Parameter classes

We divide the parameters into three classes, which is determined by their role in the estimation algorithm, namely the MCMC-SAEM. More details about the algorithm will be provided in the dedicated chapter 5. For now it suffices to understand that it is an iterative algorithm alternating between two steps. In the first step, the expectation (E-step) is approximated stochastically by sampling some variables which are called *latent*. The second step is the maximization (M-step), which consists in maximizing the likelihood over all the non-latent variables, which are called *model parameters*. In the case of the DCM, all the parameters can thus be regrouped in three categories:

- hyperparameters $(\sigma_g, \sigma_v, \sigma_\beta)$ are fixed by the user ahead of estimation. In practical implementation they are all fixed to 0.01 by default. Their role in the prior distribution of $(\bar{g}, \bar{v}, \bar{\beta})$ is to regulate the exploration of the $(\mathbf{g}, \mathbf{v}, \beta)$ parameters during the sampling

- model parameters $\theta = (\bar{g}, \bar{v}, \bar{\beta}, t_0, \sigma_\tau, \sigma_\xi, \sigma_k)$ are all maximized during the M-step in a closed form. Intuitively $(\bar{g}, \bar{v}, \bar{\beta})$ are the mean value of $(\tilde{g}, \tilde{v}, \tilde{\beta})$, and the main purpose of these priors is to control exploration. For individual parameters $(\tau, \xi, \mathbf{s})$ we have a centered gaussian diagonal structure of the random effects. Finally σ_k is the standard deviation of the noise on item k.

- latent parameters $\boldsymbol{z} = (\boldsymbol{z}_{pop}, (\boldsymbol{z}_i)_{1 \leq i \leq N})$ are sampled during the E-step. The standard mixed-effect distinction is made between population parameters $\boldsymbol{z}_{pop} = (\tilde{g}, \tilde{v}, \beta)$ and individual parameters $\boldsymbol{z}_i = (\tau_i, \xi_i, \boldsymbol{s}_i)$.

The implementation of the DCM model can be found in the open-source library Leaspy:
https://leaspy.readthedocs.io/en/stable/
The DCM is still expanding with many new variants, including all the work presented in this manuscript. The Git link to the code is:
https://gitlab.com/icm-institute/aramislab/leaspy.

In terms of vocabulary, we use the term *calibration* to refer to the estimation of the model parameters and latent parameters on a particular dataset. We use the term *personalization* to designate the estimation of only individual parameters with a fixed set of model parameters and population parameters. Typically, we use a training dataset to calibrate a DCM model, and then we use a test set for predictions. The test set is composed only of individuals that were not used during calibration, therefore we have no individual parameters for them, hence the need for the personalization. We fix the parameters of the DCM that was already calibrated, and we estimate the best individual parameters for the patients in the test set. Finally we can perform predictions, imputation, clustering or covariate analysis for these new patients.

1.1.4 Limitations

One could argue that the logistic curves model is sufficient for the needs of most neurodegenerative applications as showcased by the results obtained in Alzheimer's disease (AD)[22]. In this work the logistic curves DCM outperformed 56 other prediction methods in the TADPOLE challenge, where the goal was to predict the changes of cognitive scores and biomarkers for AD patients several years in advance. Other successful applications have been made in Huntington's disease[23] and Parkinson's disease[24].

However this approach is still very rigid in the sense that we hypothesise the shape of the trajectories. The logistic shape can be a byproduct of a biological process model, which is the case for the accumulation of prions in the brain for instance, where an accurate mathematical model of prion spread in the cortex[25] shows that the overall quantity of misfolded proteins evolves closely to a logistic over time. For cognitive scores, the logistic also unfolds from the design of the score. Indeed they are often created such that patients are normally distributed across its scale, leading to the cumulative distribution function being akin to a logistic. But for other markers the logistic hypothesis is not properly grounded, as is the case for imaging and cerebrospinal fluid markers in AD. The model would thus benefit from a wider range of possible trajectories.

Another limitation comes from the injectivity and the product manifold. Since the model is injective and is the result of a d-dimensional product of the 1-dimensional setting, the trajectory has to be injective in each dimension. This results in the trajectories being monotonous, due to their continuity. This means that we can only select markers with a monotonous progression over time. Even though usually this is no problem with a slow progressive incurable disease, some markers that are relevant to the disease might have fluctuating levels.

1.2 Geodesic Bending

These limitations led to a work questioning the flexibility of the model, and whether we could try to learn the metric instead of imposing it. The proposed solution for this metric learning problem led to the model proposed in this section. The method was developed in collaboration with Samuel Gruffaz, which led to the paper[26] from which comes the materials of the current chapter. My main contribution to this work is in the estimation algorithm, which is described in section 5.2.2. The conception of the model and the experiments as well as the analysis of the results was an equal contribution of both of us, while the practical experiments were carried out by Samuel Gruffaz. The mathematical properties of the model can be hard to understand without some notions of Riemannian geometry, so we encourage the reader to have a look at the dedicated appendix section 15.

[22] I. Koval, A. Bône, M. Louis, *et al.*, "AD Course Map charts Alzheimer's disease progression," eng, *Scientific Reports*, vol. 11, no. 1, p. 8020, Apr. 2021.

[23] I. Koval, T. Dighiero-Brecht, A. J. Tobin, *et al.*, "Forecasting individual progression trajectories in Huntington disease enables more powered clinical trials," en, *Scientific Reports*, vol. 12, no. 1, p. 18 928, Nov. 2022.

[24] R. Couronné, A. Valladier, M. Vidhaillet, *et al.*, "Modeling the progression of Parkinson's Disease: Comparison of subjects with and without Sleep Disorders," en, p. 4,

[25] J. Weickenmeier, M. Jucker, A. Goriely, *et al.*, "A physics-based model explains the prion-like features of neurodegeneration in Alzheimer's disease, Parkinson's disease, and amyotrophic lateral sclerosis," en, *Journal of the Mechanics and Physics of Solids*, vol. 124, pp. 264–281, Mar. 2019.

[26] S. Gruffaz, P.-E. Poulet, E. Maheux, *et al.*, "Learning Riemannian metric for disease progression modeling," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, *et al.*, Eds., vol. 34, Curran Associates, Inc., 2021, pp. 23 780–23 792.

1.2.1 Introduction and related work

As debated previously, the rigid setting of the DCM model comes from the metric choice which induces the shape of the trajectory curves. The two main use cases of the DCM model are the linear case and the logistic curves model. But what if we could learn the metric instead of fixing it ? This has been attempted before with the DCM model. The first approach for metric learning was proposed in[27] by learning of the metric directly as parameters of the model and later extended in[28]. The application of the DCM to MRI images, where the manifold cannot be specified, has led to the use of a variational auto-encoder (VAE) to learn the metric implicitly in[29]. In this last paper, individual trajectories were seen as transformations of straight lines evolving in a latent space learned by the auto-encoder. Although this semi-parametric representation enabled multiple progression profiles, it was not designed for disentangling space and time variability. On top of that, both approaches were computationally heavy and estimation was burdened by a high number of parameters requiring sampling.

A standard method for trajectory learning is to use the theory of shape analysis and more especially the deformation of shapes in time. Authors commonly learn the deformation as a diffeomorphism with a Reproducing Kernel Hilbert Space (RKHS), the kernel regularity encoding the smoothness of the deformation through time[30] ,[31]. We believe that this deformation method is an interesting alternative to the auto-encoders to increase model flexibility while keeping control on space and time variability. Another approach would be to rely on a push-forward method to learn a metric with pseudo-diffeomorphic transformations[32].

In this work, following[33] but taking a step back, we propose a semi-parametric method using a RKHS to learn the Riemannian metric of DCM models. We thereby retain the possibility to learn the inter-variability of patient trajectories thanks to the mixed-effect framework, keeping the advantages of the geometric and Bayesian approach presented in[34]. First, we present the model for learning the metric. The modifications of the estimation algorithm of the DCM are tackled in the dedicated chapter 5.2.2. We validate the presented method on synthetic data and on a real dataset by comparing it with previous models on the task of predicting Alzheimer's disease patient's biomarker progression. Finally, we discuss limitations and possible future works in the discussion section.

[27] M. Louis, "Computational and statistical methods for trajectory analysis in a Riemannian geometry setting," en, Ph.D. dissertation, Sorbonne Université, Oct. 2019.

[28] B. Sauty and S. Durrleman, "Riemannian Metric Learning for Progression Modeling of Longitudinal Datasets," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, Mar. 2022, pp. 1–5.

[29] B. Sauty and S. Durrleman, "Progression Models for Imaging Data with Longitudinal Variational Auto Encoders," en, in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, L. Wang, Q. Dou, P. T. Fletcher, *et al.*, Eds., ser. Lecture Notes in Computer Science, Cham: Springer Nature Switzerland, 2022, pp. 3–13.

[30] A. Bône, M. Louis, B. Martin, *et al.*, "Deformetrica 4: An Open-Source Software for Statistical Shape Analysis," Sep. 2018.

[31] S. Joshi and M. Miller, "Landmark matching via large deformation diffeomorphisms," *IEEE Transactions on Image Processing*, vol. 9, no. 8, pp. 1357–1370, Aug. 2000.

[32] G. Lebanon, *Learning Riemannian Metrics*, Oct. 2012.

[33] M. Louis, "Computational and statistical methods for trajectory analysis in a Riemannian geometry setting," en, Ph.D. dissertation, Sorbonne Université, Oct. 2019.

[34] J.-B. Schiratti, S. Allassonnière, O. Colliot, *et al.*, "A Bayesian mixed-effects model to learn trajectories of changes from repeated manifold-valued observations," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 4840–4872, Jan. 2017.

1.2.2 Method

Notations in all the following are: $|.|$ for the usual Euclidean norm, $\langle .,. \rangle$ the Euclidean scalar product, $|||L||| = \sup_{x \in (\mathbb{R}^d)^*} \frac{|L(x)|}{|x|}$ for all linear function L and $|||df|||_\infty = \sup_x |||df(x)|||$ for all differentiable function f.

In this part, $\mathcal{M}$ is an open subset of $\mathbb{R}^d$ equipped with the Riemannian metric g. We highlight objects depending on g with the notation ^{g} because g will come to be seen as a variable in the following. For instance we note the reference geodesic γ_0^g, and an individual trajectory γ_i^g.

Traditionally, the metric g is selected using empirical arguments. This metric g is the most important since the whole model depends on it: it encodes the shape of geodesics with ^{g} and the variability due to exp-parallelisation with η^g. One of the main issues of setting the metric for a 1-dimensional manifold and then taking the product is that product metrics assume that the geodesics coordinates are independent which is not necessarily true when they describe neurodegenerative bio-markers progression[35].

Thus, we assume that the noise effect $\epsilon_{ijk} = y_{ijk} - \eta_k(\psi_i(t_{ij}); \mathbf{w}_i)$ is in reality different from the real data noise $\epsilon_{ijk}^{\text{noise}}$ because of the miss-specification of the metric g, in such a way that $\epsilon_{ijk} = \epsilon_{ijk}^g + \epsilon_{ijk}^{\text{noise}}$ with ϵ_{ijk}^g the error due to the modelling. The goal is to minimize ϵ_{ijk}^g.

To state the problem, we suppose that the population and individual parameters $\theta = (\theta, (z_i)_i)$ are estimated and we want to minimize the squared reconstruction error $L(g)$ according to g:

$$L(g) := \sum_{ijk} |\epsilon_{ijk}|^2 = \sum_{ijk} |\epsilon_{ijk}^g + \epsilon_{ijk}^{\text{noise}}| = \sum_{ij} |y_{ij} - \eta_k(\psi_i(t_{ij}); \mathbf{w}_i)|^2 \tag{1.2.1}$$

To prevent the risk of overfitting since $\epsilon_{ijk}^{\text{noise}}$ is present in L, we can add a regularizing term or constrain the set of admissible metrics g. Next, we describe a method to solve this optimization problem.

The task of learning a Riemannian metric has already been addressed in the literature[36,37,38] and most of the time, it is reduced to the task of learning a diffeomorphism using the following proposition:

Proposition 1.1 (Pushforward metrics)**.** *Provided $(\mathcal{M}, g)$ a Riemannian space, N a manifold and $\phi : \mathcal{M} \to N$ a C^1 diffeomorphism, we can equip N with the Riemannian metric g^ϕ defined as:*

$$\forall\ p \in N, \quad \forall w, v \in TN_p, \quad g_p^\phi(w, v) = g_{\phi^{-1}(p)}(d\phi^{-1}(p).w, d\phi^{-1}(p).v)$$

Moreover, ϕ is an isometry, which implies that for all (m, v) in $T\mathcal{M}$ and $\gamma : (-1, 1) \to \mathcal{M}$ a differentiable curve:

$$Exp^{g^\phi}_{\phi(m)}(d\phi(m).v) = \phi \circ Exp^g_m(v), \ P^{g^\phi}_{\phi\circ\gamma, t_0, s}(d\phi(\gamma(t_0)).v) = d\phi(\gamma(s)).P^g_{\gamma, t_0, s}(v)$$

[35] B. O. Taddé, H. Jacqmin-Gadda, J.-F. Dartigues, *et al.*, "Dynamic modeling of multivariate dimensions and their temporal relationships using latent processes: Application to Alzheimer's disease," en, *Biometrics*, vol. 76, no. 3, pp. 886–899, 2020.

[36] G. Lebanon, *Learning Riemannian Metrics*, Oct. 2012.

[37] M. Louis, "Computational and statistical methods for trajectory analysis in a Riemannian geometry setting," en, Ph.D. dissertation, Sorbonne Université, Oct. 2019.

[38] S. r. Hauberg, O. Freifeld, and M. Black, "A Geometric take on Metric Learning," in *Advances in Neural Information Processing Systems*, vol. 25, Curran Associates, Inc., 2012.

Thanks to this property, we can start from an initial metric g in $\mathcal{M}$ and reach a wide variety of Riemannian metrics by considering $(\phi(\mathcal{M}), g_0^{\phi})$ with ϕ in $\mathcal{F}$ a set of C^1-diffeomorphism. Moreover, the Exp-parallelisation $\eta^{g^{\phi}}$ can be expressed in closed form as soon as ϕ and η^g are expressed in closed form as well. Nevertheless, it restricts the possibilities because the curvature is preserved. For example if g is the Euclidean metric and $\mathcal{M} = \mathbb{R}^d$, the space $(\phi(\mathcal{M}), g^{\phi})$ will be always flat (curvature equals zero) since $(\mathcal{M}, g)$ is flat.

Learning a diffeomorphism is a more common task than learning a metric especially in the field of shape analysis with the LDDMM algorithm[39] and even in deep learning with the invertible networks[40] for different applications[41,42]. Inspired from the LDDMM framework, we chose to use the following proposition to construct the set of C^1-diffeomorphism $\mathcal{F}$:

Proposition 1.2 (Sufficient condition for a diffeomorphism)**.** *If $\phi = id + f$, with f a bounded function in $C^1(\mathbb{R}^d, \mathbb{R}^d)$ such that $|||df|||_{\infty} < 1$ (H1), then ϕ is a C^1 diffeomorphism.*

In all the following we choose $\mathcal{M} = \mathbb{R}^d$ which is the most flexible choice. For example if we want to subsequently work on $(0, 1)$, consider $\phi' = \phi \circ \sigma$ with $\sigma = \frac{1}{1+\exp(-t)}$. To respect the condition of the previous proposition, we choose to take f in an RKHS denoted $\mathcal{H}$ to encode its regularity in its norm. Moreover, the RKHS functions are handy in non-parametric optimization thank to the representer theorem (for a comprehensive review of kernel methods[43]). We derived the following conditions on the kernels to fit the constraints:

Lemma 1.1. *If k is a kernel such that $k(x, y) = g(x-y)\,\mathrm{Id}$ with $g \in C^2(\mathbb{R}^d, \mathbb{R})$ and $h(x, y) = g(x-y)$ is a bounded kernel, then $\forall f \in \mathcal{H}$ s.t $||f||_{\mathcal{H}} < \frac{1}{\sqrt{|\nabla g(0)|}}$, $|||df|||_{\infty} < 1$ and f is bounded (H1).*
More specifically for the common kernels:

- *The Gaussien kernel; $g(x) := \exp(-\frac{|x|^2}{2\sigma^2})$, $\sigma > 0$, $\frac{1}{|\nabla g(0)|} = \frac{\sigma}{\sqrt{d}}$.*
- *The Sobolev kernel or generalized T-student kernel, $g(x) = \frac{1}{(1+\frac{|x|}{2\sigma^2})^a}$, $\sigma > 0, a > d$, $\frac{1}{|\nabla g(0)|} = \frac{\sigma}{a\sqrt{d}}$*

Proof.

$$\partial_i f(x).\alpha = \partial_i k(x, .)\alpha f$$

$$df(x)^T.\alpha = (\partial_i k(x, .)\alpha f)_i$$

$$||df(x)^T.\alpha||_2^2 = \sum_{i=1}^{d} \partial_i k(x, .)\alpha f^2$$

$$\forall \alpha \in S^{d-1}, ||df(x)^T.\alpha||_2^2 \leq ||f||_H^2 \sum_{i=1}^{d} ||\partial_i k(x, .)\alpha||_H^2$$

[39] S. Joshi and M. Miller, "Landmark matching via large deformation diffeomorphisms," *IEEE Transactions on Image Processing*, vol. 9, no. 8, pp. 1357–1370, Aug. 2000.
[40] J.-H. Jacobsen, A. Smeulders, and E. Oyallon, *I-RevNet: Deep Invertible Networks*, Feb. 2018.
[41] L. Younes, "Diffeomorphic Learning," *Journal of Machine Learning Research*, vol. 21, no. 220, pp. 1–28, 2020.
[42] M. A. Rana, A. Li, D. Fox, *et al.*, "Euclideanizing Flows: Diffeomorphic Reduction for Learning Stable Dynamical Systems," en, in *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, PMLR, Jul. 2020, pp. 630–639.
[43] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," *The Annals of Statistics*, vol. 36, no. 3, pp. 1171–1220, Jun. 2008.

$$\forall \alpha \in S^{d-1}, |||df(x)|||^2 \leq ||f||_H^2 \sum_{i=1}^{d} ||\partial_i k(x,.)\alpha||_H^2$$

We can enforce $||f||_H^2 \sum_{i=1}^{d} ||\partial_i k(x,.)\alpha||_H^2 < 1$ by assuming:

$$\forall \alpha \in S^{d-1}, ||f||_H^2 < \frac{1}{\sum_{i=1}^{d} ||\partial_i k(x,.)\alpha||_H^2}$$

Using the properties of a C^2 kernel, we have:

$$||\partial_i k(x,.)\alpha||_H^2 = \alpha^T \partial_{i,i}^2 k(x,.)\alpha$$

If $k(x,y) = h(x,y)I_d$, we have:

$$\alpha^T \partial_i^2 k(x,x)\alpha = ||\alpha||^2 \partial_{i,i}^2 h(x,x) = \partial_{i,i}^2 h(x,x)$$

Then

$$||f||_H^2 < \frac{1}{\Delta h(x,x)}$$

With $h(x,y) = g(x-y)$, we have

$$||f||_H^2 < \frac{1}{\Delta g(0)}$$

□

In all the following we choose a kernel k respecting the previous conditions such that for all f in the closed ball $\bar{B}_{\mathcal{H}}(0,c)$ with c a positive constant, (H1) is verified. Note that we can prevent overfitting thanks to this constraint.

Estimation overview

If the population and random effects $(\mathbf{z}_{pop}, (\mathbf{z}_i))$ are estimated, the minimization of the error of reconstruction over the metric g can be performed by solving the following problem:

$$f^* \in \underset{f \in \bar{B}_{\mathcal{H}}(0,c)}{\operatorname{argmin}} L(f), L(f) = \sum_{i,j} |y_{i,j} - \gamma_i(t_{i,j}) - f(\gamma_i(t_{i,j}))|^2 \tag{1.2.2}$$

With the representer theorem of the RKHS, we have:

$$\forall x \in, f^*(x) := \sum_{i,j} k(x_{i,j}, x) w_{i,j}, x_{i,j} = \gamma_i(t_{i,j}) \tag{1.2.3}$$

This allows us to reduce the metric optimization problem to a convex optimization problem, which can be solved. The number of weights in the sum might be too large and cause overfitting, so we restrain the basis functions to a lower number of N_c centered on control points. More details about this in the section about optimization 5.2.2. In the end we obtain a new algorithm called alternating maximization. It alternates between several iterations of the MCMC-SAEM algorithm to optimize the DCM parameters and a metric optimization. This algorithm can be performed several times in a row to compose diffeomorphisms. We start with a simple metric g_{init} (usually the Euclidean metric or the metric resulting from the push-forward of the logit), and everytime we

apply the algorithm amounts to a composition of the previously estimated diffeomorphism ϕ with the newly estimated diffeomorphism.

For N_{comp} compositions, we have $g = g^{\Phi_{N_{\text{comp}}}}$ with $\Phi_{N_{\text{comp}}} = \phi_{N_{\text{comp}}} \circ \ldots \circ \phi_1$ where $\phi_i = id + \sum_{j=1}^{N_c} k(x_j^i, .)w_j^i$, $(x_j^i)_j$ the N_c control points and $(w_j^i)_j$ their associated weights. This structure resembles the deep neural networks considering N_c as the width and N_{comp} as the depth. The total time complexity is $\mathcal{O}(n_{\text{MCMC}} d N_c N_{\text{comp}}^3)$ which reduces in practice the choice of N_{comp}. In theory, the runtime is N_{comp} times the normal runtime of a standard DCM, but this can be improved in practice. We called this method Geodesics Bending (GB).

1.2.3 Experiments

All the methods are developed in Python by extending the open-source Leaspy library:
https://leaspy.readthedocs.io
and run on a 2.80GHz CPU with 16 GB RAM.

Synthetic data

In this part, we study the learning capacity of the method and its stability on synthetic data depending on the values of the depth N_{comp}, the width N_c and the initial metric g_{init}. We study the algorithm in the context of neurodegenerative diseases. The parameters are selected to be realistic.

In all the following, to perform the metric estimation, we choose k to be the Gaussian kernel (experiments with the Sobolev kernel gives similar results). The Gaussian kernel is parametrized by its bandwidth σ, which is an hyperparameter that requires tuning. σ controls the size of the radial basis functions, so a small σ allows for sharp local variations in the diffeomorphism while a large σ only allows slow variations. Then we chose $n_{\text{MCMC}} = 200$ MCMC-SAEM iterations between metric optimization steps ($n_{\text{MCMC}} = 10000$ at the very first step since the mixed-effects are not necessarily well initialized whereas for subsequent iterations the DCM parameters starts closer to the optimum), θ_{init} is computed empirically from patients features[44]. It leaves N_{comp}, N_c and g_{init} to select according to the situation.

Data generation Once we have selected the number of patients N_{pat} and the observations' dimension d, we generate data according to the DCM model selecting the fixed effects:

$$t_0, \mathbf{v}, \mathbf{p}, g_{\text{gen}}, \sigma_\tau^2, \sigma_\xi^2, \sigma_{\text{noise}}^2$$

and sampling the random effects parameters such that:

$$\tau_i \sim \mathcal{N}(0, \sigma_\tau^2), \xi_i \sim \mathcal{N}(0, \sigma_\xi^2), \epsilon_{i,j} \sim \mathcal{N}(0, \sigma_{\text{noise}}^2), s_i^m \sim \mathcal{N}(0,1), w_i = \sum_m s_i^m A_m$$

and time parameters $\sigma_t, \text{n}_{\text{min}}^t, \text{n}_\delta^t$ to generate the patient time-points $(t_{i,j})_{1 \le j \le T_i}$ randomly:

$$T_i = n_{\text{min}}^t + \delta_i, \delta_i \sim B(n_\delta, \frac{1}{2}), t_{i,j} \sim \mathcal{N}(t_0, \sigma_t), y_{i,j} = \gamma_i^{g^\phi}(t_{i,j}) + \epsilon_{i,j]}$$

For both experiments, we set $N_{\text{pat}} = 500,\ d = 2,\ \text{n}_{\text{min}}^t = 2,\ \text{n}_\delta^t = 6,\ \sigma_t = 4.$

[44] I. Koval, "Learning Multimodal Digital Models of Disease Progression from Longitudinal Data : Methods & Algorithms for the Description, Prediction and Simulation of Alzheimer's Disease Progression," en, Ph.D. dissertation, Institut Polytechnique de Paris, Jan. 2020.

Convergence First, we experiment whether learning the metrics with kernels enables the regression of exotic trajectories. We choose $\sigma_{\text{noise}} = 0.01$, $g_{\text{gen}} = g^{\phi}$ with $\phi(x) := \exp(0.07(x+\sin(x))-1)$ and $g = \langle ., . \rangle$ for the data generation and $N_c(\sigma = 0.08) \approx 90$ for the metric estimation.

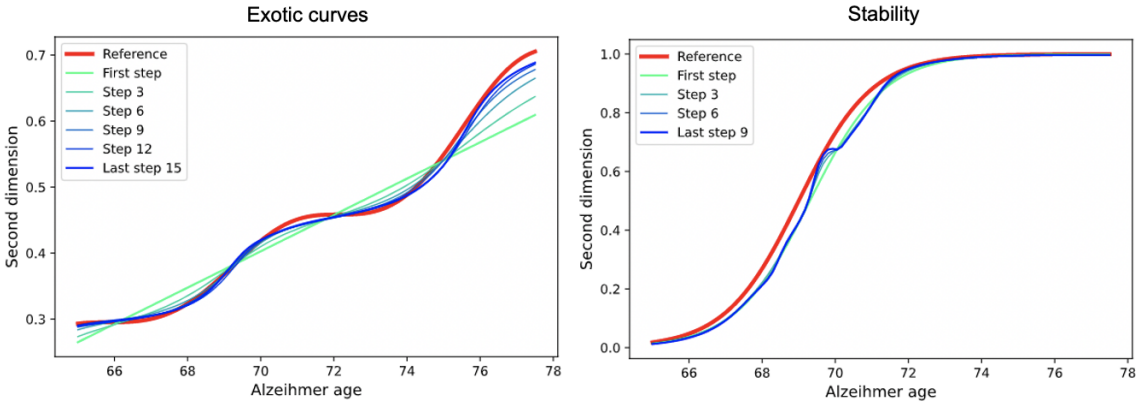

Figure 1.4: (a) The average trajectory on the second coordinate is displayed for each value of N_{comp} in the Convergence experiment. (b) The average trajectory on the second coordinate is displayed for each value of N_{comp} in the Stability experiment. Each step corresponds to a new diffeomorphism composition.

Increasing the number of compositions improves the capacity to recover the true average trajectory as we can see in Figure 1.4. The estimated noise variance $\tilde{\sigma}^2_{\text{noise}}$ decreases linearly and plateaus around the true value of noise $\sigma^2_{\text{noise}} = 0.013$ which shows that the model is a little under-fitting. The kernel regularization impacts the smoothness of the function, making it harder to fit the fast oscillations. By testing different values of the width N_c, we observed that the lower the kernel variance σ, the higher the risk to overfit. Conversely, the higher σ, the higher the risk to underfit. As for deep neural networks, there is an optimum to find regarding the hyper-parameters. In practise with datasets on neurodegenerative disease, normalized data are very noisy $\sigma^2_{\text{noise}} \approx 0.05$. To avoid over-fitting, we select the kernel variance σ in [0.1,0.5]. The choice of g_{init} influences the number of compositions required to reach a good approximation of the "true" metric.

Stability Secondly, we experiment whether the method is stable: beginning from the metric which has generated the data with g_{init}, we observe the effect of supernumerary compositions. We choose $g_{\text{gen}} = g^{\text{sig}}$ with $\text{sig}(x) := \frac{1}{1+\exp(-x)}$ and $\sigma_{\text{noise}} = 0.05$ for the data generation, $g_{init} = g^{\text{sig}}$ and $N_c(\sigma = 0.2) \approx 31$ for the metric estimation.

The compositions on ϕ produce local fluctuation on the sigmoid trend as pictured in figure 1.4. There are less observations near the inflexion point which causes over-fitting. This phenomenon is also present in datasets on neurodegenerative disease: areas of the trajectory where data are scarce can be over-fitted by GB. Nevertheless, the value of the DCM parameters are nearly constant after its first estimation in the algorithm which strengthens the model interpretability and encourages the reduction of the number of MCMC-SAEM steps n_{MCMC}.

The problem of generalization The method reveals to be flexible and quite stable provided the hyper-parameters are wisely selected. To assess whether the learning complexity of the method makes sense in practice, we propose to measure its capacity of generalization by recording its performance on prediction tasks within a 5-folds cross validation framework: running the algorithm on 4-folds, predicting on the last fold the last patient visit from its first visits, or predicting the previous visits from the last visits (as for data imputation). In the next part, these experiences are carried out on a reference dataset of neurodegenerative diseases with different types of data (cognitive scores, sub-cortical volumes from MRI, cerebro-spinal fluid biomarkers) to show the potential of a flexible method such as the one we introduced.

Real data

Introduction to ADNI The data used in this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu), 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).

First, the algorithm was tested on the subset of AD patients. For each visit, we jointly model their Mini-Mental State Examination (MMSE, a cognitive score), the volume of their hippocampi and lateral ventricles normalized by their intracranial volume (HIPP and VENTS respectively) and the concentration of amyloïd β_{1-42} in their cerebro-spinal fluid (ABETA). The goal was to understand the impact of the metric estimation on these typical AD biomarkers. Then, the model was compared to the state-of-the-art with The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) challenge[45].

An Alzheimer's disease cohort The cohort is composed of patients from all ADNI phases (ADNI1, ADNI GO, ADNI2, ADNI3), having been diagnosed with AD, with a minimum of 2 visits (4.15 visits on average and 1.5 visits for standard deviation) and $ABETA < 977pg/mL$ cutoff (so to get amyloid-positive patients only[46,47]). The dataset statistics for the selected variables are shown in Table 1.1.

Sub-cortical volumetric segmentations of MRI were performed with the Freesurfer image analysis suite v6.0.0, which is documented[48] and freely available for download online:
http://surfer.nmr.mgh.harvard.edu/.
All scores were normalized in [0,1] (to motivate the logistic prior for trajectories), using preprocessing tools from the scikit-learn library[49]: MMSE was affinely-mapped using the score bounds, other

[45] R. V. Marinescu, N. P. Oxtoby, A. L. Young, *et al.*, "TADPOLE Challenge: Accurate Alzheimer's disease prediction through crowdsourced forecasting of future data," eng, *PRedictive Intelligence in MEdicine. PRIME (Workshop)*, vol. 11843, pp. 1–10, Oct. 2019.

[46] C. R. Jack, D. A. Bennett, K. Blennow, *et al.*, "A/T/N: An unbiased descriptive classification scheme for Alzheimer disease biomarkers," eng, *Neurology*, vol. 87, no. 5, pp. 539–547, Aug. 2016.

[47] O. Hansson, J. Seibyl, E. Stomrud, *et al.*, "CSF biomarkers of Alzheimer's disease concord with amyloid- PET and predict clinical progression: A study of fully automated immunoassays in BioFINDER and ADNI cohorts," eng, *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, vol. 14, no. 11, pp. 1470–1481, Nov. 2018.

[48] B. Fischl, D. H. Salat, E. Busa, *et al.*, "Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain," eng, *Neuron*, vol. 33, no. 3, pp. 341–355, Jan. 2002.

[49] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine Learning in Python," en, *MACHINE LEARNING IN PYTHON*,

Table 1.1: Statistics for ADNI experiment. MMSE: Mini-mental state examination. MRI X ICV: intracranial volume of region X observed on MRI.

	MMSE	β-amyloïd 42	MRI Hippocampus ICV	MRI Ventricles ICV
Count	3799	1019	3395	3392
Mean	23.78	703.6	0.00379	0.0321
Standard deviation	4.70	385.6	0.00072	0.0141
Min	0	0	0.001	0.0053
25%	22	477.1	0.00327	0.0217
50%	25	612.4	0.00373	0.0301
75%	27	782.0	0.00425	0.0401
Max	30	3139.0	0.00726	0.0903

biomarkers were Box-Cox transformed[50] to un-skew their distribution, clipped between their 0.1 and 99.9 percentiles and finally affinely-mapped into [0,1]. When needed, scores were reversed such that they all increase during disease progression. The models being generative, we did not drop out missing values. We compare the DCM models using the Riemannian metric g^{sig} (baseline) and GB ($g_{\text{init}} = g^{\text{sig}}$, $N_{\text{comp}} = 6$ and $N_c(\sigma = 0.24)$), both using 2 principal directions of variation for space-shifts. Note that the baseline (DCM in yellow in Figure 1.6) is the very first iteration of GB optimization, before any diffeomorphism is applied.

Looking at the Figure 1.6, the MMSE trajectory learnt by GB has a more exponential shape compared to the DCM and the hippocampus volume trajectory has been transformed from a nearly linear curve into a piece-wise linear curve, the inter-patient variability is changed correspondingly (see Figure 1.5). It shows that there are different trajectories behaviors that DCM cannot take into account. With GB, at the beginning of the disease we observe that hippocampus volume decreases while MMSE stays constant, which is clinically coherent. We observe that the algorithm begins to stabilize after 4 steps which motivates that $N_{\text{comp}} = 6$ is a good trade-off between under-fitting and over-fitting.

Experiments: (1) Future visits prediction and (2) data imputation. For both experiments, models are trained with a 5-folds cross-validation. In the prediction experiment, we estimate, in the out-of-train fold, the individual parameters z_i on the first n_f visits of patients and predict on their very last visit. In the imputation experiment, in the out-of-train fold, we estimate the individual parameters z_i on all the patients visits except n_r visits drawn uniformly in its set of observations and predict on these n_r removed visits. We evaluate the performance of estimations with the Mean Absolute Errors (MAE) on the 5 test-sets, depending on the value of n_r and n_f varying from 1 to 2.

Results. The table 1.2 demonstrates that GB outperforms DCM except on β-amyloïd, no matter the value of n_f and n_r.

However, GB fails to generalize on β-amyloïd because of the small number of data (576 missing values out of 909 visits). When few information is available logistic prior for trajectories' shape is already almost optimal. It has been remarked that GB makes better predictions than DCM especially on patients with high scores, this fact is worth mentioning if we want to later improve

[50] G. E. P. Box and D. R. Cox, "An Analysis of Transformations," en,

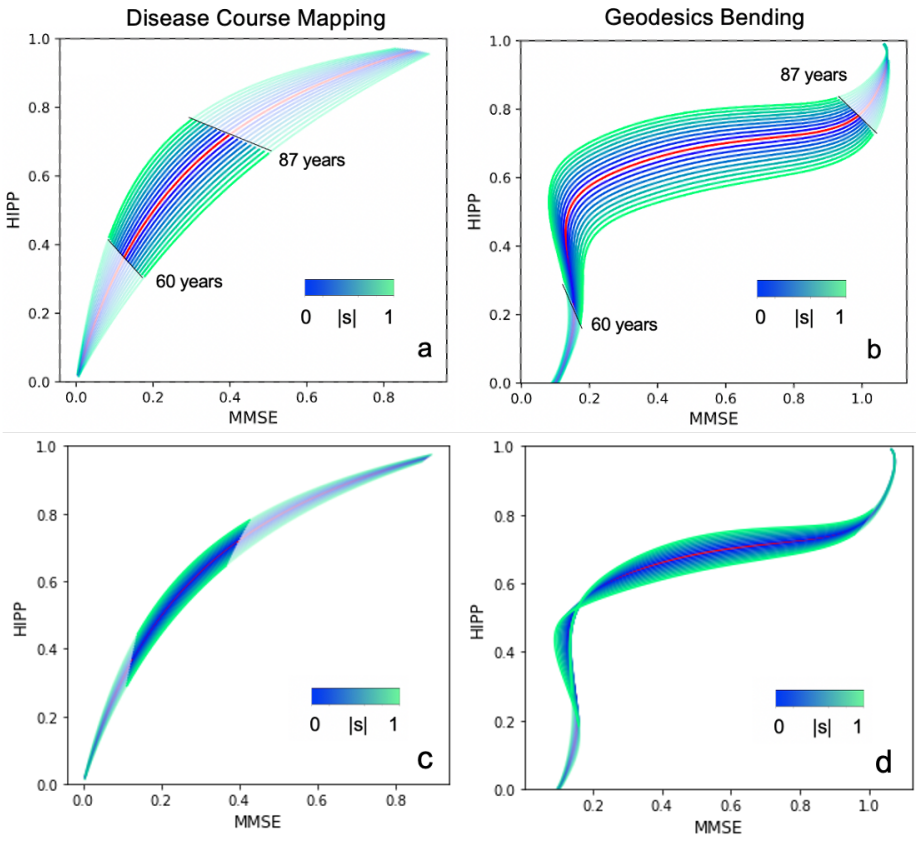

Figure 1.5: Estimated tubular coordinate system for the hippocampus volume and MMSE scores. This shows the distribution of the spatial variability around the average geodesics in red by changing the space-shifts on a the first principal direction ($w_i = sA_1, s \in [-1, 1]$) in plots **a** and **b**, and on the second direction ($w_i = sA_2, s \in [-1, 1]$) in plots **c** and **d**. 60 and 87 years are the 0.05 and 0.95 quantile for the ages ($t_{i,j}$). Notice how the second source allowed the GB model to be non-injective on the MMSE (plot **d**).

predictions with ensemble methods[51]. Now, we should assess the proposed method on a more general cohort including MCI and cognitively normal patients.

TADPOLE Challenge.[52] The TADPOLE training set is composed of data from the first three ADNI phases (ADNI 1, ADNI GO and ADNI 2). It includes approximately 1500 features acquired

[51]T. G. Dietterich, "Ensemble Methods in Machine Learning," en, in *Multiple Classifier Systems*, ser. Lecture Notes in Computer Science, Berlin, Heidelberg: Springer, 2000, pp. 1–15.

[52]R. V. Marinescu, N. P. Oxtoby, A. L. Young, *et al.*, "TADPOLE Challenge: Accurate Alzheimer's disease prediction through crowdsourced forecasting of future data," eng, *PRedictive Intelligence in MEdicine. PRIME (Workshop)*, vol. 11843, pp. 1–10, Oct. 2019.

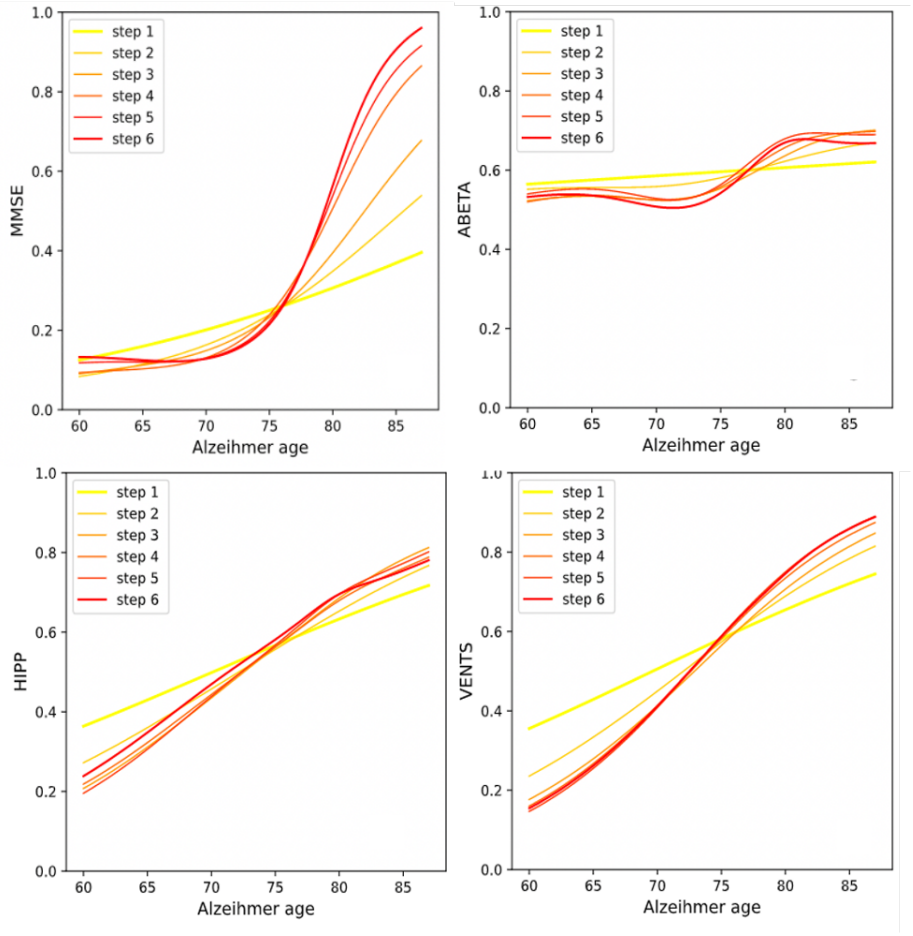

Figure 1.6: Compositions effect on the learning of each marker. The effects of each learning step on the average trajectory γ_0 is shown.

Table 1.2: Experiment performances recorded with the MAE. The proposed method is compared with the alternative using a paired, two-sided Wilcoxon signed rank test. *: $p < 0.05$, **: $p < 0.01$,***: $p < 0.001$.

	MMSE		HIPP		VENTS		ABETA	
Experiment	GB	DCM	GB	DCM	GB	DCM	GB	DCM
Pred($n_f = 1$):	0.140	0.146	**0.046*****	0.056	**0.042*****	0.056	0.124	0.114
Pred($n_f = 2$):	**0.132****	0.146	**0.037*****	0.046	**0.038*****	0.049	0.123	**0.109***
Imp($n_r = 1$):	**0.080****	0.088	**0.029*****	0.31	**0.017*****	0.020	0.128	0.123
Imp($n_r = 2$):	**0.081*****	0.089	**0.028*****	0.030	**0.023*****	0.027	0.120	**0.111****

from 1737 subjects (957 males and 780 females) during 12.741 visits for at most 22 distinct time points, between 2003 and 2017. Challengers were evaluated on their prospective predictions on enrolled individuals rolling-over for ADNI next phase, regarding the Alzheimer's Disease Assessment Scale Cognitive Score (ADAS-Cog13) and the volume of lateral ventricles normalized by the intracranial volume (VENTS), with the MAE metric. The final test set, disclosed in 2019 (time to prediction is about 2 years), is composed of 223 follow-up visits. This challenge allows for a fair comparison of the prediction performances presented here against 56 alternative methods. These methods include penalized regression, linear mixed-effect models, recurrent neural networks, and multi-task learning. Since the challenge was closed, more algorithms have been presented for the same prediction task, using the same dataset, including[53] and[54] using deep RNNs, and[55] using random forests. The forecast of ventricle volumes (VENTS) and cognitive decline (ADAS-cog) presented in this work have smaller mean absolute errors than all these competing methods.

We compared GB and DCM to the best Challenger. For the training, we select all patients having at least 2 visits and being amyloid-positive ($ABETA < 977pg/mL$[56]), it left us 765 patients with 5.5 visits on average. We use the observations of ADAS-Cog13, MMSE, hippocampus volume, ventricles volume and CSF P-TAU to calibrate the models with a 5-folds cross-validation. We apply the same transformation as previously to pre-process hippocampus volume, ventricles volume and to CSF P-TAU (Box-Cox transform, quantile clipping and min-max rescaling) as well as MMSE and ADAS-Cog13 (only affine-rescaling since scores are already bounded). For GB, we take the Gaussian kernel with $\sigma = 0.24$, $N_{\text{comp}} = 4$ and $g_{\text{init}} = g^{\text{sig}}$ with 2 principals directions for space-shifts, but 3 for DCM. This setting worked best. Nevertheless, we think that it would be interesting to select σ dimension-wise as multidimensional data may not respect the isotropy of a standard Gaussian kernel. In that way, we can choose a smaller σ for features with faster variations and conversely.

Results To know whether our results are statistically significant in the absence of other challengers' errors sample, we derive confidence intervals of DCM and GB's MAE by bootstrapping our error sample (10000 discounted draws). On figure 1.7, we observe that DCM and GB outperform the best challenger on ADAS-Cog13 and stay competitive on ventricles volume. This fact shows the potential of the geometric approach to offer interpretable and flexible models. Regarding the effect of learning the metric, GB seems to be better than DCM on ADAS-Cog13 and MMSE but not on ventricles volume. Comparing with the previous experience, it is likely that GB focuses more on the AD patients profile compared to MCI and controls. Indeed AD patients are representative of the disease progression whereas MCI and control subjects are seen by the model as slower disease progressors (delayed in time with a large time-shift τ_i and slow-paced with a low acceleration factor α_i). It highlights the limits of increasing the flexibility and encourages to combine GB with mixture models or ensemble methods in the situation of heterogeneity to control the variability.

[53] M. Nguyen, T. He, L. An, *et al.*, "Predicting Alzheimer's disease progression using deep recurrent neural networks," en, *NeuroImage*, vol. 222, p. 117 203, Nov. 2020.

[54] W. Jung, A. W. Mulyadi, and H.-I. Suk, "Unified Modeling of Imputation, Forecasting, and Prediction for AD Progression," en, in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, D. Shen, T. Liu, T. M. Peters, *et al.*, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2019, pp. 168–176.

[55] P. J. Moore, T. J. Lyons, J. Gallacher, *et al.*, "Random forest prediction of Alzheimer's disease using pairwise selection from time series data," en, *PLOS ONE*, vol. 14, no. 2, e0211558, Feb. 2019.

[56] O. Hansson, J. Seibyl, E. Stomrud, *et al.*, "CSF biomarkers of Alzheimer's disease concord with amyloid- PET and predict clinical progression: A study of fully automated immunoassays in BioFINDER and ADNI cohorts," eng,

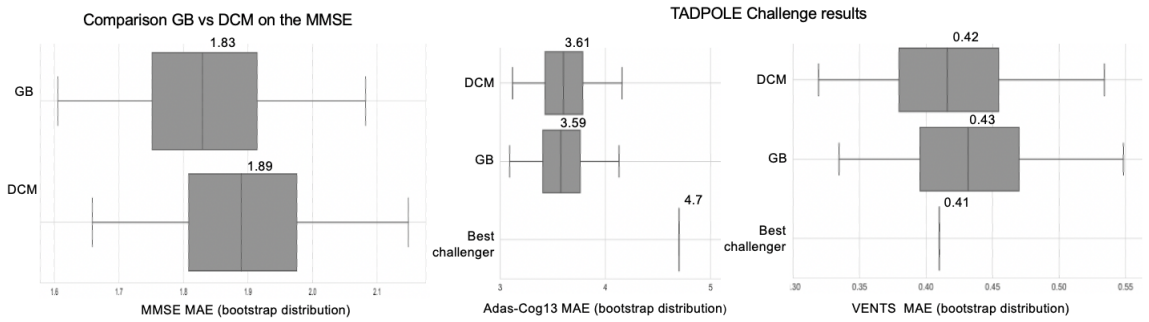

Figure 1.7: Boxplots after bootstrapping the prediction errors on the different scores. (For the boxplot : Whiskers=[2.5,97.5] %, Box=[25,50,75]%)

1.2.4 Discussion on Geodesic Bending

Riemannian metric learning applied to mixed-effect models enables to reach a sensible trade-off between flexibility and interpretability in disease progression modeling. The proposed approach allows us to disentangle time and space variability while learning the inter-patient variability and the average trajectory from the data. Without data scarcity, it proves to be efficient on homogeneous cohort for improving predictions on imaging and clinical biomarkers and suggests promising perspectives to handle heterogeneous cohorts with ensemble methods and mixture models. One of the main limitations of this work is that we now have some hyper-parameters that require careful tuning. The development of the GB model can be pursued by searching for practical guidelines for selecting hyper-parameters with empirical estimators.

1.3 Conclusion

This chapter introduced the Disease Course Mapping model, providing insights into its geometry by drawing a parallel with a simple linear mixed-effect model. We also discussed about the main limitations of the model. One limitation arises from the rigidity imposed by the choice of the metric, which determines the shapes of the possible trajectories. To address this, we introduced Geodesic Bending, an extension of the DCM where the metric is implicitly learned. With this method, we learn a diffeomorphism to compose with the starting geodesic. This results in a deformation of the initial template trajectory in order to match the observed progression patterns. The benefits of this method are sadly overshadowed by a longer optimization process and the cost of hyper-parameter tuning, which was not necessary for the DCM model. Therefore we will stick to the logistic curves DCM in the next chapters of this thesis, which already encompasses our needs for continuous modelling of the markers in neurodegenerative diseases.

Alzheimer's & Dementia: The Journal of the Alzheimer's Association, vol. 14, no. 11, pp. 1470–1481, Nov. 2018.

Chapter 2

Discrete data

Previously we focused on the geometric aspect of the model, whose aim was to better describe the progression of continuous markers on a Riemannian manifold. However features in a medical cohort are varied in nature, and often non-continuous. Symptoms and psychometric scores play a central role in neurodegenerative disease monitoring, yet they are discrete and cannot be modeled as observations on a Riemannian manifold. We will show in this chapter how to extend the Disease Course Mapping model to discrete data. First, we will provide some context on the type of discrete data that we encounter in our longitudinal cohorts. The next section will introduce the binary model and its application to symptom prediction in Parkinson's disease (PD). Then the model will be extended to ordinal data, and the experiments will showcase its potential by analysing a the most widely used clinical score in PD. Finally we will draw a parallel between the DCM extensions to discrete data and the field of Item Response Theory.

Our contributions in this chapter are the two models extending the DCM to binary and ordinal data, and both models' applications led to publications or communications. The application of the binary model was presented in an oral presentation at the Alzheimer's Disease and Parkinson's Disease conference in 2022. The ordinal model was applied to another neurodegenerative disease called ataxia, with the aim to study the Scale for the Assessment and Rating of Ataxia (SARA). This work, performed in collaboration with Paul Moulaire, delivered a clinical paper[1]. Finally the method presented in this chapter has lead to the article[2]. We included most of its content with additional experiments and explanations.

Contents of the chapter

[1] P. Moulaire, P. E. Poulet, E. Petit, *et al.*, "Temporal Dynamics of the Scale for the Assessment and Rating of Ataxia in Spinocerebellar Ataxias," en, *Movement Disorders*, vol. 38, no. 1, pp. 35–44, 2023.

[2] P.-E. Poulet and S. Durrleman, "Multivariate disease progression modeling with longitudinal ordinal data," en, *Statistics in Medicine*, vol. n/a, no. n/a,

2.1 Context: discrete data, symptoms and disease scales

Today for most neurodegenerative diseases, it is often very expensive to use sensitive early detection methods such as MRI or lumbar puncture, so these techniques are warranted mostly to confirm suspicion of diagnosis. Of course this last statement is likely to change in the next years, as research is currently heavily focused on finding reliable biomarkers for early detection that are also easy to access (mostly blood-based biomarkers). For instance in Alzheimer's disease, phospho-tau blood assays have shown promising results and will soon be implemented in clinics[3]. However for other less prevalent diseases, and even Parkinson's, it is still a long way before automatic early detection based on biomarkers is gold standard. But then, what is the first thing a general practitioner looks at when trying to diagnose a disease ? The answer is simple: symptoms. Symptoms are still at the heart of understanding a disease, because they are manifestations of the state of the patient that can be witnessed in everyday life. When talking about Alzheimer's disease, most people immediately think about the loss of memory. For Parkinson's disease, they think about tremor. Symptoms are used to characterize a disease and follow its progression. They are thus often recorded in the longitudinal follow-up of patients, under various formats. Most often they are noted on an intensity scale with a few levels, and sometimes simply written as a binary outcome present/absent.

Neurodegenerative cohorts almost always include clinical assessments of different functions such as cognition, behavior, mood or motor skills. These clinical assessments have been clinically defined specifically for each disease as surrogates for the disease timeline. The cognitive scores tend to be aggregates of subscores corresponding to several aspects of the disease. For instance in Alzheimer's disease we can mention the Alzheimer's Disease Assessment Score-Cognitive subscale[4], which is a summary of cognitive capabilities regrouped in four categories: language, praxis, memory and orientation. In Parkinson's disease, the MDS-UPDRS cognitive score[5] is composed of four parts, assessing respectively non-motor symptoms intensity, motor impact in daily life, motor assessment with tests run by the clinician and finally questions related to treatment impact. These scores are built by experts with domain knowledge.

One assessment is composed of several items, e.g. questions or graded tests. However items sometimes only cover one aspect of the disease which may not be shared by the whole population, whereas their aggregation aims at describing a unified and regular pattern of disease progression. For instance in Parkinson's disease one can identify patients with impulse control disorders[6], but this concerns only 14% of patients according to a study[7]. This information is stored as one of the many items in one of the parts of the MDS-UPDRS scale. As a more general rule, heterogeneity is better assessed through a multivariate analysis of the items before their aggregation.

The global scores aggregate responses to multiple items assessing an array of functional domains.

[3] C. E. Teunissen, I. M. W. Verberk, E. H. Thijssen, *et al.*, "Blood-based biomarkers for Alzheimer's disease: Towards clinical implementation," en, *The Lancet Neurology*, vol. 21, no. 1, pp. 66–77, Jan. 2022.

[4] J. K. Kueper, M. Speechley, and M. Montero-Odasso, "The Alzheimer's Disease Assessment Scale–Cognitive Subscale (ADAS-Cog): Modifications and Responsiveness in Pre-Dementia Populations. A Narrative Review," *Journal of Alzheimer's Disease*, vol. 63, no. 2, pp. 423–444,

[5] C. G. Goetz, B. C. Tilley, S. R. Shaftman, *et al.*, "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results: MDS-UPDRS: Clinimetric Assessment," en, *Movement Disorders*, vol. 23, no. 15, pp. 2129–2170, Nov. 2008.

[6] A. Sharma, V. Goyal, M. Behari, *et al.*, "Impulse control disorders and related behaviours (ICD-RBs) in Parkinson's disease patients: Assessment using "Questionnaire for impulsive-compulsive disorders in Parkinson's disease" (QUIP)," *Annals of Indian Academy of Neurology*, 2015.

[7] D. Weintraub, J. Koester, M. N. Potenza, *et al.*, "Impulse Control Disorders in Parkinson Disease: A Cross-Sectional Study of 3090 Patients," *Archives of Neurology*, vol. 67, no. 5, pp. 589–595, May 2010.

In contrast to it, we aim to enrich disease progression models by breaking down aggregate scores into subscores or items. In this multivariate analysis, the inter-subject variability is not only described by changes in the dynamics of progression but also by varying values for each item at a given disease stage. Disease course mapping is designed to estimate and disentangle these two sources of variability in a longitudinal data set. However, this approach has been designed for continuous measurements. The goal of this chapter is to extend this approach to binary or categorical variables. The study of such data constitutes the field of Item Response Theory (IRT), and we will draw a parallel with IRT as we proceed to the description of the model.

In IRT, a lot of the concerns of longitudinal models are not taken into account when dealing with non-continuous data. To our knowledge, there is no method in IRT which checks all the following points:

Disease timeline: the model should account for individual variability in ages. Indeed patients can have an early or late onset, and we can refer to them as slow or fast progressors. With an adequate time reparametrization, we can map all individuals to a shared disease timeline where it is easier to compare them. As mentioned before, a common disease timeline helps in understanding the standard disease duration and eventually framing disease stages.

Multivariate: the method should make use of the multiple observations in the patients' records and jointly model them. Features should not be considered independent of each other, learning the dependencies between them is a key part in understanding disease patterns.

