Robust Modeling and Prediction of Disease Progression Using Machine Learning

Mostafa Mehdipour Ghazi

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Medical Physics and Biomedical Engineering

University College London

July 31, 2020

I, Mostafa Mehdipour Ghazi, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

To my beloved mother ...

Abstract

This work studies modeling the progression of Alzheimer's disease using a parametric method robust to outliers and missing data and a nonparametric method robust to missing values and training instabilities. The proposed parametric method linearly maps the individual's age to a disease progression score (DPS) and jointly fits constrained generalized logistic functions to the longitudinal dynamics of biomarkers as functions of the DPS using M-estimation. The proposed nonparametric method applies a generalized training rule based on normalizing the input and loss to the number of available data points to the long short-term memory (LSTM) recurrent neural networks to handle missing input and target values. Moreover, a robust initialization method is developed to address the training instability in LSTM networks based on a scaled random initialization of the network weights, aiming at preserving the variance of the network input and output in the same range.

Both proposed methods are evaluated on data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) for robust modeling of volumetric magnetic resonance imaging (MRI) and positron emission tomography (PET) biomarkers, cerebrospinal fluid (CSF) measurements, as well as cognitive tests, and are compared to the state-of-the-art methods. The obtained results show that the proposed parametric model outperforms almost all state-of-the-art parametric methods in predicting biomarker values and classifying clinical status, and it generalizes well when applied to independent data from the National Alzheimer's Coordinating Center (NACC). Additionally, the proposed generalized training rule for deep neural networks achieves superior results to standard LSTMs using data imputation before training, especially when applied to data with lower rates of missing values.

A comprehensive analysis of the proposed methods in neurodegenerative disease progression modeling reveals that the proposed nonparametric method performs better than the proposed parametric method in predicting biomarker values, while the parametric method works significantly better in clinical status classification.

Impact Statement

Alzheimer's disease (AD) is the most common type of dementia and leads to progressive neurodegeneration. Nevertheless, there is no cure or effective treatment to stop the progression of AD. Hence, early diagnosis of the disease, especially in the pre-symptomatic stages, can provide time to treat symptoms and plan for the future. On the other hand, early diagnosis of AD is challenging mainly because elderly subjects can suffer from different age-related pathologies and normal aging besides AD. Therefore, methods to stage and identify at-risk individuals and important biomarkers are critical to dementia research.

Moreover, longitudinal cohorts often contain missing data points and outliers due to, for instance, dropped-out patients, unsuccessful or erroneous measurements, or different assessment patterns and modalities used for different subject groups. These issues hinder the direct application of the state-of-the-art data-driven models to the AD progression modeling and prediction. Therefore, there is a need for novel data-driven approaches that can deal with the aforementioned problems in this area.

The methods presented in this work have the potential to be used in clinical environments for a better understanding of AD for diagnostic, staging, monitorization, and prognostic purposes. The proposed robust tools can automatically analyze the complete perspective of the disease using longitudinal data in an end-to-end fashion. This is also a holistic way to implement a system suitable for both (academic) research and (industrial) clinical applications to better study, detect, and monitor AD.

The proposed methods developed to deal with heterogeneous patterns, missing data, and outliers can be applied to longitudinal studies other than AD, e.g., for

modeling the progression of COVID-19. This work has shown an impact on the scientific community through the dissemination via journals and clinical abstracts, presentation via workshops and conferences, as well as participation in relevant challenges in the field such as the Alzheimer's disease prediction of longitudinal evolution (TADPOLE) challenge.

Acknowledgements

First of all, I would like to thank my supervisors, Sébastien Ourselin and Mads Nielsen, for all their support, advice, and trust during the dementia modeling (DeMo) project, from whom I learned a lot.

My sincere gratitude goes to my co-supervisor Lauge Sørensen for providing the opportunity to work alongside him on this challenging project, who always was open for discussion and patiently provided me guidance and support throughout this journey.

I would like to extend my gratitude to Marc Modat and M. Jorge Cardoso for sharing their valuable suggestions for improving the quality of the work throughout this program.

I would like to thank as well my DeMo colleagues, Akshay, Mauricio, and Irme, who I enjoyed working and collaboration with over this study.

I also owe thanks to all the people in the Machine Learning and Image sections at the Department of Computer Science, University of Copenhagen, who all contributed to an inspiring and great working environment.