Inter-patient variability: disease heterogeneity being a major focus, the model should provide a flexible framework for individual variability. Therefore there should be individual parameters, which are expected to be expressive enough (thus multidimensional) in order to disentangle different disease mechanisms.

In this work we provide a backbone for the longitudinal analysis of binary and ordinal data with the class of mixed-effect models[8], and more specifically the disease course mapping model[9]. We will first review the literature of disease progression models and the few attempts at modelling non-continuous data. Then we will introduce the DCM model for binary and ordinal data. In the experiemnts section we showcase the interest of item modelling rather than using aggregated scores in a simple synthetic experiment. The last part of our work is dedicated to applications of the binary and ordinal models to two medical cohorts of Parkinson's disease.

To look at how binary and ordinal data is handled in disease progression models, we need to refer to the main classes of approaches. One such class includes time-to-event modelling, especially with SuStaIn[10] which is a method mixing time-to-event to define disease stages and subtyping. SuStaIn has recently been extended to deal with ordinal data as well[11], however this method is taylored for cross-sectional data and does not account for repeated measurements of individuals. The other common framework for longitudinal modelling is the one of mixed-effects models[12]. However such

[8] N. M. Laird and J. H. Ware, "Random-effects models for longitudinal data," eng, *Biometrics*, vol. 38, no. 4, pp. 963–974, Dec. 1982.

[9] J.-B. Schiratti, S. Allassonnière, O. Colliot, *et al.*, "A Bayesian mixed-effects model to learn trajectories of changes from repeated manifold-valued observations," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 4840–4872, Jan. 2017.

[10] A. L. Young, R. V. Marinescu, N. P. Oxtoby, *et al.*, "Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference," en, *Nature Communications*, vol. 9, no. 1, pp. 1–16, Oct. 2018.

[11] A. L. Young, J. W. Vogel, L. M. Aksman, *et al.*, "Ordinal SuStaIn: Subtype and Stage Inference for Clinical Scores, Visual Ratings, and Other Ordinal Data," *Frontiers in Artificial Intelligence*, vol. 4, 2021.

[12] N. M. Laird and J. H. Ware, "Random-effects models for longitudinal data," eng, *Biometrics*, vol. 38, no. 4,

models are usually built on continuous variables such as fluid biomarkers, imaging or aggregated cognitive scores. Their extension to discrete data involves using a logit[13] or a probit model[14]. For ordinal data the cumulative probit model remains one of the gold standard for longitudinal modelling of discrete markers. One recent example of such model combine with survival analysis (joint modelling) is[15].

Dealing with items has been at the heart of Item Response Theory (IRT), from which we will borrow tools to provide an extension of the disease course mapping framework. IRT has long been favoured to deal with this type of data since the first developments in the second half of the 20th century with the works of Lord and his collaborators[16]. IRT has mostly been used for cross-sectional data (even in Parkinson's disease[17]), and extensions to longitudinal data[18] ,[19] ,[20] are less common, even though neurodegeneration is best highlighted by the collection of repeated measurements. In IRT each individual i is assumed to have a latent trait θ_i, which becomes a function of time (usually linear) in the case of longitudinal IRT. In disease modelling this latent trait stands for the disease progression score. Longitudinal IRT models have been used on medical cohorts to design new composite scores[21], however the model has too few latent parameters (only one per individual) and thus does little for the understanding of heterogeneity. Multidimensional IRT models[22] use a multidimensional latent parameter θ_i for each individual, thus allowing for more heterogeneity in the model, but these have not been applied to disease progression modelling to our knowledge.

A recent attempt[23] at bridging IRT with other frameworks has been successful, where a Bayesian non-parametric approach was used to handle the dynamics of longitudinal observations. However the approach taken to model the longitudinal aspect was to discretize time and describe the evolution of the latent trait as unidimensional with an autoregressive scheme. We want to pursue this trend with the introduction of IRT to geometric mixed-effect models.

---

pp. 963–974, Dec. 1982.

[13] T. R. T. Have, A. R. Kunselman, E. P. Pulkstenis, *et al.*, "Mixed Effects Logistic Regression Models for Longitudinal Binary Response Data with Informative Drop-Out," *Biometrics*, vol. 54, no. 1, pp. 367–383, 1998.

[14] L. C. Liu and D. Hedeker, "A Mixed-Effects Regression Model for Longitudinal Multivariate Ordinal Data," en, *Biometrics*, vol. 62, no. 1, pp. 261–268, 2006.

[15] T. Saulnier, V. Philipps, W. G. Meissner, *et al.*, "Joint models for the longitudinal analysis of measurement scales in the presence of informative dropout," en, *Methods*, vol. 203, pp. 142–151, Jul. 2022.

[16] F. M. Lord, *Applications of Item Response Theory To Practical Testing Problems*. New York: Routledge, Jul. 1980.

[17] G. Gottipati, A. Berges, S. Yang, *et al.*, "Item Response Model Adaptation for Analyzing Data from Different Versions of Parkinson's Disease Rating Scales," *Pharmaceutical Research*, vol. 36, Jul. 2019.

[18] R. Gorter, J.-P. Fox, and J. W. R. Twisk, "Why item response theory should be used for longitudinal questionnaire data analysis in medical research," *BMC Medical Research Methodology*, vol. 15, no. 1, p. 55, Jul. 2015.

[19] M. Vandemeulebroecke, B. Bornkamp, T. Krahnke, *et al.*, "A Longitudinal Item Response Theory Model to Characterize Cognition Over Time in Elderly Subjects," *CPT: Pharmacometrics & Systems Pharmacology*, vol. 6, no. 9, pp. 635–641, Sep. 2017.

[20] C. Proust-Lima, V. Philipps, B. Perrot, *et al.*, "Modeling repeated self-reported outcome data: A continuous-time longitudinal Item Response Theory model," en, *Methods*, vol. 204, pp. 386–395, Aug. 2022.

[21] L. Arrington, S. Ueckert, M. Ahamadi, *et al.*, "Performance of longitudinal item response theory models in shortened or partial assessments," en, *Journal of Pharmacokinetics and Pharmacodynamics*, vol. 47, no. 5, pp. 461–471, Oct. 2020.

[22] J. C. Immekus, K. E. Snyder, and P. A. Ralston, "Multidimensional Item Response Theory for Factor Structure Assessment in Educational Psychology Research," *Frontiers in Education*, vol. 4, p. 45, 2019.

[23] A. Cremaschi, M. De Iorio, Y. Seng Chong, *et al.*, "A Bayesian nonparametric approach to dynamic item-response modeling: An application to the GUSTO cohort study," en, *Statistics in Medicine*, vol. 40, no. 27, pp. 6021–6037, 2021.

2.2 Binary model

2.2.1 Equation

Based on the logistic curves model we will now apply it to non-continuous data, starting from simple categorical data. The basic framework is when observations y_{ijk} are binary: either 0 or 1. The model adaptation is rather simple: the observations y_{ijk} are not treated as a noisy measure of the true trajectory $\eta_{ik}(\psi_i(t_{ij}))$ but as a realization of a random Bernoulli variable with probability $\eta_{ik}(\psi_i(t_{ij}))$.

$$y_{ijk} \sim \mathcal{B}\left(\eta_{ik}(\psi_i(t_{i,j}))\right) \tag{2.2.1}$$

As in the framework of IRT, if we consider the value to be the result of one item in a test as right or wrong (0 or 1), we can evaluate the observed result as a probabilistic outcome relying on a latent disease state passing through a logit model.

2.2.2 Application to symptom's prediction

The binary version of the DCM model was motivated by a specific cohort for Parkinson's disease (PD) coming from the NS-Park consortium. This dataset was recording only symptoms, most of which were encoded as 0 or 1. The data at the time of study came from 25 PD expert centers in France and gathered 18018 patients with a total of 33798 visits. Patients were included around 68 years old and followed for a mean duration of 1.8 years. The inclusion criterion was a PD diagnosis, for which the mean age was 61. It is notable that patients were included quite late in the study: about 7 years after the diagnosis. However it is known that the dopaminergic treatment in PD is a very efficient symptomatic treatment, which leads to a singular period during the first years of the disease often referred to as the "honeymoon"[24]. During this period, patients almost return to a normal life thanks to the treatment. Only a few years later does the treatment start to show signs of limited efficiency. The end of the honeymoon period after which patients start to suffer from the disease is when various symptoms start to appear, which is what the NS-Park consortium aimed at studying. Figure 2.1 summarizes the frequency of symptoms among the population.

We did not provide the raw frequency per visit since some symptoms are repeated a lot for one patient across its multiple visits while others are appearing only once and are treated afterwards (case of hallucinations for instance). This last possibility is annoying for our model since the logistic is increasing, meaning that the frequency of symptom occurrence is assumed to increase with time. As many symptoms can be treated, it would require a model being able to take treatment into account. This will be the focus of the dedicated chapter on treatment models.

Our first attempt consisted in using the full range of symptoms with a Bernoulli noise model in a prediction setting. We separated the dataset into a training set of individuals and a test set of individuals so that there is no data leakage between the two. In the test set, for each individual we divide the visits of one individual in two by setting a virtual time separation: one set constitutes the "past" visits, upon which we can perform a personalization (i.e. estimation of individual parameters), the other set constitutes the "future" visits that we try to predict.

As shown in Figure 2.2, in the second row, the predictions seem to be very good. ROC AUC range from 0.66 to 0.97 depending on the symptoms. However this is quite deceptive as in most

[24] R. Pinder, "The Honeymoon Period—And After," en, in *The Management of Chronic Illness: Patient and Doctor Perspectives on Parkinson's Disease*, R. Pinder, Ed., London: Macmillan Education UK, 1990, pp. 54–67.

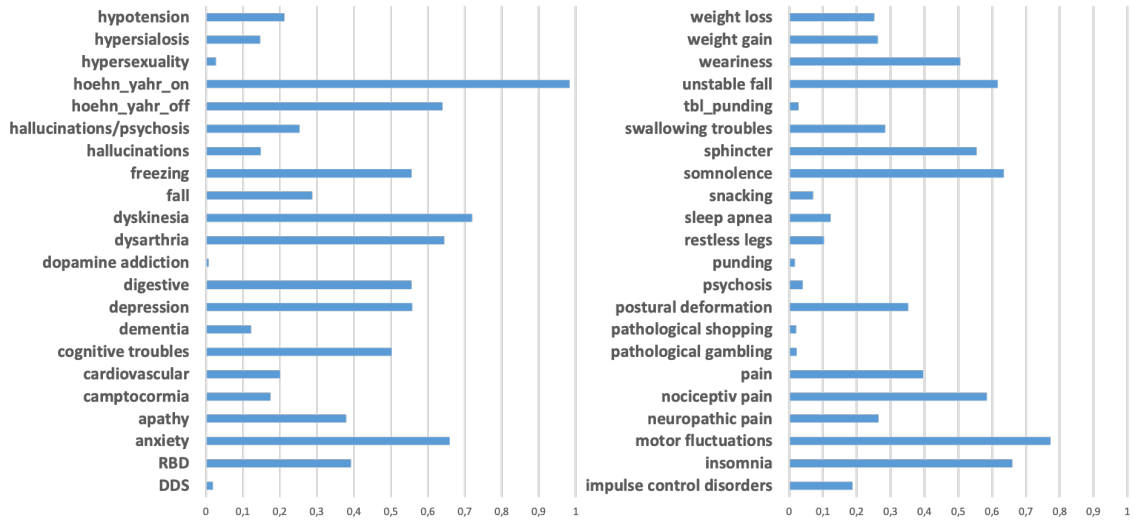

Figure 2.1: Proportion of selected individuals presenting the symptom at least once. Hoehn-Yahr is a PD severity rating scale ranging from 0 to 5 but we binarized it as ≥ 2. RBD: rapid eye movement sleep behavior disorder. DDS: dopamine dysregulation syndrome.

cases the predictions are quite easy: for most untreated symptoms, once there has been at least one occurrence before, they tend to be observed everytime afterwards. This means that once the personalization is done on an individual where one symptom has already reached 1, all the values of this symptom after are very likely to be 1 also. This explains why the AUC looks so good.

In a more realistic setting, we changed the prediction task. For each symptom, we selected patients in the test set for which we could split visits between past and future so that the symptom of interest had *never* been observed in the past. In this setting we are using the model to try and predict the first occurrence of the symptom based on all the other symptoms. This task is much more interesting for clinical practitioners as it allows them to anticipate symptom appearance and therefore adapt treatment if needed.

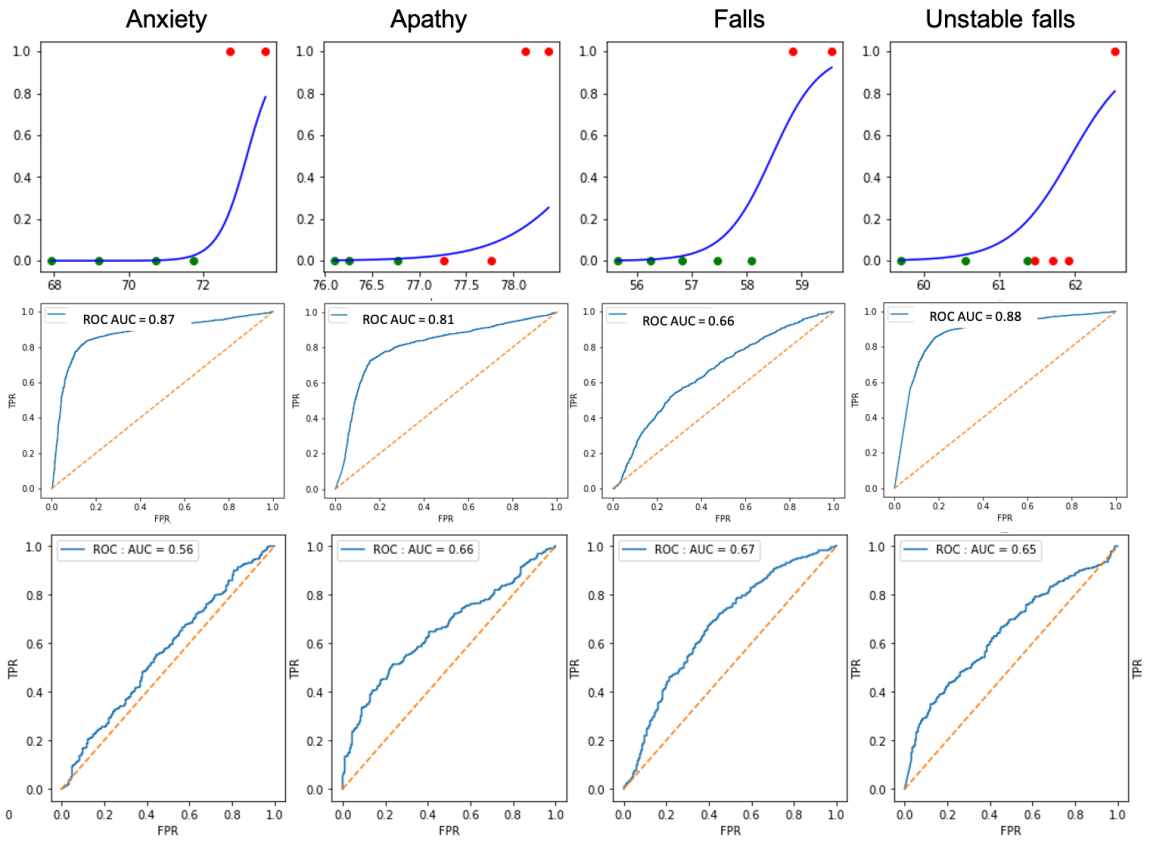

Figure 2.2: Prediction of symptoms. First row: example of prediction for one individual; green dots are past visits used to estimate individual parameters (estimation is performed with the 42 symptoms, not only the one showed) and red ones are future visits that we try to predict using the personalized DCM trajectory (blue curve). Middle row: ROC curves for this prediction task. Bottom row: ROC AUC when predicting the first occurrence of a symptom. Performances are way lower for this task.

The results are much worse as shown in the last row of Figure 2.2. The mean AUC over all symptoms is 0.64. This is to be compared to a very naive approach that would predict symptom occurrence with the median time of onset after diagnosis. Such an approach would barely have 0.54 mean AUC. We hope that these results showcase how hard this task is. We can distinguish two groups of symptoms:

- ROC AUC $\geq$ 0.6: including apathy, weariness or falls. These symptoms are partially predicted, which means that part of their variance has been captured by the model
- ROC AUC $\leq$ 0.6: including dyskinesia, RBD or postural deformation. These symptoms are un predictable using our binary DCM model without additional information

As we start to lean towards trying to detect the first occurrence of a symptom, we need to mention that this is more suited to survival analysis (see[25] for an overview of survival analysis) as we are actually dealing with events. Using the other symptoms for this task could lead to using joint modelling[26]. But survival analysis and joint modelling is outside the scope of our work in this chapter.

We propose an interesting practical approach to still use the DCM with logistic curves to predict first symptom occurrence. We first select symptoms for which this first occurrence is crucial to predict, and which is not likely to have occurred prior to study inclusion. This selection has been done after discussing the list of symptoms in depth with a Parkinson's disease expert Dr. J-C. Corvol. We retained 26 symptoms out of the 42, and then modified the data observations by setting all observations to 1 after the first occurrence. In this setting the model curves η can be interpreted as the probability of the symptom not having occurred yet. The results of the fit in this case are shown in Figure 2.3.

The average trajectory shows a clear inclusion bias (which is not enforced as showcased by the nociceptive pain curve), which is why we hypothesised that symptoms did not occur before the inclusion in the study. Otherwise we would have left censoring, i.e. we would need to take into account that the patients could have had the symptom earlier without the information whether the event occurred. These curves are similar to survival curves for each symptom. The model in this case is equivalent to fitting a logistic (but reversed) to the Kaplan-Meier curve with patients having been aligned temporally on a common disease timeline. The interpretation of the different slopes is showcased in the bottom plot of the figure. The slope defines the transition window between during which the symptom is most likely to occur for the first time, which is very different from the interpretation of $\boldsymbol{v}$ in the continuous DCM case.

To conclude on binary DCM models, we see that the application to symptom data is mitigated. Due to the variable nature of the occurrence of symptoms combined with the possibility of treatment, the DCM model has a hard time predicting the symptoms, even with a multivariate setting. Indeed there is too few information coming from binary data, which undermines the personalization step and worsens predictions. One way to improve would be to include continuous measurements along the binary ones, allowing biomarkers to guide the forecast of symptoms' first occurrence.

We need to avoid being lured into two pitfalls. The first being using a Bernoulli noise model when we are only interested in the first occurrence of a symptom. This situation warrants instead

[25] C. Kartsonaki, "Survival analysis," en, *Diagnostic Histopathology*, Mini-Symposium: Medical Statistics, vol. 22, no. 7, pp. 263–270, Jul. 2016.

[26] J. G. Ibrahim, H. Chu, and L. M. Chen, "Basic Concepts and Methods for Joint Models of Longitudinal and Survival Data," *Journal of Clinical Oncology*, vol. 28, no. 16, pp. 2796–2801, Jun. 2010.

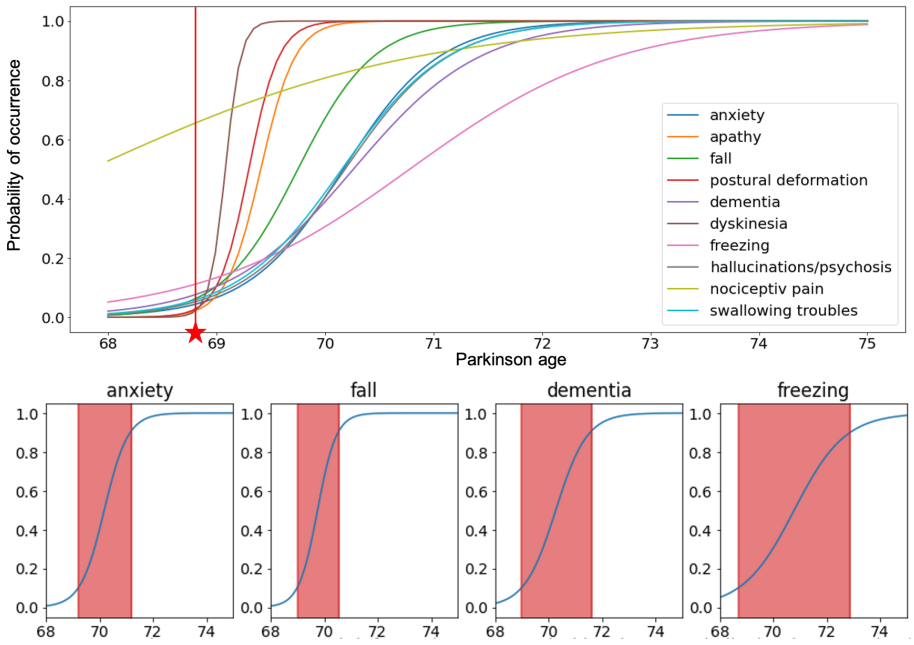

Figure 2.3: Average trajectory of the first occurrence model. Top plot: average trajectory shown for 10 of the 26 symptoms, with the inclusion date shown in red. Bottom row: average trajectory for 4 symptoms shown separately, with the "transition window" highlighted in red.

a survival model, or a joint model. The second pitfall is a computational one. When trying to optimize a model using continuous and discrete measurements at the same time, maximizing the likelihood leads to an unbalanced fit on the two types of data. Indeed, the log-likelihood of the Bernoulli leads to a binary cross-entropy while Gaussian noise leads to a mean squared error loss. These two losses are not on the same scale (one log loss and one square loss), and when looking at the gradient of those two we have the formulas .0.38 .0.39 (in appendix), where the gradient of the cross-entropy is scaled by $\frac{y-\eta}{\eta(1-\eta)}$ while the gradient of the mean squared error is scaled by $\frac{y-\eta}{\sigma^2}$. However as we mentioned, symptoms tend to be either present or absent, quickly forcing η to values very close to 0 or 1. The scaling factor thus quickly turns in favor of the binary items, meaning that the model tends to fit better the binary items. This is counter-intuitive as continuous items are more reliable and should provide more information, and thus drive the fit. This multi-loss issue currently prevents us to have decent DCM models using different data types as input. One solution would be to implement an adaptive loss scheme, using profiling weight functions (PWF) as in[27].

[27]D. Ravi, D. C. Alexander, and N. P. Oxtoby, "Degenerative Adversarial NeuroImage Nets: Generating Images that Mimic Disease Progression," en, in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, D. Shen, T. Liu, T. M. Peters, *et al.*, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2019, pp. 164–172.

2.3 Ordinal model

2.3.1 Equation

Now let's assume the measurements lie on a discontinuous space with ordered values such as for a cognitive score. Without loss of generality we assume these values are integers between 0 and L included. As for the binary case, we will again consider that the model values belonging to the manifold (which is continuous) describe a form of probability (here through the cumulative distribution function). We now specify the logistic curves model for ordinal data similar to Samejima's model (also known as cumulative logit model) from IRT[28]:

$$\begin{aligned}
\forall l \in [|1, L|], \mathbb{P}(y_{ijk} \geq l) &= \eta_{ik}(\psi_{ik}^l(t_{i,j})) \\
&= \left(1 + (\frac{1}{p_k} - 1) \exp\left(-\frac{v_k \psi_{ik}^l(t_{ij}) + w_{ik}}{p_k(1-p_k)}\right)\right)^{-1} \\
\psi_{ik}^l(t_{ij}) &= e^{\xi_i}(t_{ij} - t_0 - \tau_i) + t_0 - \sum_{m=1}^{l} \delta_k^m \\
\forall l \in [|1, L|], \delta_k^l &> 0
\end{aligned} \tag{2.3.1}$$

where ψ_{ik}^l is the time reparametrization of individual i for the item k at level l. As in the DCM model, the population parameters (p_k, v_k, t_0) define the base logistic curve (i.e. the curve for level 1 of the item) as the initial position, initial velocity and initial time respectively. τ_i and ξ_i are the time-related parameters, namely the time-shift and the log-acceleration. w_{ik} is the space-shift. δ_k^l are new parameters introduced to deal with the different levels of the item, as explained thereafter.

The idea is to model the cumulative distribution function with the logistic curves instead of the probabilities of each level directly. Each level is modelled with a logistic curve. However we do not want to overparametrize the model so we choose to enforce parallelism by setting the same velocity $\mathbf{v}$ for each level of a same item. This introduces only one parameter for each added level. The new parameter is δ_k^l, the delay in time between levels $l-1$ and l. This delay is a fixed effect, so it is shared by the whole population. Since it is added to the time reparametrization function ψ_{ik} after the affine transformation of time by τ and e^{ξ}, this time delay is a duration in the common disease timeline. This means that if one patient has an acceleration factor of 2, the expected time to jump from one level to the next is divided by 2. With the δ_k^l being positive, we also ensure that the condition $\mathbb{P}(y_{ijk} \geq l) < \mathbb{P}(y_{ijk} \geq l-1)$ is verified. From an interpretability standpoint, these delays are helpful in understanding which levels of an ordinal scale last longer than the others, or on the contrary which levels represent only a short transition.

Bayesian modelling

We will specify the probabilistic model in order to estimate the parameters by maximizing the likelihood. We set the priors as in the DCM for $\tau_i, \xi_i, \mathbf{s}_i, \mathbf{v}, \mathbf{g}, \beta$. In the ordinal model we added a new parameter: the δ parameters, which must be positive, so we set their prior distribution as a log-normal.

[28] F. Samejima, "Estimation of latent ability using a response pattern of graded scores," en, *Psychometrika*, vol. 34, no. 1, pp. 1–97, Mar. 1969.

As mentioned before, the probabilistic model allows us to deal with missing observations without needing to impute them as the likelihood is only computed for the points we observe. The robustness to the missing data as been evaluated in[29].

In terms of optimization of the log-likelihood, changing the model from a Gaussian noise to a Bernoulli variable amounts to optimizing a crossentropy between the latent model and the trajectory instead of a mean squared error:

$$\hat{\mathbf{y}}_{i,j} = \eta_i(\psi_i(t_{i,j})) \tag{2.3.2}$$

[Simplified notation]

$$\log(p(\mathbf{y}|\mathbf{z};\theta)) = -N_{tot}\log(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{N_i}||\mathbf{y}_{i,j} - \hat{\mathbf{y}}_{i,j}||^2 \tag{2.3.3}$$

[Continuous model with Gaussian noise]

$$\log(p(\mathbf{y}|\mathbf{z};\theta)) = \sum_{i=1}^{n}\sum_{j=1}^{N_i}\sum_{k=1}^{d} y_{i,j,k}\log(\hat{y}_{i,j,k}) + (1-y_{i,j,k})\log(1-\hat{y}_{i,j,k}) \tag{2.3.4}$$

[Categorical model with Bernoulli realization]

For the ordinal model, maximizing the log-likelihood is straightforward, as we can simply come back to $\mathbb{P}(y_{ijk} = l)$ by a difference $\mathbb{P}(y_{ijk} \geq l) - \mathbb{P}(y_{ijk} \geq l+1)$. One thing to note here is that the log-likelihood cannot be expressed in the curved exponential family and we are outside the scope of the theoretical guarantees for the MCMC-SAEM convergence. However, convergence has always been reached in practice.

Initialization method

Since the estimation algorithm will only converge towards a local optimum of the likelihood, we need to ensure that either we replicate the estimation with several initial points or we provide a starting point which is already a "good guess". For the population parameters of the logistic curves model, we compute the initial point for $(\mathbf{p}, \mathbf{v})$ by computing a regression on the individual progression above the first level, with $\mathbf{p}$ being deduced from the intercept and $\mathbf{v}$ from the slope. For the $(\delta_k^l)_{k,l}$, we computed them as the mean time to reach level l from level $l-1$. Indeed each δ_k^l can be understood as the time between the progression of the probabilities $\mathbb{P}(y_{ijk} \geq l)$ and $\mathbb{P}(y_{ijk} \geq l-1)$.

Implementation details

All the code for the ordinal disease course mapping model is already available on Gitlab:
https://gitlab.com/icm-institute/aramislab/leaspy

2.3.2 Experiments: Simulation study

First part: In this first experience we generate synthetic ordinal data in order to highlight two benefits of our modelling approach. On the one hand, we show how modelling items leads to a

[29] R. Couronné, M. Vidailhet, J.-C. Corvol, *et al.*, "Learning disease progression models with longitudinal data and missing values," in *ISBI 2019 - International Symposium on Biomedical Imaging*, Venice, Italy, Apr. 2019.

finer-grained description of the disease progression, as compared to modelling aggregated scores. In particular, we show that a model based on item response exhibits patients clusters that were not visible with a continuous model. On the other hand, we compare how the ordinal model fares as opposed to the logistic curves model. We show that the flexibility introduced with the δ parameters allows to learn a step function which is less rigid than an imposed template such as a logistic curve.

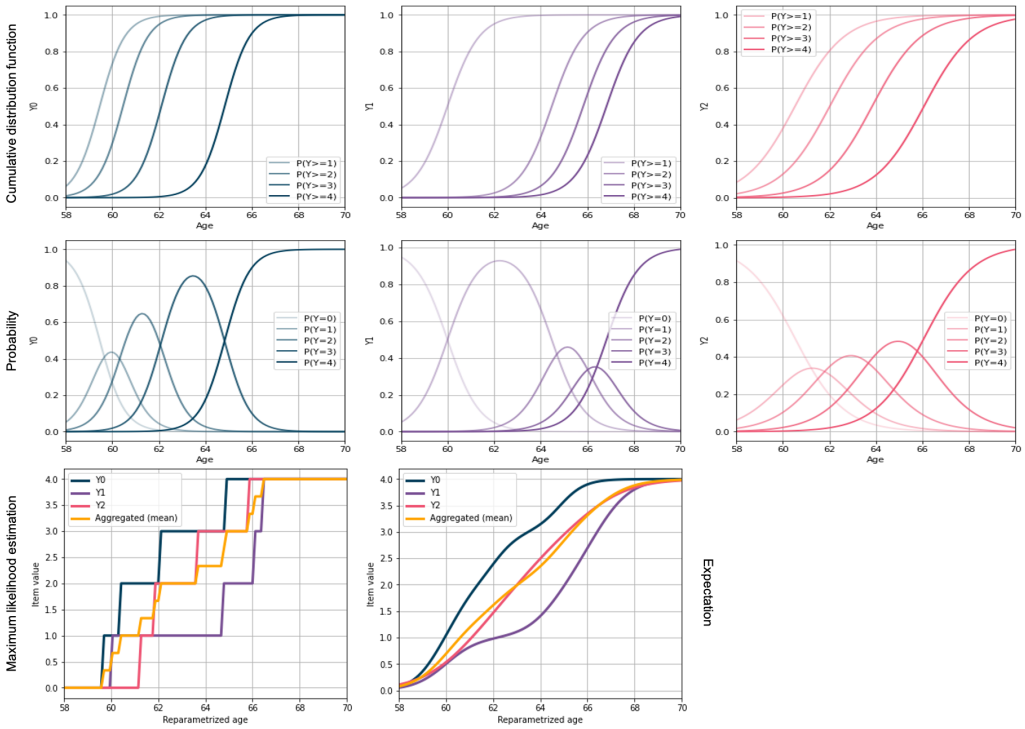

Figure 2.4: Average trajectory γ_0 for the ordinal model. Top row: for each item the cumulative distribution functions are shown. Note that δ_k^l correspond to the shift on the x-axis between the logistic curves. Middle row: raw probabilities are shown, directly obtained by difference between the logistic curves of top row. Last row: items are shown together by summarizing the multinomial distributions of items. On the left the value at each timepoint is the maximum likelihood; on the right the value is the expectation. We also plot the aggregated score computed as the mean of the items.

Generation process: We use our ordinal model in order to simulate synthetic patients and obtain their observations at randomly chosen timepoints. For the initialization, we chose the statistical model parameters θ and the population parameters $\mathbf{z}_{pop}$. The values are chosen arbitrarily in the range of plausible values for a neurodegenerative-like disease progression model. In this simulation part, we kept a low dimensionality for the data with only $d = 3$ dimensions, which is the lowest number of dimensions that allows a meaningful ICA decomposition with at least 2 sources. We then adjusted the δ_k^l so that each of the three items has an atypical progression, by which we mean it can not be easily interpolated by a linear curve or a logistic curve. Each item has $L = 4$ levels, which sums into an aggregated score ranging from 0 to 12. Figure 2.4 shows the average trajectory γ_0 of

the model, as well as the average aggregated trajectory. Once the population parameters are fixed, we randomly generate new individuals: for each individual parameter, we sample the distribution mentioned in section 2.1.2. However to generate two different clusters, we choose to randomly attribute a class 1 or 2 to a new individual. Then for the individual parameters simulation, the only difference between individuals of the two clusters is that the first source in cluster 1 follows a Gaussian distribution centered around -4 instead of 0 with a standard deviation of 1 and the source 1 in cluster 2 is oppositely centered around $+4$. This source acts as an inverter of the order of the curves. The mean trajectory in cluster 1 has the curve Y0 starting first, then Y1 follows and finally Y2, whereas for cluster 2 the order is switched: Y1 first, then Y2 and finally Y0. However the aggregated score is roughly the same in the two clusters. With an equal repartition of individuals between the two clusters, we therefore obtain a balanced dataset of individuals. Once this is done, we need to generate random timepoints for the visits of each individual. For this we decided to select the number of visits with a binomial distribution with probability $p = 0.5$ and $n = 10$ to which we add 2 to guarantee at least two visits per individual. Then each of the timepoint is sampled with a Gaussian distribution around the middle of the disease progression curve, with a time window calibrated to span about 5 years of follow-up for each subject.

We estimated four models on this synthetic dataset:

- a continuous model (equation 1.1.4) on items: to highlight the purpose of using a dedicated ordinal model for such scales, we fit a disease mapping course model with logistic curves as presented in the method section with observations considered as continuous with a Gaussian noise. The model is three dimensional with two sources
- an ordinal model (equation 2.3.1) on items: we fit the model presented in the method section, here the observations are treated as ordinal. The model is also three dimensional with two sources
- a univariate continuous model on the aggregated score: to show how the loss of item information impacts the ability of a model to produce relevant information about patients, we fit a single logistic curve model on the sum of the three items. The model is thus one-dimensional without sources
- a univariate ordinal model on the aggregated score

The first point of our synthetic experiment was to show how the gain of finer-grain information allows to learn meaningful data patterns. We extracted the individual parameters of the four models. Each individual i thus has a set of parameters $(\tau_i, \xi_i, s_i^1, s_i^2)$ in the multivariate models and (τ_i', ξ_i') in the univariate models. For visualizing the points in Figure 2.5, we decided to embed the individual in a two-dimensional space using a t-distributed stochastic neighbor embedding[30]. We see that the two clusters are clearly separated in the multivariate models, whereas the aggregated model is not able to disentangle them. A quantitative classification with a logistic regression taking in input the set of individual parameters and predicting the cluster of origin confirms this conclusion: with $(\tau_i, \xi_i, s_i^1, s_i^2)$ from either multivariate model the classifier gets close to 100% accuracy while with (τ_i', ξ_i') from the best univariate it only has 65%, which is much worse.

While the multivariate continuous model also captures individual heterogeneity by identifying the two clusters, it is not as performant in terms of pure data reconstruction. Figure 2.6 illustrates

[30] L. v. d. Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.

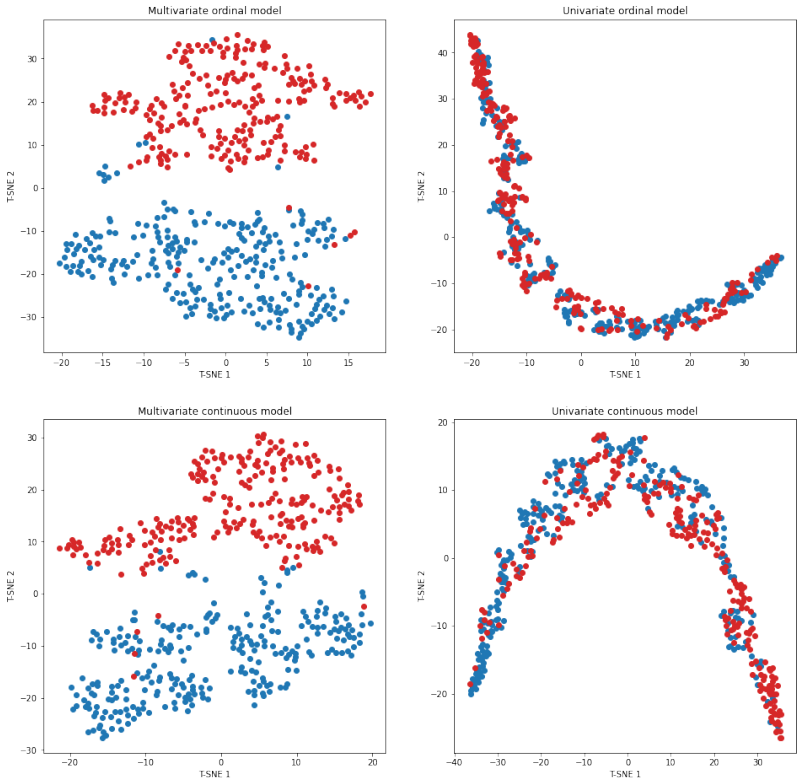

Figure 2.5: t-SNE plots for individual parameters. On the left side the parameters in the multivariate model are shown, on the right side the parameters in the univariate models. Top row shows the two ordinal models while bottom row shows the continuous models. Each color corresponds to one of the clusters (1 in blue, 2 in red).

the difference between the continuous and the ordinal model in terms of reconstruction errors on the items, with two different methods for predicting the outcome of the ordinal model. It is clear that the continuous model is worse, with a first burden being the lower bound of the noise due to the discretized scale (every 0.25 in this case). This can be mitigated in part by comparing the rounded predictions of the continuous model to the maximum likelihood predictions of the ordinal model, but even in this case we average 0.38 points of absolute error for continuous rounded predictions versus 0.34 for the ordinal maximum likelihood predictions. The second burden is the prior choice of the form of the curve (here the logistic). It is important to note that for small scale items in general there is hardly a perfect continuous curve due to the variance of item levels. For instance, a standard Likert scale for a symptom would be a $0-4$ scale with each level corresponding to a severity measure: null, mild, moderate, severe, very severe. The difference between two successive levels may vary, which highlights the role of the δ_k^l parameters.

These preliminary results encourage the use of dedicated ordinal models on ordinal scales, and this proves true even for items with more levels (especially when levels are unevenly distributed).

Second part: In the first part of the synthetic experiment we generated data using an ordinal

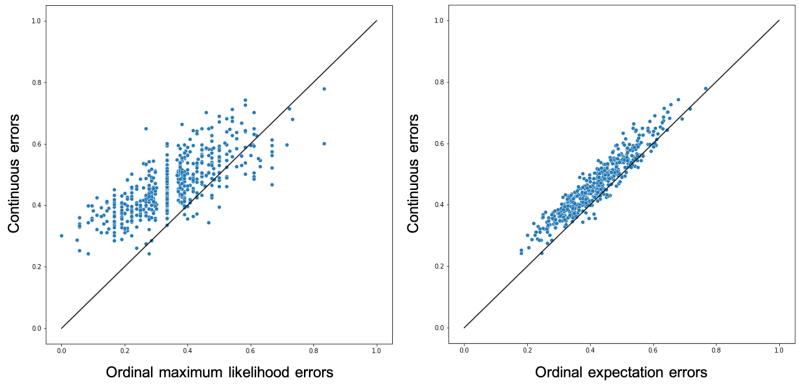

Figure 2.6: Error of reconstruction scatter plot. For each individual, we computed the mean absolute error of reconstruction over all visits and all items between the observations and the model prediction for the multivariate models. In the case of the continuous model the prediction is directly the outcome of the model, whereas for the ordinal model we computed the maximum likelihood for each value in the left plot, and the expectation in the right plot. The plots show that the ordinal model outperforms the continuous model for almost all patients in the reconstruction of the items.

model and showed that the ordinal model was better at reconstructing such data than a continuous model. It is important to check that this also holds for data not generated by the ordinal model.

We used a continuous model similar to the one used in the first part, but without the δ_k^l parameters. For each individual i generated, at each visit j for each item k we compute a noisy observation $y_{ijk} \in [0, 1]$. We then transform this data into ordinal data. We choose a Likert scale (0-4), so we multiply the observation by 4 and round it. Both multivariate models, ordinal and continuous, are calibrated on the data generated and we compute the reconstruction errors as in Figure 2.6. Results are presented in Figure 2.7 below.

We provided experimental results with data generated with a continuous model rather than the ordinal model. These results show that the ordinal model is also able to outperform a continuous model even in this setting.

2.3.3 Application to Parkinson's disease data

Data set

Data availability statement: Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database[31] (www.ppmi-info.org/access-data-specimens/download-data). For up-to-date information on the study, visit ppmi-info.org. [32]

The cohort includes mainly Parkinson's disease (PD) patients within two years of their diagnosis. There are also prodromal individuals, but we excluded those as they are a minority and some of

[31] K. Marek, D. Jennings, S. Lasch, *et al.*, "The Parkinson Progression Marker Initiative (PPMI)," en, *Progress in Neurobiology*, Biological Markers for Neurodegenerative Diseases, vol. 95, no. 4, pp. 629–635, Dec. 2011.

[32] PPMI – a public-private partnership – is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including [list the full names of all of the PPMI funding partners found at www.ppmi-info.org/about-ppmi/who-we-are/study-sponsors].

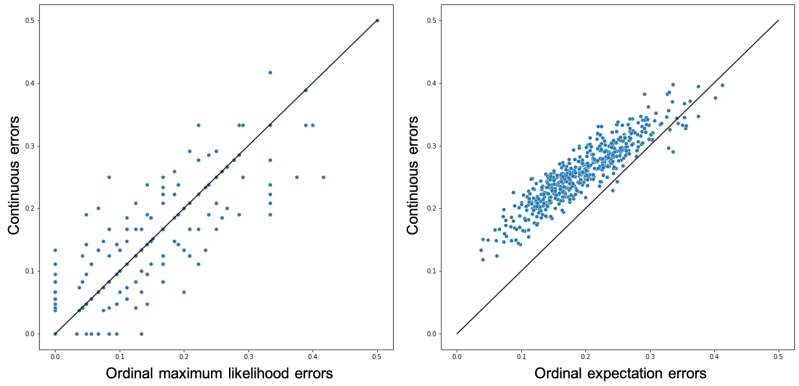

Figure 2.7: Error of reconstruction scatter plot. For each individual, we computed the mean absolute error of reconstruction over all visits and all items between the observations and the model prediction for the multivariate models. In the left plot the predictions of the ordinal model are the maximum likelihood estimation while the predictions of the continuous model are rounded. In the right plot for the continuous model the prediction is directly the outcome of the model, whereas for the ordinal model we computed the expectation. The right plots show that the ordinal model outperforms the continuous model for almost all patients in the reconstruction of the items.

them do not convert to PD. We selected all PD diagnosed patients with at least 2 visits in the study in order to have a longitudinal history for our training set, which left 900 subjects for a total of 7918 visits. From the wide range of biomarkers and clinical assessments we chose to focus on the MDS-UPDRS score. It is composed of 65 items divided in 4 sections: non-motor aspects of experiences of daily living (I) which is filled half by the clinician and half by the patient in a self-report questionnaire; motor aspects of experiences of daily living (II) which is a form only filled by the patient; motor examination (III) by the clinician; motor complications (IV) which is assessed by the clinician.

Each item is rated on a Likert scale (0: Normal, 1: Slight, 2: Mild, 3: Moderate, 4: Severe). For the third part, two versions of the test are recorded depending on the treatment effect. A version "OFF" is assessed when the patient hasn't taken medication in the last 6 hours, which is a state when the treatment should have little effect and motor symptoms will usually be observed. A version "ON" corresponds to the same motor examination shortly after the patient has taken medication, which is a state where motor complications are alleviated. Therefore we used "OFF" measures when selecting features in order to build the disease natural history with as few mitigation of treatment effect as possible. We also removed the fourth part of the MDS-UPDRS since it focuses on treatment secondary effects. In the end we obtained 59 items rated on scales from 0 to 4.

Most of the items have a low mean: 44 items have a mean between 0.4 and 1.25. Most of the observed values are 0 (66%), with 22% of 1 values, and only 12% of values above 2. For some items, the higher levels are very seldom seen so their progression cannot be modelled fully due to lack of data. The average total score on the parts I, II and III is 28.7 ± 22.9 with 90% of the values within the $[1, 60]$ range.

Qualitative experiment: a description of the MDS-UPDRS

As shown in the previous summary, patients in PPMI do not display all possible symptoms, and most of their items remain at 0 throughout their follow-up. However, items with non-zero values often occur in groups, suggesting that several items capture a similar biological phenomenon. Our goal is to unravel these trends using Independent Component Analysis (ICA) on the space-shifts.

We calibrated an ordinal DCM model on the 900 subjects using their complete set of 59 items. Missing values do not need to be imputed since the model only computes the likelihood for observed values, which is one of the advantages of using DCM. The estimation was performed with the MCMC-SAEM algorithm, using a block Gibbs sampler. The sampling was grouped for all 59 dimensions of the population parameters $\mathbf{v}, \mathbf{g}, \delta$ during the first 16,000 iterations. For the final 4,000 iterations, a more refined Gibbs sampler was used, sampling each scalar coordinate separately. The samplers employed an oscillating tempered scheme 19. The overall calibration time was 56 hours.

Due to the high number of dimensions, we chose 8 as the number of sources to reduce the dimension of individual parameters and prevent overfitting. Previous trials with fewer sources resulted in significantly lower log-likelihood at the end of the calibration, while increasing the number of sources became too computationally expensive. ICA enforces correlations within the space-shifts due to the relation $\mathbf{w}_i = \mathbf{A}\mathbf{s}_i$, with $\mathbf{A}$ being a rectangular matrix. We can analyze the correlations across individuals resulting from these space-shift values, which can be seen as a time delay in the progression of each item compared to the population's average trajectory. Figure 2.8 shows the Pearson correlation matrix between the space-shifts corresponding to each item. We performed hierarchical clustering to isolate groups of highly correlated items.

Figure 2.8: Pearson correlation matrix of space-shifts at the item level. Hierarchical clustering has been performed to order the items of the different parts of the MDS-UPDRS.

As we can see in Figure 2.8, the first group of items is comprised of tremor related items. The next two clusters gather items of the motor part (III) of the MDS-UPDRS, with one cluster containing items for the left part of the body and the other being dedicated to the right part. This is explained by the fact that Parkinson's disease very often declares itself in one side of the body before the other. Then we have a small trio of gait, freezing of gait and posture, which can be included in a larger cluster of items concerning motor aspects of daily life (essentially items of the part II related to movement and balance). Then the last large cluster includes all the elements of part I (non-motor aspects of daily life) and the last items of part II. This large cluster can be separated into two subclusters, one being the part I of the MDS-UPDRS and the second being the leftover items of part II which are more related to mouth tasks (eating, drooling...) and communication. Overall we can notice a strong correlation between items of part I and part II. These parts are both questionnaires and mostly filled by the patient, which can explain the correlation bias due to the self-assessment of one's abilities.

Parkinson's patients are commonly separated into two or three subtypes[33]: tremor dominant (TD), postural instability and gait dominant (PIGD) and mixed patients (at the limit between TD and PIGD). Here we see that the TD and PIGD tendencies are represented in the correlations of items of the first and fourth clusters.

Some items were not regrouped with others or were associated in different ways. For instance we can note that neck rigidity and facial expression were at the junction between the left and right motor clusters, suggesting that higher body motor impairment usually follows motor symptoms in one side of the body. However these two items are less correlated to motor impairment in the legs and the feet. This is very interesting as we can hypothesise that the disease implies a localised neurodegeneration of motor neurons, and then symptoms spread from their original onset, explaining the anticorrelation between top and bottom motor impairment. Speech (part III question 1) has been isolated, while spontaneity of movement (part III question 14) seems to correlate with all the motor assessments of part III, as a global appreciation of the motor state of the patient by the clinician.

We gave a specific look at the item called DDS (dopamine dysregulation syndrome) which is a disturbance of dopamine therapy with addictive patterns such as impulse control disorders (ICD) which include gambling, hypersexuality or compulsive eating[34]. This syndrome is a consequence of too much dopamine uptake (sometimes voluntarily on the patient's side). What appears is that the DDS is anticorrelated to all motor items of part III *except* for the freezing of gait (part III question 11). This backs the conclusion that freezing of gait is also a secondary effect of dopaminergic treatment rather than a disease symptom[35].

We then examined the sources individually to identify which statistically independent linear combinations of the items were found, and tried to understand if they were linked to a pathological trend. We provide a description of the 8 sources:

- source 1 (variance explained 11%): this source induces a positive delay in all motor items of the left part of the body while it pushes for a negative delay in almost all of the other items.

[33] G. T. Stebbins, C. G. Goetz, D. J. Burn, *et al.*, "How to identify tremor dominant and postural instability/gait difficulty groups with the movement disorder society unified Parkinson's disease rating scale: Comparison with the unified Parkinson's disease rating scale," en, *Movement Disorders*, vol. 28, no. 5, pp. 668–670, 2013.

[34] S. S. O'Sullivan, A. H. Evans, and A. J. Lees, "Dopamine dysregulation syndrome: An overview of its epidemiology, mechanisms and management," eng, *CNS drugs*, vol. 23, no. 2, pp. 157–170, 2009.

[35] A. Espay, A. Fasano, B. van Nuenen, *et al.*, ""On" state freezing of gait in Parkinson disease," *Neurology*, vol. 78, no. 7, pp. 454–457, Feb. 2012.

We can interpret this source as a left body delay

- source 2 (variance explained 10%): this source is almost the same as source 1 but for right body items. We call this one right body delay. These two first sources with equivalent prevalence show that Parkinson's disease is as likely to start in the left part of the body as the right
- source 3 (variance explained 12%): this source has a positive delay on all motor items linked to the limbs' movement as opposed to the other items. We name this one motor source
- source 4 (variance explained 15%): this source has a positive delay on the left part of the body and the rigidity items. This source embeds a left body and rigidity combined progression
- source 5 (variance explained 12%): this source has a positive delay on all the questions that are self-evaluated as opposed to all the motor evaluations and the non-motor assessment by the clinician. We can therefore say that this source is the patient's self report effect
- source 6 (variance explained 13%): this source regroup items of the part I (non-motor aspects of daily life) with tremor manifestations as opposed to motor issues. This source is seen as the brain-first source
- source 7 (variance explained 14%): this source is driven mostly by the motor items and more especially the gait, posture and balance items. We call this one the gait and balance source
- source 8 (variance explained 13%): this source gathers items from the part I with a strong emphasis on anxiety, depression and apathy, along with speech issues and opposed to tremor items. We coin this one the cognitive source

We observe that some sources are in practice not totally independent, for instance source 5 (self report effect) and source 4 (left body rigidity) are correlated. This suggest that the number of sources is not fully adequate. We see that sometimes a source seem to include several groups of items at the same time (source 4 is a mix of left body items and rigidity), which could be the consequence of too few dimensions to disentangle these effects.

If we come back to the classification of patients between tremor dominant and postural instability/gait dominant, we see that PIGD is solely driven by source 7, whereas tremor information is spread across sources 1 to 4.

We won't dwell on more clinical subtleties but hope this showcased the power of the ICA on the spatial effects unfolding from the Riemannian framework of the DCM. The results here could only be obtained thanks to the modelling of items with an ordinal method. The global model calibrated on the 59 items showed that some items had a very low progression profile (or could not be completely learned due to the few data available on the last levels), especially in the motor examination part. However the second part of the MDS-UPDRS scale was more reliable since it rarely stayed at zero and the progression was faster and steadier. We thus used the 13 items of the second part in the next experiment to show quantitatively how well the ordinal model performs.

Quantitative results: prediction task

In order to understand the value brought by our new take on the model, we compare the ordinal DCM to the logistic version of the DCM taking the cognitive scores as continuous values. We also

compare to a linear mixed-effect model and a no-change prediction model as our baseline. We will have two different options in order to show what the model can bring to the table: one model for the total score of one MDS part (the second part as mentioned previously) which is thus an integer between 0 and 42; and one model taking the items into account, as 42 is the sum of 13 questions with a scale from 0 to 4. In this last case we will show that taking the score as a continuous value is not valid since the noise here is lower than the step between possible values. We present the several models used in this experiment:

Continuous univariate model

The MDS UPDRS part II (/42) is normalized between 0 and 1 in order for the logistic model to work. We then maximize the log-likelihood with Gaussian noise, thus implying a fit in the least squares sense.

Continuous multivariate model

Each item (/4) of the MDS UPDRS part II is normalized between 0 and 1. As for the univariate maximize the log-likelihood with Gaussian noise, where the standard deviation of the noise is feature-dependent. For the prediction of the MDS UPDRS II total we sum the predictions of the items.

Ordinal models

In order to operate a fair comparison we use the same hyperparameters as for the continuous version, so that the ordinal model shares the same parameters as the continuous one except for the δ. We fit by maximizing the log-likelihood, and we compute the predictions using the expectation of the ordinal model: $\mathbb{E}(\eta_{ij}) = \sum_l \mathbb{P}(\hat{y}_{ij} \geq l)$. As for the continuous case, we have a univariate model built with only the MDS UPDRS II total score and a multivariate model using the 13 items. For the multivariate case, we compute the prediction as the expectation of the sum of the items.

Baseline models

We also included a baseline constant model that predicts the last known value for the patient (hypothesising no change in the future) and a linear mixed-effect model for comparison. Note that they perform very well: it is hard to beat these models when predicting in a very short future since the state of the patient evolves very slowly.

To sum up, we provide below the formulas of the models used.

The continuous univariate model computes the MDS-UPDRS part II total as:

$$y_{ij} = \left(1 + (\frac{1}{p_0} - 1)\exp\left(-\frac{v_0(e^{\xi_i}(t_{ij} - t_0 - \tau_i) + t_0)}{p_0(1-p_0)}\right)\right)^{-1} + \epsilon_{ij}$$

with $\epsilon_{ij} \sim \mathcal{N}(0, \sigma)$.

The multivariate continuous model computes the 13 items of the MDS-UPDRS part II as follows:

$$\forall k \in [|1, 13|], y_{ijk} = \left(1 + (\frac{1}{p_k} - 1)\exp\left(-\frac{v_k(e^{\xi_i}(t_{ij} - t_0 - \tau_i) + t_0) + w_{ik}}{p_k(1-p_k)}\right)\right)^{-1} + \epsilon_{ijk}$$

with $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma)$.

The ordinal univariate model computes the MDS-UPDRS part II total as:

$$\forall l \in [|1, 52|], \mathbb{P}(y_{ij} \geq l) = \left(1 + (\frac{1}{p_0} - 1)\exp\left(-\frac{v_0(e^{\xi_i}(t_{ij} - t_0 - \tau_i) + t_0 - \sum_{m=1}^{l}\delta_0^m)}{p_0(1-p_0)}\right)\right)^{-1}$$

The multivariate ordinal model computes the 13 items of the MDS-UPDRS part II as follows:

$$\forall k \in [|1,13|], \forall l \in [|1,4|], \mathbb{P}(y_{ijk} \geq l) = \left(1 + (\frac{1}{p_k} - 1)\exp\left(-\frac{v_k(e^{\xi_i}(t_{ij} - t_0 - \tau_i) + t_0 - \sum_{m=1}^{l}\delta_k^m) + w_{ik}}{p_k(1-p_k)}\right)\right)^{-1}$$

The constant prediction model for the MDS-UPDRS part II total formulates as $\hat{y}_{ij} = y_{ij_{final}}$ where $j_{final} < j$ is the last visit known for subject i in the training data.

The linear mixed-effect model for the MDS-UPDRS part II total formulates as :

$$y_{ij} = \beta + \alpha \times t_{ij} + \beta_i + \alpha_i \times t_{ij} + \epsilon_{ij}$$

with β being the fixed intercept, α the fixed slope, $\beta_i \sim \mathcal{N}(0, \sigma_\beta)$ the random intercept, $\alpha_i \sim \mathcal{N}(0, \sigma_\alpha)$ the random slope and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma)$ the Gaussian noise.

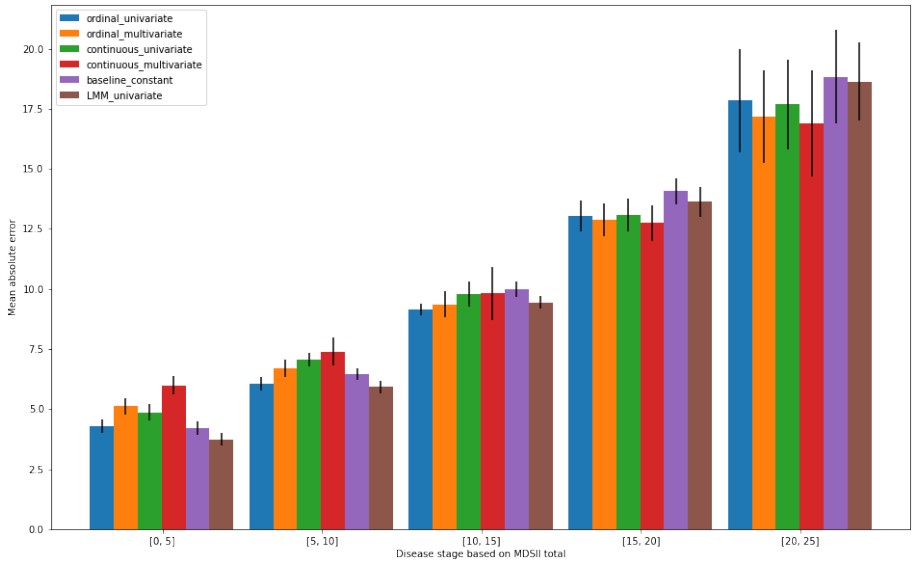

Figure 2.9: Mean absolute prediction error as a function of the current patient state. Bin ranges correspond to the MDSII value of the patient. Error bars for mean model error are estimated from cross-validation. LMM: linear mixed-effect model; continuous: disease course mapping model; ordinal: disease course mapping model, ordinal version; univariate: model learned of MDSII total directly; multivariate: model learned on MDSII items and then aggregated.

The prediction framework we use consists in two steps. First a training set is used to calibrate a model, i.e. learning the fixed effects for the mixed-effect models (linear and DCM). In a second step, we use the patients of the test set. Out of each patient's set of visits we randomly select some past visits and a visit in the future for which we want to predict. We learn the random effects of the models on the past visits and use the personalized model to compute the value at the desired future visit's timepoint. We use a repeated 8-fold cross-validation to measure the models errors. The Figure 2.9 shows the mean absolute prediction error.

As mentioned previously, we observe that constant prediction and linear mixed model perform best when the patient score is low, as they progress slowly at the beginning of the disease. However we see that the prediction from those simple models worsens as the score increases, meaning that linear and constant models are less able to capture disease progression. As a matter of fact, they largely under estimate the change for patients later in the disease.

On the ordinal side we observe a similar trend, although both ordinal models are more reliable at higher disease stages. They outperform continuous models on the low stages, due to the flexibility of the step parameters $(\delta_l)_l$. However when it comes to later disease stages continuous models catch up with ordinal models in performance. This is mainly due to the data being much sparser at these stages, leaving a larger role to the shape priors. In the long term setting it confirms that the logistic prior improves on linear and constant curve priors. Ordinal models show that their low granularity allows for better prediction at low levels when changes are small, while we lack data for higher disease stages thus leading to a mitigated performance.

Comparing univariate versus multivariate models is very interesting. At early disease stages the univariate models seem to better capture the essence of the progression. On the other hand, the higher the disease stage the better the multivariate models perform, showing that multidimensional models are better suited for disease progression prediction when the rate of change increases.

Figure 2.11 shows the average disease progression learned by univariate mixed-effect models. Spaghetti plots with wrapped individual trajectories are provided below. The Figure 2.11 shows that the logistic and the ordinal provide a somehow similar description of the aggregate score up until age 85. After the logistic model is bound to the assumption that the score will evolve to 1, which in practice is highly unlikely as it is a complete state of degeneracy and patients usually do not reach this stage. This is why the ordinal curve slows compared to the logistic at the end. The logistic model assumes that the disease progresses faster after 85 years old. On the other hand the linear mixed-effect model tends to underestimate the rate of change at this stage. Note that the shape of the step function is mostly driven by the δ values learned, and thus is more flexible. This explains why it is more accurate for the values of the MDSII total between 0 and 15 where most of the observations lie according to the histograms.

We provide a visualization of the fit of univariate models with Figure 2.10. The spaghetti plots show the individual trajectories mapped onto the average trajectories. We do this by removing the random effects, hence only leaving the noise of individual observations as a deviation from the mean trajectory, shown in black, which accounts for the fixed effects. Note that the score we are modelling is quite noisy, with some individuals experiencing very sharp and sudden surges or collapses. This is also one of the reasons for the use of an ordinal model: the noise is more flexibly modelled as a spread of probability weights across several levels rather than a noise distribution chosen in model design (here a simple Gaussian noise).

The experiment was performed again on parts I and III of the MDS-UPDRS with continuous and ordinal models, and the results were essentially the same:

- For the MDS-UPDRS part I (/42), the ordinal univariate had a mean absolute error of 4.64 while the multivariate ordinal had 4.54; continuous univariate was 4.76 and continuous multivariate was 5.41

- For the MDS-UPDRS part III (/132), the ordinal univariate had a mean absolute error of 12.8 while the multivariate ordinal had 12.3; continuous univariate was 15.8 and continuous multivariate was 13.9

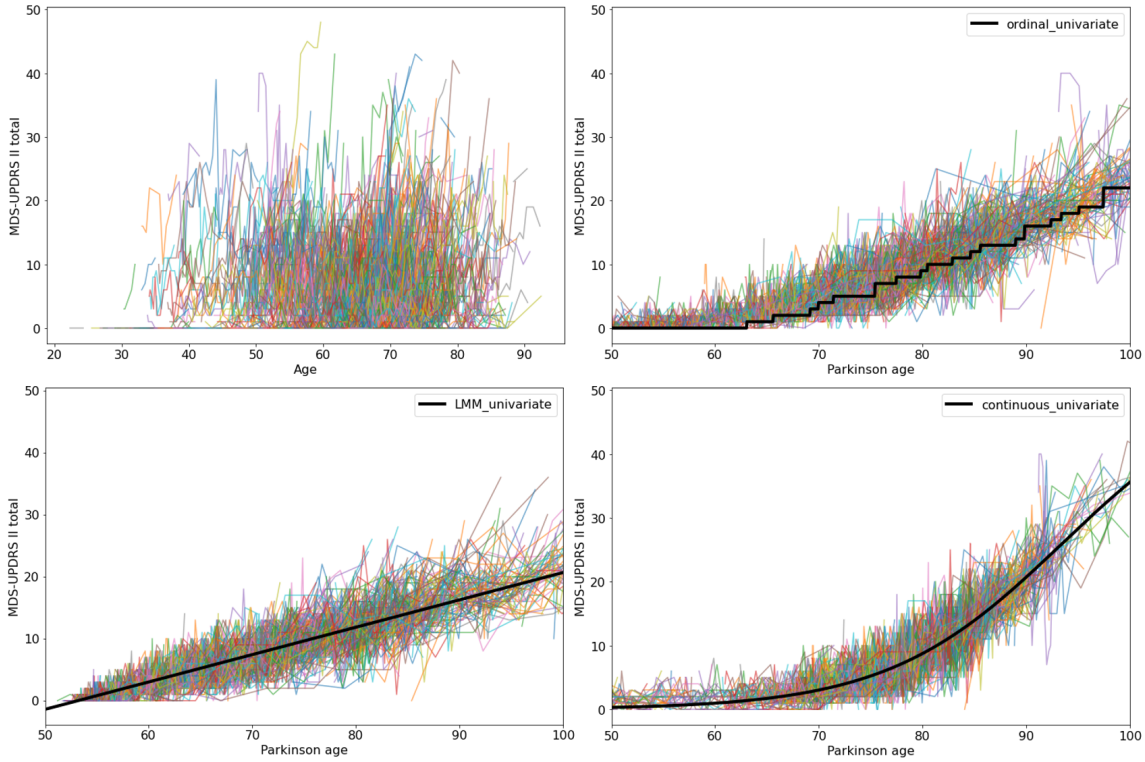

Figure 2.10: Spaghetti plots with individuals mapped onto the mean trajectory. Mapped trajectories are obtained by removing the random effects learned by each model. Top left: raw data; bottom left: linear mixed-effect model; top right: ordinal univariate model; bottom right: contiuous univariate DCM model with logistic curve.

What we seem to notice though is that the more items in a score, the better the multivariate approach performs, although it comes at a higher computational price.

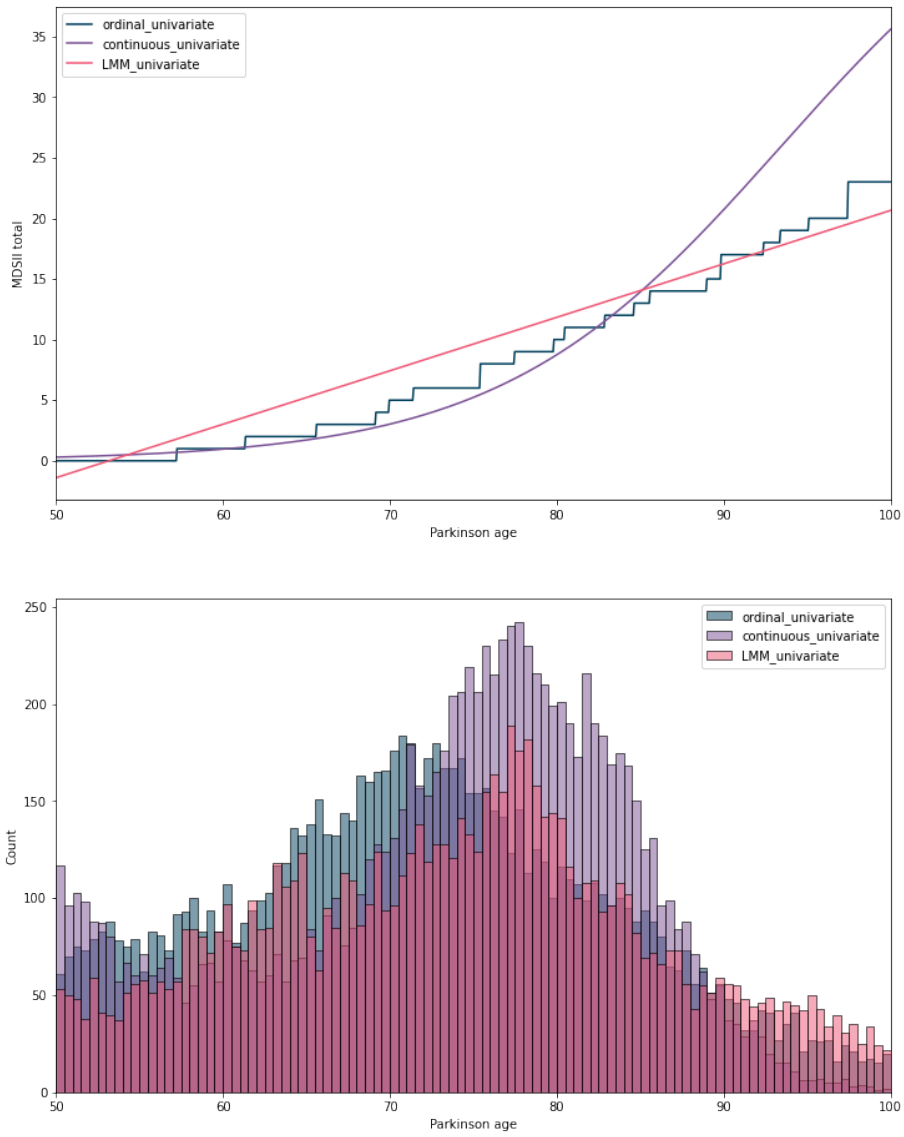

Figure 2.11: Top figure: average disease progression of the univariate mixed-effect models as a function of the age. Bottom figure: histograms of patients mapped onto the corresponding timeline of their models. Aligned with the top figure it shows where data points are distributed according to each model. Parkinson age corresponds to that mapping on the time axis, as a result of the individual random effects linked to time ((ξ_i, τ_i) for ordinal and continuous DCM, random slope and intercept for LMM). LMM: linear mixed-effect model; continuous: disease course mapping model; ordinal: disease course mapping model, ordinal version.

2.3.4 Discussion

Our work shows the potential of leveraging items with a small ordinal scale over working with an aggregated score. We built upon a non-linear mixed-effect model used mainly for continuous markers and extended it to ordinal data. This allowed to address cognitive decline by processing items of the test instead of their aggregated value. We showed how this could help understand disease trends in Parkinson's disease within the MDS-UPDRS. We believe this could be extended to other cognitive scales in other neurodegenerative scales with the aim to better understand items dynamics.

In the prediction task, which is especially hard, ordinal models had a finer prediction for early disease stages, when disease evolution is slower. Our results highlight the need to rely on multivariate data to increase prediction performance at later disease stages, when rating scores changes are of higher magnitude.

Future applications could use this model as a tool to extract information from specific items after a multivariate analysis. We believe this can be the first step to building new composite scores such as the PACC5 in Alzheimer's disease[36]. One idea would be to use the Fisher information as the metric of choice to select meaningful items in a composite score, since in IRT it is common to compute Fisher information of items for the individual parameters.

On the downside we see the limitations of the ordinal method when used on data with few values for certain levels (in our case these were the high values of the MDS UPDRS II), as the added parameters of the ordinal DCM require more data in order to be properly learned. These extra parameters also come at a computational cost within the estimation algorithm, therefore we deem the method useful when the items are rated on short scales like the Likert one.

2.4 Parallel with Item Response Theory

As mentioned in the introduction of this chapter, the study of discrete data is the focus of the Item Response Theory (IRT) field. This has been developed mostly in psychology[37], and was aimed at understanding how to best design tests to evaluate a latent trait of the patient. It makes sense that the same theory applies to cognitive tests. The other field of application of IRT has been the design of educational tests. In this context, the latent trait refers to the students' level of knowledge, which is unknown but needs to be evaluated by the test designer. By utilizing the prior knowledge of the estimated mean latent trait, the test difficulty can be tailored to minimize variance in the estimation of the students' latent trait. This theory has found extensive applications in binary and ordinal tests. The models extend to several items for multivariate tests, and can also include a multidimensional latent trait for more complex application cases. The IRT has also been adapted to longitudinal cases[38] and recently new applications to Parkinson's disease have emerged[39], but

[36] K. V. Papp, D. M. Rentz, I. Orlovsky, *et al.*, "Optimizing the preclinical Alzheimer's cognitive composite with semantic processing: The PACC5," *Alzheimer's & Dementia : Translational Research & Clinical Interventions*, vol. 3, no. 4, pp. 668–677, Nov. 2017.

[37] S. E. Embretson and S. P. Reise, *Item response theory for psychologists* (Item response theory for psychologists). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers, 2000.

[38] C. Proust-Lima, V. Philipps, B. Perrot, *et al.*, "Modeling repeated self-reported outcome data: A continuous-time longitudinal Item Response Theory model," en, *Methods*, vol. 204, pp. 386–395, Aug. 2022.

[39] S. Luo, H. Zou, G. T. Stebbins, *et al.*, "Dissecting the Domains of Parkinson's Disease: Insights from Longitudinal Item Response Theory Modeling," en, *Movement Disorders*, vol. 37, no. 9, pp. 1904–1914, 2022.

there are very few attempts at combining both into a multidimensional longitudinal latent trait[40].

Let us introduce some notations. We will keep the index i for the individual, k for the feature which is called item in this section, and j for the visit index in the longitudinal framework. The observations y_{ijk} are the results of the answers to item k for individual i at visit j, and in the case of pure cross-sectional data we drop the visit index y_{ik}. The most basic formula for an IRT model with binary items and no longitudinal modelling is as follows:

$$\mathbb{P}(y_{ik} = 1) = \eta_k(\theta_i) \tag{2.4.1}$$

$$\eta_k(\theta_i) = \frac{1}{1 + \exp(-a_k(\theta_i - b_k))} \tag{2.4.2}$$

where (a_k, b_k) are model parameters and θ_i is the latent trait of individual i. b_k refers to the level of the test, i.e. the higher b_k the lower the probability of having a right answer to the item. Ideally the designer of the test should aim for a b_k close to the mean of the individual latent traits, or in the case of an educational test, the objective should be to have b_k equal to estimated level required to pass the test. a_k is the discriminating power of the item. If b_k is required level to pass the test, then a strong discriminating power a_k will imply that individuals with sufficient knowledge should almost always pass while individuals below the threshold have almost no chance. Sometimes in the case of multiple answer questionnaire, the model takes into account a part of randomness inherent to the possibility of answering correctly by chance:

$$\mathbb{P}(y_{ik} = 1) = c_k + (1 - c_k)\eta_k(\theta_i) \tag{2.4.3}$$

Now let's complicate the model by introducing the longitudinal model. The latent trait θ_i becomes a function of time. The most simple choice for modelling it is to have a linear mixed-effect structure for the time:

$$\theta_i(t) = \alpha_i t + \beta_i \tag{2.4.4}$$

with the time slope $\alpha_i \sim \mathcal{N}(\bar{\alpha}, \sigma_\alpha^2)$ and the intercept $\beta_i \sim \mathcal{N}(\bar{\beta}, \sigma_\beta^2)$. Then if we add a multidimensional structure for θ_i we obtain: $\boldsymbol{\theta}_i \in \mathbb{R}^{N_s}$, with N_s the dimension of the latent trait, and the formula for each item now includes a scalar product $\boldsymbol{d}_k^T \boldsymbol{\theta}_i$ instead of θ_i. All in all we have the formula below:

$$\mathbb{P}(y_{ijk} = 1) = \frac{1}{1 + \exp(-a_k(\boldsymbol{d}_k^T(\boldsymbol{\alpha}_i t_{ij} + \boldsymbol{\beta}_i) - b_k))} \tag{2.4.5}$$

which is to be compared with the binary DCM model formula:

$$\mathbb{P}(y_{ijk} = 1) = \frac{1}{1 + \exp\left(\tilde{g}_k - \frac{v_k(e^{\xi_i}(t_{ij} - t_0 - \tau_i) + t_0) + w_{ik}}{p_k(1-p_k)}\right)}$$

[40] J. M. te Marvelde, C. A. W. Glas, G. Van Landeghem, *et al.*, "Application of Multidimensional Item Response Theory Models to Longitudinal Data," en, *Educational and Psychological Measurement*, vol. 66, no. 1, pp. 5–34, Feb. 2006.

We observe that even though there are subtleties in the formulas, with different identifiability conditions and distribution priors over the random effects, some parameters have essentially the same role. For instance, we see that the DCM model only has a one-dimensional acceleration, so it can be directly compared to a specific case of the latent trait formula where $\boldsymbol{\alpha}_i = \lambda_i \bar{\boldsymbol{\alpha}}$ with λ_i being an uni-dimensional factor. This boils down to the hypothesis of parallel trajectories in the DCM that does not allow a patient's acceleration to vary depending on the item k. In this case we see that e^{ξ_i} and λ_i play the same role, while $a_k \boldsymbol{d}_k^T \bar{\boldsymbol{\alpha}} \approx \frac{v_k}{p_k(1-p_k)}$ reflects the discriminating power of the item, $b_k \approx \tilde{g}_k p_k(1-p_k)$ is the difficulty of the item, and finally the space-shift can be decomposed over the sources with the k-th column of the mixing matrix $\boldsymbol{A}$ as $\frac{\boldsymbol{A}_k \boldsymbol{s}_i - v_k e^{\xi_i} \tau_i}{p_k(1-p_k)} \approx a_k \boldsymbol{d}_k \boldsymbol{\beta}_i$. The last formula shows that in multimensional IRT β_i plays a similar role to the sources in the DCM model, and we could also enforce the ICA as in the DCM by specifying the following prior: $\beta_i \sim \mathcal{N}(\bar{\beta}, \sigma_\beta^2 \boldsymbol{I}_{N_s})$.

In the ordinal case, the comparison with Samejima's model based on the cumulative logit[41] is ever so simpler:

$$\mathbb{P}(y_{ijk} \geq l) = \eta_{kl}(\theta_i(t_{ij})) \tag{2.4.6}$$

$$\eta_{kl}(\theta_i) = \frac{1}{1 + \exp(-a_k(\theta_i - b_{kl}))} \tag{2.4.7}$$

where the difficulty of the item b_{kl} now depends on the level l of the item, with increasing difficulty as levels increase. These parameters can be directly compared to the $\sum_{m=1}^{l} \delta_k^m$, which encodes this increasing difficulty of the item in the ordinal DCM model.

Overall the discrete versions of the DCM model are very similar to IRT models. The specificity lies in the hypotheses of the DCM model, namely the uni-dimensional latent acceleration random effect and the independence prior of the sources ($\approx \beta_i$), constituting the space-shift ($\approx$ intercept).

2.4.1 Fisher information

Computing Fisher information in IRT is usually the way to estimate the variance of the parameters[42] and also a way to understand which values in the ordinal scale are bringing the most information. We transpose these ideas to the longitudinal case. For the logistic curves model, we computed explicitly the Fisher information formula for both the binary and the ordinal models. These formulas can be found in the appendix 15. In IRT, computing the Fisher information for item k with the standard model 2.4 gives:

$$\mathcal{I}_k(\theta) = \mathbb{E}_{X \sim f(.;\theta)} \left(\left(\frac{\partial}{\partial \theta} log\ f(X;\theta) \right)^2 \middle| \theta \right) \tag{2.4.8}$$

$$= a_k^2 \eta_k(\theta)(1 - \eta_k(\theta)) \tag{2.4.9}$$

The main use case of this formula relies on the fact that in the asymptotic case we can have the approximation $\mathbb{V}(\hat{\theta}) = \frac{1}{\mathcal{I}(\theta)}$ where $\hat{\theta}$ denotes the maximum likelihood estimation (MLE) of θ.

[41] F. Samejima, "Estimation of latent ability using a response pattern of graded scores," en, *Psychometrika*, vol. 34, no. 1, pp. 1–97, Mar. 1969.

[42] A. Ly, M. Marsman, J. Verhagen, *et al.*, "A Tutorial on Fisher information," en, *Journal of Mathematical Psychology*, vol. 80, pp. 40–55, Oct. 2017.

Indeed, if we consider that data was generated according to the model with a true parameter θ^*, then the MLE will converge towards the true parameter θ^* as the number of samples n goes to infinity according to the following formula:

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0, \mathcal{I}(\theta^*)^{-1}) \tag{2.4.10}$$

where $\mathcal{D}$ means that the random variable on the left converges in distribution towards the distribution on the right. This formula still holds for the multidimensional θ case, where $\mathcal{I}(\theta)$ is the Fisher information matrix.

Of course, since our model uses a Bayesian framework, computing confidence intervals is not an issue as it will be a byproduct of our posterior estimation for latent parameters. The formula can still be applied to the model parameters which are maximized, and therefore do not have a posterior distribution. However we usually are more interested in individual parameters, as they allow us to personalize the model and run various analysis, including correlations with covariates, clustering, prediction, patient's stratification... For those parameters, the use of Fisher information as a mean to compute a confidence interval is of no use after the calibration by the MCMC-SAEM algorithm. The only potential application is for the personalization step: when we add new patients and try to compute their optimal individual trajectory by maximizing the likelihood over the individual parameters, we obtain only point estimates for those parameters and therefore the Fisher information formula for confidence intervals comes in handy.

But Fisher information can also be of interest in order perform item selection[43]. In our case, it could be used to evaluate the influence of items over time: we can for instance use it as a criterion to discriminate "early" versus "late" items, as well as understanding which items are most informative for each specific parameter of the model. In the case of the ordinal model we can compute the amount of information per level of each item. As the information is computed as a function of the time, we obtain profiles describing the informativeness of items at different stages of the disease.

Another interesting application is the use of this information over time to build new composite scores with a focus on specific aspects of the disease. For instance if we need to spot early motor patients in Parkinson's disease, we can look into which items provide the most early information for motor-related model parameters. We provide an example below for the Parkinson's disease ordinal model presented in the previous section.

As we see in the formula .0.58 or in the formula 2.4.8, the Fisher information of the chosen parameter depends on the model value for this parameter. However the model is function of time, so the Fisher information is varying over time. If we included a prior over the time of the visits $t_{ij} \sim \mathcal{D}$ then we could have integrated the information over time, but this is less interesting for us since we can leverage this time dependence to extract crucial insight from the items.

To illustrate our assertion, we selected one parameter of interest: τ_i the time-shift, which is a proxy for the onset age of the disease. Figure 2.12 shows the evolution in time of the Fisher information for τ for four selected items. We computed these curves as the mean over all the individual trajectories parametrized by $\boldsymbol{z}_i$. The patterns for the different items are very different, for instance we see that item 8 of part I has very low information, but also seems to be less informative as time passes. On the contrary items such as MDS-UPDRS II-1 or III-1 bring more

[43] R. J. Swartz and S. W. Choi, "A Burdened CAT: Incorporating Response Burden With Maximum Fisher's Information for Item Selection," en,

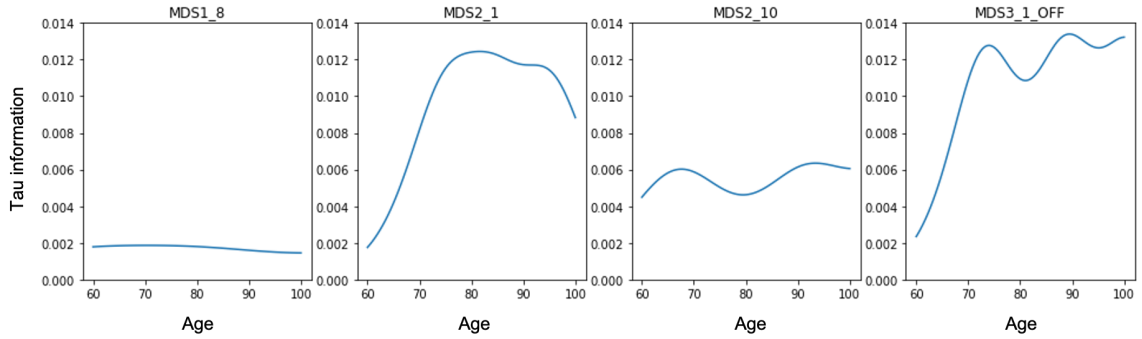

Figure 2.12: Tau information as a function of time for four selected items. The curve was obtained as the mean information over all individuals.

information as the disease progresses. Other items such as MDS-UPDRS II-10 have fluctuating information.

The information profiles can be used to design specific sub-scores targeting early disease stage for instance. We can select items that provide enough information in the early stages to form a new composite score such that the estimation of the individual parameters in a DCM model is as precise as possible.

As mentioned previously, the Fisher information can be used to approximate the variance of individual parameters estimation during personalization. We used this approximation to compute the standard deviation of τ in Figure 2.13 depending on the number of visits used for personalization. This is very useful in order to understand how reliable the estimates obtained by personalization are, and therefore whether the predictions we extrapolate based on these individualized trajectories are precise. The violin plot shows that with increasing number of visits the estimation gets better.

More importantly we observe that with less than three visits the estimated variance of τ_i is quite high, which means that the model requires multiple visits to perform well. This is paramount in the context of applications to clinical trials where a disease progression model can provide a prognosis score to be included in a PROCOVA[44] for instance, or perform patient selection based on some criterion such as $\xi_i > 1$ to only retain fast-progressors[45]. Other applications of cognitive decline predictions have been shown to allow clinical trial enrichment[46]. In this context we see that using only a baseline visit to perform the personalization is rather limiting compared to using a few follow-ups visits. The question that hinders this larger visit sampling is whether the pay-off for following patients before starting the trial is worth it given that medical examinations might be costly and that there is a very limited legal time-window between patient enrollment and clinical trial debut. The optimal application case however is when patients are enrolled from an observational cohort, where they tend to have several years of prior follow-up.

Finally the Fisher information can be leveraged to improve the design of the test[47]. We computed

[44] D. Bertolini, K. Arnemann, D. Hall, *et al.*, "Machine Learning Enables Smaller ALS Clinical Trials (P1-13.003)," en, *Neurology*, vol. 98, no. 18 Supplement, May 2022.

[45] E. Maheux, J. Ortholand, C. Birkenbihl, *et al.*, "Forecast Alzheimer's disease progression to better select patients for clinical trials," en, Jul. 2021.

[46] A. Tam, C. Laurent, S. Gauthier, *et al.*, "Prediction of Cognitive Decline for Enrichment of Alzheimer's Disease Clinical Trials," *The Journal Of Prevention of Alzheimer's Disease*, 2022.

[47] Y. Jung and I. Lee, "Optimal design of experiments for optimization-based model calibration using Fisher

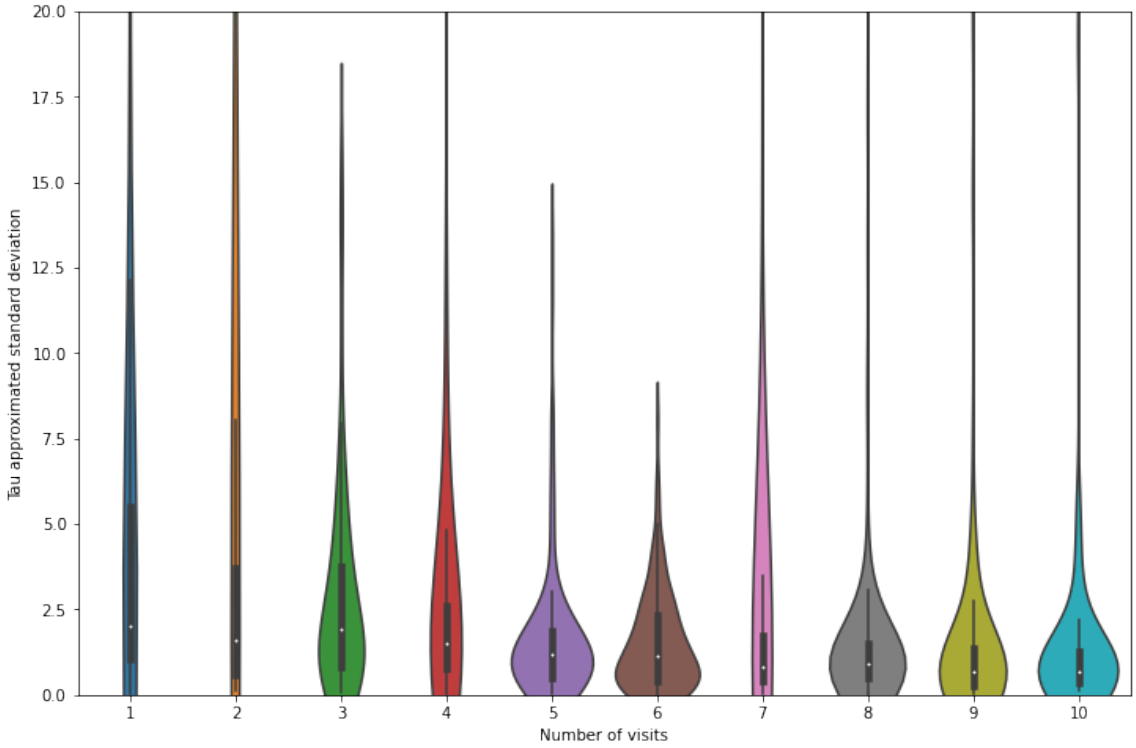

Figure 2.13: Violin plot of the approximated standard deviation of τ_i depending on the number of visits used for personalization.

the information for each individual parameter, for each item, level per level. The results are shown in Figure 2.14. The total information is obtained by summing over all visits, rather than averaging, so that it also reflects the imbalance of item level representation. The goal is to spot whether some items levels are not useful and whether the item should be reworked. We see that some items prove to be more useful for τ such as MDS-UPDRS III-2 or III-14, whereas some items bring little to nothing such as I-6 or III-11. Overall the information is mainly dominated by item level 1 with a bit level 2. Indeed the MDS-UPDRS scale is known to lean towards late disease stages sensitivity to the detriment of early sensitivity[48], which is one of its main drawbacks when used as an endpoint for clinical trials which are targeting early stages. Items with level 3 and 4 information are very seldom, meaning that the item scales are not really adapted to the distribution of patients in this cohort. However PPMI is not skewed towards early PD, so this indicates that the MDS-UPDRS scale suffers from an uninformativeness of the higher level values of its items, which are probably too severe to be observed.

information matrix," en, *Reliability Engineering & System Safety*, vol. 216, p. 107 968, Dec. 2021.

[48] M. H. Tosin, G. T. Stebbins, C. Comella, *et al.*, "Does MDS-UPDRS Provide Greater Sensitivity to Mild Disease than UPDRS in De Novo Parkinson's Disease?" en, *Movement Disorders Clinical Practice*, vol. 8, no. 7, pp. 1092–1099, 2021.

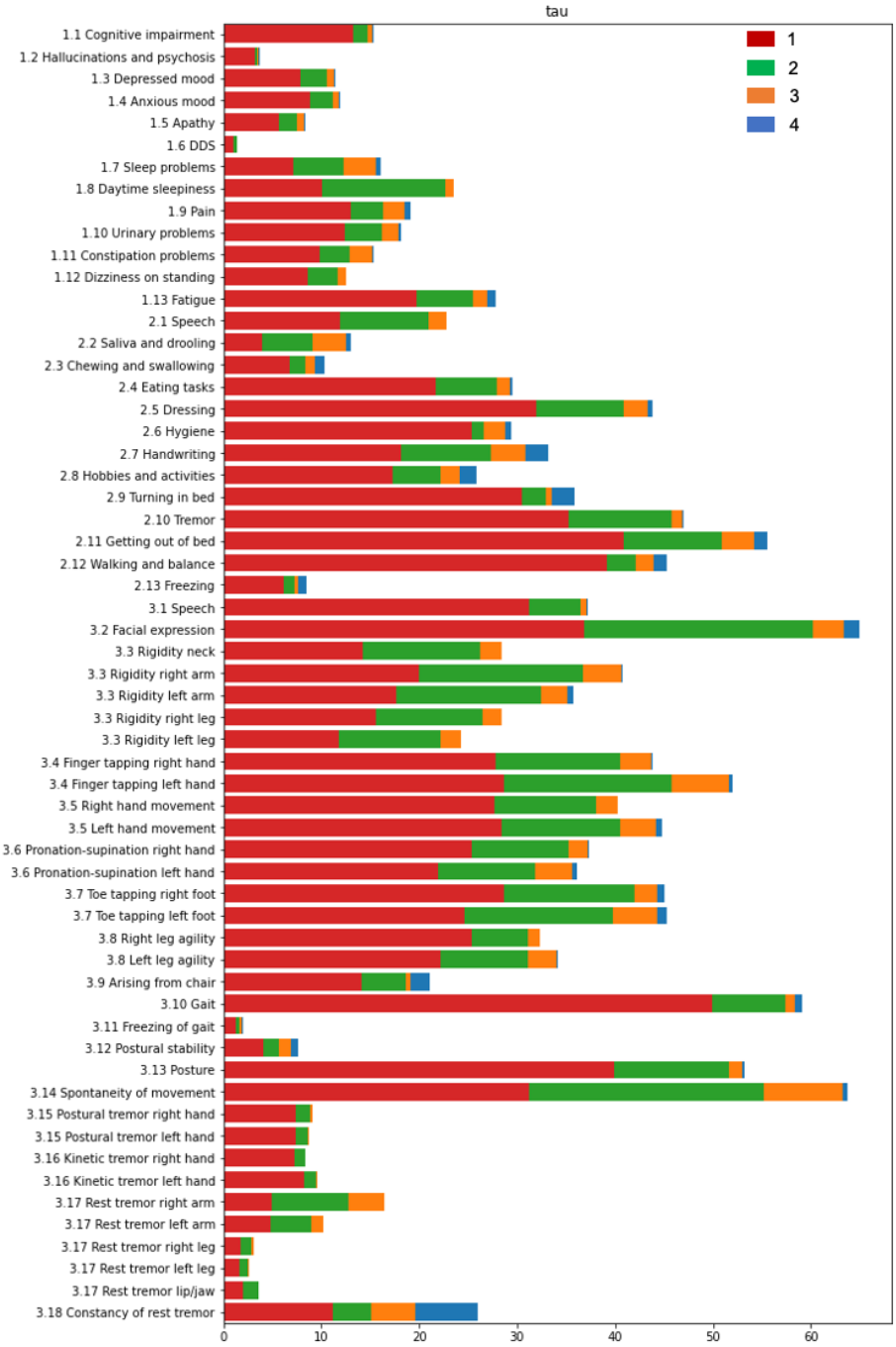

Figure 2.14: Fisher information for τ_i per item, per level. The information was summed over all visits, therefore there is an imbalance because of the item levels not being equally represented.

2.5 Conclusion

This chapter was centered around the extension of the DCM model to discrete data, i.e. binary and ordinal data. These types of data are ubiquitous in neurodegenerative cohorts where symptoms and cognitive scales are at the core of disease progression. We showed how the application of the DCM model to this type of data enlarged the horizon of potential analysis. From symptom onset prediction to fine-grained analysis of a cognitive scale, we showcased a wide array of interesting applications.

We also encountered the limitations due to modelling hypotheses and observations distribution in the cohort. Symptoms might be treated, which is in contradiction with the model's monotonic increase of the logistic. In the ordinal case, estimation of the model is very sensitive to the representation of the levels in the cohort and requires much more computational time.

The analysis of discrete data paved the way for our next chapter. Indeed, due to the nature of binary and ordinal observations, patients tend to be easily separable into groups, for instance those with a specific symptom and those without. Therefore we observed a lot of disparity in the cohorts based on those discrete features, which reflected in the estimated individual parameters of the DCM models that we estimated. The simulated data in the synthetic experiment section was purposefully designed to have the clusters in Figure 2.5. This illustration is an exaggeration of real-life data, but it shows that we now have the tools to uncover new clusters of patients. Disparities that were previously merged in aggregated scores can be disentangled with an item-level modelling. Techniques to extract those clusters are the focus of the next chapter of this thesis.

Chapter 3

Subtyping

This chapter addresses a fundamental issue of disease progression modelling. Theoretical hypotheses underlying the use of mixed-effect models with unimodal priors apply when the dataset is an homogeneous poopulation of individuals. However neurodegenerative diseases most often describe a spectrum, including more specific diseases, for instance Parkinson's disease and Lewy-body disease[1]. We therefore introduced the mixture models for Disease Course Mapping (DCM), allowing to take disease subtypes into account.

The method presented in this chapter led to the paper[2]. My contributions include the extension of the model to include latent classes, declined in two possible variants. The main novelty in this work however lies in a new heuristic to improve the estimation of the mixture, which is known to be the challenging part. This heuristic is described in the estimation chapter 5.2.1. The chapter is composed of two main parts, describing the two variants of the mixture models for DCM.

Contents of the chapter

[1] I. G. McKeith, "Spectrum of Parkinson's disease, Parkinson's dementia, and Lewy body dementia," eng, *Neurologic clinics*, vol. 18, no. 4, pp. 865–902, Nov. 2000.

[2] P.-E. Poulet and S. Durrleman, "Mixture Modeling for Identifying Subtypes in Disease Course Mapping," en, in *Information Processing in Medical Imaging*, A. Feragen, S. Sommer, J. Schnabel, *et al.*, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2021, pp. 571–582.

3.1 Context

Large medical cohorts with numerous biomarkers and cognitive scores are particularly well suited to analyse disease at a population level. Understanding how the disease progresses and what is the expected variability amongst individual is key. This variability can be driven by certain covariates, as showcased in Alzheimer's disease (AD). Indeed the genetic factor APOE-ϵ4 has major influence over the speed of the disease, and sex is also implicated in neurodegeneration, as shown in[3]. But when no prior knowledge and no specific covariate are available, the variability of individual disease patterns which is learned by the model is constrained by the specifications of the said model.

For the DCM model, as well as for standard mixed-effect models, the main limitation is the hypothesis of a unimodal distribution for random effects. The gaussian prior used in the DCM implicitly describes the population as homogeneous, and regularizes around it. Another constraint on the variability specific to the DCM model lies in the uniform speed hypothesis. Indeed the only random effect influencing the speed of the disease is the acceleration factor α_i, and it accelerates the disease timeline ψ_i which then accelerates uniformly all the markers progression η_k. This choice of modelling does not allow a patient to have a cognitive decline twice as fast as the normal progression while showing standard MRI brain atrophy for instance. Figure 3.1 illustrates how the value of $\boldsymbol{v}$ shapes the possible trajectories in the manifold $(0, 1)^2$.

These limitations are not adapted to the reality of neurodegenerative diseases, where different subtypes are often observed. For instance in AD, biological subtypes have been identified based on the tau pathology and brain atrophy[4]. In Parkinson's disease (PD), there is a commonly used classification between tremor dominant (TD) patients and postural gait instability disorder (PIGD) patients[5]. However these two groups do not clearly separate the whole population, and the question remains whether these subtypes are not actually two extremes of a continuum of variability. Moreover recent studies have shown that the subtyping is not stable in time[6], thereby questioning the classification system. We will discuss this particular case in the experiments section.

The heterogeneity in disease presentation is therefore structured around the subtype in which the individual belongs. Thus, subtyping is an essential task in disease progression modelling. However uncovering unknown clusters of individuals is a complex challenge on top of the estimation of the model parameters. It might thus be tempting to separate the clustering and the model estimation, for instance by performing *a posteriori* clustering. However this naive approach does not allow an unbiased estimation of model parameters. We will show with the DCM model to what extent ignoring subtypes in the model estimation leads to estimation errors.

The problem of subtyping has been well studied and we will focus on the literature of disease progression models. For mixed-effect models subtyping implies multimodal distributions, leading to mixture models. The mixture can be introduced at different levels. The most common is the mixture of random effects, as is implemented in longitudinal mixed-effect statistical frameworks

[3] B. Sauty and S. Durrleman, "Impact of sex and APOE-4 genotype on patterns of regional brain atrophy in Alzheimer's disease and healthy aging," *Frontiers in Neurology*, vol. 14, 2023.

[4] D. Ferreira, A. Nordberg, and E. Westman, "Biological subtypes of Alzheimer disease: A systematic review and meta-analysis," en, *Neurology*, vol. 94, no. 10, pp. 436–448, Mar. 2020.

[5] G. T. Stebbins, C. G. Goetz, D. J. Burn, *et al.*, "How to identify tremor dominant and postural instability/gait difficulty groups with the movement disorder society unified Parkinson's disease rating scale: Comparison with the unified Parkinson's disease rating scale," en, *Movement Disorders*, vol. 28, no. 5, pp. 668–670, 2013.

[6] T. Simuni, C. Caspell-Garcia, C. Coffey, *et al.*, "How stable are Parkinson's disease subtypes in de novo patients: Analysis of the PPMI cohort?" en, *Parkinsonism & Related Disorders*, vol. 28, pp. 62–67, Jul. 2016.

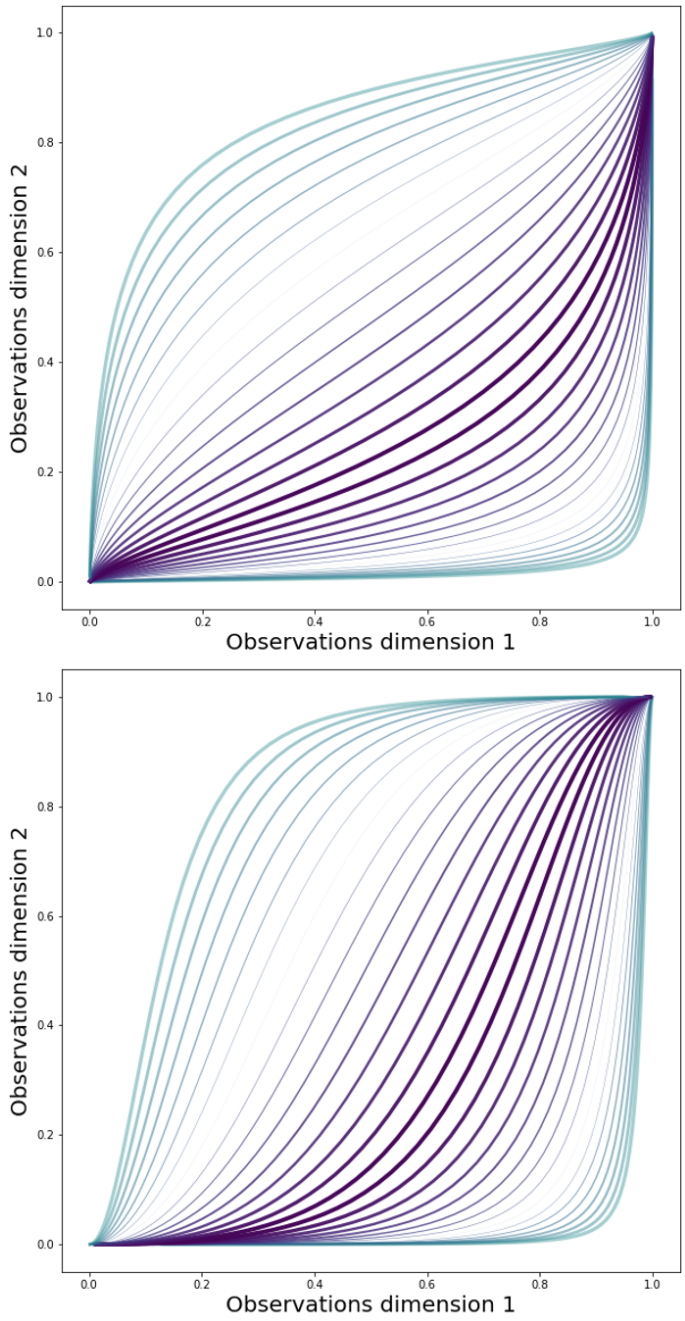

Figure 3.1: Disease Course Mapping geometric variability depending on $\boldsymbol{v}$.The bundle of trajectories is obtained by varying the values of the space shift $\boldsymbol{w}_i$ in a logistic curves DCM model (variability due to temporal random effects is not seen in the spatial representation). Top plot: $\boldsymbol{v} = (e^{-1}, e^{-1.5})$. Bottom plot: $\boldsymbol{v} = (e^{-1.5}, e^{-1})$

.

such as latent classes mixed models (LCMM)[7] or MONOLIX[8]. Another approach is to introduce a mixture at the residual noise level, which means that we can have several structural models describing the population subtypes. The challenge behind mixture models is the estimation, and approaches vary depending on the mixed-effect model chosen. The choices of algorithms will be further discussed in the dedicated chapter on optimization.

For disease subtyping many other solutions have been proposed outside of the scope of mixed-effects models. The Subtyping and Stage Inference algorithm (SuStaIn)[9] is an extension of an event-based model with a mixture and has shown very specific patterns of disease progression in AD[10]. Also in the context of AD and identification of disease subtypes in the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort, *ad hoc* methods have been developed for the classification of subtypes. The most famous one is the Murray-Dickson algorithm[11], and its adaptations[12]. More traditional models such as latent Dirichlet allocation have also been applied to the cohort[13].

Our choice to deal with this heterogeneity is thus situational, as it depends on the specification of the DCM model. The main concern for the ability of the model to accurately describe subtypes lies behind the limitation shown in figure 3.1, as we hypothesize that subtypes might lead to different *geometric* patterns rather than *temporal* patterns. This is why we first constructed the mixture model at the residual noise level, meaning that each cluster of individuals has a different DCM model, thus a different $\boldsymbol{v}$. In a second time we also describe a model with the mixture of random effects, which is more adapted to the identification of temporal clusters such as fast progressors versus slow progressors.

The work in this chapter presents materials from the published paper[14] which focused on the mixture at the residual noise level. We extend it with additional experiments as well as a comparison with a second mixture model at the individual parameters level.

3.2 Mixture of DCM models

The mixture of DCM models is a new layer atop of the hierarchical structure already built. If we write $q(\boldsymbol{y} \mid \boldsymbol{z}; \theta)$ the residual noise likelihood of the model parametrized by θ, with latent variables

[7] C. Proust-Lima, V. Philipps, and B. Liquet, "Estimation of Extended Mixed Models Using Latent Classes and Latent Processes: The R Package lcmm," en, *Journal of Statistical Software*, vol. 78, pp. 1–56, Jun. 2017.

[8] M. Lavielle and C. Mbogning, "An improved SAEM algorithm for maximum likelihood estimation in mixtures of non linear mixed effects models," *Statistics and Computing*, vol. 24, no. 5, pp. 693–707, Sep. 2014.

[9] A. L. Young, R. V. Marinescu, N. P. Oxtoby, *et al.*, "Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference," en, *Nature Communications*, vol. 9, no. 1, pp. 1–16, Oct. 2018.

[10] C. Shand, P. J. Markiewicz, D. M. Cash, *et al.*, "Heterogeneity in Preclinical Alzheimer's Disease Trial Cohort Identified by Image-based Data-Driven Disease Progression Modelling," eng, *medRxiv: The Preprint Server for Health Sciences*, p. 2023.02.07.23285572, Feb. 2023.

[11] M. E. Murray, N. R. Graff-Radford, O. A. Ross, *et al.*, "Neuropathologically defined subtypes of Alzheimer's disease with distinct clinical characteristics: A retrospective study," en, *The Lancet Neurology*, vol. 10, no. 9, pp. 785–796, Sep. 2011.

[12] S. L. Risacher, W. H. Anderson, A. Charil, *et al.*, "Alzheimer disease brain atrophy subtypes are associated with cognition and rate of decline," en, *Neurology*, vol. 89, no. 21, pp. 2176–2186, Nov. 2017.

[13] X. Zhang, E. C. Mormino, N. Sun, *et al.*, "Bayesian model reveals latent atrophy factors with dissociable cognitive trajectories in Alzheimer's disease," en, *Proceedings of the National Academy of Sciences*, vol. 113, no. 42, E6535–E6544, Oct. 2016.

[14] P.-E. Poulet and S. Durrleman, "Mixture Modeling for Identifying Subtypes in Disease Course Mapping," en, in *Information Processing in Medical Imaging*, A. Feragen, S. Sommer, J. Schnabel, *et al.*, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2021, pp. 571–582.

$\mathbf{z}$ and observations $\mathbf{y}$, then the residual noise likelihood of a mixture model with L clusters is:

$$Q(\boldsymbol{y} \mid \boldsymbol{z}; \theta^1, ..., \theta^L) = \sum_{c=1}^{L} \pi^c q(\boldsymbol{y} \mid \boldsymbol{z}; \theta^c) \tag{3.2.1}$$

where π^c denotes the probability of each cluster c with its own set of parameters $\theta^c = (\bar{\tau}^c, \bar{\xi}^c, \sigma^c_\tau, \sigma^c_\xi)$ and fixed effects $\boldsymbol{z}^c_{pop} = (\mathbf{g}^c, \mathbf{v}^c, \boldsymbol{\beta}^c)$. We assume that each individual has a true latent class c_i such that $\pi^c = \frac{\sum_{i=1}^{N} \mathbb{1}_{c_i=c}}{N}$ and a set of individual parameters $\mathbf{z}_i^{c_i} = (\tau_i^{c_i}, \xi_i^{c_i}, \mathbf{s}_i^{c_i})$ corresponding to the model attached to the cluster c_i. The total log-likelihood of the mixture model then writes:

$$\log\ Q(\boldsymbol{y}, \boldsymbol{z}, \theta^1, ..., \theta^L) = \sum_{i=1}^{N} \log\ q(\boldsymbol{y} \mid \boldsymbol{z}_i^{c_i}, \boldsymbol{z}_{pop}^{c_i}; \theta^{c_i}) \tag{3.2.2}$$

$$+ \sum_{i=1}^{N} \log\ q(\boldsymbol{z}_i^{c_i} \mid \boldsymbol{z}_{pop}^{c_i}; \theta^{c_i}) \tag{3.2.3}$$

$$+ \sum_{c=1}^{L} \log\ q(\boldsymbol{z}_{pop}^{c}; \theta^{c}) \tag{3.2.4}$$

The estimation algorithm is described in the optimization chapter. It is a mixture version of the Monte Carlo Markov Chain Stochastic Approximation Expectation Maximization (mMCMC-SAEM) algorithm, where we add the latent variables π_i^c for the probability of each individual i to be in each cluster c, and $(\boldsymbol{z}_i^c)_i$ are the individual parameters for individual i in cluster c. This allows for soft clustering throughout the estimation, avoiding one of the pitfalls of optimization with mixtures where sometimes individuals are assigned to a given cluster with hard labels.

3.3 Mixture of individual parameters

The second approach is to add the mixture at the individual parameters level. We still keep the notations for π^c the probability of cluster c and c_i the true class of individual i. The individual parameters priors become:

$$q(\begin{bmatrix}\tau_i \\ \xi_i\end{bmatrix} \mid \theta) = \sum_{c=1}^{L} \mathbb{1}_{c_i=c} \phi(\begin{bmatrix}\tau_i \\ \xi_i\end{bmatrix}; \begin{bmatrix}\bar{\tau}^c \\ \bar{\xi}^c\end{bmatrix}, \Sigma^c_{\tau,\xi}) \tag{3.3.1}$$

$$q(s_{ik} \mid \theta) = \sum_{c=1}^{L} \mathbb{1}_{c_i=c} \phi(s_{ik}; \bar{s_k}^c, \sigma^c_{s_k}) \tag{3.3.2}$$

where ϕ is the probability density function of a normal distribution (multivariate in the case of τ, ξ), and $(\bar{\tau}^c, \bar{\xi}^c, \Sigma^c_{\tau,\xi})$ are model parameters corresponding to the mean and covariance of the (τ_i, ξ_i) in cluster c. We also introduce a mean and a standard deviation for the sources. Parameters $(\boldsymbol{\beta}, \mathbf{g}, \mathbf{v})$ are shared by the clusters.

Remark: We removed the independence of individual parameters in the priors, by adding a covariance matrix for temporal parameters, while still keeping the sources independent (for the

Independent Component Analysis on the space-shifts to be relevant). In the experiment section we show why this was necessary.

Some conditions are required for the model to still be identifiable. Namely:

$$\mathbb{E}(\xi_i|\theta) = \sum_{c=1}^{L} \pi^c \bar{\xi}^c = 0 \tag{3.3.3}$$

$$\mathbb{E}(\boldsymbol{s}_i|\theta) = \sum_{c=1}^{L} \pi^c \bar{\boldsymbol{s}}^c = \mathbf{0} \tag{3.3.4}$$

$$\mathbb{V}(s_{ik}|\theta) = \sum_{c=1}^{L} \pi^c ((\sigma^c_{s_k})^2 + (\bar{s_k}^c)^2) = 1 \tag{3.3.5}$$

$$\tag{3.3.6}$$

Due to the maximization step in the EM algorithm not being explicit with the last two constraints on the sources 3.3.4 3.3.5, we only implemented a mixture for the time-related individual parameters (τ_i, ξ_i).