I must especially thank the European Union and Biomediq A/S for providing me the opportunity to work on this interesting Horizon 2020 Marie Skłodowska Curie project.

Finally, my heartfelt gratitude goes to my family, for their constant and unconditional love, support, and encouragement, and especially to my beloved mother who taught me to be strong.

Publication List

Peer-reviewed Journals Papers

[J1] **Mostafa Mehdipour Ghazi**, Mads Nielsen, Akshay Pai, M. Jorge Cardoso, Marc Modat, Sébastien Ourselin, and Lauge Sørensen. Training recurrent neural networks robust to incomplete data: application to Alzheimer's disease progression modeling. Medical Image Analysis, 53:39–46, 2019.

[J2] **Mostafa Mehdipour Ghazi**, Mads Nielsen, Akshay Pai, Marc Modat, M. Jorge Cardoso, Sébastien Ourselin, and Lauge Sørensen. Robust parametric modeling of Alzheimer's disease progression. NeuroImage, 2020.

Peer-reviewed Conference Papers

[C1] **Mostafa Mehdipour Ghazi**, Mads Nielsen, Akshay Pai, M. Jorge Cardoso, Marc Modat, Sébastien Ourselin, and Lauge Sørensen. Robust training of recurrent neural networks to handle missing data for disease progression modeling. In International Conference on Medical Imaging with Deep Learning, 2018.

[C2] **Mostafa Mehdipour Ghazi**, Mads Nielsen, Akshay Pai, Marc Modat, M. Jorge Cardoso, Sébastien Ourselin, and Lauge Sørensen. On the initialization of long short-term memory networks. In International Conference on Neural Information Processing, pages 275–286. Springer, 2019.

Peer-reviewed Abstracts

[A1] **Mostafa Mehdipour Ghazi**, Mads Nielsen, Akshay Pai, Marc Modat, M. Jorge Cardoso, Sébastien Ourselin, and Lauge Sørensen. MRI biomarkers improve disease progression modeling-based prediction of cognitive decline. In Radiological Society

of North America – Scientific Assembly and Annual Meeting, 2019.

[A2] **Mostafa Mehdipour Ghazi**, Lauge Sørensen, Akshay Pai, M. Jorge Cardoso, Marc Modat, Sébastien Ourselin, and Mads Nielsen. Disease progression modeling-based prediction of cognitive decline. In Alzheimer's Association International Conference, 2020.

Contents

1 Introduction **19**

1.1 Alzheimer's disease 19

1.1.1 AD progression 20

1.1.1.1 CSF 20

1.1.1.2 PET 21

1.1.1.3 MRI 21

1.1.1.4 Cognitive 22

1.1.2 AD subtypes 22

1.2 Disease progression modeling 23

1.2.1 Parametric DPM 24

1.2.1.1 Robust regression 25

1.2.2 Nonparametric DPM 27

1.2.2.1 Deep learning 27

1.2.2.2 Recurrent neural networks 29

1.2.2.3 LSTM networks and missing values 30

1.2.2.4 Deep neural network initialization 31

1.3 Objectives 32

1.4 Contributions 33

1.5 Thesis outline 34

2 The Proposed Parametric Method for DPM **36**

2.1 Modeling dynamics of biomarkers robust to outliers 36

2.2 The constrained logistic functions 37

2.3 The efficient optimization algorithm for model fitting 38
2.4 Biomarker value prediction 40
2.5 Clinical status classification 41
2.6 Experimental setup 41
2.6.1 Data 41
2.6.1.1 ADNI 41
2.6.1.2 NACC 42
2.6.1.3 Data filtering 43
2.6.1.4 The obtained study population 44
2.6.1.5 Data preprocessing 45
2.6.1.6 Data partitioning and bootstrapping 45
2.6.2 Evaluation metrics 46
2.6.3 Initialization of the optimization algorithm 47
2.6.4 Stopping criteria 48
2.7 Results and discussion 48
2.7.1 Biomarker modeling 48
2.7.2 Temporal ordering of biomarkers 49
2.7.3 DPS distribution versus biomarker timing 52
2.7.4 Predicting biomarker values 53
2.7.5 Classifying clinical status 54
2.7.6 Comparison with state-of-the-art results 56
2.7.7 Generalizability across cohorts 57
2.8 Conclusions 57