3.4 Experiments

3.4.1 Simulated data: Univariate model

We first show that the mixture of DCM models accurately recovers the ground truth parameters on simulated data. We generated data by initializing two unimodal disease course mapping models with chosen population parameters. We use the Kullback-Leibler divergence to determine the difference between the two model distributions. We then assume the two clusters have equal prevalence and we create 512 individuals, which consists in randomly attributing the individual to a cluster and sample individual parameters according to the distribution of the cluster. Next we arbitrarily decide the number and time of the "visits" for each individual, and we compute the values associated to these timepoints. Finally we add a small Gaussian noise to the output.

A simplified version of the model consists in using only one dimension $d = 1$, leading to an univariate model with no space-shifts. In this case, the only individual parameters left are the time-related ones: $(\xi_i)_i$ and $(\tau_i)_i$. This special case of the model converges much faster (about one minute for 2000 iterations), which allows us to initialize the model without having to use the initialization method described in the optimization section 19. We evaluated the class estimation with the area under the ROC curve (ROC AUC) metric. For the evaluation of population and individual parameters, we computed the absolute error between the ground truth parameters of each cluster and the parameters of the closest estimated cluster, this error being normalized by the standard deviation of the parameter. Figure 3.2 shows the reconstruction metrics on simulated data as a function of the Kullback-Leibler divergence between clusters.

With the ROC AUC, we are able to see that the mixture perfectly separates two clusters, once the two clusters are different enough. The estimation errors for population parameters increases slightly as the clusters get farther from each other, which is understandable since the algorithm needs to do more exploration. The mean error on individual parameters shows a very stable curve. The reconstruction is limited by the noise of the generated data. In all the univariate experiments,

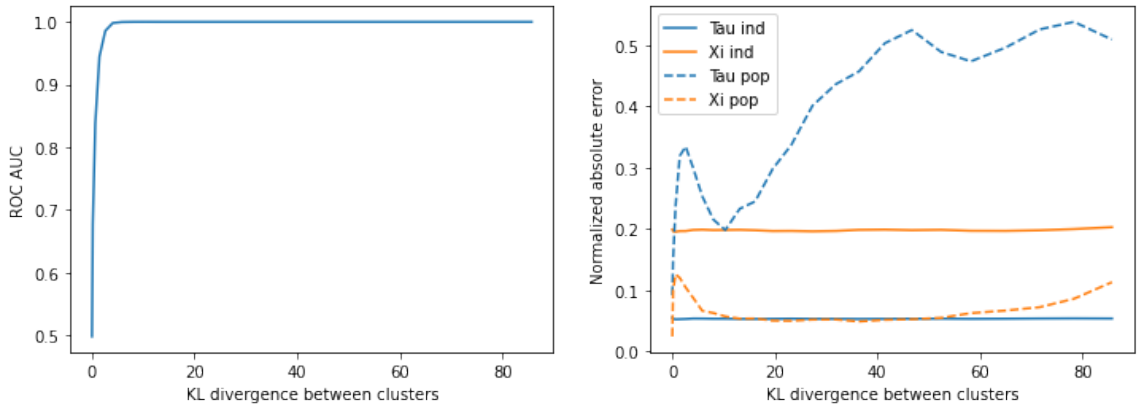

Figure 3.2: Evaluation metrics for the reconstruction of ground truth parameters. Left: ROC AUC for the estimated cluster probabilities of individuals π_i^c. Right: Absolute error between estimated parameters and ground truth, normalized by parameter standard deviation.

the estimated parameters allowed the model to reach a L2 loss close to the noise level. The consistent errors for individual parameters seemingly correspond to the range in which individual parameters differences can be mistaken for noise. For the population parameters, we see an increase in error as the clusters get farther from each other. This is likely due to the fact that the parameters need to drift farther from their initialization which hints towards an unfinished convergence. The DCM model implementation lacks a convergence criterion, a topic which will be discussed in the optimization chapter.

3.4.2 Applications to Alzheimer's disease data

We apply the method to the ADNI dataset. The data used in this chapter were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu), 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). The versions of the dataset used for this experiment were ADNI 1, 2, 3 and ADNI GO.

This application is a simple case to verify that our model works in practice, as we will later focus on Parkinson's disease which is the goal of our subtyping study. We first build a model considering only stable Alzheimer's disease (AD) patients and stable controls. This means that across all of their visits the stable controls only have a healthy control diagnosis, while stable AD must have at least two AD confirmed diagnosis. This leaves us with 618 controls and 730 AD patients, with an average of 6.5 visits per patient. The rest of the patients are considered as part of the Mild Cognitive Impairment (MCI) category.

We select 8 features that are most relevant to the progression of AD, which included cognitive scores, MRI-derived regional volumes and biomarkers level. Table 3.1 summarizes the features we used. MRI measures are intracranial volume measures. For MRI and CSF values, the data were

Table 3.1: ADNI demographics (includes controls, AD patients and MCI). MMSE: Mini-Mental State Examination, a cognitive score ranging from 0 (worst state) to 30 (perfectly healthy). ADAS-cog 13: Alzheimer's Disease Assessment Scale cognitive evaluation is a s-cognitive score ranging from 0 (healthy) to 85 (worst). RAVLT: Rey Auditory Verbal Learning Test, scaled between 0 (best) and 1 (worst). MRI measures for hyppocampus, ventricles and entorhinal are known to be sensitive to AD progression. CSF: Cerebrospinal fluid measures for proteins p-Tau and amyloid β42, which are at the heart of the biological process of neurodegeneration in AD, namely the amyloïd cascade.

	MMSE	ADAS-cog 13	AVLT	MRI Hippocampus
count	10246	9912	10131	9008
mean	26.726430	17.458136	0.529044	0.429449
std	3.931331	11.409325	0.183635	0.174020
min	0	0	0	0
25%	25	9	0.40	0.30
median	28	15	0.55	0.42
75%	29	23	0.66	0.55
max	30	85	1	1
	MRI Ventricles	MRI Entorhinal	CSF p-Tau	CSF Abeta42
count	9000	9006	2701	2702
mean	0.482749	0.490103	0.446442	0.397885
std	0.180927	0.177736	0.185596	0.199178
min	0	0	0	0
25%	0.35	0.37	0.32	0.23
median	0.48	0.47	0.44	0.40
75%	0.60	0.60	0.57	0.53
max	1	1	1	1

re-scaled between 0 and 1 so that the progression can be modelled by a logistic curve. If needed the score is reversed to be increasing instead of decreasing. We kept cognitive scores on their original scale in table 3.1, but they were applied the same re-scaling before being fed to the model, as we consider them as continuous measures due to the fine-grained scale.

We first fit a single DCM model on it. The posterior analysis of individual parameters highlights the need to take heterogeneity into account. The posterior distribution of the couple of temporal parameters (τ_i, ξ_i) is shown in Figure 3.3. The two modes correspond to the two classes of patients selected for this dataset. The point of this experiment is to emphasize the difference between the mixture of DCM models that we chose and a mixture of individual parameters model, which would only separate AD and controls in this case.

We then used the mixture of DCM. Optimal number of clusters was decided based on the Bayesian Information Criterion (BIC). The results of the mixture model after 4000 iterations of mMCMC-SAEM are shown in Figure 3.4. Figure A shows the probability of individuals to belong to cluster 1 vs cluster 2 (i.e. individuals on the left of the histogram belong to cluster 2). Overall cluster 1 gathers about 70% of the individuals. Notice that almost all controls belong to cluster 1. Figure B shows the geometric trajectories of the two clusters shown on two dimensions (the cognitive score MMSE and the MRI volume measure of the hippocampus). We chose these two dimensions as

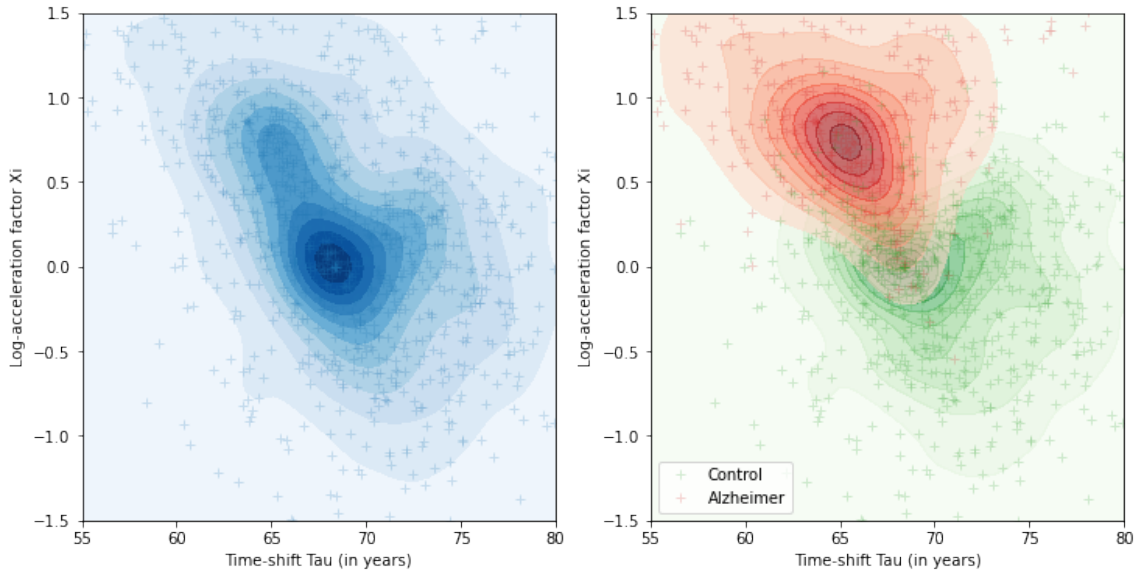

Figure 3.3: Scatter plot of individual parameters, with kernel density estimation ((KDE). Left: KDE on all AD and control patients. Right: KDE estimated separately for AD and controls.

they are probably the most important AD-related features amongst the 8. We wanted to showcase the differences in the geometric progression of biological measures against cognitive examinations. The center of the bundle is the average population trajectory γ_0, the variability shown corresponds to the curves obtained by varying the source 3 from -5 to $+5$. Notice that trajectories in cluster 2 show a decline in cognitive ability (MMSE) at a lower atrophy rate compared to cluster 1.

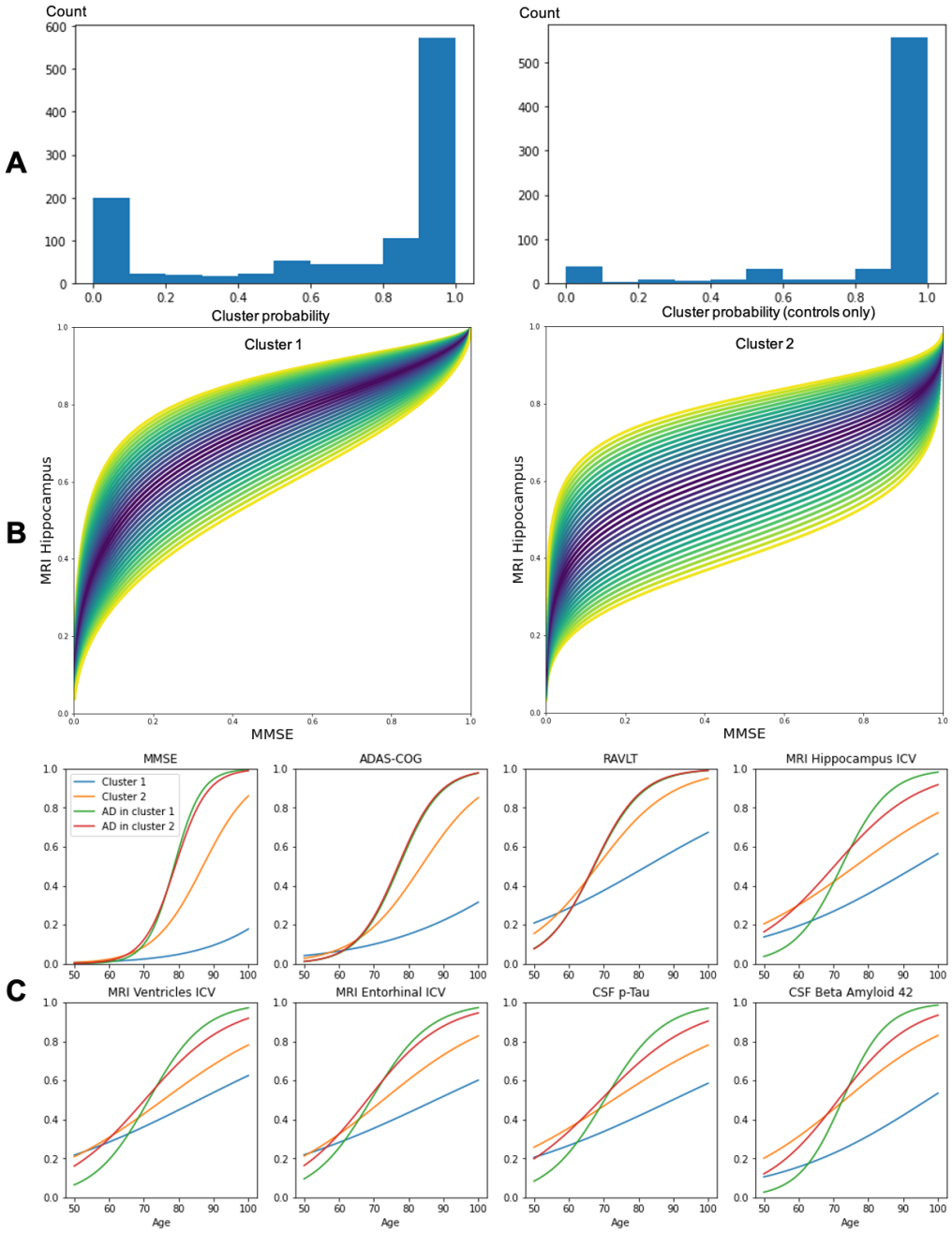

Figure 3.4: **A** Probability of individuals to belong to cluster 1 vs cluster 2. Left: all individuals. Right: controls only. **B** Geometric trajectories of the two clusters shown on two dimensions. **C** Cluster average trajectories in time.

Finally figure C shows the cluster's average trajectories in time. For each cluster we plotted the average trajectory for all patients and also for AD patients as the average of cluster 1 is skewed due to most of the controls belonging to this cluster. AD patients in both clusters have a similar cognitive evolution as shown by the cognitive scores (MMSE, RAVLT and ADAS-COG) whereas AD patients of cluster 1 have a faster progression on MRI-based and CSF biomarkers. This is a result of the different geometric patterns learned (as showcased by figure B), which is allowed by the mixture of DCM models and would not have been possible with a mixture of individual parameters only.

Several studies report an association between the atrophy rates of particular brain regions and cognitive decline[15,16], which lead to the identification of disease subtypes based on the differences in regional atrophies. Our analysis suggest that such associations are not systematic: similar pathological processes may lead to distinct pattern of cognitive decline, like similar cognitive decline may have distinct pathological processes.

Interestingly, the clustering does not correspond to AD and controls classes. Cluster 1 contains 88% of all controls and 65% of AD cases, meaning cluster 2 accounts for a minority of the data and contains mostly AD patients. To understand why the mixture model did not separate AD patients from controls, we estimated one model on AD patients only and one model on controls only. We then compute the log-likelihood of our models. Results are shown in table 3.2. They confirm that the mixture model explains the variability seen in the data better than prior categorisation based on diagnosis. We further estimated a mixture with the initial clusters being AD and controls to try and enforce the convergence towards a separation of AD and controls, and the algorithm still converged towards a similar version to the one presented in Figure 3.4.

Table 3.2: Log-likelihood of models

Model	Fit ($log\ q$)	Regularization ($log\ p$)	Total log-likelihood
Mixture	38100	-10391	27708
AD + Controls	37226	-9772	27454

As a final experiment, we performed individual personalization of MCI in the previous model, i.e. we estimated individual parameters including likelihood to belong to each cluster while keeping population parameters fixed. The distribution of MCI patients in the two clusters confirmed our hypothesis that cluster 2 is a specific subtype of AD: MCI associated to cluster 2 are mostly converters (66%) while 80% of non-converters are in cluster 1.

3.4.3 Experiment with mixture of individual parameters

We now show the difference with a mixture at the level of random effects. We first illustrate why we introduced the covariance matrix between τ and ξ in the mixture 3.3.1 with a practical experiment. We used the exact same dataset as in the previous section, i.e. ADNI with stable AD and controls. As showcased by Figure 3.3, the temporal individual parameters seem to have a bi-modal distribution, inciting us to fit a model with a mixture of individual parameters. We used

[15] X. Zhang, E. C. Mormino, N. Sun, *et al.*, "Bayesian model reveals latent atrophy factors with dissociable cognitive trajectories in Alzheimer's disease," en, *Proceedings of the National Academy of Sciences*, vol. 113, no. 42, E6535–E6544, Oct. 2016.

[16] S. L. Risacher, W. H. Anderson, A. Charil, *et al.*, "Alzheimer disease brain atrophy subtypes are associated with cognition and rate of decline," en, *Neurology*, vol. 89, no. 21, pp. 2176–2186, Nov. 2017.

the model described by equation 3.3.1, with no mixture on the sources since the implementation would be painstakingly difficult due to the constraints on the sources global distribution to keep the ICA. If we remove the covariance matrix, meaning that we keep τ and ξ independent, the model has a hard time finding two modes where the two coordinates are independent. We showcase the standard fit of two clusters in figure 3.5.

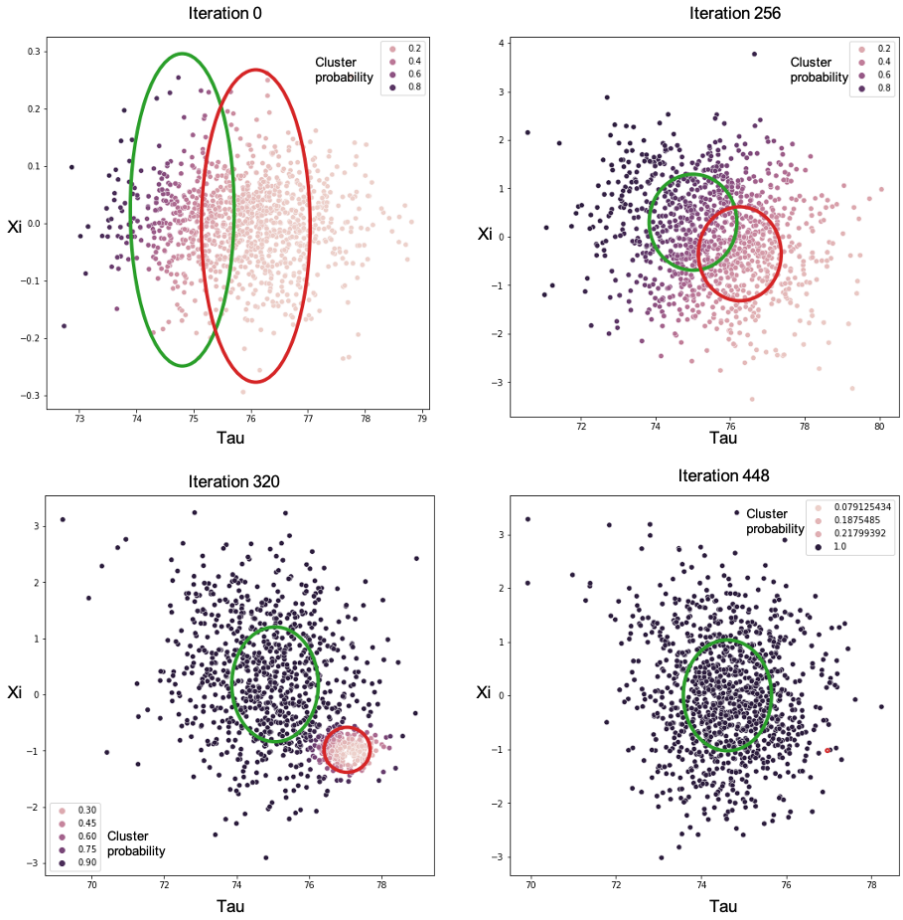

Figure 3.5: Distribution of individual parameters during the estimation (with mMCMC-SAEM). The two ellipsis correspond to the two Gaussians learned in the mixture. Each point is an individual, colored by its probability to belong to the green cluster. The red cluster progressively shrinks and eventually reaches zero probability.

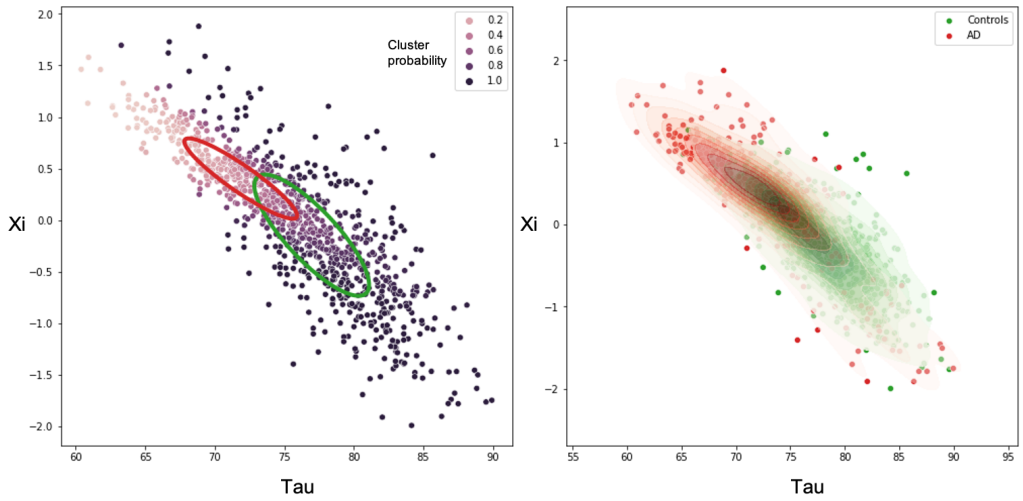

Figure 3.6: Distribution of time-related individual parameters. Left: individual parameters (τ_i, ξ_i) colored by their probability of belonging to cluster 1. Ellipsis correspond to the covariance matrix of each cluster centered around its mean. Right: individual parameters colored by their true label (AD or control), with KDE estimation.

It might be understandable that fitting two Gaussians with independent coordinates in the mixture is prone to fail, so we added the covariance matrix for more flexibility. With this model we obtain the distribution in Figure 3.6. First there is a clear distinction between this figure and Figure 3.3. The most striking part is that here we see that ξ and τ are far from independent. More specifically they tend to be anticorrelated, i.e. early onset also means faster progression. Then we notice that the two identified clusters match the KDE estimations of AD and controls, which is what was expected from the mixture of individual parameters. Finally we observe obvious facts about the clusters characteristics: AD patients (red cluster in both plots) have a lower τ and higher ξ, meaning that they start ealier and progress faster than controls.

3.4.4 Combination with DCM models on discrete data

Our goal was to be able to identify subtypes for the Parkinson's disease (PD) progression, but most of the meaningful features in PD are non-continuous variables (cognitive scores). To this end, one can combine the mixture models with the Bernoulli and ordinal noise models described in the previous chapter. For the DCM mixture model, we highlighted that one of the main perks of this mixture is to have different $\boldsymbol{v}$ in the clusters of the mixture. However in Bernoulli models, which are typically used for symptoms occurrence, the model parameter $\boldsymbol{v}$ is not as important as for continuous models. This due to the fact that in neurodegenerative disease once a symptom starts to appear, it is almost always present until it is treated. This implies that transitions from 0 to 1 tend to be sharp, and v_k becomes the parameter determining the time window of appearance of the symptom. Therefore trying to fit mixture DCM models with Bernoulli noise is not working well. On the other hand, mixture on individual parameters is a better match for Bernoulli noise models.

For ordinal noise models, the addition of the δ parameters as population parameters increases

the potential variety of models obtained with a mixture DCM models. However the increase in parameters due to the ordinal noise multiplied by the number of clusters makes the estimation almost impossible for large models, i.e. with d large (e.g. ≥ 10) or large number of levels per item. We estimated such as mixture on the PPMI ordinal data used to conduct experiments in chapter 2.3.3.

The fit ended (after about two weeks of estimation) with two clusters with probabilities 36% and 64%. The overall negative log-likelihood of the ordinal DCM model (without mixture) was 5.0×10^5 whereas the mixture with two clusters reached 4.8×10^5. This improvement is also validated but not really significant when using the Bayesian Information Criterion. We computed the BIC adpated to mixed-effects models[17], i.e. $BIC = -2 \log q + n_R \log(N) + n_F \log(N_{tot})$ where q is the likelihood, n_R the dimension of random parameters and n_F the dimension of fixed parameters, N the number of individuals and N_{tot} the total number of observations. The Figure 3.7 shows the mean trajectory in each cluster on the 15 first items of the MDS-UPDRS score. We see that cluster 1 (in blue, accounts for 36%) is mostly in advance of cluster 2, but the trajectories are not very different. The clustering does not correlate with known PD subtypes such as Tremor Dominant (TD) and Postural Instability and Gait Disorder (PIGD). It still remains to be proven that Parkinson's disease is composed of distinct subtypes or whether they are only extremes of a continuum of disease trajectories.

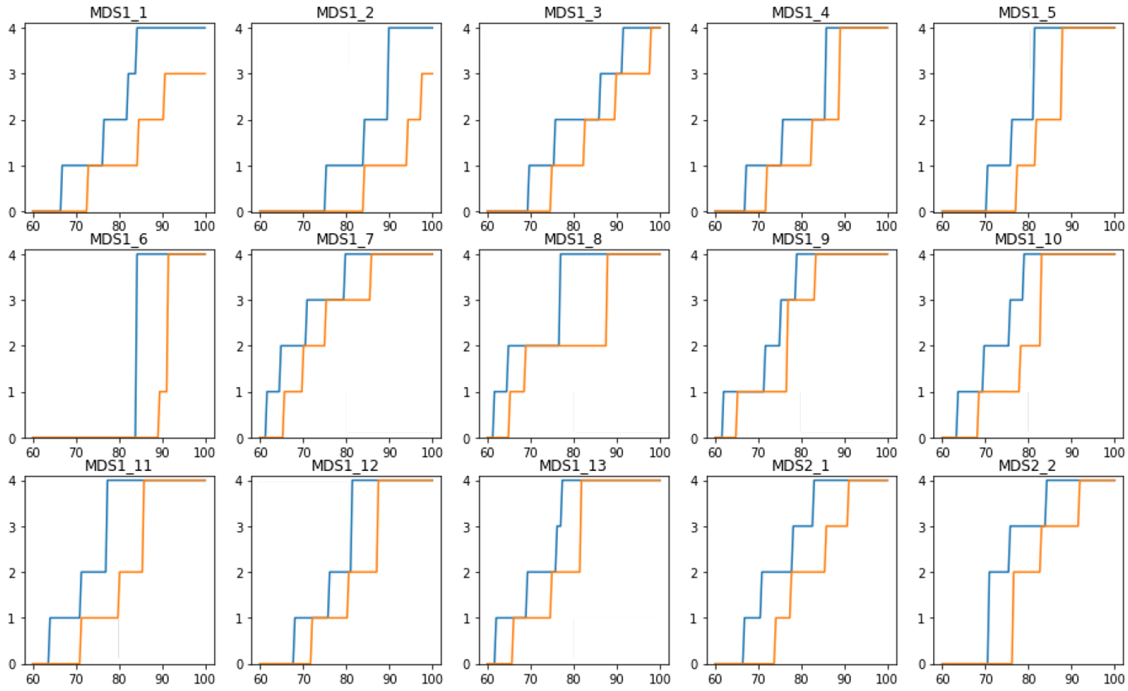

Figure 3.7: Maximum likelihood of the average trajectory in each cluster of the ordinal DCM mixture model. In blue, cluster 1. in orange, cluster 2. MDSX_Y is the item Y of part X of the MDS-UPDRS, described in section 2.3.3.

[17] M. Delattre, M. Lavielle, and M.-A. Poursat, "A note on BIC in mixed-effects models," en, *Electronic Journal of Statistics*, vol. 8, no. 1, Jan. 2014.

3.5 Discussion

The experiments here show the interest of the mixture of DCM models in deciphering different *geometric* patterns, but this is often not adapted to the obvious clusters such as controls vs patients. For this a mixture of individual parameters is more indicated. The main issue with mixture models lies in the estimation, which is tackled in the dedicated section 5. This is also where the main limitations of the method are highlighted, since the convergence is very complicated in practice despite the theoretical guarantees.

When it comes to potential applications, the two mixture possibilities have very different prospects. For the mixture of individual parameters:

- The possibility of identifying new subtypes is limited, since the clusters are designed on temporal markers only. This means that subtypes are characterized by their differences in onset age and speed of progression. These two features of disease progression do not require such a complex model as a DCM to be extracted from a longitudinal cohort, and clustering can be inferred without burdening the estimation of the mixed-effect model. Therefore this mixture is more adapted if subtypes are already known to affect the temporality of the disease. In this case the mixture is used to improve the model parameter estimation and to confirm the disease subtypes
- The model scales way better with the number of clusters than the mixture of DCM models, making it easier to fit in more complex settings than just a handful of clusters

In the case of the mixture of DCM models:

- This version is more indicated when looking for subtypes with specific disease patterns. This is very useful when trying to find new trends in a longitudinal dataset that has often been through plenty of clustering methods, as the clusters found are not based on easily observed features such as speed of progression. The different clusters found rely on differences in geometric trajectories, therefore separating different disease courses or pathogenicity mechanisms.
- However it will not separate temporal clusters or simpler variability patterns that can be handled by individual parameters (acceleration, time-shift and space-shift). The results are thus more "cryptic".
- The mixture has a very hard time scaling with the number of clusters as the number of parameters grows linearly and estimation is already very complicated with only 2 or 3 clusters

The code and implementation of these two mixture models has been done with the open-source library Leaspy https://gitlab.com/icm-institute/aramislab/leaspy, and will be released in the near future.

3.6 Conclusion

Neurodegenerative diseases are known to be very heterogeneous, with subtypes having very distinct disease manifestations. One of the challenges is to find methods able to separate the individuals in classes corresponding to *biologically* meaningful subtypes. One pitfall among many is to mistake a

continuum of variability for clusters, leading to characterization of extreme patients when the "average" patient does not actually belong to any cluster. In mixed-effect model terms, we should not try to enforce clusters with mixture when a random effect already captures the needed variability.

The need to subtype and find biologically meaningful markers of variability amongst patients is fueled by the hope of treatment. Pharmaceutical companies and laboratories have faced many failures in trying to develop neurodegenerative cures, especially for Alzheimer's disease. Their interest is going now towards more personalized approaches, by trying to target a specific population ("find the right treatment for the right person at the right time" - the mantra of personalized medicine). Hence the need to cluster patients. Treatment development are motivating most of the research in neurodegenerative diseases, and this conveniently allows us to transition to the next chapter of this thesis centered around treatment modelling.

Chapter 4

Treatment models

Preceding chapters have been driven by the will to improve a disease progression model. Learning the perfect curve to match the observed progression, allowing the model to take more types of data as input, jointly estimating latent classes for individuals, all of our work contributed to enhancing the Disease Course Mapping (DCM) model. We are able to achieve a higher performance in predicting the evolution of the patient's state. The question that arises is: what comes next ? How do we leverage this progression model into clinical application ? We need to bear in mind that the ultimate goal of our research is to cure the neurodegenerative diseases. Therefore we should try to gear towards improving the development of a treatment, by assessing its efficacy or identifying target populations. One route that has already been taken is the enrichment of clinical trials[1] ,[2]. In this chapter we will propose two methods to assess treatment effect. The first method is a generic use of any disease progression model to estimate treatment effect without requiring placebo or control patients. The second one is a theoretical extension of the DCM model to disease-modifying treatment effect.

Our contributions in this chapter are various. We defined a theoretical framework to perform treatment effect estimation using only treated patients and a disease progression model. We then validated this framework and applied it to the estimation of dopaminergic treatment effect in Parkinson's disease. This method has received the best poster award at the 2022 International Society for Clinical Biostatistics conference. Our final contribution is the introduction of a new model for disease-modifying treatment effects, called Piece-wise Geodesic model, and an analysis of its viability on synthetic data.

Contents of the chapter

[1] A. Tam, C. Laurent, S. Gauthier, *et al.*, "Prediction of Cognitive Decline for Enrichment of Alzheimer's Disease Clinical Trials," *The Journal Of Prevention of Alzheimer's Disease*, 2022.

[2] E. Maheux, J. Ortholand, C. Birkenbihl, *et al.*, "Forecast Alzheimer's disease progression to better select patients for clinical trials," en, Jul. 2021.

4.1 Introduction to treatment modelling

Modelling long term evolution of a process such as a neurodegenerative disease is a complex problem, which is the stem challenge of disease progression models. Different approaches have been developed to tackle this issue, either using mixed-effect models, event-based models or dynamic models based on ODE. All these models have proven to be quite potent proponents in the task of disease history description. However their potential in generalisation tasks such as prediction often relates to the (explicit or implicit) priors : the "template" family of curves for mixed-effect models or the class of functions modelling the dynamics for the ODE.

These approaches all rely on the fact that the disease progression is a common pattern, shared amongst the population. This is quite convenient in some observational studies, but this doesn't match the clinical trial framework. Because the treatment impacts the disease progression directly or through the mitigation of symptoms and other manifestations, the observations are impacted in a way that causes a rupture with the previous models : the future observations are not a simple "mechanical" consequence of the current state of the patient, but they take into account the treatment effect as well. Even though disease progression models perform very well when describing the disease's evolution without external interference, which we will call the natural disease progression, they do not adequately deal with a treatment acting over the disease progression. A successful curative treatment can even revert the progression or decrease some biomarkers' observed levels, making it difficult to use smooth models. Thus very few attempts have been made to include treatment in longitudinal models describing disease progression[3].

Two kinds of treatment are commonly acknowledged : symptomatic treatments and disease-modifying ones. Symptomatic treatments only affect the consequences of the disease by alleviating the burden of symptoms (hence the name). However they do not cure the root of the dysfunction-ment, whereas disease-modifying drugs impact the disease much earlier in the pathology pathways. These last ones are thus more desirable, provided that it is not too late to cure the patient (or even better, that the neurodegenerative process can be reverted). Those treatment effects are depicted in Figure 4.1.

The methodology to identify if a treatment (drug or therapy) is disease-modifying in clinical trials is debated, but it stands out that a simple placebo arm vs treatment arm is not sufficient. One has to either add delayed treatment start, or use "staggered withdrawal" strategies, in order to determine whether patients who followed treatment during a certain period of time catch up with placebo patients in disease state by the end of the study.

The question now is about modelling, or more especially what should we include in the models to allow treatment modification on the outcomes, whether it be symptomatic or disease-modifying. We will now present several ideas, within different models and contexts.

4.1.1 Modelling treatment effect

Traditionally a treatment is evaluated through a randomized controlled trial, which often has a small size. On the other hand, one can exploit the high amount of data in observational cohorts to better understand the treatment effect on disease progression[4]. One of the known issues when

[3] S. Chib and B. H. Hamilton, "Semiparametric Bayes analysis of longitudinal data treatment models," en, *Journal of Econometrics*, vol. 110, no. 1, pp. 67–89, Sep. 2002.

[4] S. L. Silverman, "From Randomized Controlled Trials to Observational Studies," en, *The American Journal of Medicine*, vol. 122, no. 2, pp. 114–120, Feb. 2009.

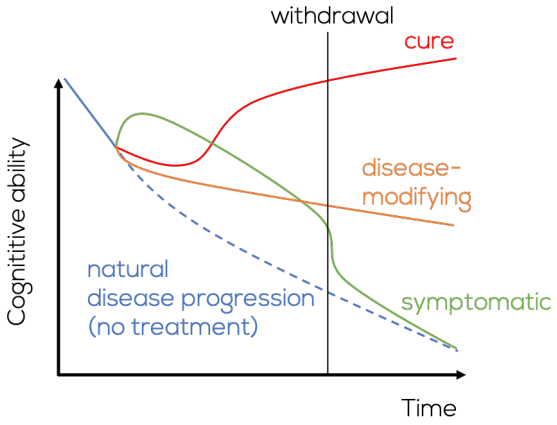

Figure 4.1: Scheme of treatment effect possible types. The Y-axis represents an observable marker for the cognition (could be another observable disease progression marker). The blue curve represents the natural progression of the disease without treatment. Three types of treatment effect are shown, where a cure is a specific disease-modifying treatment. The main difference to be noted is how the progression is modified after stopping the treatment.

using observational studies is that the data is not subject to randomization[5]. Furthermore, when a treatment has already been proven efficient in a disease, observational cohorts are only composed of treated subjects as everyone is prescribed the treatment. Our application focuses on Parkinson's disease, for which dopaminergic therapy is a known symptomatic treatment[6,7]. In the event that a cohort is only composed of patients without controls or placebo, we still would like to assess the treatment influence over the disease progression. This problem has never been addressed to our knowledge yet we believe it has many implications. We aim to learn meaningful information about the treatment without controls using a disease progression model.

But how do we model treatment effect traditionally ? This field is well documented in biostatistics[8], however, this is often done outside of the longitudinal scope. Treatment effect models often apply to clinical trials, where the treatment effect is assessed as the mean difference in the designated endpoint of the trial between the placebo arm and the treated arm. This difference is known as the average treatment effect (ATE), and is measured at a precise time-point, e.g. a fixed number of weeks from trial start. To add some useful information about variability in treatment effect, one can assess the treatment effect within certain groups defined by a covariate. This quantity is known as conditional average treatment effect. The problem of measuring treatment

[5]P. C. Austin and A. Laupacis, "A tutorial on methods to estimating clinically and policy-meaningful measures of treatment effects in prospective observational studies: A review," eng, *The International Journal of Biostatistics*, vol. 7, no. 1, p. 6, 2011.

[6]R. M. A. de Bie, C. E. Clarke, A. J. Espay, *et al.*, "Initiation of pharmacological therapy in Parkinson's disease: When, why, and how," eng, *The Lancet. Neurology*, vol. 19, no. 5, pp. 452–461, May 2020.

[7]C. Lungu, J. M. Cedarbaum, T. M. Dawson, *et al.*, "Seeking progress in disease modification in Parkinson disease," en, *Parkinsonism & Related Disorders*, vol. 90, pp. 134–141, Sep. 2021.

[8]P. C. Austin and A. Laupacis, "A tutorial on methods to estimating clinically and policy-meaningful measures of treatment effects in prospective observational studies: A review," eng, *The International Journal of Biostatistics*, vol. 7, no. 1, p. 6, 2011.

effect conditionally on some covariates is a standard regression problem[9], for which the current state-of-the-art is Bayesian Additive Regression Trees (BART)[10]. This method provides a flexible and agnostic framework without requiring too much data. However the longitudinal aspect of the treatment effect is often not accounted for, and treatment effect models with progression are still not standard[11]. One can still add time as a covariate for the CATE regression model, but this would leave the regression model to learn both the disease progression dynamics as well as the treatment effect. This is not efficient as we could use disease progression models to alleviate the learning task. One could add the treatment in a mixed-effect model as time-dependent covariate, but this would only estimate the treatment effect without its interaction with other covariates. Having to specify the potential interactions of the treatment with other variables is possible, but tedious and will introduce more complexity in the model.

We will show in the additive treatment model section 4.2 how we can do this. In this section, we propose a two-step approach to address these challenges by combining disease progression models with treatment effects. Our method is fully Bayesian. A longitudinal model describes the latent disease progression and a regression model learns the treatment effect as the difference between the latent disease state and the observed disease state. We will tailor our method for an application to Parkinson's disease, but it can be generalized fairly straightforwardly. Based on the Parkinson's Progression Markers Initiative (PPMI) observational cohort[12], we will use the model to describe the dopaminergic therapy's effect on clinical assessments, including the motor evaluation, quality of life and imaging data.

4.1.2 Disease-modifying treatment effect

Another drawback of ATE and CATE models is that they often ignore the difference between disease-modifying and symptomatic effects. As they are defined as a mean difference between a treated arm and a placebo arm, they do not account for the longitudinal aspect. As noted in[13], the disease-modifying treatment effect is poorly accounted for in most models. Often it is assessed as a change in the rate of disease progression, but this is limited to linear models. Most often disease progression models are used separately on the treated arm and the placebo arm, and the disease-modifying effect is assessed as the difference between the two progression models[14,15].

Some classes of models are better suited to model disease-modifying mechanisms, namely dynamical models[16] or generative models where we could simulate treatment effect as an intervention

[9] J. Abrevaya, Y.-C. Hsu, and R. P. Lieli, "Estimating Conditional Average Treatment Effects," *Journal of Business & Economic Statistics*, vol. 33, no. 4, pp. 485–505, Oct. 2015.

[10] D. P. Green and H. L. Kern, "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees," *The Public Opinion Quarterly*, vol. 76, no. 3, pp. 491–511, 2012.

[11] L. L. Raket, "Progression models for repeated measures: Estimating novel treatment effects in progressive diseases," en, *Statistics in Medicine*, vol. 41, no. 28, pp. 5537–5557, 2022.

[12] K. Marek, D. Jennings, S. Lasch, *et al.*, "The Parkinson Progression Marker Initiative (PPMI)," en, *Progress in Neurobiology*, Biological Markers for Neurodegenerative Diseases, vol. 95, no. 4, pp. 629–635, Dec. 2011.

[13] P. Chan and N. Holford, "Drug Treatment Effects on Disease Progression," *Annual Review of Pharmacology and Toxicology*, vol. 41, no. 1, pp. 625–659, 2001.

[14] S. F. Cook and R. R. Bies, "Disease Progression Modeling: Key Concepts and Recent Developments," en, *Current Pharmacology Reports*, vol. 2, no. 5, pp. 221–230, Oct. 2016.

[15] C. S. Venuto, N. B. Potter, E. Ray Dorsey, *et al.*, "A review of disease progression models of Parkinson's disease and applications in clinical trials," en, *Movement Disorders*, vol. 31, no. 7, pp. 947–956, 2016.

[16] B. O. Taddé, H. Jacqmin-Gadda, J.-F. Dartigues, *et al.*, "Dynamic modeling of multivariate dimensions and their temporal relationships using latent processes: Application to Alzheimer's disease," en, *Biometrics*, vol. 76, no. 3, pp. 886–899, 2020.

on the system[17]. On the other hand, non-linear mixed-effect model lack a general framework for such effect, and we will discuss about a possible choice of model in the disease-modifying model section.

4.2 Additive treatment effect model

Would it be possible to measure the effect of a treatment using a longitudinal observational cohort that lacks a control arm? To answer this, we propose a novel application of disease progression to assess the treatment effect in neurodegenerative diseases. In this section we provide a method to infer treatment effects by leveraging patient data in the absence of a placebo/control arm.

This model was motivated by the PPMI cohort, and more especially the way L-dopa treatment impacts Parkinson's disease. PD is characterized by the neurodegeneration of dopaminergic neurons, starting from the substantia nigra, and the subsequent loss of dopamine production in the brain. Dopamine is essential to the motor neurons, explaining the symptoms related to movement and tremor. The widely adopted treatment is the levodopa or L-dopa, the precursor of the dopamine. It provides the material needed to the brain of the patient in order to produce the lacking dopamine. This essentially results in a disappearing of motor issues, at least during a certain phase of the disease.

This treatment is purely symptomatic, as it does not help prevent further neurodegeneration. However it is highly efficient during the first years after diagnosis. Still, this yields a more simple task of modelling, since we do not have to take disease-modifying effects into account. Therefore the treatment effect can be directly measured at each timepoint, without requiring any longitudinal modelling.

Furthermore, PPMI provides us the very data we need for it: "OFF" and "ON" measures. Baseline visits are required to be clean of treatment, and enrolled patients should not have followed a long treatment prior to the baseline. Then patients can undergo dopaminergic therapy, the choice being to the discretion of the clinician and the patient. Values of dopaminergic treatment dose (L-dopa, dopamine agonists and others) are reported as L-dopa equivalent. For each subsequent visit, patient undergoing therapy will have two measures: one measure "OFF" deprived of treatment effect (in practice, patients are required to be clean of treatment for more than 6 hours) and a measure "ON" when the drug is effective (approximately one hour after taking medication). Even though the OFF state is not exactly the same as untreated PD, studies use it as a proxy to the state of a patient without any treatment[18]. Treatment effect can then be approximated as the raw difference between the ON state and the OFF state.

For this, we propose to leverage disease progression models. Our proposed approach combines them with a treatment model. The disease progression model computes the latent natural disease trajectory under the no-treatment hypothesis. The treatment model is a regression on the difference between the observed trajectory and the predicted one. One can interpret the natural disease progression predictions as a digital control twin of the patient who took the treatment.

Let's introduce notations and the mathematical formulation for this. Let $\mathbf{y}_{ij} \in \mathbb{R}^d$ be the d-dimensional observation of patient i at age t_{ij}, accompanied by a set of covariates $\mathbf{X}_{ij}$. Each patient i has n_i observations. We use the index k for the dimension coordinates y_{ijk}. We introduce the

[17] Y. Iturria-Medina, F. M. Carbonell, R. C. Sotero, *et al.*, "Multifactorial causal model of brain (dis)organization and therapeutic intervention: Application to Alzheimer's disease," en, *NeuroImage*, vol. 152, pp. 60–77, May 2017.

[18] R. Cilia, E. Cereda, A. Akpalu, *et al.*, "Natural history of motor symptoms in Parkinson's disease and the long-duration response to levodopa," *Brain*, vol. 143, no. 8, pp. 2490–2501, Aug. 2020.

random variable Y_{ijk} for the observations, whose distribution generated the observations. We will use the terms "ON state" for the trajectory of the patient under treatment, while "OFF state" will be acknowledged as the natural disease history state (which is not exactly true). The ON and OFF states are indicated by the treatment variable T: $T_{ij} = 1$ for the ON state or $T_{ij} = 0$ for the OFF state, and the realizations of the variable Y_{ijk} will be superscripted with "ON" and "OFF" accordingly. Conditional average treatment effect in the most general case can therefore be written as the following :

$$CATE_k(\mathbf{x}) \quad = \quad \mathbb{E}(Y|T=1, \mathbf{X}=x) - \mathbb{E}(Y|T=0, \mathbf{X}=\mathbf{x}) \tag{4.2.1}$$

If we had infinite samples for ON and OFF measures, we could simply estimate the $CATE$ via the difference between the means of samples matching the covariate x. But this is not always possible, for instance if x is a continuous variable. Instead, we will use a regression of $CATE_k(x)$ as a function of x, which corresponds to the covariates $\mathbf{X}_{ij}$. These covariates can include all types of internal variables, such as the OFF state for instance, and external cofactors such as sex or genetic risk factors. The covariates can be time-dependent as well.

In our observational cohort, we have visits for $T_{ij} = 1$ but rarely with $T_{ij} = 0$ as measure as the disease worsens. In the case where we have a model for disease progression, the OFF state is not a data, but the output of the model:

$$y_{ij}^{OFF} = \eta(\mathbf{z}_{ij}; \psi) + \epsilon_{ij} \tag{4.2.2}$$

where η is the disease progression model, parametrized by ψ and depending on individual time-dependent covariates $\mathbf{z}_{ij}$. We call η the *latent* disease progression model. The noise structure ϵ_{ij} can be chosen depending on the structure of the data. In the continuous case we choose:

$$\epsilon_{ij} \sim \mathcal{N}(\mathbf{0}_d, \Sigma_d)$$

with Σ diagonal for the sake of simplicity, but more complex structures can be introduced : time-dependent noise with Gaussian processes, a different noise covariance matrix... We then can infer the approximated treatment effect values:

$$\Delta_{ijk} = y_{ijk}^{ON} - \eta_k(t_{ij}, x; \mathbf{z}_i, \theta) \tag{4.2.3}$$

We will now regress the $CATE$ using these approximated values. What we call the *treatment effect model* writes:

$$\phi(\mathbf{X}_{ij}; \omega) = \Delta_{ij} + \epsilon'_{ij} \tag{4.2.4}$$

where ω are the parameters of the regression and ϵ' is the residual noise. ϕ is the difference between the value of the marker observed under treatment and the *latent disease progression model* which is the expected value under no treatment. Our method is pictured in Figure 4.2.

Plugging (4.2.2) and (4.2.4) together we obtain :

$$\mathbf{y}_{ij}^{ON} = \eta(\mathbf{z}_{ij}; \psi) + \phi(\mathbf{x}; \omega) + \epsilon_{ij} - \epsilon'_{ij} \tag{4.2.5}$$

Since we consider ψ and ϵ_{ij} to be learned already, we can assume that ϵ'_{ij} will summarize the residual noise of both the disease progression model η and the treatment effect model f. In the

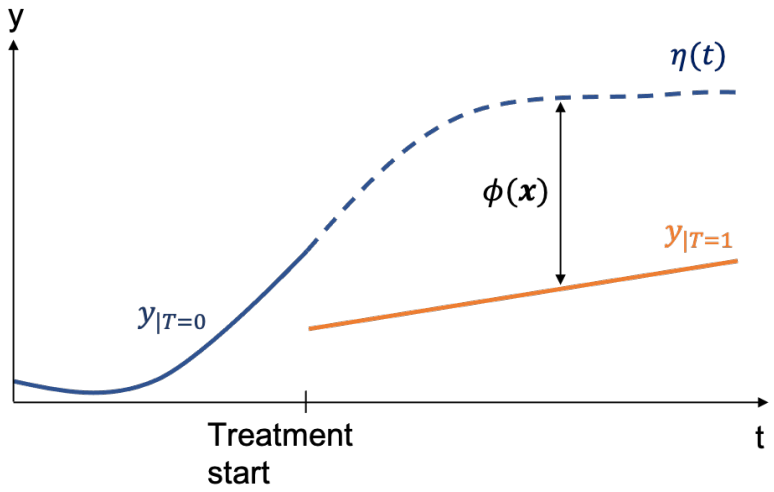

Figure 4.2: Model scheme. The plain curve is the progression of one individual, in blue before treatment starts and in orange after. The dashed blue curve is the predicted progression of the disease without treatment. Treatment effect ϕ can be estimated as the difference between the dashed blue line and the orange line, as a function of variables $\mathbf{x}$.

case where ϵ_{ij} has a specific bias (for instance the model has a hard time fitting young individuals), we can decompose this noise as the sum of a structured bias and pure noise. In this case the bias will be included in the learning of ϕ. The condition for ϕ to be an unbiased estimator of $CATE$ is then:

$$\forall x, \mathbb{E}(Y_{ijk}|T_{ij} = 0, \mathbf{X}_{ij} = \mathbf{x}) - \eta_k(\mathbf{z}_{ij}; \psi) = 0 \tag{4.2.6}$$

In practice, this means that the predictions of our progression model should not be biased when conditioned on covariates. Therefore we need to verify after the calibration of the η model that the residuals are independent of the covariates used in the treatment effect model ϕ. It is often easy to validate this condition for covariates that are external to the progression model, such as sex or ethnicity. However the model predictions are usually very dependent on other covariates, such as age or time since diagnosis. Indeed it is well known that prediction models have a hard time predicting a long time ahead. DCM for instance is known to make long-term predictions that are prone to under-estimation.

However we can also understand condition 4.2.6 locally: for all x for which the disease progression models are unbiased, then $\phi(\mathbf{x})$ can be considered as an unbiased estimate of $CATE(\mathbf{x})$. This is less constraining as we can empirically select the range of values for x for which we make accurate predictions.

We now present the models used in our application for η and ϕ.

4.2.1 Latent disease progression model

Any longitudinal model is suitable in this application, this is not specific to the DCM. Indeed the latent disease progression model can be any disease progression model as long as it provides an unbiased estimation of $E(Y_{ijk}|T_{ij} = 0, \mathbf{x}_{ij} = x)$, so that the inferred values $\tilde{y}_{ijk}$ can be unbiased estimators of CATE. Therefore we will make several applications: two with the DCM model, and one with another disease progression model.

For the second disease progression model, we will use a dynamical model fitted using the penalty method. This application was done in collaboration with Pr. Bruno Jedynak and PhD Michael Wells from Portland State University who developed this method. The penalty method is an iterative method for fitting a system of Ordinary Differential Equations (ODE) to data. Specifically, we fit an ODE of the form:

$$\frac{\partial}{\partial t}\eta_k(t) = f(\eta_k(t)) \tag{4.2.7}$$

It minimizes a cost over a set of discretized trajectories and functions f in a space of smooth functions, specifically a reproducing kernel Hilbert space. The cost function has three terms: a data term, a penalty term, and a regularization term. The data term keeps the trajectories close to the data. The penalty term ensures the trajectories follow the ODE (4.2.7). The regularization term corrzsponds to the norm of the function f, called the field, in the RKHS and it is meant to prevent overfitting. The penalty term is a loose constraint on the ODE, namely $||\frac{\partial \eta_k}{\partial t} - f(\eta_k)|| \leq \frac{1}{\lambda}$, where λ is increased at each iteration. The optimal fit is found after several iterations when the algorithm converges. Details and validation of the algorithm are provided in[19].

4.2.2 Treatment effect model

We chose the treatment effect model $\phi : \mathbf{x} \mapsto \phi(\mathbf{x})$ to be Bayesian. Indeed we are interested in a descriptive model for the treatment effect, and a Bayesian framework provides reliable estimates with credibility intervals. We used three treatment effect models:

- Bayesian linear regression (BLR): a model able to capture linear relationships between predictors and outcome, but which does not capture interactions between two input variables x_l and $x_{l'}$, thus limiting the descriptive power of the model

- Bayesian additive regression trees (BART): the state-of-the-art model for CATE regressions, which is flexible enough to capture any relationship between the covariates $\mathbf{x}$ and the outcome. However, in practice the trees did not embed significant interactions between input variables so we added a third model

- Bayesian kernel machine regression (BKMR): a multivariate kernel regression model allows for a lot of flexibility due to the choice of the kernel and the hyperparameters associated with it. This model, although harder to fit and more dependent on variable selection, has the advantage of modeling interactions between predictors which is very valuable from a descriptive point of view

[19] K. Lahouel, M. Wells, V. Rielly, *et al.*, *Learning nonparametric ordinary differential equations from noisy data*, Feb. 2023.

BLR and BART allow the computation of variable importance as a percentage of the explained outcome variance. Since BKMR is less performant when the dimension of $\mathbf{x}$ increases, we used BLR and BART for variable selection to reduce dimensions to a meaningful set of covariates $\mathbf{x}_{|S}$. More details about each method are provided next.

Bayesian linear regression

A deeper introduction to BLR models can be found in[20]. With our notations the model takes the following form (in a single dimension for the sake of simplicity, extension to multidimensional output is straightforward):

$$\phi(\mathbf{x}; \alpha, \beta) = \alpha^T \mathbf{x} + \beta \tag{4.2.8}$$

with $\alpha \in \mathbf{R}^m$ with m being the dimension of the covariates $\mathbf{x}_{ij}$ and $\beta \in \mathbf{R}$. The BLR model describes α and β as random variables for which we provide normal priors $\alpha_l \sim \mathcal{N}(0, \sigma_l), \forall l \in [|1, m|]$ and $\beta \sim \mathcal{N}(0, \sigma_\beta)$ where $(\sigma_l)_l$ and σ_β are hyperparameters. We decided to fix those hyperparameters to values depending on the training data: $\sigma_\beta = Var(y_{ij})$, $\sigma_l = \frac{Cov(x_{ijl}, y_{ij})}{Var(x_{ijl})}$ (which would be the slope coefficient in a non-Bayesian setting).

Bayesian kernel machine regression

First we will describe standard kernel regression. Here we use a reproducing kernel Hilbert space $\mathcal{H}$, $f \in \mathcal{H}$ with a kernel k. In the applications we chose the standard gaussian kernel $k(\mathbf{x}, \mathbf{x}') = exp(-\frac{||\mathbf{x}-\mathbf{x}'||^2}{\sigma^2})$. With the representer theorem, ϕ can be expressed as a function of the observations :

$$\phi(\mathbf{x}) = \sum_{ij} w_{ij} k(\mathbf{x}, \mathbf{x}_{ij}) \tag{4.2.9}$$

In order to keep the dimension of the problem within a reasonable range and to avoid ill-conditioned kernel matrices, we select only a few control points $(\mathbf{x}_i^c)_{i=1...n_c}$. These points can be placed on a grid in $\mathbb{R}^m$ if m is the dimension of $\mathbf{x}$ for a regular spacing, or we can chose to subsample the data $(\mathbf{x}_{ij})_{i=1...n, j=1..n_i}$ using the Kmeans++ initialization method[21]. Hence :

$$\phi(\mathbf{x}) = \sum_{i} w_i k(\mathbf{x}, \mathbf{x}_i^c) \tag{4.2.10}$$

The least squares problem can be formulated as an optimization problem (every uppercase corresponds to the vectorization of the lowercase variable):

$$\min_{i=1...n_c, w_i \in \mathbb{R}^d} \sum_{ij} ||\mathbf{y}_{ij}^{ON} - \eta(\mathbf{z}_{ij}; \psi) - \phi(\mathbf{x}_{ij})||^2 + \lambda ||f||_{\mathcal{H}}^2 \tag{4.2.11}$$

$$\min_{W \in \mathbb{R}^{n_c \times d}} ||Y^{ON} - \eta - KW||^2 + \lambda \sum_i W_{:,i}^T K^c W_{:,i} \tag{4.2.12}$$

[20] A. Gelman, J. B. Carlin, H. S. Stern, *et al.*, "Bayesian Data Analysis," en, Chapman and Hall/CRC, Nov. 2013.
[21] D. Arthur and S. Vassilvitskii, "K-means++: The Advantages of Careful Seeding," en,

where $\mathbf{y}_{ij}^{ON}$ refer to the observations in the ON state, $K = (k(\mathbf{x}_{ij}, \mathbf{x}_i^c)) \in \mathbb{R}^{\sum_i n_i \times n_c}$ and $K^c = (k(\mathbf{x}_i^c, \mathbf{x}_i^c)) \in \mathbb{R}^{n_c \times n_c}$.

The solution is explicit as this problem is convex and the gradient can be computed to look for the critical point, which yields:

$$W^* = (K^T K + \lambda K^c)^{-1}(Y^{ON} - \eta) \tag{4.2.13}$$

BKMR[22] is the result of adding a Bayesian framework to kernel regression. This is very similar to Gaussian processes, but here we present the model from the kernel regression point of view so we refer the reader to[23] for equivalences. We describe a single-dimension setting, as the extension to multidimensional output is straightforward.

The function ϕ is still approximated in a reproducing kernel Hilbert space (RKHS) where the basis functions are $\{k(\mathbf{x}, \cdot), \mathbf{x} \in \mathbf{R}\}$ for a chosen kernel $k : \mathbf{R}^m \times \mathbf{R}^m \to \mathbf{R}$ which is symmetric positive-definite. The representer theorem implies that the minimizing function in the RKHS is of the following form $\sum_{ij} \alpha_{ij} k(\mathbf{x}_{ij}, \cdot)$. We add a random intercept to the model which yields:

$$\phi(\mathbf{x}; \alpha, \beta) = \sum_{ij} \alpha_{ij} k(\mathbf{x}_{ij}, \mathbf{x}) + \beta \tag{4.2.14}$$

with $\alpha \in \mathbf{R}^N$ with N being the total number of observations and $\beta \in \mathbf{R}$. In the Bayesian setting, we add priors to those parameters: $\alpha_{ij} \sim \mathcal{N}(0, \sigma_\alpha), \forall i \in [|1, n|], \forall j \in [|1, n_i|]$, $\sigma_\alpha \sim \mathcal{N}(0, \bar{\sigma_\alpha})$ and $\beta \sim \mathcal{N}(0, \sigma_\beta)$.

As for BLR we define the hyperparameter $\sigma_\beta = Var(y_{ij})$ and fix $\bar{\sigma_\alpha} = 10^3$. The parameter σ_α which corresponds to the regularization strength is automatically adjusted thanks to the Bayesian setting.

We chose the gaussian kernel, which formulates as:

$$k(\mathbf{x}, \mathbf{x'}) = \prod_{l=1}^{m} \exp\left(\frac{(x_l - x_l')^2}{\sigma_l^2}\right) \tag{4.2.15}$$

with σ_l being the bandwidth for the l-th covariate. These bandwidths are hyperparameters. All hyperparameters were optimized using nested cross-validation.

However, due to the non-continuous aspect of some of the covariates used in our application, one can argue that the Gaussian kernel is not always adapted. Some ideas to go further would be to adapt kernels to the type of data used in the covariates, for instance a specific kernel for discrete data.

Bayesian additive regression trees

BART was introduced in[24]. The model consists in a sum-of-trees:

[22] J. F. Bobb, L. Valeri, B. Claus Henn, *et al.*, "Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures," eng, *Biostatistics (Oxford, England)*, vol. 16, no. 3, pp. 493–508, Jul. 2015.

[23] M. Kanagawa, P. Hennig, D. Sejdinovic, *et al.*, *Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences*, Jul. 2018.

[24] H. A. Chipman, E. I. George, and R. E. McCulloch, "BART: Bayesian additive regression trees," *The Annals of Applied Statistics*, vol. 4, no. 1, pp. 266–298, Mar. 2010.

$$\phi(\mathbf{x}; \mathcal{T}, \mu) = \sum_{l=1}^{m} g(\mathbf{x}; \mathcal{T}_l, \mu_l) \tag{4.2.16}$$

where m is the number of trees, $\mathcal{T}_l$ is the structure of the l-th tree (with interior nodes decision rules) and μ_l its set of leaf values, and g is the function returning the output of a given tree with specified parameters applied to an input $\mathbf{x}$. Specific priors are put on $\mathcal{T}_l$ and μ_l while the number of trees m and the parameter controlling depth α are hyperparameters. We refer the reader to the source paper for more details.

4.2.3 Estimation method

End-to-end estimation of the two models η and ϕ would be too complicated since the values for the regression are inferred from the latent disease progression model. Therefore we used a two-step estimation.

First, the latent disease progression model needs to be calibrated on data coming from untreated individuals. In our case, this data corresponds to the observations of the patients prior to the beginning of the treatment. In most other applications where this data is unavailable, one can estimate the model parameters on a different dataset. For instance, in a new clinical trial, one could use a model estimated thanks to the control arm of a previous trial or thanks to an observational cohort of the disease progression. For the DCM, calibration is performed with the Monte-Carlo Markov chain stochastic approximation expectation maximization algorithm (MCMC-SAEM)[25]. It results in estimating the population parameters $\mathbf{z}_{pop}$ on a training set. From there we can compute the individual parameters $\mathbf{z}_i$ of a new patient based on its inclusion visit or any amount of points before the beginning of the treatment while keeping $\mathbf{z}_{pop}$ fixed. Individual parameters are considered covariates that can be fed into the regression.

Since DCM is a Bayesian method, we have posterior distributions for parameters and predicted observations $\eta_k(t_{ij}, x; \mathbf{z}_i, \mathbf{z}_{pop})$. However, computing the regression based on distributions instead of point estimates is too expensive computationally as it would require many Monte-Carlo samples of those variables. We opted to use the maximum a posteriori (MAP) estimates for both $\mathbf{z}_i$ and $\eta_k(t_{ij}, x; \mathbf{z}_i, \mathbf{z}_{pop})$.

For the ODE model, the penalty method described previously is applied for the fit.

The Bayesian regression models were all estimated through Monte-Carlo sampling, with a No U-Turn sampler (NUTS)[26] for BLR and BKMR, and an ad-hoc sampler for BART[27].

The models were implemented as described next using the probabilistic library PyMC3 in Python for Bayesian models[28]. We used an open-source implementation of BART in PyMC3[29]. The penalty method was implemented in Python with the TensorFlow library.

[25] E. Kuhn and M. Lavielle, "Coupling a stochastic approximation version of EM with an MCMC procedure," fr, *ESAIM: Probability and Statistics*, vol. 8, pp. 115–131, 2004.

[26] M. D. Homan and A. Gelman, "The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1593–1623, Jan. 2014.

[27] H. A. Chipman, E. I. George, and R. E. McCulloch, "BART: Bayesian additive regression trees," *The Annals of Applied Statistics*, vol. 4, no. 1, pp. 266–298, Mar. 2010.

[28] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck, "Probabilistic programming in Python using PyMC3," en, *PeerJ Computer Science*, vol. 2, e55, Apr. 2016.

[29] M. Quiroga, P. G. Garay, J. M. Alonso, *et al.*, *Bayesian additive regression trees for probabilistic programming*, Oct. 2022.

4.2.4 Experiments

Simulations

We already know that the disease progression models that we use are validated in previous works ([30] for the DCM and[31] for the ODE model). Therefore we just need to validate the second part of the method proposed above. We performed some simulations to better understand how well the treatment effect was estimated under various scenarios.

Scenario 1: True predictors In this first scenario, we generated 10000 random observations X of dimension 2 and assumed that the first dimension was the natural disease measure (OFF state). Then we computed the treatment effect as a deterministic function f of X, mimicking the situation where the treatment effect would be increasing as the disease increases and as another covariate increases (which could be thought of as the treatment dose). We decided that f would be linear so that the Bayesian linear regression would be able to fit the outcome: $f(x_1, x_2) = 10x_1 + x_2$. Finally we add a Gaussian noise to the output $Y = f(X) + \epsilon_Y$. In this very simple case we first recorded the R squared metric for the different regressions on a test set while progressively increasing the noise level ϵ_Y in the training set. Results are shown in Table 4.1 below. This simply shows that regression models perform very well until a certain amount of noise, which is around the standard deviation of the treatment effect Y.

Noise level	0.0	0.1	1.0	10.0	100
BLR	1.0	1.0	1.0	0.98	0.66
BKMR	0.99	0.98	0.98	0.94	0.41
BART	1.0	1.0	0.99	0.97	0.61

Table 4.1: R-squared for the regression models on simulated data depending on the noise level of the outcome. The noise level is multiplicative factor between the standard deviation of the noise and the standard deviation of the input data X.

Scenario 2: Noisy disease progression model In this second scenario, we use the same data as previously, with a noise level on the outcome of 0.1 allowing the regressions to find the true $CATE$ if the predictors are true. Except that the predictors coming from the disease progression model are noisy, that is we add a Gaussian centered noise on x_1: $\tilde{x_1} = x_1 + \epsilon_X$. We can think of this as the uncertainty of the predictions of the disease progression model. Notice that in this case the predictors are still unbiased. We then have the true treatment effect $Y = f(X) + \epsilon_Y = y^{ON} - x_1$ and the inferred treatment effect which is based on the disease model predictions $\hat{Y} = y^{ON} - \tilde{x_1}$ (cf equation (4.2.3)). The regressions are trained to predict $\hat{Y}$ from $(\tilde{x_1}, x_2)$, but we test them on their ability to reconstruct the real treatment effect Y. Table 4.2 records the R-squared as we increase the noise level ϵ_X on the predictors. The results show that under a certain threshold, this noise is not an issue and the regression still perform very well.

[30] I. Koval, A. Bône, M. Louis, *et al.*, "AD Course Map charts Alzheimer's disease progression," eng, *Scientific Reports*, vol. 11, no. 1, p. 8020, Apr. 2021.

[31] K. Lahouel, M. Wells, V. Rielly, *et al.*, *Learning nonparametric ordinary differential equations from noisy data*, Feb. 2023.

Noise level	0.0	0.1	1.0	10.0	100
BLR	1.0	1.0	0.99	0.80	0.03
BKMR	0.98	0.98	0.98	0.80	0.05
BART	1.0	1.0	0.99	0.83	0.10

Table 4.2: R-squared for the regression models on simulated data depending on the noise level of the predictors.

Scenario 3: Biased disease progression model In this last scenario, we add a noise to the disease progression model x_1 so that its predictions are biased. We tried several types of bias. The first case is a constant bias: $\tilde{x_1} = x_1 + C + \epsilon_X$ with a certain constant C. Then we add a bias conditionally to the other covariate x_2 used in the regression: $\tilde{x_1} = x_1 + 0.1x_2 + \epsilon_X$. The average marginal effect of each predictor is then estimated with its 95% credibility interval and we compare it to the ground truth in Figure 4.3.

The first and third rows of Figure 4.3 highlight the bias effect directly on the marginal effect of x_1. In both cases the regressions have learned the bias. This is clearly visible in the second row, where we see that the marginal effect on x_2 has learned a constant bias (which is exactly equal to the constant C). We also observe the high fluctuations of the estimated effect for x_2 on the BKMR and BART models, which is due to the smaller effect of x_2 relative to x_1 in Y, therefore the noise is felt more for this predictor.

What these simulations show is an overall high robustness to the noise, while being susceptible to the bias in the disease progression model predictions. Even in this case, the regressions learn *with* the bias, which means that we can still interpret the results if we have an idea of where the bias come from.

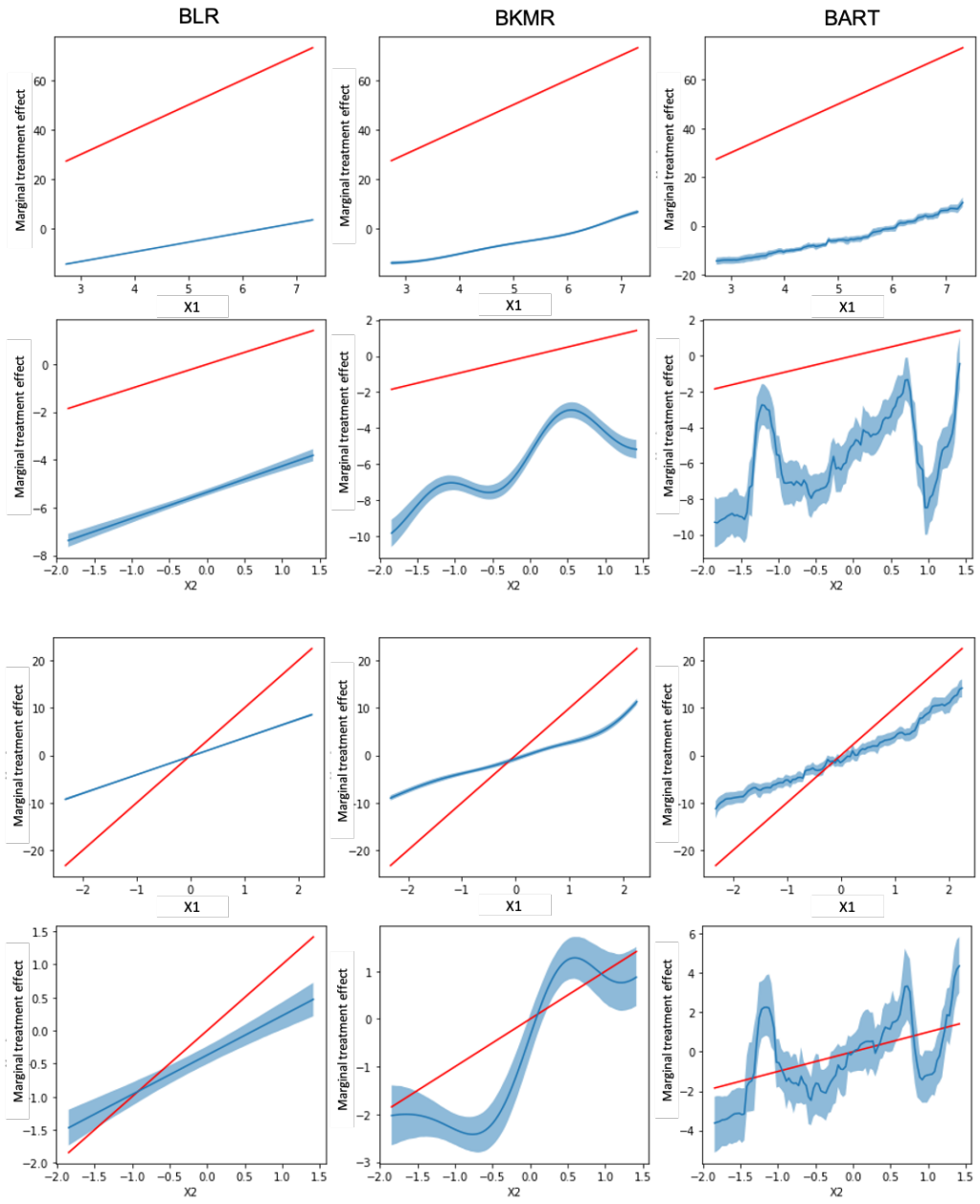

Figure 4.3: Marginal treatment effect compared to ground truth. First two rows correspond to the constant bias $\tilde{x_1} = x_1 + C + \epsilon_X$ and last two rows correspond to the linear bias $\tilde{x_1} = x_1 + 0.1x_2 + \epsilon_X$. Each column correspond to a regression model. The regression mean posterior value is plotted in plain blue line with the 95% credibility interval while the ground truth is plotted in red.

PPMI dataset

Our application concerns Parkinson's disease (PD), the second most prevalent neurodegenerative disease. It is characterized by the loss of neurons responsible for dopamine production in the brain, especially in the substantia nigra. Lack of dopamine leads mainly to motor symptoms, but this can be treated by a dopamine replacement strategy. Dopaminergic therapy has shown to be an efficient symptomatic treatment, even if no disease-modifying effect has been found[32]. All patients in the cohort follow this treatment.

Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database[33] (www.ppmi-info.org/access-data- specimens/download-data). For up-to-date information on the study, visit ppmi-info.org. [34]

The dataset comprises PD patients included within two years of their diagnosis. We selected all patients with at least two visits in the study for the longitudinal history, which resulted in 900 subjects for a total of 7918 visits. We included basic covariates such as sex, age, weight, BMI, treatment category, and dose and a clinical assessment: the MDS-UPDRS score. It is subdivided in 4 sections: non-motor aspects of experiences of daily living (I); motor aspects of experiences of daily living (II); motor examination (III) by the clinician; motor complications (IV) which measure the strength of secondary effects due to the dopaminergic treatment. As mentionned earlier, the third part is measured twice for a patient under treatment: an "OFF" version is assessed when the treatment effect has worn off; an "ON" version is measured shortly after the patient has taken its medication. This particularity of the cohort allows us to have direct access to the values of $Y_{ijk}|T_{ij} = 0$.

Dataset demographics can be found in Table 4.3.

Clinical imaging biomarkers are extracted from DAT-scan with single-photon emission computed tomography (SPECT) which traces the dopamine transporter molecule. The values are recorded for the areas where the signal is the most important in PD, which is the striatum. The caudate and putamen are subparts of the striatum. We refer to them not as left and right but ipsilateral or contralateral depending on the side of onset of the motor symptoms. 1704 visits were recorded with DAT-scan and but only 135 patients had imaging data prior to the start of treatment.

Experiment on motor score

Our first experiment uses the MDS-UPDRS part III (motor score) as the outcome, which ranges from 0 to 132. We calibrated a DCM model on observations pre-treatment as our latent disease progression model. We then predicted the evolution of this motor score under no treatment with the DCM model. Since the cohort recorded the OFF-state values of the MDS-UPDRS part III, we can measure the prediction performance. We obtain a mean absolute error of 6.9, which has to be compared to the standard measurement error of this clinical score (the inherent noise), which according to[35] is at 4.31. We estimated a BLR and a BART model on the set of observed covariates

[32] C. Lungu, J. M. Cedarbaum, T. M. Dawson, *et al.*, "Seeking progress in disease modification in Parkinson disease," en, *Parkinsonism & Related Disorders*, vol. 90, pp. 134–141, Sep. 2021.

[33] K. Marek, D. Jennings, S. Lasch, *et al.*, "The Parkinson Progression Marker Initiative (PPMI)," en, *Progress in Neurobiology*, Biological Markers for Neurodegenerative Diseases, vol. 95, no. 4, pp. 629–635, Dec. 2011.

[34] PPMI – a public-private partnership – is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including [list the full names of all of the PPMI funding partners found at www.ppmi-info.org/about-ppmi/who-we-are/study-sponsors].

[35] P. Martinez-Martin, C. Rodriguez-Blazquez, M. Alvarez-Sanchez, *et al.*, "Expanded and independent validation of the Movement Disorder Society–Unified Parkinson's Disease Rating Scale (MDS-UPDRS)," en, *Journal of*

name	MDS-UPDRS part I	MDS-UPDRS part II	MDS-UPDRS part III	MDS-UPDRS part IV	TD score
count	7897	7897	7519	7922	7922
mean	7.998607	8.729264	25.512701	1.222545	5.447740
std	5.562359	6.387908	12.429672	2.494955	4.544258
min	0	0	0	0	0
median	7	7	24	0	5
max	38	48	100	17	30
name	PIGD score	Treament dose	Cumulative treatment dose	Medication category	Sex
count	7922	5757	7922	6862	7922
mean	2.086342	546.896595	885.235715	-	55%M/45%F
std	2.239677	808.732513	1660.498211	-	-
min	0	0	0	0	-
median	1	421.2	325.0	2	-
max	16	24165	42400	7	-
name	Age since baseline	Weight	Body Mass Index		
count	7922	3732	3728		
mean	2.30794	78.506645	26.698898		
std	2.07245	16.847416	4.677159		
min	0	37.3	13.818027		
median	1.625	77.8	26.178115		
max	9.125	162	61.728395		

Table 4.3: Demographics table for PPMI variables. Count is the number of visits where the variable is available. MDS-UPDRS: Movement Disorder Society - Unified Parkinson's Disease Rating Scale. TD: Tremor Dominant. PIGD: Postural Instability and Gait Disorder. Medication category: 0 = No medication, 1 = Levodopa, 2 = Dopamin agonist, 3 = Other, 4 = Levodopa and other, 5 = Levodopa and dopamin agonist, 6 = Dopamin agonist and other, 7 = Levodopa, dopamin agonist and other.

to perform variable selection since BKMR does not fare well with too many covariates.

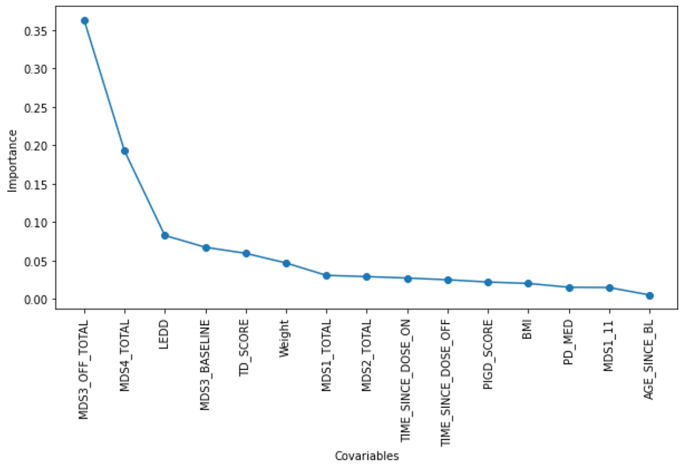

Figure 4.4: Variable importance in BART regression of ground truth treatment effect (ON measure minus OFF measure). MDS[X]: part X of the MDS-UPDRS. The baseline score corresponds to the last value of the score before treatment starts. LEDD: levodopa equivalent daily dose (levodopa is the main dopaminergic medication). TD_SCORE: tremor dominant score. PIGD_SCORE: postural imbalance and gait disorder score. TIME_SINCE_DOSE_[ON/OFF]: time since the last medication at the moment of the examination [ON/OFF]. BMI: body mass index. PD_MED: type of dopaminergic treatment. MDS1_11: gastrointestinal complications. AGE_SINCE_BL: time since baseline.

Our variable selection relied on Figure 4.4. Using the whole range of covariates the model highlighted the MDS-UPDRS parts III and IV as the most important variables. Unfortunately, part IV only records the secondary effects of treatment, and a disease progression model trained on pre-treatment data will only predict zeros for the value of this covariate under the no-treatment hypothesis. However, this score is quite correlated to the treatment dose (LEDD) and the fact that LEDD is the third most important variable suggests that this variable be used instead for the regression. The R^2 with the full set of covariates reached 0.86 while reducing to those two only dropped to 0.82. The dose is interesting all the more so as we can *act* on this covariate, which means that we can *control* treatment effect. We thus reduced the covariates to the treatment dose and the MDS-UPDRS part III OFF-state representing the current disease state. This MDS-UPDRS part III is either observed directly, using the OFF measures or estimated, with the DCM.

We verified that the DCM model predictions were not biased conditionally to the selected covariates. There was no bias for the treatment dose, but a small bias for the predictions conditioned on the predicted value, as the model tends to under estimate at large values. However it does not affect the range of values where most of the patients are.

We estimated the three treatment effect models with two different settings. First, we regressed the observed treatment effect ϕ_{GT}, which is the measured ON value minus the measured OFF value. We then reconstructed the ON score $\hat{y}_{ON} = y_{OFF} + \phi_{GT}(\mathbf{x}_{GT})$. Using the ground truth treatment effect and predictor provides an upper bound on the best metric scores we can achieve using inferred

Neurology, vol. 260, no. 1, pp. 228–236, Jan. 2013.

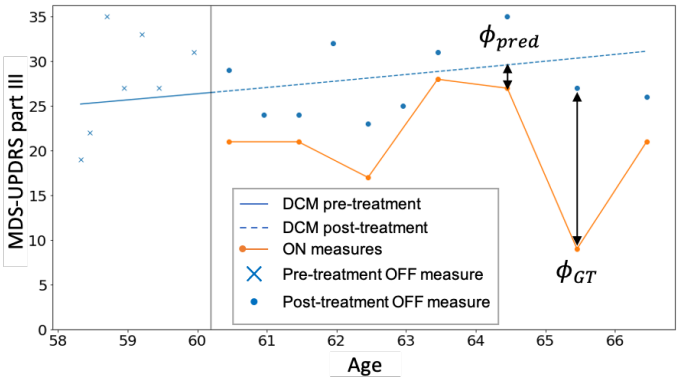

Figure 4.5: Model illustration on data from a PPMI subject. DCM model is personalized with pre-treatment data (crosses) to predict post-treatment latent disease progression (dashline). Setting 1 (ϕ_{GT}) models the difference between OFF and ON measures. Setting 2 (ϕ_{pred}) models the difference between predicted OFF and ON measures.

values for treatment effect and predictors, which was our second setting. In this second part, the regressions ϕ_{pred} used the predicted OFF value and the treatment dose as input and the inferred treatment effect (measured ON value minus predicted OFF value) as output. The reconstructed outcome becomes $\hat{y}_{ON} = \eta(t) + \phi_{pred}(\mathbf{x}_{pred})$. Figure 4.5 illustrates these two settings on actual data from PPMI.

We also provided a third setting for comparison without using the latent disease progression model predictions. We use the common approach used in treatment models, using baseline variables. This provides a lower bound error target on the reconstruction metrics. Here we estimated the regressions to predict the ON score directly. As predictors, we used the treatment dose, the baseline OFF score, i.e., the last measure without treatment, and the time since treatment started. We reported the prediction metrics on the reconstructed ON score in Table 4.4.

Compared to setting 1 using ground truth treatment effect and predictors, setting 2 is worse as it is expected since it uses inferred predictors and treatment effect. But setting 2 only lost around 2.3 points in mean absolute error while using the score inferred with DCM which had about 7 points of absolute error. It shows that the treatment effect models compensated for the DCM prediction errors. Comparing settings 2 and 3 we see that leveraging the information from the longitudinal predictive latent disease progression model improved the metrics by a significant margin, yet still far from the ground truth.

Overall BART performed best out of the three treatment effect models. The difference in maximum a posteriori predictions was low, but the log-likelihood was much better because BART has a much tighter credible interval for its predictions. However, the cost was paid in runtime.

Figure 4.6 shows the posterior distributions of the treatment effect models. The top and middle rows correspond to the treatment effect ϕ learned in settings 1 and 2 respectively. The treatment effect in inference setting 2 was underestimated by a shift of approximately 2 points, but it captured the same trends. The 2D interaction plot shows how the dose affects the treatment effect depending on the disease state. Specifically, the best treatment effect is obtained with an increasing dose as the disease progresses toward later stages.

Table 4.4: Reconstruction metrics for the ON-state MDS-UPDRS part III computed with four folds. Setting 1: ground truth (GT) treatment effect and ground truth covariates are used. This is the best possible performance. Setting 2: inferred treatment effect and predicted (pred) covariates are used. This is our method. Setting 3: baseline (BL) predictors are used with the time since baseline. This is the standard method. BLR: Bayesian Linear Regression. BKMR: Bayesian Kernel Machine Regression. BART: Bayesian Additive Regression Trees. R^2: R-squared coefficient of maximum a posteriori (MAP) prediction. MAE: Mean Absolute Error of MAP prediction. LL: posterior log-likelihood.

Setting	Regression	R^2	MAE	LL	Runtime (s)
1 (GT)	BLR	0.81±0.02	5.19±0.10	-6150±527	37.9±2.1
	BKMR	0.81±0.03	5.11±0.09	-501±38	94.9±3.4
	BART	0.82±0.03	5.11±0.10	-114±13	5240±294
2 (pred)	BLR	0.55±0.02	7.47±0.14	-7170±954	38.1±2.6
	BKMR	0.56±0.02	7.42±0.13	-535±21	108±13
	BART	0.57±0.02	7.37±0.19	-117±18	5190±110
3 (BL)	BLR	0.50±0.04	8.1±0.1	-5770±749	41.7±9.4
	BKMR	0.48±0.03	8.2±0.2	-188±23	166±28
	BART	0.52±0.03	7.98±0.16	-98±11	4550±143

Figure 4.6 only showed the mean effect of the dose on the treatment effect. Figure 4.7 includes the MDS-UPDRS part III (used as a proxy for the disease state). Notice how the effect seems to be proportional. This suggests an improvement of the model in this specific case of the MDS-UPDRS part III: measuring the treatment effect not as a difference between ON and OFF states, but as a difference relative to OFF score. The slope indicates that we observe a mean 35% improvement of the motor score under treatment.

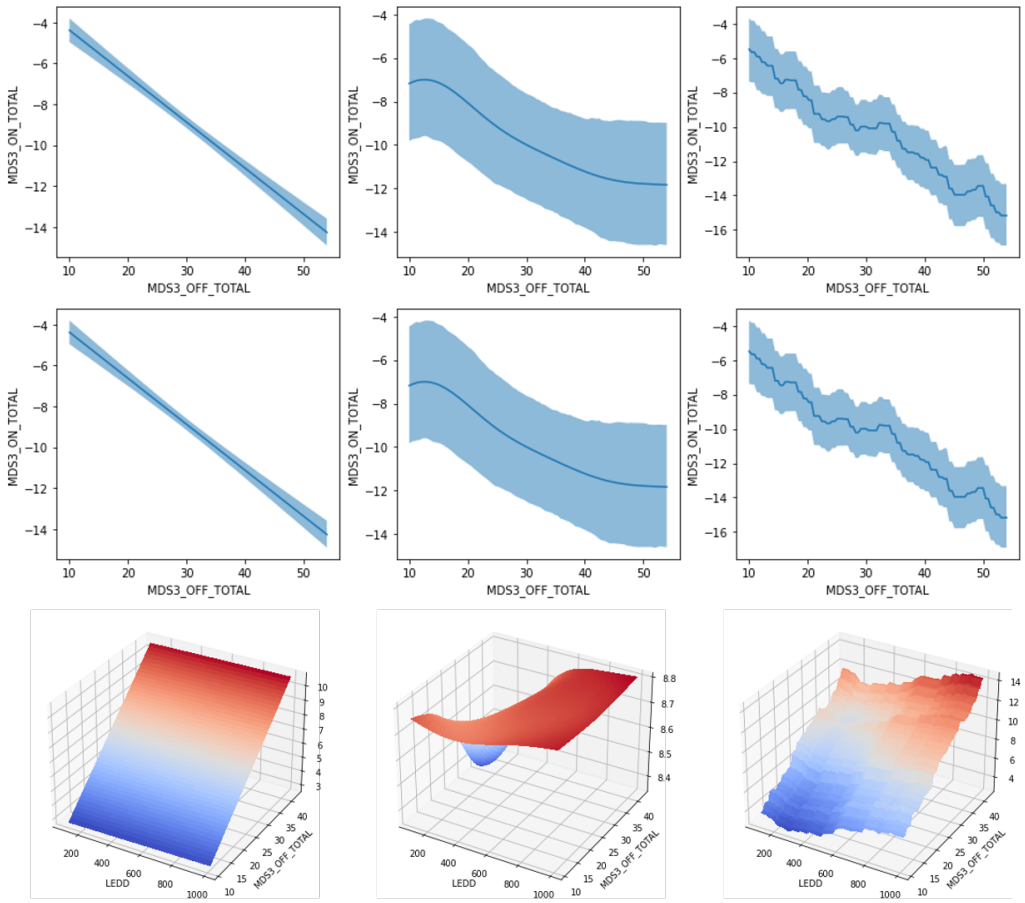

Figure 4.6: Top row: posterior distribution of dose on real treatment effect (ON minus OFF) according to the three treatment effect models. The plain line is the mean, the colored area is the 95% credibility interval. Middle row: same plot but with inferred treatment effect (ON minus predicted OFF). Bottom row: interaction plot, displays the mean treatment effect learned by each regression as a function of the predicted disease state (predicted MDS-UPDRS part III OFF) and the dose.

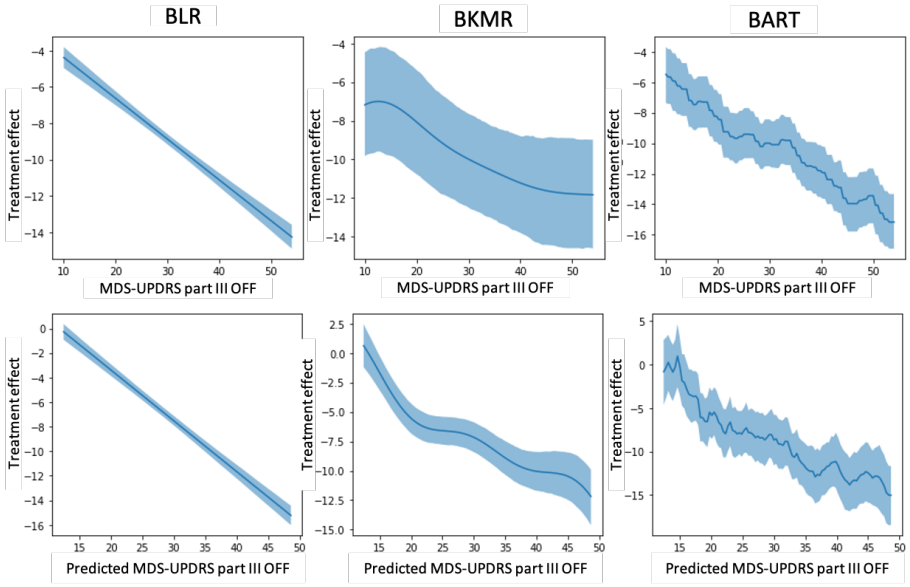

Figure 4.7: Top row: posterior distribution of disease state on real treatment effect (ON minus OFF) according to the three treatment effect models. The plain line is the mean, the colored area is the 95% credibility interval. Bottom row: same plot but with inferred treatment effect (predicted ON minus OFF).

Experiment on quality of life score

Our second experiment focused on the MDS-UPDRS part II, which ranges from 0 to 42. This score evaluates patients' quality of life (QoL), and QoL improvement is becoming one of the main criteria for treatment acceptance. However, there was no OFF measure for this score so the treatment effect could only be inferred. We thus replicated settings 2 and 3 from the last experiment. The best treatment effect model was BART in setting 2, with R^2:0.83 and MAE:2.56 on the reconstructed MDS-UPDRS part II in setting 2. This is much better than the best regression in setting 3 with R^2:0.71 and MAE:3.31. Table 4.5 reports full metrics on the reconstruction of the MDS-UPDRS part II.

The interest of the model, besides being able to predict the score of patients, is to understand how covariates influence the treatment effect on QoL. Our analysis of the full set of covariates revealed that the most important variables were the QoL without treatment as predicted by the DCM and the PIGD score (postural imbalance and gait disorder) as PIGD patients have a worsening of their QoL under treatment as opposed to TD (tremor dominant) patients. PIGD and TD are the two main sub-groups in Parkinson's disease, and TD patients tend to respond better to dopaminergic treatment than PIGD patients[36].

Table 4.5: Reconstruction metrics for the MDS-UPDRS part II computed on a separate test set.

Setting	Regression	R^2	MAE	LL	Runtime (s)
2 (pred)	BLR	0.82	2.72	-1.64e+04	59.4
	BKMR	0.80	2.65	-7.55e+02	989
	BART	**0.83**	**2.56**	**-2.04e+02**	5e+03
3 (BL)	BLR	0.68	3.45	-1.69e+04	54.9
	BKMR	0.63	3.56	-4.78e+02	913
	BART	0.71	3.31	-2e+02	5e+03

Experiment on imaging

We use the penalty method to fit an ODE to the imaging biomarkers. We use pre-treatment data only so that this model extrapolates the latent disease progression without the influence of treatment. The data and the fitted ODE curves are represented for the putamen in Figure 4.8. This figure shows that imaging data is quite noisy, so the model provides a smoothing of the trajectory.

We now have the predicted latent disease progression curve for each patient, which we will consider as our synthetic placebo arm. We can then compare it to the values observed under treatment. However, since imaging measures are noisy and sparse, we use a second ODE model to impute the trajectories of our treatment arm. This model uses the full set of points of patients, thus smoothing out the biomarkers and minimizing the noise. The difference between the treatment arm model and the placebo arm model is shown in Figure 4.9. The difference is consistent across individuals, which hints towards an effect of the dopaminergic treatment over the dopaminergic transporters in the putamen.

[36] C. Marras, K. R. Chaudhuri, N. Titova, *et al.*, "Therapy of Parkinson's Disease Subtypes," en, *Neurotherapeutics*, vol. 17, no. 4, pp. 1366–1377, Oct. 2020.

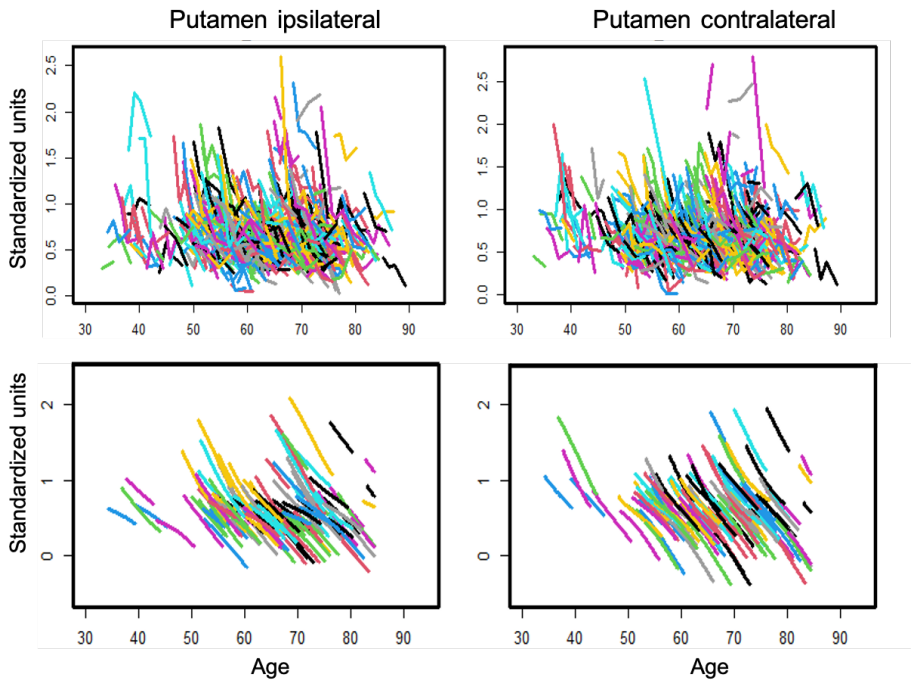

Figure 4.8: Spaghetti plot for raw data (top) and model fit (bottom) on Putamen. Left column is ipsilateral (putamen on the side of motor signs onset), and right column is contralateral.

In order to understand how the treatment effect is impacted by potential covariates we estimate the regression models. The $CATE$ outcome is approximated as the difference between the treatment arm model and the placebo arm model as shown in figure 3. We started with all the regressors mentioned in the previous experiments plus the values of the imaging markers predicted by the placebo arm model.

Variable selection was first performed by computing variable importance with BART as shown in Figure 4.10. The selected variables which explained the most the treatment effect are the predicted values of the imaging biomarkers. We also included the treatment dose and the cumulative treatment dose as covariates of interest due to the possibility to control for those values.

The metrics for the regressions' fit are shown in Table 4.6. As we can see BART outperforms the other methods at the cost of a much higher runtime while the linear regression is not able to reconstruct the outcome as well, indicating that the interactions are mostly non-linear. To validate our approach, we also computed the reconstruction error between real values y_{ijk} and model predictions $\eta_k(t_{ij}, x; \mathbf{z}_i, \theta) + \phi(\mathbf{x}_{ij}; \omega)$. For the putamen contralateral (resp. ipsilateral) the best method (BART) obtained a mean absolute error of 0.10 (resp. 0.17), which is good when compared to the noise of the real imaging markers.

To understand the interactions of these covariates on treatment effect we plot their average effect evaluated as the mean of the posterior of the models. Figure 4.11 shows the mean effect for some variables of interest. The first row tells us that the treatment effect is higher for low values of the putamen, which corresponds to later stages of the disease as the value for this marker decreases

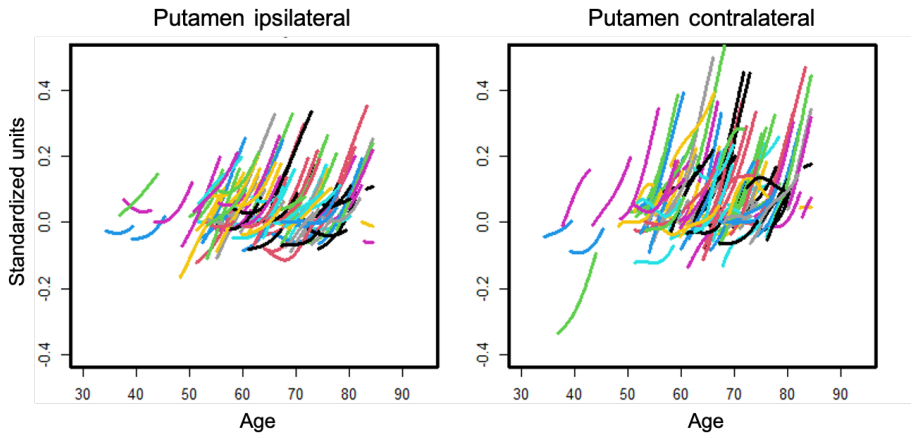

Figure 4.9: Difference between the treatment arm model and the simulated placebo arm model. Each line is a separate individual. The plots show that the treatment seems to improve the dopaminergic transporter signal over time.

Table 4.6: Prediction metrics for the regression models computed with four folds. R^2: R-squared coefficient of maximum a posteriori (MAP) prediction. MAE: Mean Absolute Error of MAP prediction. LL: posterior log-likelihood.

Outcome	Regression	R^2	MAE	LL	Runtime (s)
Putamen Ipsilateral	BLR	0.61±0.02	0.053±0.003	-1.48±0.30×10^3	37±3
	BKMR	0.72±0.04	0.047±0.003	-39±13	118±23
	BART	0.77±0.02	0.044±0.003	-14.5±4.4	3.2±0.2×10^3
Putamen Contralateral	BLR	0.69±0.03	0.061±0.006	-1.5±0.3×10^3	53±20
	BKMR	0.77±0.04	0.050±0.005	-61±22	192±72
	BART	0.82±0.02	0.046±0.005	-16±3	1.7±0.2×10^3

with time. On the second row, we see that the treatment effect seems to be correlated with the cumulative treatment dose, although the credibility interval is large due to the averaging over all the over covariates effect. However, it is much more interesting to look at the mean treatment effect as a function of both these two covariates as their effect is dependent on each other. The BART and BKRM in the last row show that at early disease stages (Putamen values around 1) the cumulative treatment dose has almost no influence as the treatment effect is close to 0. However, at later stages (when the putamen values are closer to 0.3) the treatment effect is correlated with the cumulative treatment dose. Overall the regression models suggest that the dopaminergic treatment is reducing the loss of dopaminergic transporters in the putamen compared to the natural disease progression. In practice, it might imply that the treatment, although it is thought to be purely symptomatic, might have a slight disease-modifying effect.

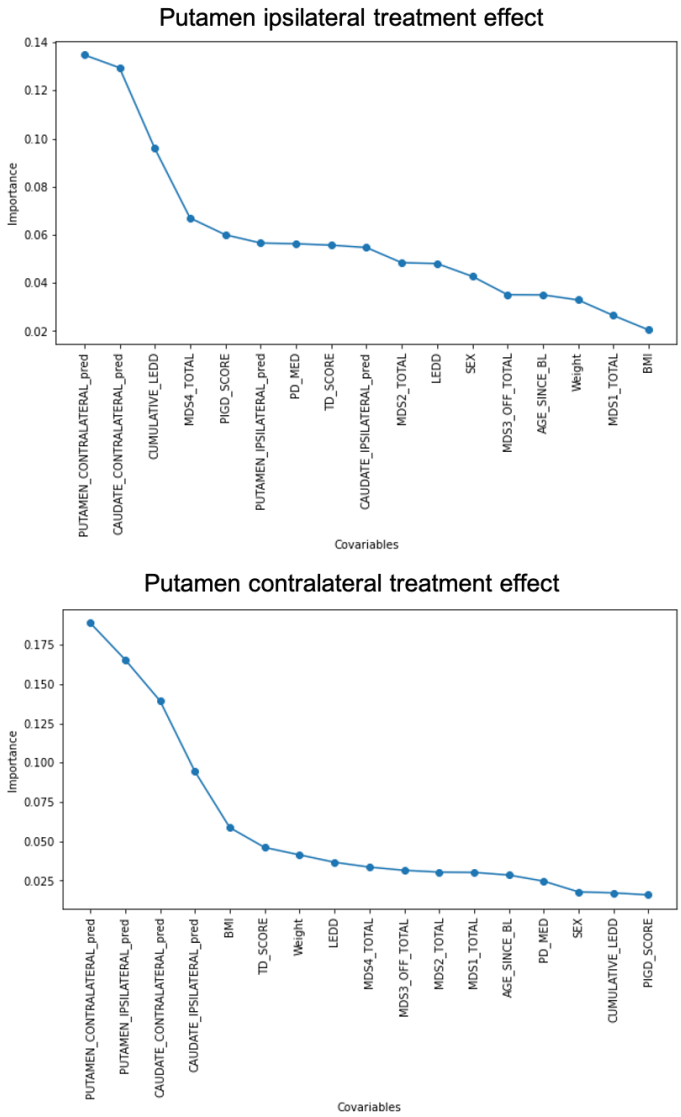

Figure 4.10: Variable importance computed with BART on full set of covariates. Variables predicted by the disease progression model are noted with "_pred". MDSX_TOTAL: Movement Disorder Society - Unified Parkinson's Disease Rating Scale, part X. TD: Tremor Dominant. PIGD: Postural Instability and Gait Disorder. PD_MED: medication category. LEDD: Levodopa equivalent daily dose (=treatment dose). AGE_SINCE_BL: age since baseline. BMI: Body Mass Index.

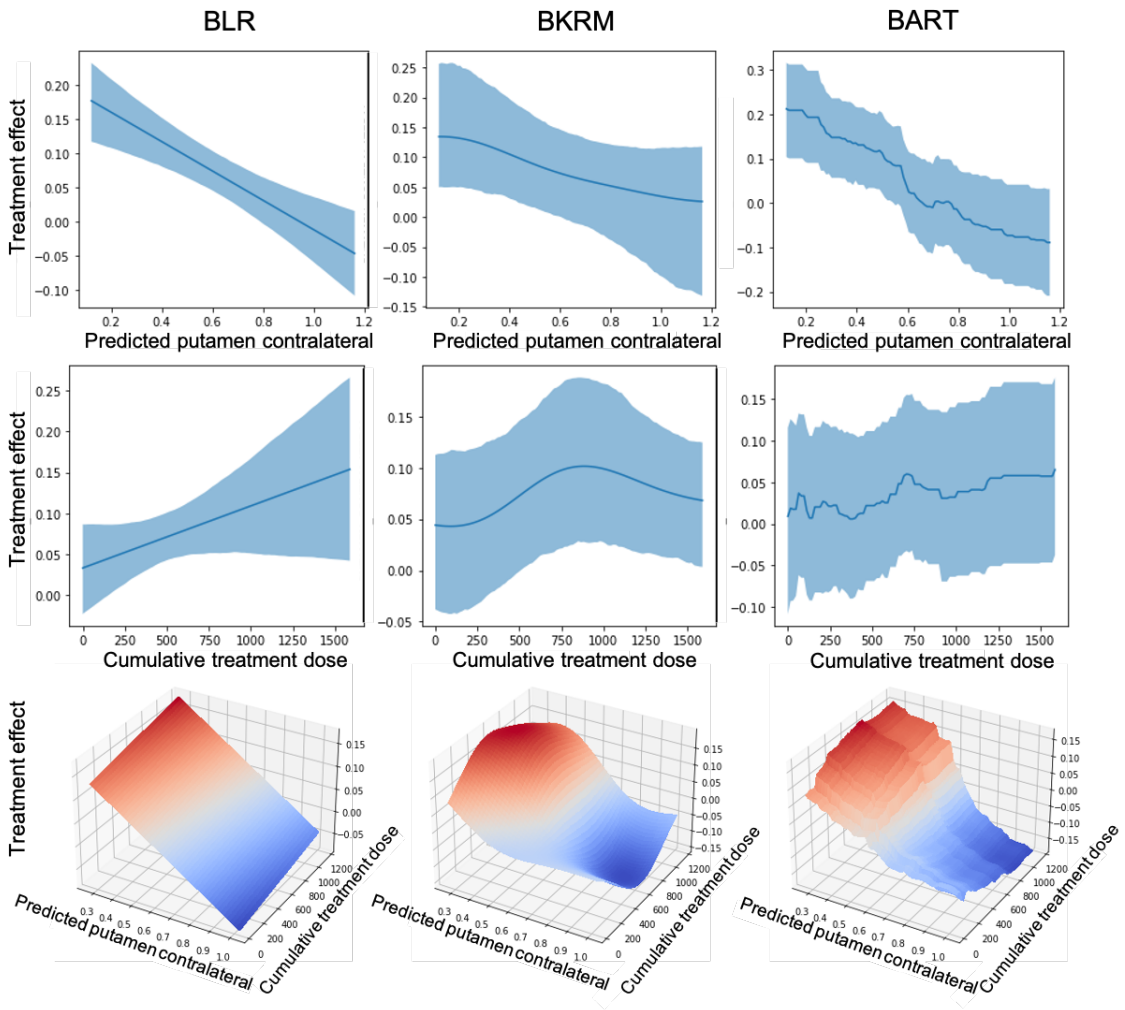

Figure 4.11: Average treatment effect with 95% credibility interval. Top row is the average effect conditioned on predicted putamen value. Treatment effect is stronger as the putamen loses dopamine transporters. Middle row is the average effect conditioned on the cumulative treatment dose. Bottom row shows the interaction with both. At low putamen values the treatment effect correlates with the cumulative treatment dose.

4.2.5 Discussion

We proposed a fully Bayesian method to infer treatment effects in observational cohorts without a control arm. Our approach relies on a disease progression model to infer the unobserved disease state of a patient under no treatment. The treatment effect is then modeled with a regression using various predictors, including disease markers predicted by the disease progression model. A subject disease state can then be predicted by adding the treatment effect to the latent disease progression.

In the simulation experiment we showcased the strength of the method while highlighting the main defects, namely the learning of the latent disease progression model's bias by the treatment effect model. This does not impact predictions of the patients under treatment, but this negates the main interest of this method which is to disentangle the conditional treatment effect $CATE$ from the natural disease history.

We then showed that the model proved to be reliable in a real-life setting, with close to ground-truth data for ON and OFF states. Our approach also outperformed the baseline approach in predicting a patient's disease state in the second experiment.

In the last experiments, we extended the method to scores that were not observed without treatment effect, limiting our validation possibilities. We still obtained high predicting performances on the MDS-UPDRS part II (quality of life score) and on imaging data. More specifically, we applied our method to SPECT data for dopamine transporters in the striatum in Parkinson's disease. In this case, our results suggest that the dopaminergic treatment affects the disease progression. There is an indication that the cumulative treatment dose slows down the loss of dopamine transporters in the putamen.

The treatment effect model can be useful for understanding and optimizing treatment strategies for new patients. Another possible application concerns new treatments, as many are being tested for neurodegenerative diseases, with high expectations. The standard way to validate a new treatment is a clinical trial. However, a trial able to demonstrate a meaningful effect in a very slowly progressive disease is costly, and the statistical power is often very weak. To reduce the size of the trial one could leverage a latent disease progression model to create a synthetic control arm and estimate the treatment effect as we did in the observational study.

Our approach has inherent limits based on the latent disease progression model quality. The better the predictions for the latent disease progression, the better the treatment effect estimation. A proper theoretical study about the properties of this estimator remains to be done, especially if we wish this method to be applied to one-arm clinical trials with the disease progression model providing a synthetic control arm.

Finally we have seen that this approach is suited for simple treatment effects as symptomatic ones, but including disease-modifying effects requires to add time-dependent covariates with history information (cumulative treatment dose for instance). This is not the best way to model such an effect, so we will focus on a new model, more adequate for disease-modifying treatment effects.

4.3 Piecewise-geodesic model

We have seen a "simple" additive model, based on the hypothesis that a treatment is purely symptomatic. This allowed us to avoid any complex longitudinal modelling. Let's now use the ODE formulation in order to understand what is at stake with disease-modifying treatments :

$$\frac{\partial \mathbf{y}}{\partial t}(t) = f(\mathbf{y}(t)) \tag{4.3.1}$$

f is the function we want to learn. It is possible to parametrize it as was done with the penalty method in the second experiment previously. It is notable that the introduction of covariates (such as the treatment T) in the ODE equation is not as easy as with mixed-effect models. In these mechanical models it is quite complex to introduce time-dependent covariates since the equations' integration won't be tractable. However we can choose a certain form for these time-dependent covariates in order to keep the properties of the ODE. Here we postulate that the treatment effect is a function of the position $\mathbf{y}$ and the treatment dosage $T(t)$ (which is supposed to be known at every timepoint). This echoes the work in the previous model where we saw that the most important variables for treatment effect models were the treatment dose and the latent disease state. The ODE equation becomes :

$$\frac{\partial \mathbf{y}}{\partial t}(t) = f(\mathbf{y}(t)) + g(\mathbf{y}(t), T(t)) \tag{4.3.2}$$

where g is the treatment effect such that $g(\cdot, 0) = 0$. Note that if $T(t)$ is not fully known, for instance if we don't have prior knowledge about how treatment will be prescribed in future visits, or if T is a control process which is latent but not directly observed, then the above equation cannot be solved. We would require further hypothesis about T, e.g a template curve or an interpolation method in between visits.

Another important remark: for symptomatic effects we can assume that $||g|| >> ||f||$, and we have convergence to a stable point for the ODE $\frac{\partial \mathbf{y}}{\partial t}(t) = c + g(\mathbf{y}(t), T(t))$. Let me illustrate in the case of Parkinson's disease: L-dopa does not act instantly on the organism, it operates within one hour approximately, during which the dynamic of the symptoms changes until it stabilizes to a certain point. However one hour can be considered very small compared to the disease progression scale, which can span up to decades. Therefore the effect of the L-dopa as an "additive" symptomatic treatment can be seen as time-limit approximation of a very fast dynamic effect. However it is often not easy to learn on different time scales, therefore if the treatment has a symptomatic *and* a disease-modifying effect then it is easier to learn the two parts separately rather than the total effect on the dynamics directly.

4.3.1 Method

For this section it is preferable to understand the basics of Riemannian geometry. If needed, the reader can refer to the introduction to Riemannian geometry in the appendix 15.

The classical way to handle treatment in statistical models is to treat it as a time-dependent covariate. If we go back to a general formulation of a mixed-effect model, we write :

$$\mathbf{y}_{ij} = \eta(t_{ij}, \mathbf{X}_{ij}; \mathbf{z}_{pop}, \mathbf{z}_i) + \epsilon_{ij} \quad \text{[Mixed-effect model with time-dependent covariates]} \tag{4.3.3}$$

$$\mathbf{z}_{ij} = f(T_{ij}, \mathbf{z}_{ij'}, j' < j) \quad \text{[Covariate structure]} \tag{4.3.4}$$

The main idea is that the time-dependent covariate $\mathbf{z}_{ij}$ linked to the treatment is a function of the current dose but also of the history of the treatment. If we only used a function of the treatment dose, we would end up with a symptomatic model since when the treatment has a null dose the function f with no history needs to be null too (no effect when no treatment).

All of the modelling task resides in choosing how to integrate previous treatment history into those covariates. Here we will specify our proposed approach with the DCM model in its generic framework. In this context we will need to specify T over time as well. Indeed we want to avoid any integration as in a dynamic approach, so we will assume that T is piece-wise constant. This is a fair hypothesis since at each visit the patient will be given a treatment prescription with a certain dose which should be constant until the next visit. The only drawback of this hypothesis is that it stops us from making future predictions without a new treatment policy function. Writing this hypothesis for one individual i yields:

$$T_i(t) = \sum_{j=1}^{n_i} a_{ij} \mathbb{I}_{t \in [b_{i,j}, b_{i,j+1}]} \tag{4.3.5}$$

with $0 < b_{i,1} < ... < b_{i,j} < b_{i,j+1} < ... < b_{i,n_i+1} = \infty$ being the dates of the changes in treatment, called breakpoints. Note that $T = 0$ before $b_{i,1}$. Let's write the set of individual parameters in the absence of treatment (or before $b_{i,1}$) as $(\tau_i^0, \xi_i^0, \mathbf{w}_i^0)$. The main idea of our model is that on each interval $[b_{i,j}, b_{i,j+1}]$ the trajectory will follow a geodesic in the manifold (or more precisely be an Exp-parallelization of a geodesic). Hence the name: Piece-wise Geodesic model.

We will build a reference curve that is piece-wise geodesic, such that the trajectory of individual i is obtained by Exp-parallelization 10 from it, as in the DCM. However, since each individual has now its own treatment function T with different breakpoints, the reference curve will be different for each individual. We will denote γ_i the reference curve for individual i. We will use the ODE characterization of the geodesics. Therefore we have $t_0, \mathbf{p}_0, \mathbf{v}_0$ as the parameters defining the geodesic γ_0. Before $b_{i,1}$, the Piece-wise Geodesic γ_i is equal to γ_0. At $b_{i,1}$, the triplet $(b_{i,1}, \gamma_0(b_{i,1}), \dot{(\gamma_0)}(b_{i,1}))$ defines the same geodesic γ_0. The treatment applies a perturbation on the trajectory by changing the value of the derivative only, yielding a new triplet $(b_{i,1}, \gamma_0(b_{i,1}), \mathbf{v}_1)$ which defines a new geodesic passing by $\gamma_0(b_{i,1})$. The reference curve γ_i follows this new geodesic until the next breakpoint $b_{i,2}$, where the same process is reiterated.

More formally: we fix $j \in [|1, n_i|]$. The piece of geodesic followed during the interval $(b_{i,j}, b_{i,j+1})$ is parametrized by $(b_{i,j}, \mathbf{p}_{i,j}, \mathbf{v}_{i,j})$ where $\gamma_i(b_{i,j}^+) = \mathbf{p}_{i,j}$ and $\dot{\gamma}_i(b_{i,j}^+) = \boldsymbol{v}_{i,j}$. At time $b_{i,j+1}$, $\gamma_i(b_{i,j+1}^-) = \mathbf{p}_{i,j+1}$ (we assume continuity of the trajectory) and $\dot{\gamma}_i(b_{i,j+1}^-) = \tilde{\boldsymbol{v}}_{i,j}$. The perturbation is encoded as a function f such that:

$$\dot{\gamma}_i(b_{i,j+1}^+) = \boldsymbol{v}_{i,j+1} = f(\mathbf{p}_{i,j+1}, \tilde{\boldsymbol{v}}_{i,j}, T_i(b_{i,j}^-), T_i(b_{i,j}^+)) \tag{4.3.6}$$

We call f the *treatment function*. We could also include covariates in this function, such as the weight (patients with a high weight often need a higher dose for a similar treatment effect) or treatment responsiveness, but we will keep it simple for now.

The individual trajectory then results from an Exp-parallelization 10 of the reference curve γ_i with the space-shift $\mathbf{w}_i$. Exp-parallelisation means that the model will be "locally" following the same modification as γ_i upon the curve $\eta^{\mathbf{w}_i}$. It is well defined thanks to the continuity of γ_i and the fact that the Exp-parallelization is well defined on each interval. Figure 4.12 illustrates this procedure.

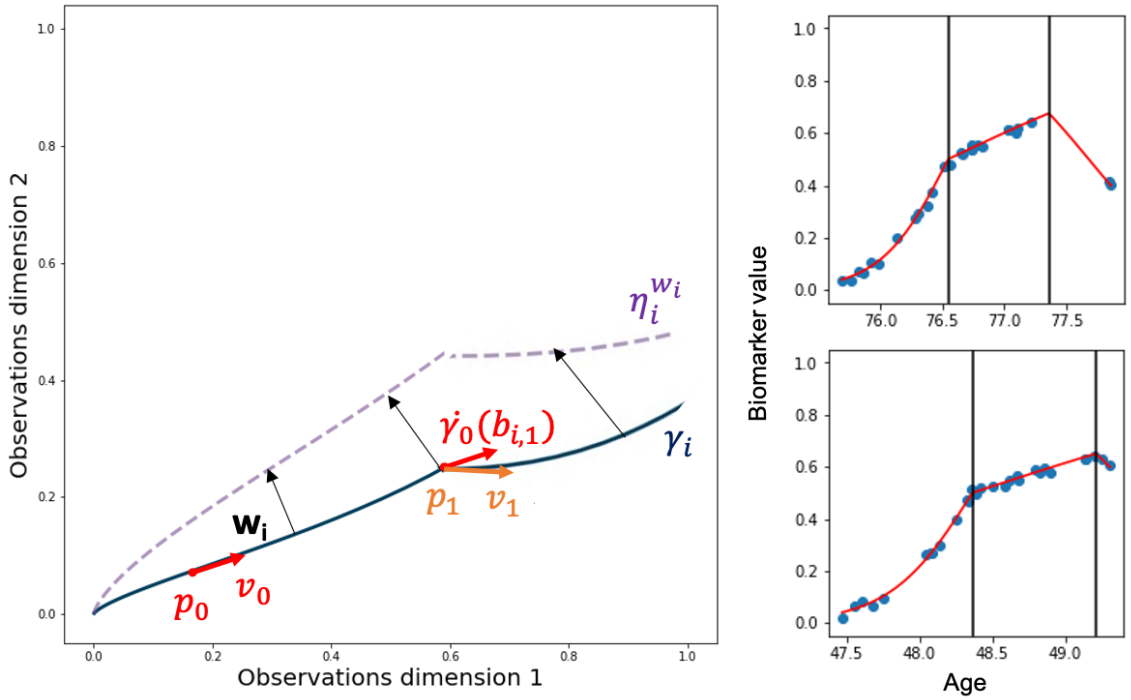

Figure 4.12: Left: Piewe-wise geodesic model scheme. γ_i is the reference curve for individual i and $\eta_i^{\mathbf{w}_i}$ is the Exp-parallelization of γ_i with $\mathbf{w}_i$. Right: samples of data generated following a Piece-wise Geodesic model on a logistic curves manifold. The black vertical lines correspond to the breakpoints, the individual trajectory after Exp-parallelization is shown in red and data points are in blue.

One could argue that the model could be discontinuous at the breakpoints, and that we could learn the perturbation on the position as well. This is very sensible, as some treatments may have radical changes on the disease trajectory, for instance a surgery operation. However it becomes tricky to define the Exp-parallelization in a discontinuous setting: how do we transport $\mathbf{w}_i$ from $\gamma_i(b_{i,j}^-)$ to $\gamma_i(b_{i,j}^+)$? We could do the transport along the geodesic between these two points, which would require to compute the Riemannian logarithm. This would lead to heavy computations and a lot of new parameters, which would be too much considering that the current choice of model already has issues with the estimation.

Estimation With this choice of model, treatment effect understood as the parametrization of f is a population parameter in the mixed-effect model. Estimation of the parameters of f is performed during the MCMC-SAEM algorithm, as described in the next chapter. The method has been implemented in Python, under the Leaspy library.

4.3.2 Experiments

We only performed simulation experiments with this model for two reasons: the model is very data-hungry which is not a problem in simulations, and we have not yet a large neurodegenerative dataset with a disease-modifying treatment. We show that the model is able to recover the ground truth in a simplified setting.

We want to parametrize the treatment effect function f as a function of only the position p for simplification. In order to still keep a dependence on the velocity, we will write the perturbation as a function g with a small norm on top of the identity: $f(\mathbf{p}, \mathbf{v}) = \mathbf{v} + g(\mathbf{p})$.

We then need to parametrize g. As often in this thesis, we resorted to a Reproducing Kernel Hilbert Space (RKHS) with kernel k, but one could use a simple linear function if needed. Using the representer theorem, the function g writes:

$$g(\mathbf{p}) = \sum_{ij} \omega_{i,j} k(\mathbf{p}, \mathbf{p}_{i,j}) \tag{4.3.7}$$

where the parameters are the weights $\omega_{i,j}$. They are as numerous as there are breakpoints in the dataset, which is very variable depending on the cohort. In a clinical trial for instance, we would only have one breakpoint per patient corresponding to the beginning of the treatment.

Data is generated by randomly sampling individual parameters, then for each individual we generate breakpoints and compute the trajectory. Finally we generate random timepoints for the visits, take the trajectory's value at the visits and add a small noise. We will adjust the settings with the number of breakpoints and the number of visits per individual to evaluate the limits of the learning capabilities of our model.

Univariate setting In a first setting, we generated each visit to be a breakpoint. Intuitively, the changes in $\mathbf{v}$ depending on $\mathbf{p}$ can be understood as the practitioner adjusting the treatment of the patient at each visit $b_{i,j}$ based on the current observed state $\mathbf{p}_{i,j}$. We now generate data using the Euclidean setting, where the geodesics are straight lines. This simplified setting amounts to the model being able to capture the change of slope at each breakpoint. The generating function for the change of slope is a heavyside. We chose the Gaussian kernel for the learning of g, which means that g will be a linear combination of radial basis functions. We optimized the kernel bandwidth by random search.

Figure 4.13 shows the results of the simulation of 1024 individuals, 10 visits per individual, and the model obtained after convergence. We observe that the ground truth function is quite accurately modelled, the wavelets owning to the Gaussian kernel.

We repeated the experiment with other generating functions for g, such as linear or sinusoidal functions. In a setting with 1024 individuals and 10 visits per individual, the model is able to learn the target treatment function quite quickly, as shown in Figure 4.14. We also performed experiments with only one breakpoint per patient. In these experiments the ability of the model to learn the proper treatment function scaled with the number of individuals, i.e. the number of breakpoints, rather than the number of visits.

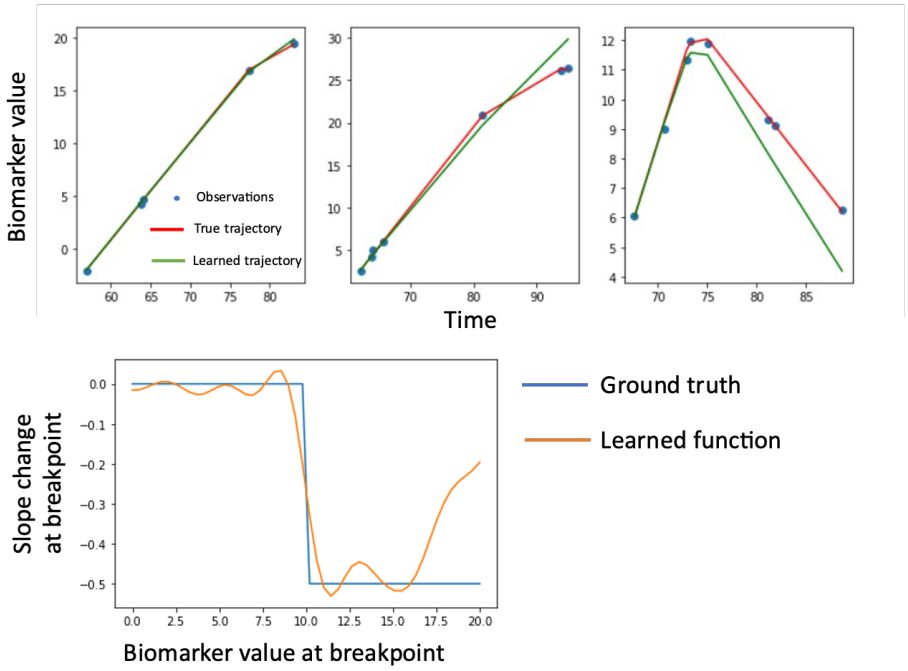

Figure 4.13: Piece-wise Geodesic model results on simulated data. The results are shown after 300 iterations of MCMC-SAEM, with an optimization of the weights $\omega_{i,j}$ every 10 iterations. The first row corresponds to three individuals, where the true trajectory is shown in red with the observations in blue, while the learned model is shown in green. On the bottom plot we represent the function g learned by the Piece-wise Geodesic model against the ground truth.

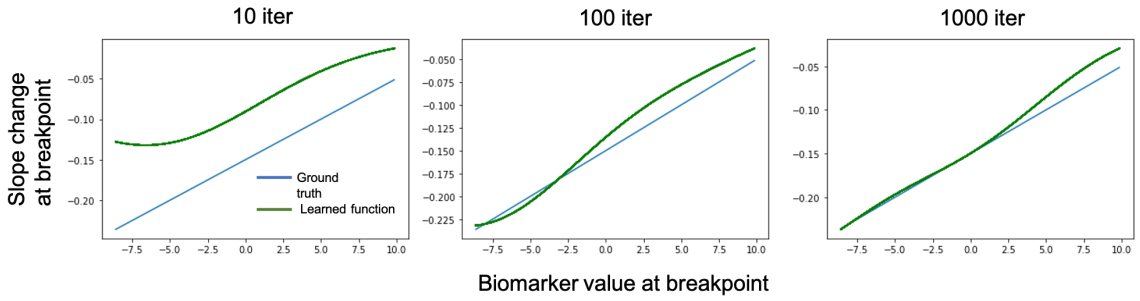

Figure 4.14: Convergence of the learned function g during the iterations of the MCMC-SAEM. The results are shown after 10, 100 and 1000 iterations of MCMC-SAEM, with an optimization of the weights $\omega_{i,j}$ every 10 iterations.

Multivariate setting We extended the previous experiment to a two-dimensional setting. This means that the function g is more complex as both its input and output are two-dimensional now. The results are shown in Figure 4.15. On the individual level the trajectories learned are quite good. However the learned function is not matching the ground truth. This is due to the breakpoints being concentrated in the space, which does not allow radial basis functions to cover enough space.

To remedy this problem, we changed the centers of the radial basis functions from the data points themselves to regularly spaced points on a grid. This is a relaxation of the representer theorem, but allows for more flexibility in the learning of the function. This also avoids ill-conditioning that sometimes happen when centers of the radial-basis functions are too close to each other. With a regular grid, the learning is much better, and we can even approximate complicated functions, as shown in Figure 4.16.

4.3.3 Discussion

These experiments show that in a simple linear case we can perform a reliable estimation of complex treatment function for the change of slope. The extension to the case where the function also takes the treatment dose and over covariates as input is straightforward. However, if we increase the number of input variables in the learned treatment function, or if the observations dimension increases, the RKHS model requires an exponential amount of points to achieve proper learning. This is very limiting as in practice the amount of data used to effectively learn the treatment function is the total number of breakpoints over the population. As our targeted application concerns clinical trials, this number will be limited to the number of patients in the treated arm, which is far less than the ten thousand points used in our simulated experiments.

This burden can be alleviated by choosing other forms for the treatment function, which is the next step towards a practical application of the Piece-wise Geodesic model. The most simple choice would be a linear function. Any parametric family of functions with a low number of parameters can be used as long as we can compute the gradient of the function with respect to its parameters. This is necessary to perform the gradient descent, as described in the estimation algorithm for the Piece-wise Geodesic model in the next chapter. On this occasion we will also discuss the other limitation of the model: the difficulty to learn the treatment function in a non-Euclidean manifold, namely the logistic curves manifold.

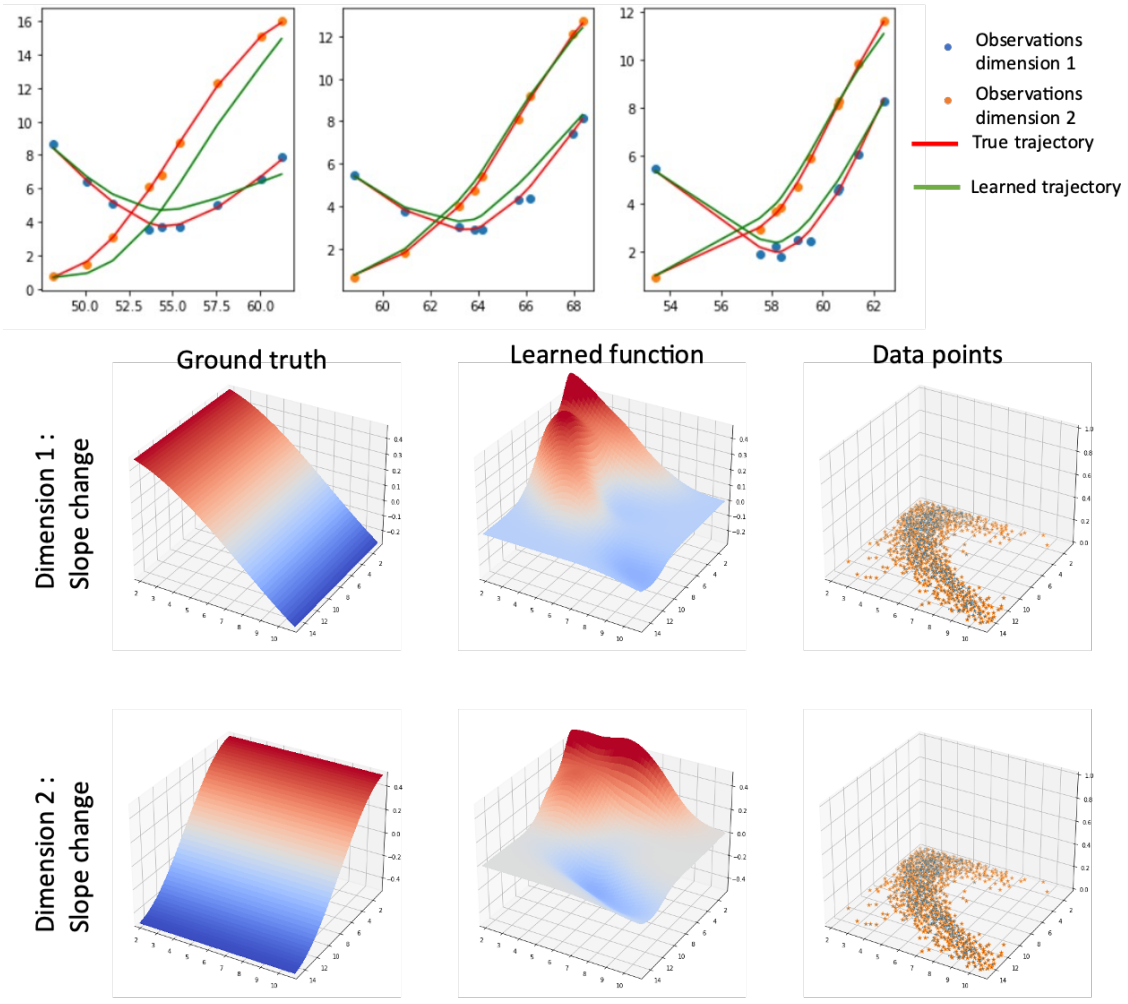

Figure 4.15: Piece-wise Geodesic model results on simulated data in dimension 2. The results are shown after 1000 iterations of MCMC-SAEM, with an optimization of the weights $\omega_{i,j}$ every 10 iterations. The first row corresponds to three individuals. Both biomarkers are shown on the same plot, the true trajectory is shown in red with the observations in blue and orange, while the learned model is shown in green. On the bottom plot we represent the function g learned by the Piece-wise Geodesic model against the ground truth with the repartition of data points, which are the centers of the radial basis functions composing the learned g function.

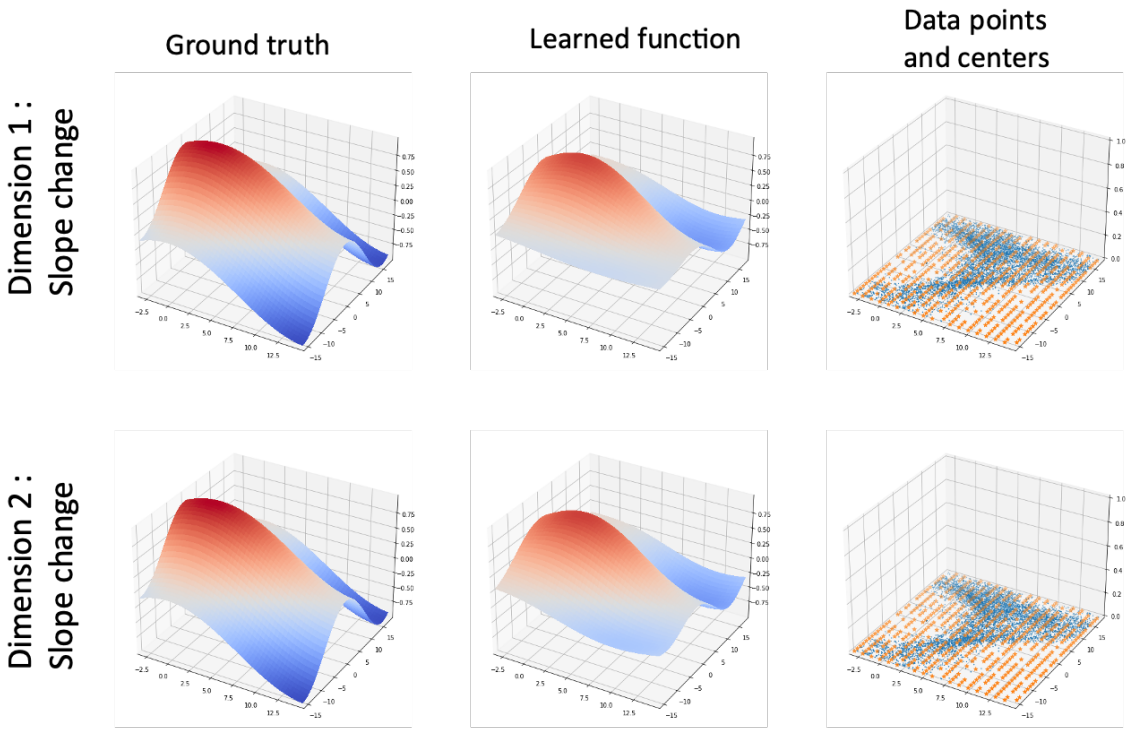

Figure 4.16: Piece-wise Geodesic model results on simulated data in dimension 2. We plotted the function g learned by the Piece-wise Geodesic model against the ground truth with the repartition of data points (in blue) and the centers of the RBF (in orange).

4.4 Conclusion

Treatment effect modelling introduces a new layer of complexity atop the natural disease history modelling task. Being able to approximate the state of the patient, had he not followed a treatment, is permitted by the disease progression models. Based on this latent progression, which can be seen as a digital twin of the patient, we showed that we were able to reconstruct the treatment effect under some conditions on the disease progression model. We believe that this technique has the potential to one day replace placebo arms in some clinical trials, thereby resolving the ethical dilemma of not giving a potentially life-saving drug to half the participants in the trial. We do not claim to have proven the reliability of the method yet, there is still plenty of work to be done. But we hope that we have laid the foundations for future research in this direction.

Current research in neurodegenerative diseases is hoping to find cures or disease-modifying treatments to slow the progression. Treatments with long-term modification of a patient's trajectory call for other types of model than the regressions used before. Estimating if a treatment reduced the rate of progression is easily assessed in a linear setting, less so in the non-linear case. We proposed a new model based on the DCM, called Piece-wise Geodesic, where we introduced a treatment function accounting for the changes of trajectory after every new prescription. We showed that this function could be learned with synthetic data. We have not yet been able to apply this model to real-life data because there is no such treatment for the most common neurodegenerative diseases, not mentioning the fact that the study needs to provide sufficient data for the model to be estimated. However β-amyloïd targeting therapies in Alzheimer's disease have made significant progress in the last two years, with the first drug having recently been approved by the US Food and Drug Administration (FDA)[37]. We can thus expect data from cohorts following patients over the next few years. The Piece-wise Geodesic model is thus an anticipation of the needs of the field, and we hope that it will be applicable in a near future. In the meantime, we provide some ideas for additional developments: simplifying the parametrization of the treatment function, allowing the model to have discontinuities at the breakpoints and introduce an individual latent variable measuring how well a patient responds to treatment.

[37] O. o. t. Commissioner, *FDA Converts Novel Alzheimer's Disease Treatment to Traditional Approval*, en, Jul. 2023.

Chapter 5

Estimation

This chapter is structured differently from the other chapters of this thesis. Indeed this is not the presentation of a complete work on a particular subject with some applications to clinical data. Instead it is a patchwork of small parts from all the models presented in the previous chapters. It is however unified by the theme: the estimation algorithm for the DCM and all its proposed variants. The goal was to regroup all the questions about optimization and parameters estimation in the same chapter, so that everything is simpler to understand.

The main algorithm is a variant of the Expectation-Maximization algorithm. We will first present this algorithm, which was chosen for the estimation of the DCM model in the original work[1]. Then we will present the modifications that we brought to it to adapt to each particular model. In order we will describe variants for the Geodesic Bending model, for the mixture models, and finally for the piece-wise geodesic model.

Contents of the chapter

[1] J.-B. Schiratti, S. Allassonnière, A. Routier, *et al.*, "A Mixed-Effects Model with Time Reparametrization for Longitudinal Univariate Manifold-Valued Data," en, in *Information Processing in Medical Imaging*, S. Ourselin, D. C. Alexander, C.-F. Westin, *et al.*, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2015, pp. 564–575.

5.1 The MCMC-SAEM algorithm

The main goal of the estimation algorithm is to find the model parameters θ that maximize the likelihood q:

$$q(\mathbf{y}, \theta \mid \theta_{hyper}) = \int q(y, \mathbf{z}, \theta \mid \theta_{hyper}) d\mathbf{z} \tag{5.1.1}$$

$$= \int q(y \mid \mathbf{z}, \theta, \theta_{hyper}) q(\mathbf{z}, \theta \mid \theta_{hyper}) d\mathbf{z} \tag{5.1.2}$$

where $\mathbf{z} = (\mathbf{z}_{pop}, (\mathbf{z}_i)_i)$ are the latent parameters, composed of the population and individual parameters. However this integral is not always tractable since the latent variables are unknown. However the Expectation-Maximization (EM) algorithm, as introduced by[2], was designed for such cases. This iterative algorithm alternates between an estimation of the latent variables, thus allowing to compute the expected likelihood, and a maximization of the model parameters based on this expectation. The EM algorithm thereby produces a maximum likelihood estimate or, in the case of a Bayesian model, a maximum a posteriori. The algorithm is described in Algorithm 1.

Algorithm 1: Expectation-Maximization

$\theta^{(0)} \leftarrow \theta_0$ $k \leftarrow 0$ **while** $k \leq n$ **do**
  $k \leftarrow k + 1$
  Expectation step : Compute $Q(\theta, \theta^{(k)}) = \mathbb{E}_{p(z|y;\theta^{(k)})} (\log p(y, z; \theta))$
  Maximization step : Update $\theta^{(k+1)} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta, \theta^{(k)})$
end
Result: return $\theta^{(k)}$

At iteration k, the E-step computes the expectation of the likelihood over the latent variables with the model parameters $\theta^{(k)}$ being fixed. Then the expectation being a function of the model parameters θ, the M-step finds the best parameters.

In linear cases the EM is sufficient, as in[3]. However in most non-linear cases, the expectation is intractable due to the integration over the latent variables. This is where the Stochastic Approximation EM shines[4]. The expectation in the E-step is not computed explicitly, but approximated by drawing samples of latent variables $z^{(k)} \sim p(z|y; \theta^{(k)})$. The algorithm only requires one sample for $\mathbf{z}$ per step, which avoids computing a full Monte-Carlo approximation of the integral in the expectation formula. The convergence towards the true integral is ensured by a Robbins-Monro scheme[5], where after drawing the sample $z^{(k)}$ and computing $\log p(y, z^{(k)}; \theta)$, one computes the approximated expectation $Q_k(\theta) = Q_{k-1}(\theta) + \epsilon_k(\log p(y, z^{(k)}; \theta) - Q_{k-1})$, with $\sum_k \epsilon_k = \infty$ and $\sum_k \epsilon_k^2 < \infty$.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data Via the EM Algorithm," en, *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[3] G. Verbeke and E. Lesaffre, "A Linear Mixed-Effects Model With Heterogeneity in the Random-Effects Population," *Journal of the American Statistical Association*, vol. 91, no. 433, pp. 217–221, 1996.

[4] B. Delyon, M. Lavielle, and E. Moulines, "Convergence of a Stochastic Approximation Version of the EM Algorithm," *The Annals of Statistics*, vol. 27, no. 1, pp. 94–128, 1999.

[5] H. Robbins and S. Monro, "A Stochastic Approximation Method," *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, Sep. 1951.

However an additional difficulty lies in the fact that $p(z \mid y; \theta^{(k)})$ might be unknown. The Bayes rule gives us that:

$$p(z \mid y; \theta^{(k)}) = \frac{p(y \mid z; \theta^{(k)}) p(z; \theta^{(k)})}{p(y; \theta^{(k)})} = \frac{p(y \mid z; \theta^{(k)}) p(z; \theta^{(k)})}{\int p(y \mid z; \theta^{(k)}) p(z; \theta^{(k)}) dz} \tag{5.1.3}$$

where $p(y \mid z; \theta^{(k)})$ is given by the model itself, $p(z; \theta^{(k)})$ is a known prior on the latent variables and the denominator is a constant. We thus know the distribution of the latent variables conditioned on the observations up to the normalization constant. The Metropolis-Hastings algorithm[6] can then be used to sample from this distribution. The Metropolis-Hastings algorithm is based on a Markov chain, hence the name Monte-Carlo Markov Chain SAEM. This algorithm is proven to converge to a local maximum of the observed likelihood $p(y|\theta)$ (see[7],[8],[9]). However the convergence has only been proven for the curved exponential family, which is the family of distributions for which the log-likelihood can be written as: $\forall \theta \in \Theta, log\ q(\mathbf{y}, \mathbf{z}; \theta) = -\Phi(\theta) + \langle S(\mathbf{y}, \mathbf{z}), g(\theta) \rangle$ where Φ and g are smooth functions, S are a function of $\mathbf{z}$, $\mathbf{y}$ called the sufficient statistics. To be noted, all the models presented in this work fall into the curved exponential family, except for the ordinal model.

The sufficient statistics are to be understood as a summary of the required information from the latent variables $\mathbf{z}$ and the observations $\mathbf{y}$. Once the summary statistics are computed, one can derive the log-likelihood linearly from it by definition. Therefore the Robbins-Monro convergence scheme for the computation of the expectation is equivalent to applying the Robbins-Monro scheme directly on the sufficient statistics, as described in the algorithm below. Then the maximization step consists in finding the argmax of the expected log-likelihood based on the sufficient statistics. The maximization in the base DCM model formulation is direct, so the new parameters $\theta^{(k)}$ are computed explicitly in closed form.

The pseudo-code is given in the complete Algorithm 2.

π is a proposition law that is known and from which we can easily sample. Typically we use a Gaussian centered on $z^{(k-1)}$ with an adaptive variance[10], i.e. we automatically scale the variance so that the acceptance rate stays around 30%[11]. In practice we use a Metropolis-Hastings within Gibbs sampler, which corresponds to a sequential sampling of the coordinates of $\mathbf{z}$.

The algorithm requires a burn-in phase, which corresponds to a certain number of iterations $n_{burn-in}$ during which $\forall k \leq n_{burn-in}, \epsilon_k = 1$. This corresponds to a memory-less state, allowing first the Markov chain to converge. Once the Markov chain is stable, the sampling step corresponds to drawing the latent variables from their posterior distribution. The Robbins-Monro scheme in the burn-out phase only allows for the convergence of the model parameters to a maximum a posteriori.

We refer the reader to the appendix 15 for complete formulas of the log-likelihood. We did not

[6] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, Apr. 1970.

[7] E. Kuhn and M. Lavielle, "Coupling a stochastic approximation version of EM with an MCMC procedure," fr, *ESAIM: Probability and Statistics*, vol. 8, pp. 115–131, 2004.

[8] S. Allassonnière and E. Kuhn, "Stochastic algorithm for Bayesian mixture effect template estimation," en, *ESAIM: Probability and Statistics*, vol. 14, pp. 382–408, Jan. 2010.

[9] S. Allassonnière and E. Kuhn, "Convergent stochastic Expectation Maximization algorithm with efficient sampling in high dimension. Application to deformable template model estimation," en, *Computational Statistics & Data Analysis*, vol. 91, pp. 4–19, Nov. 2015.

[10] H. Haario, E. Saksman, and J. Tamminen, "An Adaptive Metropolis Algorithm," *Bernoulli*, vol. 7, no. 2, pp. 223–242, 2001.

[11] S. Chib and E. Greenberg, "Understanding the Metropolis-Hastings Algorithm," *The American Statistician*, vol. 49, no. 4, pp. 327–335, Nov. 1995.

Algorithm 2: Monte Carlo Markov Chain Stochastic Approximation Expectation-Maximization

$\theta^{(0)} \leftarrow \theta_0$ $S_0 \leftarrow 0$
$z^{(0)} \leftarrow z_0$
$(\epsilon_k)_{k \geq 0}$ (for the Robbins-Monro scheme)
$k \leftarrow 0$
while $k \leq n$ **do**
  $k \leftarrow k + 1$
  Simulation step: Given $(y, \theta^{(k)})$*, sample* $z^{(k)}$:
    Draw $z^* \sim \pi(z^{(k-1)})$
    Compute acceptance ratio: $\alpha = \max(\frac{q(y,z^*|\theta)}{q(y,z^{(k-1)}|\theta)}, 1)$
    Accept with probability α*:* $z^{(k)} \leftarrow z^*$
    or reject: $z^{(k)} \leftarrow z^{(k-1)}$
  Stochastic Approximation step : $S_k \leftarrow S_{k-1} + \epsilon_k \left(S(y, z^{(k)}) - S_{k-1}\right)$
  Maximization step : $\theta_k = \underset{\theta}{\operatorname{argmax}} \left(-\log \Phi(\theta) + \langle \tilde{S}_k, g(\theta) \rangle\right)$
end
Result: *return* $\theta^{(k)}$

derive the sufficient statistics here as it has already been done previously in[12] and[13]. We will give one example to understand how it works:

The log-likelihood term for the prior of the time-shifts is $\log q_{prior}((\tau_i)_i \mid \theta) = -N \log(\sigma_\tau \sqrt{2\pi}) - \frac{1}{2\sigma_\tau^2} \sum_{i=1}^{N} (\bar{\tau} - \tau_i)^2$. The summary statistics associated to the model parameters $\bar{\tau}$ and σ_τ are $S_1 = \sum_{i=1}^{N} \tau_i$ and $S_2 = \sum_{i=1}^{N} \tau_i^2$. From these sufficient statistics we can derive the explicit maximization rules: $\bar{\tau} \leftarrow \frac{1}{N} S_1$ and $\sigma_\tau \leftarrow \frac{1}{N}(S_2 - 2S_1 + \bar{\tau})$.

In its implementation, there is no convergence criterion for the algorithm. We currently fix an arbitrary number of iterations which is large enough for the sampling Markov chain to stabilize. However some improvements could be made by using a stopping criterion such as the Gelman and Rubin convergence criterion for MCMC[14].

We will now delve into the difference brought by the several models.

[12] I. Koval, "Learning Multimodal Digital Models of Disease Progression from Longitudinal Data : Methods & Algorithms for the Description, Prediction and Simulation of Alzheimer's Disease Progression," en, Ph.D. dissertation, Institut Polytechnique de Paris, Jan. 2020.

[13] J.-B. Schiratti, "Methods and algorithms to learn spatio-temporal changes from longitudinal manifold-valued observations," en, Ph.D. dissertation, Université Paris Saclay (COmUE), Jan. 2017.

[14] A. Gelman and D. B. Rubin, "Inference from Iterative Simulation Using Multiple Sequences," *Statistical Science*, vol. 7, no. 4, pp. 457–472, Nov. 1992.

5.2 Variants of the MCMC SAEM

5.2.1 MMCMC SAEM for mixture models

We saw that the MCMC SAEM was the base estimation algorithm for the DCM model. Moreover the EM algorithm is also a staple for mixture model estimation. Combining the estimation of a mixture model and the non-linear mixed-effect models has already been studied[15]. However the theoretical properties are often not enough as the MCMC SAEM tends, in practice, to suffer from local maxima attraction. In particular, changing cluster assignments is known to be challenging in such methods; it is a problem referred to as trapping states. Adding a tempering scheme has been shown to alleviate this issue by flattening the target distribution and thus easing the exploration of the parameter space[16]. A mixture model on top of a disease progression model was also proposed in[17], the clustering during estimation was handled with hard labels and a probability for individuals to switch from one cluster to another.

We remind the reader of our formulation of the mixture of DCM models:

$$Q(\boldsymbol{y} \mid \boldsymbol{z}; \theta^1, ..., \theta^L) = \sum_{c=1}^{L} \pi^c q(\boldsymbol{y} \mid \boldsymbol{z}; \theta^c)$$

and the mixture of individual parameters within the DCM model:

$$q(\begin{bmatrix}\tau_i \\ \xi_i\end{bmatrix} \mid \theta) \quad = \quad \sum_{c=1}^{L} \mathbb{1}_{c_i=c} \phi(\begin{bmatrix}\tau_i \\ \xi_i\end{bmatrix}; \begin{bmatrix}\bar{\tau}^c \\ \bar{\xi}^c\end{bmatrix}, \Sigma^c_{\tau,\xi})$$

which have both the same log-likelihood for the observations once we introduce the latent classes of the individuals c_i:

$$q(\boldsymbol{y}_i \mid \boldsymbol{z}_i; \theta^1, ..., \theta^L) = \sum_{c=1}^{L} \mathbb{1}(c_i = c) q(\boldsymbol{y}_i \mid \boldsymbol{z}_i; \theta^c) \tag{5.2.1}$$

In the estimation, we compute the latent variables $\pi_i^c = \mathbb{P}(c_i = c \mid \mathbf{z}_i, \mathbf{z}_{pop}, \theta)$. Contrarily to the other latent variables, they can be directly computed in the expectation step, as:

$$\pi^c_{K+1,i} = \frac{\pi^c_K q(y_i|\theta^c_K, z^c_{K,i}, z^c_{K,pop}))p(z^c_{K,i}|\theta^c_K, z^c_{K,pop}))}{\sum_j \pi^j_K q(y_i|\theta^j_K, z^j_{K,i}, z^j_{K,pop}))p(z^j_{K,i}|\theta^j_K, z^j_{K,pop}))} \tag{5.2.2}$$

Contrarily to[18] where the latent variable is directly c_i, forcing individuals to attach to a cluster, our use of soft cluster labelling avoids cluster freeze and trapping states. We provide below the algorithm for the mixture of DCM models.

[15] M. Lavielle and C. Mbogning, "An improved SAEM algorithm for maximum likelihood estimation in mixtures of non linear mixed effects models," *Statistics and Computing*, vol. 24, no. 5, pp. 693–707, Sep. 2014.

[16] S. Allassonnière and J. Chevallier, "A new class of stochastic EM algorithms. Escaping local maxima and handling intractable sampling," en, *Computational Statistics & Data Analysis*, vol. 159, p. 107 159, Jul. 2021.

[17] V. Debavelaere, S. Durrleman, S. Allassonnière, *et al.*, "Learning the Clustering of Longitudinal Shape Data Sets into a Mixture of Independent or Branching Trajectories," en, *International Journal of Computer Vision*, vol. 128, no. 12, pp. 2794–2809, Dec. 2020.

[18] V. Debavelaere, S. Durrleman, S. Allassonnière, *et al.*, "Learning the Clustering of Longitudinal Shape Data Sets into a Mixture of Independent or Branching Trajectories," en, *International Journal of Computer Vision*, vol. 128, no. 12, pp. 2794–2809, Dec. 2020.

Algorithm 3: M-MCMC-SAEM estimation

Initialization: $\pi_0, (\theta_0^c), (\mathbf{z}_{0,i}^c), (\pi_{0,i}^c)$
for $K = 0...N$ **do**
  for $c = 1...L$ **do**
    • **E step**
    Compute probability of individual i to belong to cluster c: $\pi_{K+1,i}^c$
    Population parameters estimation
    Sample $z_{*,pop}^c$
    Compute acceptance ratio $\alpha = 1 \wedge \frac{q(\mathbf{y}|\theta_K^c,(z_{K,i}^c)_i,z_{*,pop}^c)q((z_{K,i}^c)_i|\theta_K^c,z_{*,pop}^c)}{q(\mathbf{y}|\theta_K^c,(z_{K,i}^c)_i,z_{K,pop}^c)q((z_{K,i}^c)_i|\theta_K^c,z_{K,pop}^c}$
    Set $z_{K+1,pop}^c = z_{*,pop}^c$ with probability α else $z_{K+1,pop}^c = z_{K,pop}^c$
    Individual parameters estimation
    Sample $(z_{*,i}^c)_i$
    Compute acceptance ratios $\alpha_i = 1 \wedge \frac{q(y_i|\theta_K^c,z_{*,i}^c,z_{K+1,pop}^c))q(z_{*,i}^c|\theta_K^c,z_{K+1,pop}^c)}{q(y_i|\theta_K^c,z_{K,i}^c,z_{K+1,pop}^c))q(z_{K,i}^c|\theta_K^c,z_{K+1,pop}^c)}$
    $\forall i$, set $z_{K+1,i}^c = z_{*,i}^c$ with probability α_i else $z_{K+1,i}^c = z_{K,i}^c$
    Compute sufficient statistics $(S^c(y_i, \mathbf{z}_{K+1}))_i$
    • **M step**
    Update $\theta_{K+1}^c = \underset{\theta\in\Theta}{argmax}\ log\ q(\mathbf{y}, z_{K+1,pop}^c, (z_{K+1,i}^c)_i, \theta)$
    Compute $\pi_K^c = \frac{1}{n}\sum_i \pi_{K,i}^c$
  end
end

The challenging part is the update of model parameters within each cluster. In the case of a single model, the updates of the parameters θ of the model are computed in a closed form and are only a function of the total sufficient statistic, which writes $S(\mathbf{y}, \mathbf{z}_{K+1}) = \sum_i S(y_i, \mathbf{z}_{K+1})$ since the observations $(y_i)_i$ are supposed independent. When mixtures are involved, the log-likelihood changes to $log\ q(\mathbf{y}, \mathbf{z}^c, \theta^c) = \sum_i log\ (\sum_c \pi_i^c q(y_i, z_i^c, z_{pop}^c, \theta^c))$. However since the EM computes an expectation of the log-likelihood, we can condition on the true latent class π_i of each individual i such that $\pi_i^c = \mathbf{P}(\pi_i = c)$:

$$\begin{aligned}
\mathbf{E}_\mathbf{z}\Big(\log\big(\sum_c \pi_i^c q(y_i, z_i^c, z_{pop}^c, \theta^c)\big)\Big) &= \mathbf{E}_\pi\Big(\mathbf{E}_\mathbf{z}\big(\log\big(\sum_c \mathbf{1}_{\pi_i=c} q(y_i, z_i^c, z_{pop}^c, \theta^c)\big)|\pi_i\big)\Big) \\
&= \mathbf{E}_\pi\Big(\mathbf{E}_\mathbf{z}\big(\sum_c \mathbf{1}_{\pi_i=c} \log\big(q(y_i, z_i^c, z_{pop}^c, \theta^c)\big)|\pi_i\big)\Big) \\
&= \mathbf{E}_\mathbf{z}\big(\sum_c \pi_i^c \log\ (q(y_i, z_i^c, z_{pop}^c, \theta^c))\big)
\end{aligned}$$

Thus we obtain the total sufficient statistics for each cluster $S^c(\mathbf{y}, \mathbf{z}_{K+1}) = \sum_i \pi_i^c S(y_i, \mathbf{z}_{K+1})$ which can be used to update θ^c. This formula is intuitive: each sufficient statistic for each cluster weights the contribution of the individual data by the probability of this individual being in the cluster. This result also proves that the mixture model still belongs to the curved exponential family. Therefore, the convergence of this algorithm is guaranteed by the MCMC-SAEM proof[19].

[19] E. Kuhn and M. Lavielle, "Coupling a stochastic approximation version of EM with an MCMC procedure," fr, *ESAIM: Probability and Statistics*, vol. 8, pp. 115–131, 2004.

In the case of the individual parameters mixture model, the main difference is that the sampling of the population parameters is shared between clusters. The maximization of $\bar{\tau}^c, \bar{\xi}^c, \Sigma^c_{\tau,\xi}$ is done separately for each clusters using the weighted sufficient statistics $S^c(\mathbf{y}, \mathbf{z}_{K+1})$, but is shared for the rest of the model parameters. The constraint on the ξ to be centered 3.3.3 (for identifiability with $\mathbf{v}$) is solved in the M-part by shifting $\tilde{\boldsymbol{v}}$ by the mean of the ξ_i and substracting it from the ξ_i. Intuitively we just scale the average speed of the trajectory so that the acceleration factor is still centered around 1.

Tempered scheme

Even with theoretical guarantees, the convergence might be very slow in practice. The estimation of individual parameters in a non-linear single model may be challenging. Building a mixture on top of it further adds to the difficulty. Practical experiences, described in the results section, show a high reliability on initialization. One of the bottleneck is the amount of cluster regularization contained in the term $p(z_i^c|\theta^c, z^c_{pop})$ in the likelihood. Until the cluster population parameters stabilize, we do not want to restrict the exploration of the individual parameters space.

We thus propose to use a tempered scheme for the Gibbs sampler, mimicking simulated annealing. Tempered MCMC-SAEM has been shown to converge[20]. In a tempered scheme, inverse temperature comes as a multiplier of the log-likelihood of the model. We realized our model had overly constraining regularization, outweighing the data term in the likelihood. Thus we applied temperature only to the regularization term of the log-likelihood in the sample. In practice, we propose to simply replace q by $\tilde{q} = q^{1/T}$.

For the temperature scheme, we opted for a sinusoidal pattern with a decreasing hull (sine cardinal) following[21]:

$$T(\kappa) = 1 + b\frac{sin(\kappa)}{\kappa},\ \kappa = \Delta + 2\pi\frac{K}{p} \tag{5.2.3}$$

where b can be seen as the amplitude of the oscillations, Δ as a phase delay, and p as a period, K is the iteration number. The use of this tempered scheme allows for an alternation between exploratory phases (high temperature) and exploitation phases (low temperature). The values of the hyperparameters were empirically set to $b = 1, \Delta = 0, p = N_{iter}/100$ after several attempts.

Initialization method

With the same practical considerations, we had to find an initialization method which would improve the odds of convergence towards the global optimum. Random or manual initialization of the model parameters $(\theta_0^c)_c$ was not satisfying, as we will see in the experiments.

We opted for an initialization as close to the clusters as possible. We first fit a disease course mapping model (without mixture) on the whole data. This yields a set of parameters $(\theta_{init}, z_{pop,init}, (z_{i,init}))$. Then we use a Gaussian mixture model (GMM) on the estimated individual parameters $z_{i,init}$. The GMM allows us to identify L clusters within the individual data. The parameters of each mode c in the GMM combined with the population parameters θ_{init} produce new population parameters for a disease course mapping model which represent the mode: for instance

[20] S. Allassonnière and J. Chevallier, "A new class of stochastic EM algorithms. Escaping local maxima and handling intractable sampling," en, *Computational Statistics & Data Analysis*, vol. 159, p. 107 159, Jul. 2021.

[21] S. Allassonnière and J. Chevallier, "A new class of stochastic EM algorithms. Escaping local maxima and handling intractable sampling," en, *Computational Statistics & Data Analysis*, vol. 159, p. 107 159, Jul. 2021.

the new t_0^c is the shift of $t_{0,init}$ by the mean value of τ_i in the mode. This provides initialization parameters for as many disease course mapping models as there are modes in the GMM, and we use these initialization parameters for the clusters of the mixture model.

Please note that the use of a *post hoc* clustering on the estimated parameters is not equivalent to the joint estimation in the mixture model, so this initialization method does not directly give the right cluster parameters. Indeed, we can only produce clusters such that $\forall c, \mathbf{v}^c \in Span(\mathbf{v}_{init})$, which is a strong limit to the posterior analysis of a one-class model.

Simulated data: Multivariate model

In the previous experiments 3.4.1, the results of the mixture DCM model showed consistent performance on a simplified univariate model. However when we add multiple sources, the number of parameters increases drastically and convergence might be more challenging in practice. We generated a dataset as in the 3.4.1 experiment but with two sources and three-dimensional observations. We first fitted a single disease course mapping model on the generated dataset, yielding model parameters $(\theta_{init}, z_{pop,init}, (z_{i,init})_i)$. We then fitted several mixture of DCM models to illustrate our practical improvements. We compared the performance of different initialization methods. In each case, we fitted a model with and without the tempered scheme. Final results are shown in Table 5.1. Each model estimation takes about 5 minutes for 4,000 iterations.

Table 5.1: Performances of models. Single: single disease course mapping model; Random: random initialization; Init: initialization for both clusters at $(\theta_{init}, z_{pop,init})$; GMM: the initialization method described previously using a GMM on $(z_{i,init})_i$; True: perfect initialization at the ground truth parameters.

Model	ROC AUC	Fit ($log\ q$)	Regularization ($log\ p$)
Single	-	42288	-3662
Random non-tempered	0.50	25922	-1439
Random tempered	0.50	36813	-3622
Init non-tempered	0.53	36523	-3637
Init tempered	0.56	40034	-3882
GMM non-tempered	**0.99**	**45787**	**-2914**
GMM tempered	**0.99**	**46871**	**-3068**
True non-tempered	1.00	53613	-2735
True tempered	1.00	53571	-2280

The table shows that the tempered version allows for a better fit at the cost of regularization overall. Random initialization does not work. "Init" shows a better fit but is not able to cluster the individuals. "GMM" initialization is the best method in absence of better heuristics.

Practical applications in multivariate settings also emphasized the need for a pertinent initialization. Otherwise the mixture of DCM models is prone to diverging.

5.2.2 Alternating maximization algorithm

In the Geodesic Bending model, we have a new model parameter: the metric g, or more specifically its parametrization. The metric g was optimized via the learning of a diffeomorphism f in a reproducing kernel hilbert space with kernel k. The maximization step can be stated as an optimization

problem. If the population and random effects $\theta = (\theta_0, (z_i))$ are estimated, the minimization of the error of reconstruction (in the gaussian noise case) can be performed by solving the following problem:

$$f^* \in \underset{f \in \bar{B}_{\mathcal{H}}(0,c)}{\operatorname{argmin}} L(f), L(f) = \sum_{i,j} |y_{i,j} - \gamma_i(t_{i,j}) - f(\gamma_i(t_{i,j}))|^2 \tag{5.2.4}$$

Thanks to the representer theorem, we have:

$$\forall x \in, f^*(x) := \sum_{i,j} k(x_{i,j}, x) w_{i,j}$$

where $x_{i,j} = \gamma_i(t_{i,j})$. With this parametric form for f^*, the previous problem (5.2.4) is equivalent to:

$$\underset{W}{minimize} \qquad ||U - K_X W||^2 \tag{5.2.5a}$$

$$\text{subject to} \qquad W^T K_X W \leq c_0, \tag{5.2.5b}$$

where the previous variables are vectorized:

$$W := (w_{i,j})_{i,j},\ U := (y_{i,j} - \gamma_i(t_{i,j}))_{i,j},\quad K_X := (k(x_{i,j}, x_{k,l}))_{((i,j),(k,l))}$$

and c_0 is the constant ensuring that f is a diffeomorphism.

The minimization of a quadratic function under constraints (5.2.5a) can be solved numerically within a reasonable time as soon as the matrix K_X is not in high dimensions. The total number of visits in a medical cohort can reach up to dozens of thousands and the points $x_{i,j}$ are close to each other, which often results in the matrix being ill-conditioned. To overcome these obstacles, we assume that $y_{i,j} \in [0.1]^d$ and we create a grid of control points in $[0,1]^d$ with step-size σ if σ is the kernel variance (there exists other methods[22], but this one is more close to the LDDMM framework[23] and gives satisfying results[24]). We only keep the grid points which are σ-close to the observations, thus removing the useless ones. We note N_c the number of control points, which can increase exponentially with the dimension according to the previous construction: $N_c(\sigma) \approx \frac{1}{\sigma^d}$. If d is too large, we can choose another method to obtain a reasonable number of control points, a simple way to do so is to subsample the points $(x_{i,j})$. In all the following, the grid method is used because of its practical upsides (no randomness, no risk of ill-conditioning of the matrix K_X).

Even though we can compute the argmax for f, it would be too costly to do so at every iteration of the MCMC SAEM. Therefore we only maximize f every n_{MCMC} iterations. This initially was called the alternating maximization algorithm, as the procedure alternates between MCMC SAEM iterations that only optimize the base DCM model parameters, and the metric optimization. This metric optimization is then repeated for each diffeomorphic composition. In the original work[25],

[22] F. Liu, X. Huang, Y. Chen, *et al.*, "Random Features for Kernel Approximation: A Survey on Algorithms, Theory, and Beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 7128–7148, Oct. 2022.

[23] S. Joshi and M. Miller, "Landmark matching via large deformation diffeomorphisms," *IEEE Transactions on Image Processing*, vol. 9, no. 8, pp. 1357–1370, Aug. 2000.

[24] T. Dao and C. D. Sa, "Gaussian Quadrature for Kernel Features," en,

[25] S. Gruffaz, P.-E. Poulet, E. Maheux, *et al.*, "Learning Riemannian metric for disease progression modeling," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, *et al.*, Eds., vol. 34, Curran Associates, Inc., 2021, pp. 23 780–23 792.

the number of composition is not selected on any precise criterion, but one could simply compute the Bayesian Information Criterion to decide if the added parameters are worth it. The algorithm is described below.

Algorithm 4: Geodesics Bending (Alternating maximization algorithm)

Initialization: g_{init}, θ_{init}, N_{comp}, k, N_c, n_{MCMC}
$g \leftarrow g_{\text{init}}$
$\phi \leftarrow id$
$\theta \leftarrow \theta_{\text{init}}$
for $l = 1$ *to* N_{comp} **do**
  $\phi_l^0 \leftarrow id$
  for $k = 1$ *to* N_c **do**
    $g_l^k \leftarrow g^{\phi_l^k \circ \phi}$
    Run the MCMC-SAEM for n_{MCMC} iterations with metric g to estimate θ_l $\theta \leftarrow \theta_l$
    Solve the optimization problem (5.2.5a) with parameters θ to estimate f_l^k
    $\phi_l^0 \leftarrow id + f_l^k$
  end
  $\phi \leftarrow \phi_l^{N_c} \circ \phi$
  $g \leftarrow g^{\phi}$
end
Result: return θ, g

We start with a simple metric g_{init} (usually the Euclidean metric or the metric resulting from the push-forward of the logit), and we estimate the mixed-effect parameters θ with the MCMC-SAEM algorithm. Then, with θ fixed, we estimate the metric by numerically solving the optimization problem. We repeat these two steps N_c time to obtain one diffeomorphism ϕ_1 that we compose with the identity. Then we loop on this diffeomorphism learning, and each new one is composed with the previous one. We repeat this loop N_{comp} times. The hyperparameters k, N_{comp} and N_c are tuned to make a compromise between loss minimization, computation time and number of parameters.

For N_{comp} compositions, we have $g = g^{\Phi_{N_{\text{comp}}}}$ with $\Phi_{N_{\text{comp}}} = \phi_{N_{\text{comp}}} \circ \ldots \circ \phi_1$ where $\phi_i = +\sum_{j=1}^{m} k(x_j^i, .)w_j^i$, $(x_j^i)_j$ the m control points and $(w_j^i)_j$ their associated weights. This structure resembles the deep neural networks considering m as the width and N_{comp} as the depth. The total time complexity is $\mathcal{O}(n_{\text{MCMC}} d N_c m N_{\text{comp}}^3)$ which reduces in practice the choice of N_{comp}.

In the end the estimation algorithm for Geodesic Bending is a MCMC-SAEM with us taking some liberties in the frequency of the maximization rule.

5.2.3 Gradient maximization step

In the piece-wise geodesic model we introduced a new function, the treatment function f, which we will generally parametrize with θ_f. These parameters become model parameters if we maximize them during the EM, or population parameters if we prefer to sample them. As a general rule of thumb, we prefer avoid sampling parameters as a stochastic search is less optimal than a concrete formula. A model with a higher number of sampled parameters will require more overall MCMC SAEM steps to converge. This is particularly true when the choice of parametrization of f involves

a high number of parameters, such as the RKHS formulation used in 4.3.2. Sampling with a simple Metropolis-Hastings algorithm in such high dimensional cases could be complicated due to the energetic landscape being mostly flat.

Then the question becomes: can we solve the maximization problem ?

$$\theta_f^* = \underset{\theta_f}{\operatorname{argmax}}\ Q(\theta_f, \mathbf{y}, \theta, \mathbf{z}) \tag{5.2.6}$$

Let q_{prior} be a Gaussian prior added to the weights θ_f. If we are in the continuous setting, with a Gaussian noise model, maximizing the likelihood becomes equivalent to the regularized least-square problem:

$$\theta_f^* = \underset{\theta_f}{\operatorname{argmin}} \sum_{i,j,k} \frac{1}{2\sigma_k^2}(y_{ijk} - \eta_k(\psi_i(t_{ij}); \mathbf{w}_i))^2 + \lambda||\theta_f||^2 \tag{5.2.7}$$

We actually used a mixed L1 and L2 regularization in practice, which is equivalent to a mix between a Gaussian and a Laplace prior.

The problem (5.2.7) is not trivial. One simple solution would be to be able to compute the gradient of the objective function and to use it with a gradient descent for example. However the gradient cannot be formulated explicitly. Here is why. After every j-th breakpoint of individual i $b_{i,j}$, the reference curve follows a trajectory that we can generally write with the Riemannian Exponential as : $t \mapsto Exp_{p_{i,j}}(t\mathbf{v}_{i,j})$. The next point $p_{i,j+1}$ is therefore equal to $Exp_{p_{i,j}}((b_{i,j+1} - b_{i,j})\mathbf{v}_{i,j})$, and the next velocity $v_{i,j+1}$ is $f(p_{i,j+1}, \tilde{v}_{i,j+1}, X_{i,j+1})$ where $\tilde{v}_{i,j+1}$ is the derivative at point $\mathbf{p}_{i,j+1}$, and thus a function of $p_{i,j}$ and $v_{i,j}$. So if we assume f to be a function of the position $p_{i,j+1}$, and of the velocity $v_{i,j+1}$, then the velocity after the next break $v_{i,j+1}$ depends on θ_f, $p_{i,j}$ and $v_{i,j}$. And the breakpoint after that will add another layer of complexity by applying f one more time on a position that already depends on f. And so on and so forth.

Every model prediction at time t_{ij} will thus depend on the number of breakpoints before t_{ij}. For every n breakpoints before t_{ij} the function f would have been applied n times in the graph of operations leading to $\eta_k(\psi_i(t_{ij}); \mathbf{w}_i)$. This is reminiscent of recurrent neural networks where the same function is applied several times in a row. In the same way we will use backpropagation to compute the gradient of the objective with respect to the treatment function parameters θ_f, and then perform gradient descent to find a local optimum.

Performing a full gradient descent would be way too long at each iteration of the MCMC-SAEM. We thus relax the maximization step by only performing one gradient descent step during the maximization. This only improves the expected likelihood Q instead of maximizing it, but this is still sufficient to guarantee the convergence of the algorithm. This version of the EM is called Generalized EM, and is proven to converge in[26]. As for Geodesic Bending, the optimization operation in the M-step is still costly so we only perform the gradient step every n_{MCMC} MCMC SAEM iterations.

Computing the gradient via backpropagation however suffers from the same issues as in the field of Deep Learning with vanishing gradients. This is the case when using non-linear "link" functions, for instance in the logistic curves case. The logistic has almost a null gradient on both extremities, which leads to these vanishing gradient issues. This problem is the reason why all the experiments in the piece-wise geodesic model section 4.3.2 were performed in the linear case.

[26] R. Harpaz and R. Haralick, "TR-2006001: The EM Algorithm As a Lower Bound Optimization Technique," *Computer Science Technical Reports*, Jan. 2006.

Another possibility that we did not explore but which is very promising is to allow the model to learn the breakpoints b_{ij} by setting them as latent variables, and sampling them during the MCMC-SAEM.

5.2.4 Discussion

The MCMC SAEM is a powerful algorithm to estimate model parameters while also providing sampling for the latent variables. This proves to be very useful as we are in a Bayesian setting, the Markov chain samplers converge towards the posterior distribution of the latent variables, thus we can easily access it.

Even if the Metropolis-Hastings algorithm guarantees the sampler to converge towards the target distribution, in practice we can run into computational issues. We could easily end up in a problematic case where the proposal is never accepted due to the high dimension of $\mathbf{z}$, even though we have an adaptive variance proposal. Indeed the chance to sample in the right direction to increase the log-likelihood becomes exponentially smaller in high dimensions, making the Metropolis-Hastings almost useless in practice. This is why we use the Gibbs variation, where each coordinate of the latent variables is sampled separately. However the process can be very slow, as the progress is still dictated by a random walk search.

The Metropolis-Hastings is simple to implement but has long been surclassed by other samplers in the field of MCMC sampling. Namely the Hamiltonian Monte-Carlo (HMC) sampler[27] uses a dynamical approach to speed-up sampling using gradient information. The state-of-the-art sampler currently is a variant of the HMC called the No-U-turn sampler[28]. This last one is more complicated to implement.

Another sampler using the gradient to improve sampling is the Metropolis Adjusted Langevin Algorithm (MALA) and its Riemannian improvements. In the MALA, the proposal is informed by the local gradient of the parameters to improve the convergence towards the mode of the distribution. This proves to be very useful in high dimensional settings where the random walk has a hard time finding its way to the mode. In[29] the authors propose to improve both the MALA and the HMC samplers by using the natural gradient instead of the standard gradient. The natural gradient corresponds to the gradient in a Riemannian manifold where the metric is the Fisher information matrix of the parameters we wish to sample. Both samplers show an increased convergence rate compared to their non-Riemannian counterpart. Since we already computed the gradients and the Fisher information of the DCM models 15, it would be a simple future improvement to implement these samplers instead of the Metropolis-Hastings within Gibbs.

[27] M. Betancourt, *A Conceptual Introduction to Hamiltonian Monte Carlo*, Jul. 2018.

[28] M. D. Homan and A. Gelman, "The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1593–1623, Jan. 2014.

[29] M. Girolami and B. Calderhead, "Riemann manifold Langevin and Hamiltonian Monte Carlo methods," en, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 2, pp. 123–214, 2011.

Conclusion and perspectives

The work presented in this thesis has pushed the boundaries of a previous disease progression model, namely the Disease Course Mapping model. Building upon a geometric model designed to disentangle spatial and temporal effects from a longitudinal dataset, we provided some necessary extensions to apply the model to a wider range of neurodegenerative diseases' cohorts. The model is able to provide insights at the global population level, thereby improving our understanding of the shared patterns of progression across all the individual trajectories. Additionally the hierarchical structure allows for a personalization of the trajectories at an individual level, which is crucial for prognosis or clinical trial enhancement.

We introduced Geodesic Bending as a method to palliate the rigidity in the original DCM model. This technique, inspired from the large deformation diffeomorphic metric mapping field, allows the model to adapt the shape of the trajectory to the data by learning an implicit Riemannian metric. In the applications to Alzheimer's disease, we have shown that this method outperforms the state-of-the-art in predicting the progression of cognitive decline. The average trajectory learned by the model was more in adequacy with the known progression of the disease compared to the logistic curves model. More interestingly the deformations broke the monotonic paradigm of the DCM model. The trajectory remains injective in the multivariate setting, but each biormarker's progression profile can now be learned without the monotonous assumption. This greatly expands the application possibilities of disease progression modelling while retaining the benefits of the interpretability unfolding from the spatio-temporal description.

We then transcended the continuous framework that was inherent to the design of the DCM model with the trajectories on the Riemannian manifold. We used the logistic curves formulation of the DCM to describe the evolution of a latent progression while a Bernoulli and a cumulative logit model described the discrete observations. Even though the model improvement is very simple, the implications in practice are manifold. Symptoms occurrence, disease staging and rating scales are now available in the DCM model. This opens the door to symptom prediction or staging diagnosis. Furthermore, the possibility to model items of a psychometric scale directly instead of the aggregated score allows to describe the disease at the finer-grained level, yielding even more individual variability.

Across our experiments with neurodegenerative data we have felt the urge to improve the model to deal with the heterogeneity of patients' profiles. The random effects of the mixed-effect model structure only account for a natural variability, *random* as implied by the name. However this variability can stem from non-random sources, which can be described by external covariates or by disease subtypes. Focusing on the latter we introduced a mixture option in the DCM model, either at the individual parameters level or at the global progression level. Both rely on the use of latent classes for the individuals. The mixture model aims at jointly estimating these latent classes

while learning the average disease progression of each class. The two variants each have their own preferred use case, allowing to uncover new patterns of variability across the data.

In the last part of our work we focused on what we believe to be the next area of research for disease progression models. Treatment modelling's importance is going to surge as new therapeutic approaches are progressively delivering results. Applications of disease progression models are plural, including in clinical trials. We proposed one such application, which paves the way for the design of clinical trials with only one treated arm by leveraging a synthetic control arm. In our proposed framework, a disease progression model is used to predict the natural disease progression, i.e. under no treatment, allowing a treatment model to regress the difference with the observed progression of the treated patient. We showcased our approach using Parkinson's disease data to analyse the effect of the dopaminergic treatment over motor symptoms, quality of life and the levels of dopamine transporter in the putamen. Finally we introduced the Piece-wise Geodesic model, a Riemannian based model to account for disease-modifying treatment effect over the trajectories. This model has been tested on synthetic data and still awaits the proper dataset to be applied on real-life data.

All of the design of those models and extensions is subject to the estimation condition. We had to design them so that identifiability was ensured, while making sure that the new parameters could be learned properly. In the estimation chapter, we explained our modelling choices from the point of view of the MCMC-SAEM algorithm. We provided insights into why we chose certain parameters to be latent variables, which ones we sampled and which ones were directly maximized, or simply updated following a gradient step in the case of the Piece-wise Geodesic model. We identified the practical issues preventing convergence and discussed about empirical solutions.

Our improvements to the DCM framework all contribute to design better, larger, more comprehensive models for neurodegenerative diseases. By doing so, we indirectly increase our understanding of disease progression, we embrace the heterogeneity and help in the development of a cure. This has been highlighted in our many applications, where we tried to stress the clinical implications of the results. During this thesis we have striven to communicate with clinicians, on the occasion of conferences 15 or by exchanging with neurologist from our institute, so that they benefit from our work and that we could improve our model based on their feedback.

Limitations and perspectives

Precise limitations have been highlighted at the end of each dedicated chapter. We will try to summarize them:

- Geodesic Bending: the flexibility of the model comes at a higher computational cost. The method is more data-hungry due to a higher risk of overfitting on few samples.
- Binary model: models using purely symptom features do not perform well in terms of prediction, the model should include for informative features, i.e. continuous markers, to guide the trajectory.
- Ordinal model: the model also requires more data, specifically all the levels of the ordinal items should be observed. As the ordinal models tend to use more features, their complexity increases, especially computationally.
- Mixture model: estimation of multivariate mixture models remains complex, and convergence is hard to reach in some cases. The use of better samplers should improve in this matter

- Additional treatment effect framework: the treatment effect estimation relies on the predicting accuracy of the disease progression model. However training such a model on pre-treatment observations only is usually prone to fail long-term predictions. The solution would be to use another cohort to learn the natural disease history
- Piece-wise Geodesic model: the current parametrization makes the model very data dependent. The model also needs to be confronted to real-life data.

The main challenge often resides in the estimation of those methods. As we increase the complexity more data is required to properly estimate the models. Some sparsity in the parametrization of the learned function, for instance in Geodesic Bending or in the Piece-wise Geodesic model, could help regarding this issue. We believe that one major improvement could be the implementation of better samplers.

Some further investigation into the compatibility of the models should be done. For instance crossing Geodesic Bending with the mixture model could lead to learning different shapes of progression for the subtypes. We mentioned that symptoms were difficult to model because many were influenced by treatments, so it would make sense to use the treatment model to assess what would have been the progression without treatment and how much the treatment affected the symptom. Finally we hope that our first steps in the field of treatment modelling will be followed as this will prove to be useful in the years to come.

Publications and scientific communication

Scientific articles

Journal articles

- Pierre-Emmanuel Poulet, Stanley Durrleman. Multivariate disease progression modelling with longitudinal ordinal data. Statistics in Medicine, 2023; 1- 20. doi: 10.1002/sim.9770

- Moulaire, P., Poulet, P.E., Petit, E., Klockgether, T., Durr, A., Ashisawa, T., Tezenas du Montcel, S. (2023), Temporal Dynamics of the Scale for the Assessment and Rating of Ataxia in Spinocerebellar Ataxias. Mov Disord, 38: 35-44. https://doi.org/10.1002/mds.29255

Conference proceedings papers

- Poulet, PE., Durrleman, S. (2021). Mixture Modeling for Identifying Subtypes in Disease Course Mapping. In: Feragen, A., Sommer, S., Schnabel, J., Nielsen, M. (eds) Information Processing in Medical Imaging. IPMI 2021. Lecture Notes in Computer Science(), vol 12729. Springer, Cham. https://doi.org/10.1007/978-3-030-78191-0_44. Poster presentation at the virtual conference event

- Samuel Gruffaz, Pierre-Emmanuel Poulet, Etienne Maheux, Bruno Jedynak, Stanley Durrleman. Learning Riemannian metric for disease progression modeling. NeurIPS 2021 - Thirty-fifth Conference on Neural Information Processing Systems, Dec 2021, link

- (in preparation) Pierre-Emmanuel Poulet, Bruno Jedynak, Stanley Durrleman. A Bayesian symptomatic treatment model using latent disease progression, 2023

Conference communications

- Pierre-Emmanuel Poulet, Jean-Christophe Corvol, Stanley Durrleman. A latent mixed-effects model for longitudinal categorical data in Parkinson disease. CompAge 2020 - Computational approaches for ageing and age-related diseases 2020, Sep 2020, Paris, France. Poster presentation at the virtual conference event

- Pierre-Emmanuel Poulet, Stanley Durrleman, Jean Christophe Corvol. Predicting symptom onset in Parkinson's disease with latent mixed-effect model. AD/PD 2021 - 15th International Conference on Alzheimer's Parkinson's Diseases, Mar 2021, Barcelone, Spain. Oral presentation at the virtual conference event
- Pierre-Emmanuel Poulet, Bruno Jedynak, Stanley Durrleman. Learning treatment effect in neurodegenerative diseases with a Bayesian mixed-effect model. 43rd Annual Conference of the International Society for Clinical Biostatistics (ISCB), Aug 2022, Newcastle upon Tyne, United Kingdom. Poster presentation, award for best poster.
- Pierre-Emmanuel Poulet, Bruno Jedynak, Zoltan Mari, Stanley Durrleman. Non-parametric modeling for dopaminergic treatment effect in Parkinson's disease. AD/PD 2023 - 17th International Conference on Alzheimer's Parkinson's Diseases, Mar 2023, Goteborg, Sweden. Poster presentation
- (in preparation) Pierre-Emmanuel Poulet, Bruno Jedynak, Stanley Durrleman. Piece-wise Geodesic: a longitudinal disease-modifying treatment effect model. 44th Annual Conference of the International Society for Clinical Biostatistics (ISCB), Aug 2023, Milan, Italy. Poster presentation.

Teaching

- Teaching mission, Sorbonne university, Computational department, Jussieu. List of courses:
  - 2020: "Elements of programming" License 1 (equivalent: bachelor 1) introductory course to Python, practical sessions
  - 2021: "Elements of programming" License 1, practical sessions
  - 2022: "Elements of programming" License 1, group classes and practical sessions
  - 2022: "Data science" License 1 introductory course to data science, practical sessions
- "Disease course mapping with longitudinal data" practical session workshop at AI4health Virtual Winter school, with Arnaud Valladier, Igor Koval, Etienne Maheux, Juliette Ortholand and Stanley Durrleman, Jan 2021
- "Disease course mapping with longitudinal data" practical session workshop at AI4health Winter school, with Némo Fournier, Etienne Maheux and Juliette Ortholand, Jan 2022
- "Disease course mapping with longitudinal data" practical session workshop at AI4health Summer school, with Sofia Kaisaridi and Juliette Ortholand, Jul 2023

Software development

Our code contributions are all published in the Python open-source library Leaspy:
https://leaspy.readthedocs.io/en/stable/
The library is still expanding with many new variants, including all the work presented in this manuscript. The Gitlab link to the code is:
https://gitlab.com/icm-institute/aramislab/leaspy.
The Leaspy library was released in 2017 under the GNU GPLv3 license, and benefits from the work of various contributors. All the code is peer-reviewed by the developers before being released in the stable version. During my thesis I contributed as a developer of the library, with commits corresponding to the discrete data models in the `master` branch, the collaborative implementation of the`GeodesicBending` branch with Samuel Gruffaz, the whole `mixture_model` branch and the whole `treatment_modelling` branch.

The stable version of Leaspy (i.e. the `pip` release and the `master` branch of the Git) already include the binary and ordinal DCM models, which I developed and implemented. A minimal example of code usage is given below:

```
# load packages
import pandas as pd
from leapy import Leaspy, Data, AlgorithmSettings

# load data
df = pd.read_csv("path_to_my_data.csv", index_col=["ID", "TIME"])

# Data format:
# ID    TIME    Y1  Y2
# 01    60.     0.  1.
# 01    65.     ... ...

leaspy_data = Data.from_dataframe(df)

# Instantiating a logistic model
leaspy = Leaspy("logistic", noise_model="gaussian") # for continuous data
# or
leaspy = Leaspy("logistic", noise_model="bernoulli") # for binary data
# or
leaspy = Leaspy("logistic", noise_model="ordinal") # for ordinal data

# Calibration
algo_settings = AlgorithmSettings('mcmc_saem', n_iter=1024)
leaspy.fit(leaspy_data, algo_settings)

# Personalization
```

```
settings_personalization = AlgorithmSettings("scipy_minimize", use_jacobian=True)
individual_parameters = leaspy.personalize(leaspy_data, settings_personalization)

# Prediction
values = leaspy.estimate(timepoints, individual_parameters)
```

An introduction to Leaspy with hands-on tutorial notebooks can be found at:
https://gitlab.com/icm-institute/aramislab/disease-course-mapping-tutorials/-/tree/master/challenges

Our personal code for Geodesic Bending, mixture models and Piece-wise geodesic model is still waiting for a stable release. Leaspy is undergoing a deep refactoring for a second version as of the time of writing this thesis, and will probably be released in the year following this work.

Appendix

Riemannian geometry

This section is an introduction to the concepts of Riemannian geometry used in this thesis. We don't claim it to be a comprehensive overview of Riemannian geometry, for this we refer the reader to the excellent book by Manfredo Do Carmo: *Riemannian Geomerty*[30]. Riemannian geometry is a powerful tool to study structured manifolds. The applications are ubiquitous: for instance medical images only span a subset of all the possible images, and such subset might be of a much lower dimension with a specific geometry. We will therefore often see the manifold $\mathcal{M}$ as an open subset of $\mathbb{R}^D$. We will first provide all the definitions required to reach the notion of Riemannian manifold.

Definition 1 (Manifold)**.** *A topological space $\mathcal{M}$ is a d-dimensional manifold iff, $\forall p \in \mathcal{M}, \exists U(p)$ a neighbourhood and a homeomorphism $x : U(p) \to x(U(p)) \subset \mathbb{R}^d$.*

(U, x) is called a local coordinate chart. A collection of charts which covers $\mathcal{M}$ is called an atlas. A manifold is called *smooth* iff for any two charts $(U, x), (V, y)$ in an atlas, $y \circ x^{-1}$ is C^∞. The dimension d of an embedded manifold is smaller than the space in which it is embedded $d \leq D$.

Intuitively, a smooth manifold of dimension d is a topological space where each point can be assigned coordinates, and these coordinates are differentiable.

Definition 2 (Tangent space)**.** *The tangent space at p in a smooth manifold $\mathcal{M}$ is $\{v / \exists a curve c : (-\epsilon, \epsilon) \to \mathcal{M} / v = \dot{c}(0) and p = c(0)\}$. It is noted $T\mathcal{M}_p$.*

The tangent space is therefore the vector space of all possible derivatives of curves passing through p. This space is distinct in each point of the manifold, but if we write $(x_1, ..., x_d)$ the local coordinates in the neighbourhood of p, we can define the basis $(\frac{\partial}{_1}(p), ..., \frac{\partial}{_d}(p))$ of the tangent space $T\mathcal{M}_p$.

Definition 3 (Vector field)**.** *A vector field X is a smooth function mapping points to vectors in the tangent space $X : p \in \mathcal{M} \mapsto X(p) \in T\mathcal{M}_p$.*

In a local chart coordinates, if the coordinates of a point p are $(x_1, ..., x_d)$, we can write the vector field as:

$$X(p) = \sum_{i=1}^{d} X(x_1, ..., x_d) \frac{\partial}{_i}(p)$$

We can now define:

[30] M. Do Carmo, *Riemannian Geometry*, en.

Definition 4 (Riemannian manifold)**.** *If $\forall p \in \mathcal{M}$, $g_p(.,.)$ is a scalar product on $T\mathcal{M}_p$ and $: p \to g_p$ is smooth, then $(\mathcal{M}, g)$ is a Riemannian manifold. g is called the Riemannian metric.*

One way to construct a metric is to use push-forward metrics, as is done in the Geodesic Bending model:

Definition 5 (Pushforward metrics)**.** *Provided $(\mathcal{M}, g)$ a Riemannian space, N a manifold and $\phi : \mathcal{M} \to N$ a C^1 diffeomorphism, we can equip N with the Riemannian metric g^ϕ defined as:*

$$\forall\ p \in N, \quad \forall w, v \in TN_p, \quad g^\phi_p(w, v) = g_{\phi^{-1}(p)}(d\phi^{-1}(p).w, d\phi^{-1}(p).v)$$

The use of push-forward metrics allows to carry a metric from a manifold to another through a diffeomorphism. In the most simple case the starting manifold is the Euclidean space equipped with the Euclidean metric, from which we simply apply a diffeomorphism.

An important notion in Riemannian geometry is the geodesic, which is the "shortest" path (in the sense of the Riemannian metric) between two points of the manifold. The interest of defining push-forward metrics is that the geodesics then are obtained by simply applying the diffeomorphism to the geodesics in the starting manifold.

Let's define the geodesics:

Definition 6 (Distance and geodesic)**.** *Given a Riemannian manifold, we define the Riemannian distance d such that*

$$\forall p_1, p_2 \in \mathcal{M},\ d(p_1, p_2) = \inf_{\gamma:[0,1]\to\mathcal{M}, \gamma(0)=p_1, \gamma(1)=p_2} \int_0^1 \sqrt{g_{\gamma(t)}(\dot\gamma(t), \dot\gamma(t))}dt \tag{.0.8}$$

If γ^* minimize the previous problem, γ^* is called a geodesic.

In the Euclidean space, geodesics are straight lines such that $\ddot\gamma = 0$. However we cannot yet compute second order derivatives in the Riemannian manifold since the tangent spaces are *different* in every point. Indeed, if we have a trajectory $\gamma :]-1, 1[\to \mathcal{M}$, we can compute the first order derivative for each time $\dot\gamma(t) = \lim_{\epsilon\to 0} \frac{\gamma(t+\epsilon)-\gamma(t)}{\epsilon}$ and this element belongs to the tangent space $T\mathcal{M}_{\gamma(t)}$ by definition. However we cannot compute $\ddot\gamma(t) = \lim_{\epsilon\to 0} \frac{\dot\gamma(t+\epsilon)-\dot\gamma(t)}{\epsilon}$ since $\dot\gamma(t+\epsilon)$ and $\dot\gamma(t)$ belong to two possibly different vector spaces. In Euclidean geometry this is not an issue as the tangent spaces are all the same.

This is where the notion of connection comes into play:

To do it, we reduce the problem to the vectors of the basis of each tangent space ($\frac{\partial}{i}(p)$). A connection defines how the basis is varying locally. Precisely:

Definition 7 (Connection)**.** *An affine connection ∇ maps two vector fields $v : p \in \mathcal{M} \to v(p) \in T\mathcal{M}_p$, $w : p \in \mathcal{M} \to w(p) \in T\mathcal{M}_p$ to a single vector field $\nabla_v w$ with the following properties:*

1. *[Left-linearity] $\forall X_1, X_2$ and Y vector fields:*

$$\forall f, g \in C^\infty(\mathcal{M}, \mathbb{R}), \forall p \in \mathcal{M}, \nabla_{fX_1+gX_2}Y(p) = f(p)\nabla_X Y(p) + g(p)\nabla_{X_2}Y(p)$$

2. *[Right-Leibniz-linearity] $\forall X, Y$ f-vector fields:*

$$\forall f \in C^\infty(\mathcal{M}, \mathbb{R}), \forall p \in \mathcal{M}, \nabla_X(fY)(p) = f(p)\nabla_{X_1}Y(p) + X(f)(p)Y(p)$$

where $X(f)(p) = \lim_{\epsilon\to 0} \frac{f(\gamma(\epsilon))-f(p)}{\epsilon}$ and $\gamma :]-1, 1[\to \mathcal{M}$ is a diffentiable curve such that $\gamma(0) = p, \dot\gamma(0) = X(p)$.

3. [R-right-linearity] $\forall X, Y_1, Y_2$ *vector fields:*

$$\forall a, b \in \mathbb{R}, \forall p \in \mathcal{M}, \nabla_X(aY_1 + bY_2)(p) = a\nabla_X Y_1(p) + b\nabla_X Y_2(p)$$

The connection can be defined by its action on the basis $(\frac{\partial}{i}(p))$. If we write $\nabla_{\frac{\partial}{i}(p)}\frac{\partial}{j}(p)(p) = \sum_k \Gamma^k_{i,j}(p)\frac{\partial}{k}(p)$, the coefficients $\Gamma^k_{i,j}(p) \in \mathbb{R}$ are characterizing the connection thanks to the linearity properties. They are called the Christoffel's symbols.

We can now compute the second order derivative by considering that $\nabla_{\dot\gamma}\dot\gamma$ is the analogue of $\ddot\gamma$. There exists a unique specific connection such that the solutions of $\nabla_{\dot\gamma}\dot\gamma = 0$ are exactly the geodesics. This connection is called the Levi-Civita connection, and it depends on the metric g of the manifold. We will admit these results here. More details on this connection in the book[31].

We decompose the coordinates of $\dot\gamma$ as $\dot\gamma(t) = \sum_i \dot\gamma_i(t)e_i(\gamma(t))$. Then:

$$\nabla_{\dot\gamma}\dot\gamma(\gamma(t)) = \sum_k \left(\ddot\gamma_k(t) + \sum_{i,j} \Gamma^k_{i,j}(\gamma(t))\dot\gamma_i(t)\dot\gamma_j(t))e_k(\gamma(t) \right)$$

where $\ddot\gamma_k(t) = \lim_{\epsilon\to 0} \frac{\dot\gamma_k(t+\epsilon)-\dot\gamma_k(t)}{\epsilon}$.

In the end the geodesics equation can be reformulated as:

$$\ddot\gamma_k(t) + \sum_{i,j} \Gamma^k_{i,j}(\gamma(t))\dot\gamma_i(t)\dot\gamma_j(t) = 0 \tag{.0.9}$$

This second order Ordinary Differential Equations (ODE) has solutions under some regularity conditions on the Christoffel's symbols. Geodesics can thus be reduced to the Cauchy conditions in 0 ($\gamma(0) = p$, $\dot\gamma(0) = v$).

We can now define an essential tool of Riemannian geometry:

Definition 8 (Riemannian Exponential)**.** *The Riemannian Exponential is defined for* $p \in \mathcal{M}$, $t_0 \in \mathbb{R}$ *and* $v \in T\mathcal{M}_p$ *as the mapping:*

$$Exp_{p,t_0}(v)(.) : \left\{ \begin{array}{ccc} I & \to & \mathcal{M} \\ t & \mapsto & Exp_{p,t_0}(v)(t) \end{array} \right.$$

where $Exp_{p,t_0}(v)(t)$ *is the point reached at time* t *by the geodesic starting at time* t_0 *at point* p *with velocity* v*.*

We will often reduce the Exponential to an abusive notation ${}_p(v)$ where $t_0 = 0$ and $t = 1$ for simplification. This operation of starting at point p and following the direction v for a unit time is called *shooting*.

We say that a Riemannian space is complete if the map : ${}_p : T\mathcal{M} \to \mathcal{M}$ is well-defined for all point p in the manifold $\mathcal{M}$ (the ODE admits solutions defined on the open interval $\mathbb{R}$). Note that in the Euclidean case ${}_p(v) = p + v$.

In order to finally reach the last notion of this section, which is the Exp-parallelization, we require one last tool: parallel transport. It will enable us to define parallel vectors in tangents space along a trajectory according to a given connection ∇.

[31] M. Do Carmo, *Riemannian Geometry*, en.

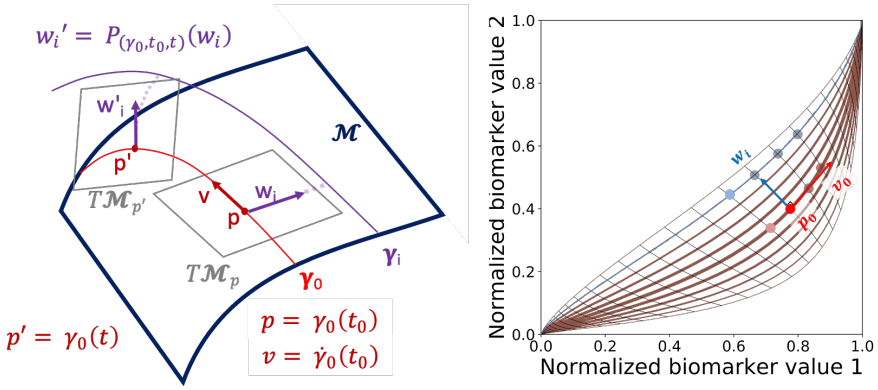

Figure 1: Left: Sketch of the Exp-paralellisation in an abstract manifold. Right: Exp-parallelisation for the unit square as a manifold with the metric $g_{(p_1,p_2)}((x_1,x_2),(y_1,y_2)) := \frac{x_1 y_1}{p_1^2(1-p_1)^2} + \frac{x_2 y_2}{p_2^2(1-p_2)^2}$. The average trajectory in red and an individual trajectory in blue or purple. Reproduced from[32]

Definition 9 (Parallel transport)**.** *Let $\gamma : t \mapsto \gamma(t)$ be a smooth curve on $\mathcal{M}$ and X a vector field on $\gamma_(t)$. The vector field X is said to be parallel along $\gamma_(t)$ iff $\nabla_{\gamma_(t)} X = 0$.*

The parallel transport is unique: given a curve γ and a vector v_0 tangent to $\gamma(t_0)$, there exists a unique vector field X parallel along γ such that $X(\gamma(t_0)) = v_0$. We denote $P_{\gamma,t_0,t}(v_0)$ the isometry mapping v_0 to $X(\gamma(t))$.

Remark that in the Euclidean space $X(\gamma(t)) = v_0$ (parallel vector are colinear).

Definition 10 (Exp-parallelization)**.** *Let $(\mathcal{M}, g)$ be a geodesically complete Riemannian manifold equipped with its Levi-Civita connection ∇^g. Let $\gamma : I \to \mathcal{M}$ be a curve on $\mathcal{M}$, $t_0 \in I$ and $w \in T\gamma(\sqcup_\prime), w \neq 0$. The Exp-parallelization of γ according to w at t_0 is the curve:*

$$\eta_w(\gamma, \dot{}) : \left\{ \begin{array}{ccc} I & \to & \mathcal{M} \\ t & \mapsto & Exp_{\gamma(t_0)}(P_{(\gamma,t_0,t)}) \end{array} \right.$$

Intuitively, the curve $\eta_w(\gamma, \dot{})$ is a shift from γ in the direction w at t_0. At t_0, the point constructed is simply by shooting with w starting from $\gamma(t_0)$. For the rest of the curve, the only thing changing is that we don't take directly w (as it lives in $T\mathcal{M}_{\gamma(t_0)}$), but the result of the transport of this vector w parallel to the curve γ.

Disease Course Mapping formulas

Introduction

This section is a summary of all the formulas for logistic disease course mapping (DCM) model: basic formulation of the model, parameters introduction, gradients and finally Fisher information

matrix. We provide the formulas *as they are implemented*, which includes some simplifications over the theoretical model introduced in[33].

Model description

Observations and dataset

The input of the model is a longitudinal dataset, for which we introduce the following notations :

- a number of patients N, that will be indexed with i,
- for each patient a certain number n_i of visits, that will be indexed with j,
- for each visit of a patient, a time $t_{i,j}$ is associated and a set of features $y_{i,j,k}$ with $1 \leq k \leq d$ being the index of the feature

Model equation

The framework has been first developed in[34] in a general formulation.

The age of the patient is mapped onto the common disease timeline through time reparametrization $\psi_i(t)$:

$$\psi_i(t) = \alpha_i(t - \tau_i) \tag{.0.10}$$

where α_i is the acceleration factor of patient i and τ_i is its time-shift. The acceleration is reparametrized as $\alpha_i = e^{\xi_i}$ to enforce its positivity. The theoretical formulation is $\psi_i(t) = \alpha_i(t - t_0 - \tau_i) + t_0$ with $\bar{\tau} = 0$, but this is equivalent to equation (.0.10) above with $t_0 = \bar{\tau}$ while the $+t_0$ term is absorbed by other parameters.

The trajectory is then obtained by computing an exp-parallellization of a geodesic in the Riemannian manifold. In the logistic case the geodesic is:

$$\gamma_k(t) = \left(1 + (\frac{1}{p_k} - 1) \times \exp(-\frac{v_k}{p_k(1-p_k)}(t - t_0)\right)^{-1} \tag{.0.11}$$

where p_k and v_k are the k-th coordinate of the position and velocity (derivative of the curve) at time $t_0 = \bar{\tau}$ of the average trajectory. After exp-parallelization you get :

$$\forall k \in [|1, d|], \quad \eta_k(\psi_i(t_{ij}); \mathbf{w}_i) = \left(1 + (\tfrac{1}{p_k} - 1) \exp\left(-\tfrac{v_k \psi_i(t_{ij}) + w_{ik}}{p_k(1-p_k)}\right)\right)^{-1} \tag{.0.12}$$

with w_{ik} being the k-th coordinate of the space-shift $\mathbf{w}_i$. Since we will use this notation very often, we introduce the abbreviation $\eta_k(\psi_i(t_{ij}); \mathbf{w}_i) = \eta$ which we will call the "model".

The space-shifts are decomposed along N_s sources ($0 \leq N_s \leq d - 1$):

[33] J.-B. Schiratti, S. Allassonnière, A. Routier, *et al.*, "A Mixed-Effects Model with Time Reparametrization for Longitudinal Univariate Manifold-Valued Data," en, in *Information Processing in Medical Imaging*, S. Ourselin, D. C. Alexander, C.-F. Westin, *et al.*, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2015, pp. 564–575.

[34] J.-B. Schiratti, S. Allassonnière, A. Routier, *et al.*, "A Mixed-Effects Model with Time Reparametrization for Longitudinal Univariate Manifold-Valued Data," en, in *Information Processing in Medical Imaging*, S. Ourselin, D. C. Alexander, C.-F. Westin, *et al.*, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2015, pp. 564–575.

$$\boldsymbol{w}_i = \boldsymbol{A}\boldsymbol{s}_i \tag{.0.13}$$

where $\boldsymbol{A} \in \mathbb{R}^{d \times N_s}$ is called the mixing matrix and is composed of N_s columns orthogonal to v_0 for identifiability. This is done by computing an orthogonal basis $(\mathcal{B}_m)_{1 \leq m \leq d-1}$ of $Span(\boldsymbol{v})^{\perp}$ with the Householder method. Then the columns of $\boldsymbol{A}$ are linear combinations of this basis $\mathcal{B}$:

$$\boldsymbol{A}_{:,l} = \sum_{m=1}^{d-1} \beta_{ml} \mathcal{B}_m \tag{.0.14}$$

where the coefficients $(\beta_{ml}, 1 \leq m \leq d-1, 1 \leq l \leq N_s)$ are the parameters which are learned in the model.

Some more notations are introduced below:

$$p_k = \frac{1}{1+g_k} \tag{.0.15}$$
$$g_k = e^{\tilde{g_k}} \tag{.0.16}$$
$$v_k = e^{\tilde{v_k}} \tag{.0.17}$$

A calculation that comes back often is $\frac{1}{p_k(1-p_k)} = \frac{(1+g_k)^2}{g_k}$.

Bayesian model

The disease course mapping framework adds a probabilistic model over the geometrical model (which is not structurally identifiable as there can be several triplets (t_0, p_0, v_0) describing the same set of trajectories). Some parameters are introduced to have priors over individual parameters.

Priors

Here is the complete prior list:

[Population parameters (fixed effects)]

$$\begin{aligned} \tilde{g_k} &\sim \mathcal{N}(\bar{g_k}, \sigma_g^2) && (.0.18)\\ \tilde{v_k} &\sim \mathcal{N}(\bar{v_k}, \sigma_v^2) && (.0.19)\\ \beta_{ml} &\sim \mathcal{N}(\bar{\beta_{ml}}, \sigma_\beta^2) && (.0.20) \end{aligned}$$

[Individual parameters (random effects)]

$$\begin{aligned} \tau_i &\sim \mathcal{N}(\bar{\tau}, \sigma_\tau^2) && (.0.21)\\ \xi_i &\sim \mathcal{N}(0, \sigma_\xi^2) && (.0.22)\\ s_{il} &\sim \mathcal{N}(0, 1) && (.0.23) \end{aligned}$$

Parameter classes

Let's divide the parameters into categories:

- hyperparameters: these parameters are fixed by the user ahead of estimation. These include $\sigma_g, \sigma_v, \sigma_\beta$ which are all fixed to 0.01 by default. They appear in priors involving the same scheme: $x \sim \mathcal{N}(\bar{x}, \sigma_x^2)$. These priors are not really constraining. Intuitively they only provide the exploration strength during training for the population parameters. They do not shape the posterior of related parameters as the $\bar{g}, \bar{v}, \bar{\beta}$ are updated to maximize the likelihood at each step of the EM algorithm which amounts to setting them to g, v, β respectively.

- model parameters: $\theta = (\bar{g}, \bar{v}, \bar{\beta}, \bar{\tau}, \sigma_\tau, \sigma_\xi)$ these are parameters which appear as the mean of gaussian priors. Intuitively they are the mean value of the related parameters, which for population parameters (g, v, β) is the best value to use in the model formula (.0.12). For individual parameters (τ, ξ) they imply that the distribution of such random effects is expected to be gaussian and $(\bar{\tau}, 0)$ is the mean of this distribution. These parameters are all maximized during the M-step of the EM algorithm used for estimation.

- latent parameters: $\boldsymbol{z} = (\boldsymbol{z}_{pop}, (\boldsymbol{z}_i)_{1 \leq i \leq N})$ which are sampled during the E-step of the EM. We distinguish between population parameters $\boldsymbol{z}_{pop} = (\tilde{g}, \tilde{v}, \beta)$ which involve the whole dataset (all the population) in order to be updated during estimation, and individual parameters $\boldsymbol{z}_i = (\tau_i, \xi_i, \boldsymbol{s}_i)$ which only require data points of individual i to be updated.

From model to data

Depending on the type of data of the observations, the link between the model η and the data values y_{ijk} is different. Here we provide the expressions for continuous data modelled with Gaussian noise, binary data modelled with Bernoulli noise and ordinal data modelled with a cumulative probit model.

Gaussian noise The data points are continuous values, supposedly normalized between 0 and 1, and are supposed to follow a normal law:

$$y_{ijk} \sim \mathcal{N}(\eta_k(\psi_i(t_{ij}); \mathbf{w}_i), \sigma_k^2) \tag{.0.24}$$

where σ_k is the standard deviation of the noise of item k. This parameter is learned and is a model parameter, which means it is updated during the M-step of the EM.

Bernoulli noise The data points are binary values, either 0 or 1, and we model them with a Bernoulli:

$$y_{ijk} \sim \mathcal{B}(\eta_k(\psi_i(t_{ij}); \mathbf{w}_i)) \tag{.0.25}$$

There is no supplementary parameter to learn for this type of data.

Ordinal noise Ordinal data is a bit more complicated: observations y_{ijk} are supposed to be integers between 0 and a given maximum level L_k. The model η describes the cumulative distribution function instead of the probability distribution function. Due to the logistic curve going up to 1 as time passes, this implies that the ordinal scale should be ordered such that individuals start at 0 and eventually reach the maximum score in the infinite time horizon. Here is how it formulates:

$$\forall k \in [|1, d|], \forall l \in [|1, L_k|], \mathbb{P}(y_{ijk} \geq l) = \eta_k(\psi_i^{k,l}(t_{ij}); \mathbf{w}_i) \tag{.0.26}$$

$$\psi_i^{k,l}(t) = e^{\xi_i}(t - \tau_i) - \sum_{m=1}^{l} \delta_k^m \tag{.0.27}$$

with $\delta_k^m > 0$ being the average time delay between levels $m-1$ and m for item k. These $(\delta_k^m)_{1 \leq k \leq d, 1 \leq m \leq L_k}$ parameters are new population parameters, thus latent, i.e. sampled during the E-step of the algorithm. This adds the following prior:

$$\delta_k^m \sim \mathcal{N}(\bar{\delta_k^m}, \sigma_\delta^2) \tag{.0.28}$$

with $(\bar{\delta_k^m})_{1 \leq k \leq d, 1 \leq m \leq L_k}$ being the corresponding new model parameters and σ_δ is a new hyperparameter fixed at 0.01.

Complete equation (.0.26) with logistic curves model formulates as:

$$\eta_k(\psi_i^{k,l}(t_{ij}); \mathbf{w}_i) = \left(1 + (\frac{1}{p_k} - 1) \exp\left(-\frac{v_k(e^{\xi_i}(t_{ij} - \tau_i) - \sum_{m=1}^{l} \delta_k^m) + w_{ik}}{p_k(1 - p_k)}\right)\right)^{-1}$$

Likelihood

Estimating the model is done by maximizing the likelihood. Model parameters θ are not random, which is the meaning of the ";" separator before including θ in the next formulas. We usually compute the log-likelihood formulas. The total log-likelihood of the model writes:

$$\log\ q(\boldsymbol{y}, \boldsymbol{z}; \theta) = \log\ q(\boldsymbol{y} \mid \boldsymbol{z}; \theta) \tag{.0.29}$$

$$+ \log\ q(\boldsymbol{z}_{pop}; \theta) \tag{.0.30}$$

$$+ \sum_{i=1}^{N} \log\ q(\boldsymbol{z}_i \mid \boldsymbol{z}_{pop}; \theta) \tag{.0.31}$$

We will detail each term below.

Data attachment

The first term, referred to as *attachment*, is measuring how the model fits the data with the parameters θ and $\boldsymbol{z}$. This term depends on the current noise model, thus on the data type:

$$\log(q(\mathbf{y}|\mathbf{z};\theta)) = -\sum_{k=1}^{d}\left(N_{tot}\log(\sigma_k\sqrt{2\pi}) + \tfrac{1}{2\sigma_k^2}\sum_{i=1}^{N}\sum_{j=1}^{N_i}(y_{i,j,k} - \eta_k(\psi_i(t_{ij});\mathbf{w}_i))^2\right) \quad \text{[Gaussian noise]}$$

$$\log(q(\mathbf{y}|\mathbf{z};\theta)) = \sum_{i=1}^{N}\sum_{j=1}^{N_i}\sum_{k=1}^{d} y_{i,j,k}\log(\eta_k(\psi_i(t_{ij});\mathbf{w}_i)) + (1-y_{i,j,k})\log(1-\eta_k(\psi_i(t_{ij});\mathbf{w}_i)) \quad \text{[Bernoulli noise]}$$

$$\log(q(\mathbf{y}|\mathbf{z};\theta)) = \sum_{i=1}^{N}\sum_{j=1}^{N_i}\sum_{k=1}^{d}\sum_{l=1}^{L_k}\mathbb{1}(y_{i,j,k}=l)\log(\eta_k(\psi_i^{k,l}(t_{ij});\mathbf{w}_i) - \eta_k(\psi_i^{k,l+1}(t_{ij});\mathbf{w}_i)) \quad \text{[Ordinal noise]}$$

with $\eta_k(\psi_i^{k,0}(t_{ij}),\mathbf{w}_i) = 1$ and $\eta_k(\psi_i^{k,L_k+1}(t_{ij});\mathbf{w}_i) = 0$ in the ordinal case (i.e. $\mathbb{P}(y_{ijk} \geq 0) = 1$ and $\mathbb{P}(y_{ijk} \geq L_k) = 0$), since they were not defined by equation (.0.26). For the Gaussian noise equation, $N_{tot} = \sum_{i=1}^{N} N_i$ is the total number of observations. Data could be missing, with different features being observed or not at various timepoints so the formula above needs to be adjusted by masking out missing timepoints (in this case the N_{tot} would vary depending on k and would be the total of observations in feature k). Notice that we obtain mean squared error (MSE) in the Gaussian noise and a binary crossentropy (BCE) in the Bernoulli noise, which in Machine Learning are the standard losses to deal with continuous and binary data respectively. The next two prior log-likelihood terms correspond to the regularity term:

Population parameters prior

The second term is the log-likelihood of the population parameters' prior. This includes the Gaussian prior for $(\tilde{g},\tilde{v},\tilde{\beta})$:

$$\begin{aligned}
\log\ q(\boldsymbol{z}_{pop};\theta) &= -d\log(\sigma_g\sqrt{2\pi}) - \frac{1}{2\sigma_g^2}||\bar{\boldsymbol{g}} - \tilde{\boldsymbol{g}}||^2 && (.0.32)\\
&\quad -d\log(\sigma_v\sqrt{2\pi}) - \frac{1}{2\sigma_v^2}||\bar{\boldsymbol{v}} - \tilde{\boldsymbol{v}}||^2 && (.0.33)\\
&\quad -(d-1)N_s\log(\sigma_\beta\sqrt{2\pi}) - \frac{1}{2\sigma_\beta^2}\sum_{m=1}^{d-1}\sum_{l=1}^{N_s}(\bar{\beta_{ml}} - \tilde{\beta}_{ml})^2 && (.0.34)
\end{aligned}$$

Individual parameters prior

The last term is the log-likelihood of the individual parameters' prior. We have thus:

$$\begin{aligned}
\log\ q(\boldsymbol{z}_i \mid \boldsymbol{z}_{pop};\theta) &= -\log(\sigma_\tau\sqrt{2\pi}) - \frac{1}{2\sigma_\tau^2}(\bar{\tau} - \tau_i)^2 && (.0.35)\\
&\quad -\log(\sigma_\xi\sqrt{2\pi}) - \frac{1}{2\sigma_\xi^2}\xi_i^2 && (.0.36)\\
&\quad -N_s\log(\sqrt{2\pi}) - \frac{1}{2}||\boldsymbol{s}_i||^2 && (.0.37)
\end{aligned}$$

Gradients

We provide here the formulas for the partial derivatives of the log-likelihood with respect to the different latent parameters. We only compute gradients for latent parameters as the other parameters are either fixed or maximized during the M-step, thus not requiring gradient information during optimization. On the other hand, latent parameter sampling can be improved with gradient. We can also mention the personalization process of the DCM which involves the search of the *maximum a posteriori* for individual latent parameters while all other parameters are fixed. A local optimum can be obtained by an optimization solver, the performance of which is greatly improved with the gradient being provided.

Noise gradient

We first derive the data attachment term of formula (.0.29). The formulas for Gaussian and Bernoulli noise allow us to chain derivatives, so it is easier to first derive the log-likelihood with respect to the model and then derive the model η with respect to the different parameters (where η is the short notation for $\eta_k(\psi_i(t_{ij}); \mathbf{w}_i)$). Since we can rewrite $\log\ q(\boldsymbol{y} \mid \boldsymbol{z}; \theta) = f(\eta(x))$ for any parameter x, we can chain rule the gradient for any list of parameters X:

$$\nabla_X \log\ q(\boldsymbol{y} \mid \boldsymbol{z}; \theta) \quad = \quad \frac{\partial \log\ q(\boldsymbol{y} \mid \boldsymbol{z}; \theta)}{\partial \eta} \nabla_X \eta \tag{.0.38}$$

and for each different noise model we have:

$$\begin{aligned}
\frac{\partial \log\ q(\boldsymbol{y} \mid \boldsymbol{z}; \theta)}{\partial \eta} &= -\sum_k \frac{1}{\sigma_k^2} \sum_{i,j} (\eta - y_{i,j,k}) && \text{[Gaussian noise]} \\
\frac{\partial \log\ q(\boldsymbol{y} \mid \boldsymbol{z}; \theta)}{\partial \eta} &= \sum_{i,j,k} \frac{y_{i,j,k} - \eta}{\eta(1-\eta)} && \text{[Bernoulli noise]}
\end{aligned}$$

For the ordinal noise this is a little different due to the likelihood term depending on $\eta_k(\psi_i^{k,l}(t_{ij}); \mathbf{w}_i)$ and $\eta_k(\psi_i^{k,l+1}(t_{ij}); \mathbf{w}_i)$ for which the simplified notations are η_k^l and η_k^{l+1} respectively. The gradient for any list of parameters X thus becomes:

$$\nabla_X \log\ q(\boldsymbol{y} \mid \boldsymbol{z}; \theta) = \sum_{i,j,k,l} \mathbb{1}(y_{i,j,k} = l) \frac{\nabla_X \eta_k^l - \nabla_X \eta_k^{l+1}}{\eta_k^l - \eta_k^{l+1}} \quad \text{[Ordinal noise]}$$

Model gradients

We list here all the derivatives with respect to the different parameters. Let h denote the term $\tilde{g} - \frac{v_k \psi_i(t_{ij}) + w_{ik}}{p_k(1-p_k)}$ so that equation (.0.12) becomes $\eta = \frac{1}{1+\exp(h)}$. We can chain the gradients for simpler calculus.

$$\frac{\partial \eta}{\partial h} = -\exp(h)\eta^2 = \eta(\eta - 1) \tag{.0.39}$$

and then by computing the derivatives of h w.r.t. the parameters we obtain:

[Individual parameters (random effects)]

$$\frac{\partial \eta}{\partial \tau_i} = \eta(\eta-1)\frac{v_k e^{\xi_i}}{p_k(1-p_k)} \quad (.0.40)$$

$$\frac{\partial \eta}{\partial \xi_i} = \eta(\eta-1)\left(-\frac{v_k \psi_i(t_{ij})}{p_k(1-p_k)}\right) \quad (.0.41)$$

$$\frac{\partial \eta}{\partial s_{il}} = \eta(\eta-1)\left(-\frac{\beta_{kl}}{p_k(1-p_k)}\right) \quad (.0.42)$$

[Population parameters (fixed effects)]

$$\frac{\partial \eta}{\partial \beta_{ml}} = \eta(\eta-1)\left(-\frac{s_{il}\mathcal{B}_{mk}}{p_k(1-p_k)}\right) \quad (.0.43)$$

$$\frac{\partial \eta_l}{\partial \delta_k^m} = \eta(\eta-1)\frac{\mathbb{1}(l>m)v_k}{p_k(1-p_k)} \text{ [Ordinal added parameters]} \quad (.0.44)$$

For $\boldsymbol{v}$ and $\boldsymbol{g}$ computations are trickier. Indeed, the space-shift vector $\boldsymbol{w}_i = \sum_{l=1}^{N_s}\sum_{m=1}^{d-1}\beta_{ml}s_{il}\mathcal{B}_m$ is a complicated function of $\boldsymbol{v}$ and $\boldsymbol{p}$ since the orthonormal basis $\mathcal{B}$ of $Span(\boldsymbol{v})^{\perp}$ is obtained with a Householder method. The basis is orthonormal with the inner product of the Riemannian manifold, which depends on the metric $\boldsymbol{G}$ of the current point $\boldsymbol{p}$: $\langle \boldsymbol{x}, \boldsymbol{y}\rangle_p = \langle \boldsymbol{x}, \boldsymbol{G}(\boldsymbol{p})y\rangle$ where $\langle , \rangle$ is the Euclidean inner product. In the logistic curves case, the metric is diagonal due the product of manifolds and the diagonal elements are: $\boldsymbol{G}(\boldsymbol{p})_{ii} = \frac{1}{p_i^2(1-p_i)^2} = \frac{(1+\exp(\tilde{g}_i))^4}{\exp(\tilde{g}_i)^2}$. The Householder method arbitrarily selects a dimension, which here is selected as the first one without loss of generality, and then computes the orthonormal basis according to the Euclidean inner product with the following steps:

$$\boldsymbol{\omega} = \boldsymbol{G}(\boldsymbol{p})\boldsymbol{v} \quad (.0.45)$$

$$\kappa = \text{sign}(\omega_1)||\boldsymbol{\omega}|| \quad (.0.46)$$

$$\boldsymbol{u} = \boldsymbol{\omega} - \kappa e_1 \quad (.0.47)$$

$$\tilde{\boldsymbol{u}} = \frac{\boldsymbol{u}}{||\boldsymbol{u}||} \quad (.0.48)$$

$$\boldsymbol{B} = \boldsymbol{I}_d - 2\tilde{\boldsymbol{u}}\tilde{\boldsymbol{u}}^T \quad (.0.49)$$

$$\mathcal{B} = \boldsymbol{B}_{2:d} \quad (.0.50)$$

where $\boldsymbol{e}$ is the canonical basis, $\boldsymbol{I}_d$ the identity matrix and $\mathcal{B}$ is the matrix $\boldsymbol{B}$ stripped of the first column. This first column is colinear to $\boldsymbol{G}(\boldsymbol{p})\boldsymbol{v}$, so the other columns are orthogonal to it in the canonical inner product, thus orthogonal to $\boldsymbol{v}$ in the manifold inner product.

Differentiating through this algorithm yields the following gradients:

$$\frac{\partial \eta}{\partial \tilde{g}_{k'}} = \eta(\eta-1) \left(1 - \delta_{kk'}(v_k\psi_i(t_{ij}) + w_{ik})(e^{\tilde{g}_k} - e^{-\tilde{g}_k}) - \frac{1}{p_k(1-p_k)} \sum_{l=1}^{N_s} \sum_{m=1}^{d-1} \beta_{ml} s_{il} \frac{\partial \mathcal{B}_{mk}}{\partial \tilde{g}_{k'}}\right) \quad (.0.51)$$

$$\frac{\partial \eta}{\partial \tilde{v}_{k'}} = \eta(\eta-1) \frac{1}{p_k(1-p_k)} \left(\delta_{kk'} v_k \psi_i(t_{ij}) + \sum_{l=1}^{N_s} \sum_{m=1}^{d-1} \beta_{ml} s_{il} \frac{\partial \mathcal{B}_{mk}}{\partial \tilde{v}_{k'}}\right) \quad (.0.52)$$

$$\frac{\partial \mathcal{B}_{mk}}{\partial \tilde{g}_{k'}} = 2(e^{\tilde{g}_{k'}} + 2 + e^{-\tilde{g}_{k'}})(e^{\tilde{g}_{k'}} - e^{-\tilde{g}_{k'}}) \frac{\partial \mathcal{B}_{mk}}{\partial \tilde{\omega}_{k'}} \quad (.0.53)$$

$$\frac{\partial \mathcal{B}_{mk}}{\partial \tilde{v}_{k'}} = \frac{1}{p_k^2(1-p_k)^2} \frac{\partial \mathcal{B}_{mk}}{\partial \tilde{\omega}_{k'}} \quad (.0.54)$$

$$\frac{\partial \mathcal{B}_{mk}}{\partial \tilde{\omega}_{k'}} = -2(\tilde{u}_m \frac{\partial \tilde{u}_k}{\partial \omega_{k'}} + \tilde{u}_k \frac{\partial \tilde{u}_m}{\partial \omega_{k'}}) \quad (.0.55)$$

$$\frac{\partial \tilde{u}_k}{\partial \omega_{k'}} = \frac{1}{||\boldsymbol{u}||}(\delta_{kk'} - \delta_{k1}\text{sign}(\omega_1)\frac{\omega_k'}{||\boldsymbol{\omega}||}) - \frac{u_k}{||\boldsymbol{u}||^3}\left(u_{k'} + \omega_{k'}(1 - \frac{|\omega_1|}{||\boldsymbol{\omega}||})\right) \quad (.0.56)$$

where δ is the classical Kronecker symbol. Notice that even though η stands for the model value of individual i for dimension k, its derivative w.r.t to population parameters $g_{k'}$ and $v_{k'}$ with $k' \neq k$ is not null, whereas for individual parameters such as $\tau_{i'}, \xi_{i'}$ the gradient is null if $i' \neq i$.

Priors gradient

For each of the individual and population parameters we also have a gradient coming from the prior (the regularity term). As we only have Gaussian priors, for any parameter x:

$$\frac{\partial \log\ q(x;\theta)}{\partial x} = -\frac{x - \bar{x}}{\sigma_x^2} \quad (.0.57)$$

where $\bar{x}$ and σ_x are the mean and standard deviation of the prior of x.

Fisher information

The Fisher information is the variance of the partial derivative of the log-likelihood w.r.t. a given parameter X. In a multidimensional setting we have the Fisher information matrix: $\mathcal{I}(X) = \mathbb{E}\left(\nabla_X \log q(\boldsymbol{y}; X) \nabla_X \log q(\boldsymbol{y}; X)^T \mid X\right)$. The formula requires a distribution for the observations, however in the DCM model we provided the noise model as a distribution for $\boldsymbol{y}$ *given* the time t. We never assumed a distribution for the times t_{ij}. We do not want to specify this distribution as it varies a lot depending on the disease and dataset and integrating over such a distribution is intractable. However we can specify the information *given* the time of the observation t. Then we compute the empirical Fisher information by summing over the observed times t_{ij}. We provide below the formulas for all the latent parameters $\boldsymbol{z} = (\boldsymbol{z}_{pop}, \boldsymbol{z}_i)$ of the model:

$$\mathcal{I}(t,\boldsymbol{z}) = \sum_{k=1}^{d}\left(\frac{1}{\sigma_k^2}\nabla_{\boldsymbol{z}}\eta\nabla_{\boldsymbol{z}}\eta^T\right)+R \quad \text{[Gaussian noise]} \quad (.0.58)$$

$$\mathcal{I}(t,\boldsymbol{z}) = \sum_{k=1}^{d}\left(\frac{1}{\eta(1-\eta)}\nabla_{\boldsymbol{z}}\eta\nabla_{\boldsymbol{z}}\eta^T\right)+R \quad \text{[Bernoulli noise]} \quad (.0.59)$$

$$\mathcal{I}(t,\boldsymbol{z}) = \sum_{k=1}^{d}\sum_{l=1}^{L_k}\left(\frac{(\nabla_{\boldsymbol{z}}\eta_l-\nabla_{\boldsymbol{z}}\eta_{l+1})(\nabla_{\boldsymbol{z}}\eta_l-\nabla_{\boldsymbol{z}}\eta_{l+1})^T}{\eta_l-\eta_{l+1}}\right)+R \quad (.0.60)$$

[Ordinal noise]

$$\mathcal{I}(\boldsymbol{z}) = \sum_{i=1}^{N}\sum_{j=1}^{N_i}\mathcal{I}(t_{ij},\boldsymbol{z}) \quad \text{[Empirical Fisher information]} \quad (.0.61)$$

where R is the information matrix of the priors of $\boldsymbol{z}$, i.e. $R_{a,b}=\frac{(a-\bar{a})(b-\bar{b})}{\sigma_a^2\sigma_b^2}$ for the coefficients corresponding to two latent variables a,b with Gaussian priors.

In the ordinal case, for any individual parameter x, we have $\frac{\partial\eta_l}{\partial x}=\eta_l(\eta_l-1)\frac{\partial h_l}{\partial x}=\eta_l(\eta_l-1)\frac{\partial h}{\partial x}$ which does not depend on the level l anymore, thus simplifying the formula into: $\mathcal{I}(t,\boldsymbol{z}_i)=\sum_{k=1}^{d}\sum_{l=1}^{L_k}\frac{(\eta_l(\eta_l-1)-\eta_{l+1}(\eta_{l+1}-1))^2}{\eta_l-\eta_{l+1}}\nabla_{\boldsymbol{z}_i}h\nabla_{\boldsymbol{z}_i}h^T+R$.

The Fisher information formulas for model parameters θ have not been explicitly computed here as we were mostly interested in latent parameters (cf sections 2.4.1). Note that these model parameters do not appear in the model expression η. In practice this means that the Fisher information for model parameters is not varying with the times of observations t_{ij}, so we cannot perform the same analysis as in section 2.4.1. As for the use of the Fisher information matrix for the natural gradient in Riemannian samplers[35], this is only relevant for sampled parameters, i.e. latent variables. However the Fisher information can still be used to approximate the variance of the estimation of the model parameters, using the fact that $\sqrt{n}(\hat{\theta}-\theta^*)$ converges in distribution towards $\mathcal{N}(0,\mathcal{I}(\theta^*)^{-1})$ as the number of observations n grows to infinity, where $\hat{\theta}$ is the estimator for θ and θ^* is the true parameter.

[35] M. Girolami and B. Calderhead, "Riemann manifold Langevin and Hamiltonian Monte Carlo methods," en, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 2, pp. 123–214, 2011.

Bibliography

[1] C. Abi Nader, N. Ayache, P. Robert, M. Lorenzi, and Alzheimer's Disease Neuroimaging Initiative, "Monotonic Gaussian Process for spatio-temporal disease progression modeling in brain imaging data," eng, *NeuroImage*, vol. 205, p. 116 266, Jan. 2020.

[2] J. Abrevaya, Y.-C. Hsu, and R. P. Lieli, "Estimating Conditional Average Treatment Effects," *Journal of Business & Economic Statistics*, vol. 33, no. 4, pp. 485–505, Oct. 2015.

[3] S. Allassonnière and J. Chevallier, "A new class of stochastic EM algorithms. Escaping local maxima and handling intractable sampling," en, *Computational Statistics & Data Analysis*, vol. 159, p. 107 159, Jul. 2021.

[4] S. Allassonnière and E. Kuhn, "Stochastic algorithm for Bayesian mixture effect template estimation," en, *ESAIM: Probability and Statistics*, vol. 14, pp. 382–408, Jan. 2010.

[5] S. Allassonnière and E. Kuhn, "Convergent stochastic Expectation Maximization algorithm with efficient sampling in high dimension. Application to deformable template model estimation," en, *Computational Statistics & Data Analysis*, vol. 91, pp. 4–19, Nov. 2015.

[6] L. Arrington, S. Ueckert, M. Ahamadi, S. Macha, and M. O. Karlsson, "Performance of longitudinal item response theory models in shortened or partial assessments," en, *Journal of Pharmacokinetics and Pharmacodynamics*, vol. 47, no. 5, pp. 461–471, Oct. 2020.

[7] D. Arthur and S. Vassilvitskii, "K-means++: The Advantages of Careful Seeding," en,

[8] P. C. Austin and A. Laupacis, "A tutorial on methods to estimating clinically and policy-meaningful measures of treatment effects in prospective observational studies: A review," eng, *The International Journal of Biostatistics*, vol. 7, no. 1, p. 6, 2011.

[9] L. A. Beckett, "Maximum Likelihood Estimation in Mallows's Model Using Partially Ranked Data," en, in *Probability Models and Statistical Analyses for Ranking Data*, M. A. Fligner and J. S. Verducci, Eds., ser. Lecture Notes in Statistics, New York, NY: Springer, 1993, pp. 92–107.

[10] A. L. Benabid, "Deep brain stimulation for Parkinson's disease," en, *Current Opinion in Neurobiology*, vol. 13, no. 6, pp. 696–706, Dec. 2003.

[11] D. Bertolini, K. Arnemann, D. Hall, and J. Walsh, "Machine Learning Enables Smaller ALS Clinical Trials (P1-13.003)," en, *Neurology*, vol. 98, no. 18 Supplement, May 2022.

[12] M. Betancourt, *A Conceptual Introduction to Hamiltonian Monte Carlo*, Jul. 2018.

[13] R. M. A. de Bie, C. E. Clarke, A. J. Espay, S. H. Fox, and A. E. Lang, "Initiation of pharmacological therapy in Parkinson's disease: When, why, and how," eng, *The Lancet. Neurology*, vol. 19, no. 5, pp. 452–461, May 2020.

[14] M. Bilgel, J. L. Prince, D. F. Wong, S. M. Resnick, and B. M. Jedynak, "A multivariate nonlinear mixed effects model for longitudinal image analysis: Application to amyloid imaging," en, *NeuroImage*, vol. 134, pp. 658–670, Jul. 2016.

[15] J. F. Bobb, L. Valeri, B. Claus Henn, D. C. Christiani, R. O. Wright, M. Mazumdar, J. J. Godleski, and B. A. Coull, "Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures," eng, *Biostatistics (Oxford, England)*, vol. 16, no. 3, pp. 493–508, Jul. 2015.

[16] A. Bône, "Learning adapted coordinate systems for the statistical analysis of anatomical shapes. Applications to Alzheimer's disease progression modeling," en, Ph.D. dissertation, Sorbonne Université, Jan. 2020.

[17] A. Bône, M. Louis, B. Martin, and S. Durrleman, "Deformetrica 4: An Open-Source Software for Statistical Shape Analysis," Sep. 2018.

[18] G. E. P. Box and D. R. Cox, "An Analysis of Transformations," en,

[19] H. Braak, K. D. Tredici, U. Rüb, R. A. I. de Vos, E. N. H. Jansen Steur, and E. Braak, "Staging of brain pathology related to sporadic Parkinson's disease," en, *Neurobiology of Aging*, vol. 24, no. 2, pp. 197–211, Mar. 2003.

[20] P. Chan and N. Holford, "Drug Treatment Effects on Disease Progression," *Annual Review of Pharmacology and Toxicology*, vol. 41, no. 1, pp. 625–659, 2001.

[21] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud, *Neural Ordinary Differential Equations*, Dec. 2019.

[22] S. Chib and E. Greenberg, "Understanding the Metropolis-Hastings Algorithm," *The American Statistician*, vol. 49, no. 4, pp. 327–335, Nov. 1995.

[23] S. Chib and B. H. Hamilton, "Semiparametric Bayes analysis of longitudinal data treatment models," en, *Journal of Econometrics*, vol. 110, no. 1, pp. 67–89, Sep. 2002.

[24] H. A. Chipman, E. I. George, and R. E. McCulloch, "BART: Bayesian additive regression trees," *The Annals of Applied Statistics*, vol. 4, no. 1, pp. 266–298, Mar. 2010.

[25] R. Cilia, E. Cereda, A. Akpalu, F. S. Sarfo, M. Cham, R. Laryea, V. Obese, K. Oppon, F. Del Sorbo, S. Bonvegna, A. L. Zecchinelli, and G. Pezzoli, "Natural history of motor symptoms in Parkinson's disease and the long-duration response to levodopa," *Brain*, vol. 143, no. 8, pp. 2490–2501, Aug. 2020.

[26] O. o. t. Commissioner, *FDA Converts Novel Alzheimer's Disease Treatment to Traditional Approval*, en, Jul. 2023.

[27] S. F. Cook and R. R. Bies, "Disease Progression Modeling: Key Concepts and Recent Developments," en, *Current Pharmacology Reports*, vol. 2, no. 5, pp. 221–230, Oct. 2016.

[28] R. Couronné, "Progression models for Parkinson's Disease," en, Ph.D. dissertation, Sorbonne Université, Sep. 2021.

[29] R. Couronné, A. Valladier, M. Vidhaillet, J. C. Corvol, S. Lehéricy, and S. Durrleman, "Modeling the progression of Parkinson's Disease: Comparison of subjects with and without Sleep Disorders," en, p. 4,

[30] R. Couronné, M. Vidailhet, J.-C. Corvol, S. Lehéricy, and S. Durrleman, "Learning disease progression models with longitudinal data and missing values," in *ISBI 2019 - International Symposium on Biomedical Imaging*, Venice, Italy, Apr. 2019.

[31] A. Cremaschi, M. De Iorio, Y. Seng Chong, B. Broekman, M. J. Meaney, and M. Z. L. Kee, "A Bayesian nonparametric approach to dynamic item-response modeling: An application to the GUSTO cohort study," en, *Statistics in Medicine*, vol. 40, no. 27, pp. 6021–6037, 2021.

[32] T. Dao and C. D. Sa, "Gaussian Quadrature for Kernel Features," en,

[33] V. Debavelaere, S. Durrleman, S. Allassonnière, and for the Alzheimer's Disease Neuroimaging Initiative, "Learning the Clustering of Longitudinal Shape Data Sets into a Mixture of Independent or Branching Trajectories," en, *International Journal of Computer Vision*, vol. 128, no. 12, pp. 2794–2809, Dec. 2020.

[34] M. Delattre, M. Lavielle, and M.-A. Poursat, "A note on BIC in mixed-effects models," en, *Electronic Journal of Statistics*, vol. 8, no. 1, Jan. 2014.

[35] B. Delyon, M. Lavielle, and E. Moulines, "Convergence of a Stochastic Approximation Version of the EM Algorithm," *The Annals of Statistics*, vol. 27, no. 1, pp. 94–128, 1999.

[36] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data Via the EM Algorithm," en, *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[37] T. G. Dietterich, "Ensemble Methods in Machine Learning," en, in *Multiple Classifier Systems*, ser. Lecture Notes in Computer Science, Berlin, Heidelberg: Springer, 2000, pp. 1–15.

[38] M. Do Carmo, *Riemannian Geometry*, en.

[39] M. C. Donohue, H. Jacqmin-Gadda, M. Le Goff, R. G. Thomas, R. Raman, A. C. Gamst, L. A. Beckett, C. R. Jack, M. W. Weiner, J.-F. Dartigues, and P. S. Aisen, "Estimating long-term multivariate progression from short-term data," *Alzheimer's & dementia : the journal of the Alzheimer's Association*, vol. 10, no. 0, S400–S410, Oct. 2014.

[40] J. Dumurgier and C. Tzourio, "Epidemiology of neurological diseases in older adults," en, *Revue Neurologique*, vol. 176, no. 9, pp. 642–648, Nov. 2020.

[41] S. Durrleman, X. Pennec, A. Trouvé, J. Braga, G. Gerig, and N. Ayache, "Toward a Comprehensive Framework for the Spatiotemporal Statistical Analysis of Longitudinal Shape Data," en, *International Journal of Computer Vision*, vol. 103, no. 1, pp. 22–59, May 2013.

[42] R. S. Eisinger, C. W. Hess, D. Martinez-Ramirez, L. Almeida, K. D. Foote, M. S. Okun, and A. Gunduz, "Motor subtype changes in early Parkinson's disease," en, *Parkinsonism & Related Disorders*, vol. 43, pp. 67–72, Oct. 2017.

[43] A. Elbaz, L. Carcaillon, S. Kab, and F. Moisan, "Epidemiology of Parkinson's disease," en, *Revue Neurologique*, Neuroepidemiology, vol. 172, no. 1, pp. 14–26, Jan. 2016.

[44] S. E. Embretson and S. P. Reise, *Item response theory for psychologists* (Item response theory for psychologists). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers, 2000.

[45] A. Espay, A. Fasano, B. van Nuenen, M. Payne, A. Snijders, and B. Bloem, ""On" state freezing of gait in Parkinson disease," *Neurology*, vol. 78, no. 7, pp. 454–457, Feb. 2012.

[46] D. Ferreira, A. Nordberg, and E. Westman, "Biological subtypes of Alzheimer disease: A systematic review and meta-analysis," en, *Neurology*, vol. 94, no. 10, pp. 436–448, Mar. 2020.

[47] C. R. Fields, N. Bengoa-Vergniory, and R. Wade-Martins, "Targeting Alpha-Synuclein as a Therapy for Parkinson's Disease," *Frontiers in Molecular Neuroscience*, vol. 12, 2019.

[48] B. Fischl, D. H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. van der Kouwe, R. Killiany, D. Kennedy, S. Klaveness, A. Montillo, N. Makris, B. Rosen, and A. M. Dale, "Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain," eng, *Neuron*, vol. 33, no. 3, pp. 341–355, Jan. 2002.

[49] R. A. Fisher, "XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance.," en, *Earth and Environmental Science Transactions of The Royal Society of Edinburgh*, vol. 52, no. 2, pp. 399–433, Jan. 1919.

[50] T. Fletcher, "Geodesic Regression on Riemannian Manifolds," en, Sep. 2011, p. 75.

[51] H. M. Fonteijn, M. Modat, M. J. Clarkson, J. Barnes, M. Lehmann, N. Z. Hobbs, R. I. Scahill, S. J. Tabrizi, S. Ourselin, N. C. Fox, and D. C. Alexander, "An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease," eng, *NeuroImage*, vol. 60, no. 3, pp. 1880–1889, Apr. 2012.

[52] N. Fournier and S. Durrleman, "Covariate-Aware Longitudinal Modelling for Neurodegenerative Diseases," en,

[53] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, "Bayesian Data Analysis," en, Chapman and Hall/CRC, Nov. 2013.

[54] A. Gelman and D. B. Rubin, "Inference from Iterative Simulation Using Multiple Sequences," *Statistical Science*, vol. 7, no. 4, pp. 457–472, Nov. 1992.

[55] M. Gilat, N. D'Cruz, P. Ginis, W. Vandenberghe, and A. Nieuwboer, "Freezing of gait and levodopa," English, *The Lancet Neurology*, vol. 20, no. 7, pp. 505–506, Jul. 2021.

[56] M. Girolami and B. Calderhead, "Riemann manifold Langevin and Hamiltonian Monte Carlo methods," en, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 2, pp. 123–214, 2011.

[57] C. G. Goetz, B. C. Tilley, S. R. Shaftman, G. T. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M. B. Stern, R. Dodel, B. Dubois, R. Holloway, J. Jankovic, J. Kulisevsky, A. E. Lang, A. Lees, S. Leurgans, P. A. LeWitt, D. Nyenhuis, C. W. Olanow, O. Rascol, A. Schrag, J. A. Teresi, J. J. van Hilten, and N. LaPelle, "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results: MDS-UPDRS: Clinimetric Assessment," en, *Movement Disorders*, vol. 23, no. 15, pp. 2129–2170, Nov. 2008.

[58] R. Gorter, J.-P. Fox, and J. W. R. Twisk, "Why item response theory should be used for longitudinal questionnaire data analysis in medical research," *BMC Medical Research Methodology*, vol. 15, no. 1, p. 55, Jul. 2015.

[59] G. Gottipati, A. Berges, S. Yang, C. Chen, M. Karlsson, and E. Plan, "Item Response Model Adaptation for Analyzing Data from Different Versions of Parkinson's Disease Rating Scales," *Pharmaceutical Research*, vol. 36, Jul. 2019.

[60] D. P. Green and H. L. Kern, "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees," *The Public Opinion Quarterly*, vol. 76, no. 3, pp. 491–511, 2012.

[61] S. Gruffaz, P.-E. Poulet, E. Maheux, B. Jedynak, and S. DURRLEMAN, "Learning Riemannian metric for disease progression modeling," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, Eds., vol. 34, Curran Associates, Inc., 2021, pp. 23 780–23 792.

[62] H. Haario, E. Saksman, and J. Tamminen, "An Adaptive Metropolis Algorithm," *Bernoulli*, vol. 7, no. 2, pp. 223–242, 2001.

[63] O. Hansson, J. Seibyl, E. Stomrud, H. Zetterberg, J. Q. Trojanowski, T. Bittner, V. Lifke, V. Corradini, U. Eichenlaub, R. Batrla, K. Buck, K. Zink, C. Rabe, K. Blennow, L. M. Shaw, Swedish BioFINDER study group, and Alzheimer's Disease Neuroimaging Initiative, "CSF biomarkers of Alzheimer's disease concord with amyloid- PET and predict clinical progression: A study of fully automated immunoassays in BioFINDER and ADNI cohorts," eng, *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, vol. 14, no. 11, pp. 1470–1481, Nov. 2018.

[64] J. A. Hardy and G. A. Higgins, "Alzheimer's disease: The amyloid cascade hypothesis," eng, *Science (New York, N.Y.)*, vol. 256, no. 5054, pp. 184–185, Apr. 1992.

[65] R. Harpaz and R. Haralick, "TR-2006001: The EM Algorithm As a Lower Bound Optimization Technique," *Computer Science Technical Reports*, Jan. 2006.

[66] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, Apr. 1970.

[67] S. r. Hauberg, O. Freifeld, and M. Black, "A Geometric take on Metric Learning," in *Advances in Neural Information Processing Systems*, vol. 25, Curran Associates, Inc., 2012.

[68] T. R. T. Have, A. R. Kunselman, E. P. Pulkstenis, and J. R. Landis, "Mixed Effects Logistic Regression Models for Longitudinal Binary Response Data with Informative Drop-Out," *Biometrics*, vol. 54, no. 1, pp. 367–383, 1998.

[69] B. He and S. Luo, "Joint modeling of multivariate longitudinal measurements and survival data with applications to Parkinson's disease," eng, *Statistical Methods in Medical Research*, vol. 25, no. 4, pp. 1346–1358, Aug. 2016.

[70] M. M. Hoehn and M. D. Yahr, "Parkinsonism: Onset, progression, and mortality," en, *Neurology*, vol. 17, no. 5, pp. 427–427, May 1967.

[71] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," *The Annals of Statistics*, vol. 36, no. 3, pp. 1171–1220, Jun. 2008.

[72] M. D. Homan and A. Gelman, "The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1593–1623, Jan. 2014.

[73] J. Horsager, K. B. Andersen, K. Knudsen, C. Skjærbæk, T. D. Fedorova, N. Okkels, E. Schaeffer, S. K. Bonkat, J. Geday, M. Otto, M. Sommerauer, E. H. Danielsen, E. Bech, J. Kraft, O. L. Munk, S. D. Hansen, N. Pavese, R. Göder, D. J. Brooks, D. Berg, and P. Borghammer, "Brain-first versus body-first Parkinson's disease: A multimodal imaging case-control study," *Brain*, vol. 143, no. 10, pp. 3077–3088, Oct. 2020.

[74] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," en, *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, Jun. 2000.

[75] J. G. Ibrahim, H. Chu, and L. M. Chen, "Basic Concepts and Methods for Joint Models of Longitudinal and Survival Data," *Journal of Clinical Oncology*, vol. 28, no. 16, pp. 2796–2801, Jun. 2010.

[76] J. C. Immekus, K. E. Snyder, and P. A. Ralston, "Multidimensional Item Response Theory for Factor Structure Assessment in Educational Psychology Research," *Frontiers in Education*, vol. 4, p. 45, 2019.

[77] Y. Iturria-Medina, F. M. Carbonell, R. C. Sotero, F. Chouinard-Decorte, and A. C. Evans, "Multifactorial causal model of brain (dis)organization and therapeutic intervention: Application to Alzheimer's disease," en, *NeuroImage*, vol. 152, pp. 60–77, May 2017.

[78] C. R. Jack, D. A. Bennett, K. Blennow, M. C. Carrillo, H. H. Feldman, G. B. Frisoni, H. Hampel, W. J. Jagust, K. A. Johnson, D. S. Knopman, R. C. Petersen, P. Scheltens, R. A. Sperling, and B. Dubois, "A/T/N: An unbiased descriptive classification scheme for Alzheimer disease biomarkers," eng, *Neurology*, vol. 87, no. 5, pp. 539–547, Aug. 2016.

[79] C. R. Jack, D. S. Knopman, W. J. Jagust, R. C. Petersen, M. W. Weiner, P. S. Aisen, L. M. Shaw, P. Vemuri, H. J. Wiste, S. D. Weigand, T. G. Lesnick, V. S. Pankratz, M. C. Donohue, and J. Q. Trojanowski, "Update on hypothetical model of Alzheimer's disease biomarkers," *Lancet neurology*, vol. 12, no. 2, pp. 207–216, Feb. 2013.

[80] C. R. Jack, D. S. Knopman, W. J. Jagust, L. M. Shaw, P. S. Aisen, M. W. Weiner, R. C. Petersen, and J. Q. Trojanowski, "Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade," English, *The Lancet Neurology*, vol. 9, no. 1, pp. 119–128, Jan. 2010.

[81] J.-H. Jacobsen, A. Smeulders, and E. Oyallon, *I-RevNet: Deep Invertible Networks*, Feb. 2018.

[82] J. Jankovic, M. McDermott, J. Carter, S. Gauthier, C. Goetz, L. Golbe, S. Huber, W. Koller, C. Olanow, and I. Shoulson, "Variable expression of Parkinson's disease: A base-line analysis of the DATATOP cohort. The Parkinson Study Group," eng, *Neurology*, vol. 40, no. 10, pp. 1529–1534, Oct. 1990.

[83] B. M. Jedynak, A. Lang, B. Liu, E. Katz, Y. Zhang, B. T. Wyman, D. Raunig, C. P. Jedynak, B. Caffo, and J. L. Prince, "A Computational Neurodegenerative Disease Progression Score: Method and Results with the Alzheimer's Disease Neuroimaging Initiative Cohort," *NeuroImage*, vol. 63, no. 3, pp. 1478–1486, Nov. 2012.

[84] S. Joshi and M. Miller, "Landmark matching via large deformation diffeomorphisms," *IEEE Transactions on Image Processing*, vol. 9, no. 8, pp. 1357–1370, Aug. 2000.

[85] W. Jung, A. W. Mulyadi, and H.-I. Suk, "Unified Modeling of Imputation, Forecasting, and Prediction for AD Progression," en, in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2019, pp. 168–176.

[86] Y. Jung and I. Lee, "Optimal design of experiments for optimization-based model calibration using Fisher information matrix," en, *Reliability Engineering & System Safety*, vol. 216, p. 107 968, Dec. 2021.

[87] M. Kanagawa, P. Hennig, D. Sejdinovic, and B. K. Sriperumbudur, *Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences*, Jul. 2018.

[88] C. Kartsonaki, "Survival analysis," en, *Diagnostic Histopathology*, Mini-Symposium: Medical Statistics, vol. 22, no. 7, pp. 263–270, Jul. 2016.

[89] V. Kmetzsch, E. Becker, D. Saracino, D. Rinaldi, A. Camuzat, I. Le Ber, O. Colliot, and f. t. P.-D. s. group for the, "Disease Progression Score Estimation From Multimodal Imaging and MicroRNA Data Using Supervised Variational Autoencoders," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 12, pp. 6024–6035, Dec. 2022.

[90] I. Koval, "Learning Multimodal Digital Models of Disease Progression from Longitudinal Data : Methods & Algorithms for the Description, Prediction and Simulation of Alzheimer's Disease Progression," en, Ph.D. dissertation, Institut Polytechnique de Paris, Jan. 2020.

[91] I. Koval, A. Bône, M. Louis, T. Lartigue, S. Bottani, A. Marcoux, J. Samper-González, N. Burgos, B. Charlier, A. Bertrand, S. Epelbaum, O. Colliot, S. Allassonnière, and S. Durrleman, "AD Course Map charts Alzheimer's disease progression," eng, *Scientific Reports*, vol. 11, no. 1, p. 8020, Apr. 2021.

[92] I. Koval, T. Dighiero-Brecht, A. J. Tobin, S. J. Tabrizi, R. I. Scahill, S. Tezenas du Montcel, S. Durrleman, and A. Durr, "Forecasting individual progression trajectories in Huntington disease enables more powered clinical trials," en, *Scientific Reports*, vol. 12, no. 1, p. 18 928, Nov. 2022.

[93] J. K. Kueper, M. Speechley, and M. Montero-Odasso, "The Alzheimer's Disease Assessment Scale–Cognitive Subscale (ADAS-Cog): Modifications and Responsiveness in Pre-Dementia Populations. A Narrative Review," *Journal of Alzheimer's Disease*, vol. 63, no. 2, pp. 423–444,

[94] E. Kuhn and M. Lavielle, "Coupling a stochastic approximation version of EM with an MCMC procedure," fr, *ESAIM: Probability and Statistics*, vol. 8, pp. 115–131, 2004.

[95] K. Lahouel, M. Wells, V. Rielly, E. Lew, D. Lovitz, and B. M. Jedynak, *Learning nonparametric ordinary differential equations from noisy data*, Feb. 2023.

[96] N. M. Laird and J. H. Ware, "Random-effects models for longitudinal data," eng, *Biometrics*, vol. 38, no. 4, pp. 963–974, Dec. 1982.

[97] M. Lavielle and C. Mbogning, "An improved SAEM algorithm for maximum likelihood estimation in mixtures of non linear mixed effects models," *Statistics and Computing*, vol. 24, no. 5, pp. 693–707, Sep. 2014.

[98] G. Lebanon, *Learning Riemannian Metrics*, Oct. 2012.

[99] M. J. Lindstrom and D. M. Bates, "Newton—Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data," *Journal of the American Statistical Association*, vol. 83, no. 404, pp. 1014–1022, Dec. 1988.

[100] F. Liu, X. Huang, Y. Chen, and J. A. K. Suykens, "Random Features for Kernel Approximation: A Survey on Algorithms, Theory, and Beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 7128–7148, Oct. 2022.

[101] L. C. Liu and D. Hedeker, "A Mixed-Effects Regression Model for Longitudinal Multivariate Ordinal Data," en, *Biometrics*, vol. 62, no. 1, pp. 261–268, 2006.

[102] G. Livingston, J. Huntley, A. Sommerlad, D. Ames, C. Ballard, S. Banerjee, C. Brayne, A. Burns, J. Cohen-Mansfield, C. Cooper, S. G. Costafreda, A. Dias, N. Fox, L. N. Gitlin, R. Howard, H. C. Kales, M. Kivimäki, E. B. Larson, A. Ogunniyi, V. Orgeta, K. Ritchie, K. Rockwood, E. L. Sampson, Q. Samus, L. S. Schneider, G. Selbæk, L. Teri, and N. Mukadam, "Dementia prevention, intervention, and care: 2020 report of the Lancet Commission," English, *The Lancet*, vol. 396, no. 10248, pp. 413–446, Aug. 2020.

[103] F. M. Lord, *Applications of Item Response Theory To Practical Testing Problems*. New York: Routledge, Jul. 1980.

[104] M. Lorenzi, M. Filippone, G. B. Frisoni, D. C. Alexander, and S. Ourselin, "Probabilistic disease progression modeling to characterize diagnostic uncertainty: Application to staging and prediction in Alzheimer's disease," en, *NeuroImage*, Mapping diseased brains, vol. 190, pp. 56–68, Apr. 2019.

[105] M. Lorenzi, X. Pennec, G. B. Frisoni, and N. Ayache, "Disentangling normal aging from Alzheimer's disease in structural magnetic resonance images," en, *Neurobiology of Aging*, Novel Imaging Biomarkers for Alzheimer's Disease and Related Disorders (NIBAD), vol. 36, S42–S52, Jan. 2015.

[106] M. Louis, "Computational and statistical methods for trajectory analysis in a Riemannian geometry setting," en, Ph.D. dissertation, Sorbonne Université, Oct. 2019.

[107] C. Lungu, J. M. Cedarbaum, T. M. Dawson, E. R. Dorsey, C. Faraco, H. J. Federoff, B. Fiske, R. Fox, A. M. Goldfine, K. Kieburtz, E. A. Macklin, H. Matthews, G. Rafaloff, R. Saunders-Pullman, N. F. Schor, M. A. Schwarzschild, B.-A. Sieber, T. Simuni, D. J. Surmeier, A. Tamiz, M. H. Werner, C. B. Wright, and R. Wyse, "Seeking progress in disease modification in Parkinson disease," en, *Parkinsonism & Related Disorders*, vol. 90, pp. 134–141, Sep. 2021.

[108] S. Luo, H. Zou, G. T. Stebbins, M. A. Schwarzschild, E. A. Macklin, J. Chan, D. Oakes, T. Simuni, C. G. Goetz, and P. S. G. S.-P. Investigators, "Dissecting the Domains of Parkinson's Disease: Insights from Longitudinal Item Response Theory Modeling," en, *Movement Disorders*, vol. 37, no. 9, pp. 1904–1914, 2022.

[109] A. Ly, M. Marsman, J. Verhagen, R. P. P. P. Grasman, and E.-J. Wagenmakers, "A Tutorial on Fisher information," en, *Journal of Mathematical Psychology*, vol. 80, pp. 40–55, Oct. 2017.

[110] L. v. d. Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.

[111] E. Maheux, J. Ortholand, C. Birkenbihl, E. Thibeau-Sutre, M. Sood, D. Archetti, V. Bouteloup, I. Koval, and S. Durrleman, "Forecast Alzheimer's disease progression to better select patients for clinical trials," en, Jul. 2021.

[112] K. Marek, D. Jennings, S. Lasch, *et al.*, "The Parkinson Progression Marker Initiative (PPMI)," en, *Progress in Neurobiology*, Biological Markers for Neurodegenerative Diseases, vol. 95, no. 4, pp. 629–635, Dec. 2011.

[113] R. V. Marinescu, A. Eshaghi, M. Lorenzi, A. L. Young, N. P. Oxtoby, S. Garbarino, S. J. Crutch, and D. C. Alexander, "DIVE: A spatiotemporal progression model of brain pathology in neurodegenerative disorders," en, *NeuroImage*, vol. 192, pp. 166–177, May 2019.

[114] R. V. Marinescu, N. P. Oxtoby, A. L. Young, E. E. Bron, A. W. Toga, M. W. Weiner, F. Barkhof, N. C. Fox, P. Golland, S. Klein, and D. C. Alexander, "TADPOLE Challenge: Accurate Alzheimer's disease prediction through crowdsourced forecasting of future data," eng, *PRedictive Intelligence in MEdicine. PRIME (Workshop)*, vol. 11843, pp. 1–10, Oct. 2019.

[115] C. Marras, K. R. Chaudhuri, N. Titova, and T. A. Mestre, "Therapy of Parkinson's Disease Subtypes," en, *Neurotherapeutics*, vol. 17, no. 4, pp. 1366–1377, Oct. 2020.

[116] P. Martinez-Martin, C. Rodriguez-Blazquez, M. Alvarez-Sanchez, T. Arakaki, A. Bergareche-Yarza, A. Chade, N. Garretto, O. Gershanik, M. M. Kurtis, J. C. Martinez-Castrillo, A. Mendoza-Rodriguez, H. P. Moore, M. Rodriguez-Violante, C. Singer, B. C. Tilley, J. Huang, G. T. Stebbins, and C. G. Goetz, "Expanded and independent validation of the Movement Disorder Society–Unified Parkinson's Disease Rating Scale (MDS-UPDRS)," en, *Journal of Neurology*, vol. 260, no. 1, pp. 228–236, Jan. 2013.

[117] J. M. te Marvelde, C. A. W. Glas, G. Van Landeghem, and J. Van Damme, "Application of Multidimensional Item Response Theory Models to Longitudinal Data," en, *Educational and Psychological Measurement*, vol. 66, no. 1, pp. 5–34, Feb. 2006.

[118] P. McCullagh, *Generalized Linear Models*, en. Routledge, Oct. 2018.

[119] I. G. McKeith, "Spectrum of Parkinson's disease, Parkinson's dementia, and Lewy body dementia," eng, *Neurologic clinics*, vol. 18, no. 4, pp. 865–902, Nov. 2000.

[120] P. J. Moore, T. J. Lyons, J. Gallacher, and f. t. A. D. N. Initiative, "Random forest prediction of Alzheimer's disease using pairwise selection from time series data," en, *PLOS ONE*, vol. 14, no. 2, e0211558, Feb. 2019.

[121] P. Moulaire, P. E. Poulet, E. Petit, T. Klockgether, A. Durr, T. Ashisawa, S. Tezenas du Montcel, and f. t. R. Consortium, "Temporal Dynamics of the Scale for the Assessment and Rating of Ataxia in Spinocerebellar Ataxias," en, *Movement Disorders*, vol. 38, no. 1, pp. 35–44, 2023.

[122] C. Murphy, "Olfactory and other sensory impairments in Alzheimer disease," en, *Nature Reviews Neurology*, vol. 15, no. 1, pp. 11–24, Jan. 2019.

[123] M. E. Murray, N. R. Graff-Radford, O. A. Ross, R. C. Petersen, R. Duara, and D. W. Dickson, "Neuropathologically defined subtypes of Alzheimer's disease with distinct clinical characteristics: A retrospective study," en, *The Lancet Neurology*, vol. 10, no. 9, pp. 785–796, Sep. 2011.

[124] T. Nedelec, B. Couvy-Duchesne, F. Monnet, T. Daly, M. Ansart, L. Gantzer, B. Lekens, S. Epelbaum, C. Dufouil, and S. Durrleman, "Identifying health conditions associated with Alzheimer's disease up to 15 years before diagnosis: An agnostic study of French and British health records," English, *The Lancet Digital Health*, vol. 4, no. 3, e169–e178, Mar. 2022.

[125] M. Nguyen, T. He, L. An, D. C. Alexander, J. Feng, and B. T. T. Yeo, "Predicting Alzheimer's disease progression using deep recurrent neural networks," en, *NeuroImage*, vol. 222, p. 117 203, Nov. 2020.

[126] S. S. O'Sullivan, A. H. Evans, and A. J. Lees, "Dopamine dysregulation syndrome: An overview of its epidemiology, mechanisms and management," eng, *CNS drugs*, vol. 23, no. 2, pp. 157–170, 2009.

[127] N. P. Oxtoby, A. L. Young, D. M. Cash, T. L. S. Benzinger, A. M. Fagan, J. C. Morris, R. J. Bateman, N. C. Fox, J. M. Schott, and D. C. Alexander, "Data-driven models of dominantly-inherited Alzheimer's disease progression," *Brain*, vol. 141, no. 5, pp. 1529–1544, May 2018.

[128] K. V. Papp, D. M. Rentz, I. Orlovsky, R. A. Sperling, and E. C. Mormino, "Optimizing the preclinical Alzheimer's cognitive composite with semantic processing: The PACC5," *Alzheimer's & Dementia : Translational Research & Clinical Interventions*, vol. 3, no. 4, pp. 668–677, Nov. 2017.

[129] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, and D. Cournapeau, "Scikit-learn: Machine Learning in Python," en, *MACHINE LEARNING IN PYTHON*,

[130] R. C. Petersen, P. S. Aisen, L. A. Beckett, M. C. Donohue, A. C. Gamst, D. J. Harvey, C. R. Jack, W. J. Jagust, L. M. Shaw, A. W. Toga, J. Q. Trojanowski, and M. W. Weiner, "Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization," en, *Neurology*, vol. 74, no. 3, pp. 201–209, Jan. 2010.

[131] R. Pinder, "The Honeymoon Period—And After," en, in *The Management of Chronic Illness: Patient and Doctor Perspectives on Parkinson's Disease*, R. Pinder, Ed., London: Macmillan Education UK, 1990, pp. 54–67.

[132] R. B. Postuma, A. E. Lang, J. F. Gagnon, A. Pelletier, and J. Y. Montplaisir, "How does parkinsonism start? Prodromal parkinsonism motor changes in idiopathic REM sleep behaviour disorder," *Brain*, vol. 135, no. 6, pp. 1860–1870, Jun. 2012.

[133] P.-E. Poulet and S. Durrleman, "Mixture Modeling for Identifying Subtypes in Disease Course Mapping," en, in *Information Processing in Medical Imaging*, A. Feragen, S. Sommer, J. Schnabel, and M. Nielsen, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2021, pp. 571–582.

[134] P.-E. Poulet and S. Durrleman, "Multivariate disease progression modeling with longitudinal ordinal data," en, *Statistics in Medicine*, vol. n/a, no. n/a,

[135] *Projections de population à l'horizon 2060 - Insee Première - 1320*.

[136] C. Proust-Lima, V. Philipps, and B. Liquet, "Estimation of Extended Mixed Models Using Latent Classes and Latent Processes: The R Package lcmm," en, *Journal of Statistical Software*, vol. 78, pp. 1–56, Jun. 2017.

[137] C. Proust-Lima, V. Philipps, B. Perrot, M. Blanchin, and V. Sébille, "Modeling repeated self-reported outcome data: A continuous-time longitudinal Item Response Theory model," en, *Methods*, vol. 204, pp. 386–395, Aug. 2022.

[138] M. Quiroga, P. G. Garay, J. M. Alonso, J. M. Loyola, and O. A. Martin, *Bayesian additive regression trees for probabilistic programming*, Oct. 2022.

[139] L. L. Raket, "Statistical Disease Progression Modeling in Alzheimer Disease," *Frontiers in Big Data*, vol. 3, 2020.

[140] L. L. Raket, "Progression models for repeated measures: Estimating novel treatment effects in progressive diseases," en, *Statistics in Medicine*, vol. 41, no. 28, pp. 5537–5557, 2022.

[141] M. A. Rana, A. Li, D. Fox, B. Boots, F. Ramos, and N. Ratliff, "Euclideanizing Flows: Diffeomorphic Reduction for Learning Stable Dynamical Systems," en, in *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, PMLR, Jul. 2020, pp. 630–639.

[142] D. Ravi, D. C. Alexander, and N. P. Oxtoby, "Degenerative Adversarial NeuroImage Nets: Generating Images that Mimic Disease Progression," en, in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2019, pp. 164–172.

[143] S. L. Risacher, W. H. Anderson, A. Charil, P. F. Castelluccio, S. Shcherbinin, A. J. Saykin, A. J. Schwarz, and For the Alzheimer's Disease Neuroimaging Initiative, "Alzheimer disease brain atrophy subtypes are associated with cognition and rate of decline," en, *Neurology*, vol. 89, no. 21, pp. 2176–2186, Nov. 2017.

[144] H. Robbins and S. Monro, "A Stochastic Approximation Method," *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, Sep. 1951.

[145] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck, "Probabilistic programming in Python using PyMC3," en, *PeerJ Computer Science*, vol. 2, e55, Apr. 2016.

[146] F. Samejima, "Estimation of latent ability using a response pattern of graded scores," en, *Psychometrika*, vol. 34, no. 1, pp. 1–97, Mar. 1969.

[147] T. Saulnier, V. Philipps, W. G. Meissner, O. Rascol, A. Pavy-Le Traon, A. Foubert-Samier, and C. Proust-Lima, "Joint models for the longitudinal analysis of measurement scales in the presence of informative dropout," en, *Methods*, vol. 203, pp. 142–151, Jul. 2022.

[148] B. Sauty and S. Durrleman, "Progression Models for Imaging Data with Longitudinal Variational Auto Encoders," en, in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds., ser. Lecture Notes in Computer Science, Cham: Springer Nature Switzerland, 2022, pp. 3–13.

[149] B. Sauty and S. Durrleman, "Riemannian Metric Learning for Progression Modeling of Longitudinal Datasets," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, Mar. 2022, pp. 1–5.

[150] B. Sauty and S. Durrleman, "Impact of sex and APOE-4 genotype on patterns of regional brain atrophy in Alzheimer's disease and healthy aging," *Frontiers in Neurology*, vol. 14, 2023.

[151] J.-B. Schiratti, S. Allassonnière, A. Routier, O. Colliot, and S. Durrleman, "A Mixed-Effects Model with Time Reparametrization for Longitudinal Univariate Manifold-Valued Data," en, in *Information Processing in Medical Imaging*, S. Ourselin, D. C. Alexander, C.-F. Westin, and M. J. Cardoso, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2015, pp. 564–575.

[152] J.-B. Schiratti, "Methods and algorithms to learn spatio-temporal changes from longitudinal manifold-valued observations," en, Ph.D. dissertation, Université Paris Saclay (COmUE), Jan. 2017.

[153] J.-B. Schiratti, S. Allassonnière, O. Colliot, and S. Durrleman, "A Bayesian mixed-effects model to learn trajectories of changes from repeated manifold-valued observations," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 4840–4872, Jan. 2017.

[154] J. Serroyen, G. Molenberghs, G. Verbeke, and M. Davidian, "Nonlinear Models for Longitudinal Data," *The American Statistician*, vol. 63, no. 4, pp. 378–388, Nov. 2009.

[155] C. Shand, P. J. Markiewicz, D. M. Cash, D. C. Alexander, M. C. Donohue, F. Barkhof, N. P. Oxtoby, and Alzheimer's Disease Neuroimaging Initiative, "Heterogeneity in Preclinical Alzheimer's Disease Trial Cohort Identified by Image-based Data-Driven Disease Progression Modelling," eng, *medRxiv: The Preprint Server for Health Sciences*, p. 2023.02.07.23285572, Feb. 2023.

[156] A. Sharma, V. Goyal, M. Behari, A. Srivastva, G. Shukla, and D. Vibha, "Impulse control disorders and related behaviours (ICD-RBs) in Parkinson's disease patients: Assessment using "Questionnaire for impulsive-compulsive disorders in Parkinson's disease" (QUIP)," *Annals of Indian Academy of Neurology*, 2015.

[157] S. L. Silverman, "From Randomized Controlled Trials to Observational Studies," en, *The American Journal of Medicine*, vol. 122, no. 2, pp. 114–120, Feb. 2009.

[158] T. Simuni, C. Caspell-Garcia, C. Coffey, S. Lasch, C. Tanner, and K. Marek, "How stable are Parkinson's disease subtypes in de novo patients: Analysis of the PPMI cohort?" en, *Parkinsonism & Related Disorders*, vol. 28, pp. 62–67, Jul. 2016.

[159] G. T. Stebbins, C. G. Goetz, D. J. Burn, J. Jankovic, T. K. Khoo, and B. C. Tilley, "How to identify tremor dominant and postural instability/gait difficulty groups with the movement disorder society unified Parkinson's disease rating scale: Comparison with the unified Parkinson's disease rating scale," en, *Movement Disorders*, vol. 28, no. 5, pp. 668–670, 2013.

[160] R. J. Swartz and S. W. Choi, "A Burdened CAT: Incorporating Response Burden With Maximum Fisher's Information for Item Selection," en,

[161] B. O. Taddé, H. Jacqmin-Gadda, J.-F. Dartigues, D. Commenges, and C. Proust-Lima, "Dynamic modeling of multivariate dimensions and their temporal relationships using latent processes: Application to Alzheimer's disease," en, *Biometrics*, vol. 76, no. 3, pp. 886–899, 2020.

[162] A. Tam, C. Laurent, S. Gauthier, and C. Dansereau, "Prediction of Cognitive Decline for Enrichment of Alzheimer's Disease Clinical Trials," *The Journal Of Prevention of Alzheimer's Disease*, 2022.

[163] C. E. Teunissen, I. M. W. Verberk, E. H. Thijssen, L. Vermunt, O. Hansson, H. Zetterberg, W. M. van der Flier, M. M. Mielke, and M. del Campo, "Blood-based biomarkers for Alzheimer's disease: Towards clinical implementation," en, *The Lancet Neurology*, vol. 21, no. 1, pp. 66–77, Jan. 2022.

[164] M. H. Tosin, G. T. Stebbins, C. Comella, C. G. Patterson, D. A. Hall, and t. S. S. Group, "Does MDS-UPDRS Provide Greater Sensitivity to Mild Disease than UPDRS in De Novo Parkinson's Disease?" en, *Movement Disorders Clinical Practice*, vol. 8, no. 7, pp. 1092–1099, 2021.

[165] M. Vandemeulebroecke, B. Bornkamp, T. Krahnke, J. Mielke, A. Monsch, and P. Quarg, "A Longitudinal Item Response Theory Model to Characterize Cognition Over Time in Elderly Subjects," *CPT: Pharmacometrics & Systems Pharmacology*, vol. 6, no. 9, pp. 635–641, Sep. 2017.

[166] C. S. Venuto, N. B. Potter, E. Ray Dorsey, and K. Kieburtz, "A review of disease progression models of Parkinson's disease and applications in clinical trials," en, *Movement Disorders*, vol. 31, no. 7, pp. 947–956, 2016.

[167] G. Verbeke and E. Lesaffre, "A Linear Mixed-Effects Model With Heterogeneity in the Random-Effects Population," *Journal of the American Statistical Association*, vol. 91, no. 433, pp. 217–221, 1996.

[168] J. Weickenmeier, M. Jucker, A. Goriely, and E. Kuhl, "A physics-based model explains the prion-like features of neurodegeneration in Alzheimer's disease, Parkinson's disease, and amyotrophic lateral sclerosis," en, *Journal of the Mechanics and Physics of Solids*, vol. 124, pp. 264–281, Mar. 2019.

[169] D. Weintraub, J. Koester, M. N. Potenza, A. D. Siderowf, M. Stacy, V. Voon, J. Whetteckey, G. R. Wunderlich, and A. E. Lang, "Impulse Control Disorders in Parkinson Disease: A Cross-Sectional Study of 3090 Patients," *Archives of Neurology*, vol. 67, no. 5, pp. 589–595, May 2010.

[170] E. Yang, M. Farnum, V. Lobanov, T. Schultz, R. Verbeeck, N. Raghavan, M. N. Samtani, G. Novak, V. Narayan, A. DiBernardo, and Alzheimer's Disease Neuroimaging Initiative, "Quantifying the pathophysiological timeline of Alzheimer's disease," eng, *Journal of Alzheimer's disease: JAD*, vol. 26, no. 4, pp. 745–753, 2011.

[171] L. Younes, "Diffeomorphic Learning," *Journal of Machine Learning Research*, vol. 21, no. 220, pp. 1–28, 2020.

[172] A. L. Young, R. V. Marinescu, N. P. Oxtoby, M. Bocchetta, K. Yong, N. C. Firth, D. M. Cash, D. L. Thomas, K. M. Dick, J. Cardoso, J. van Swieten, B. Borroni, D. Galimberti, M. Masellis, M. C. Tartaglia, J. B. Rowe, C. Graff, F. Tagliavini, G. B. Frisoni, R. Laforce, E. Finger, A. de Mendonça, S. Sorbi, J. D. Warren, S. Crutch, N. C. Fox, S. Ourselin, J. M. Schott, J. D. Rohrer, and D. C. Alexander, "Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference," en, *Nature Communications*, vol. 9, no. 1, pp. 1–16, Oct. 2018.

[173] A. L. Young, J. W. Vogel, L. M. Aksman, P. A. Wijeratne, A. Eshaghi, N. P. Oxtoby, S. C. R. Williams, D. C. Alexander, and, "Ordinal SuStaIn: Subtype and Stage Inference for Clinical Scores, Visual Ratings, and Other Ordinal Data," *Frontiers in Artificial Intelligence*, vol. 4, 2021.

[174] W. J. Zetusky, J. Jankovic, and F. J. Pirozzolo, "The heterogeneity of Parkinson's disease: Clinical and prognostic implications," eng, *Neurology*, vol. 35, no. 4, pp. 522–526, Apr. 1985.

[175] X. Zhang, E. C. Mormino, N. Sun, R. A. Sperling, M. R. Sabuncu, B. T. T. Yeo, and t. A. D. N. Initiative, "Bayesian model reveals latent atrophy factors with dissociable cognitive trajectories in Alzheimer's disease," en, *Proceedings of the National Academy of Sciences*, vol. 113, no. 42, E6535–E6544, Oct. 2016.