3 The Proposed Nonparametric Method for DPM 60

3.1 The basic LSTM architecture 60
3.2 The proposed training algorithm for handling missing values 61
3.3 Feedforward 62
3.4 Backpropagation through time 63
3.5 Momentum batch gradient descent 64
3.6 The proposed initialization for efficient training of LSTMs 65

3.6.1 The nonlinear activation functions in LSTM training 68
3.7 Experiments and results 69
3.7.1 Data 69
3.7.2 Experimental setup 70
3.8 Results and discussion 71
3.8.1 Modeling biomarkers 71
3.8.2 Classifying clinical status 71
3.8.3 Robustness to missing values 73
3.9 Conclusions 74

4 Comparison of the Two Proposed Methods for DPM **76**
4.1 Biomarker value prediction 76
4.2 Clinical status classification 76
4.3 The effects of different modalities on the performance 78
4.4 Robustness to missing values 79
4.5 Cognitive decline prediction using few visits 79
4.6 Conclusions 81

5 Conclusion **85**
5.1 Summary 85
5.2 Discussion 86
5.3 Future work 87

Bibliography **89**

List of Figures

1.1 Different stages of AD and the symptoms. Adapted from [1]. 20

1.2 A hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. Adapted from [2]. 24

1.3 An illustration of the AD progression modeling method proposed in [3, 4]. Left: A Sigmoid function is fitted to the biomarker measurements of each subject. Middle: The biomarker trajectories are aligned by linearly transforming subject age to DPS. Right: The aligned biomarker fit is obtained for all subjects. 25

1.4 An illustration of how the proposed method (red curves) tackles the existing biomarker curve-fitting problems using simulated data generated based on logistic functions and additive white Gaussian noise. Left: A flexible function is used to fit the asymmetric shape of the simulated data points. Middle: A constrained function is utilized to estimate the exact dynamic range of the biomarker. Right: A robust estimator is applied to fit a curve to the simulated data contaminated with outliers. 27

2.1 Estimated curves per bootstrap (in gray) for the ADNI biomarkers using the modified Stannard function and the logistic loss. The average of the bootstrapped curves per biomarker is shown as the black curve. 50

2.2 Estimated curves per bootstrap (in gray) for the ADNI biomarkers using the modified Stannard function and the logistic loss. The average of the bootstrapped curves per biomarker is shown as the black curve. 51

2.3 The average of the normalized curves of the ADNI biomarkers across 100 bootstraps. 52

2.4 Temporal ordering of the ADNI biomarkers in the disease course obtained using inflection points and quantified through 100 bootstraps. The values in the matrix represent the frequency of occurrences (probabilities) and the units in the x-axis indicate the relative ordering of the biomarkers. 53

2.5 Estimated class-conditional likelihoods using the DPSs obtained from 100 ADNI-trained bootstraps. The box plots indicate the 25th to 75th percentiles of the estimated inflection points per biomarker, centrally marked with the median, and they are extended to the most extreme non-outlier inflection points using dashed lines. 54

2.6 Estimated curves per bootstrap (in gray) for the NACC biomarkers using the modified Stannard function and the logistic loss. The average of the bootstrapped curves per biomarker is shown as the black curve. The last subfigure shows the average of the normalized curves of the NACC biomarkers across 100 bootstraps. 58

3.1 An illustration of a vanilla LSTM unit with peephole connections in red. The solid and dashed lines show weighted and unweighted connections, respectively. 61

3.2 Illustration of how the normalization factors are related to the input and output of an unfolded RNN. Assume an RNN with three consecutive time points $\{t-1, t, t+1\}$, three input nodes, four hidden nodes, and two output nodes. Missing data for an instance subject j is illustrated as black nodes. We wish to weight the input vector and loss function according to the number of available data points in the input and output nodes. In this example, the subject j has only one input measurement at time t and one data point in the m-th output node. Hence, the input signal and loss function are weighted by 1/3 and 1, respectively. 62

3.3 Test modeling performance of different methods for various amounts of data. 73

3.4 Test diagnostic performance of different methods for various amounts of data. 74

4.1 Test modeling performance of different methods for various amounts of data. 80

4.2 Test diagnostic performance of different methods for various amounts of data. 80

4.3 Cognitive test prediction results for the test subjects per visit using the nonparametric (left) and parametric (right) methods. The error bars are calculated based on a 95% confidence interval for population standard deviation per visit. 82

4.4 Cognitive test prediction results for the test subjects per visit using the nonparametric (left) and parametric (right) methods. The error bars are calculated based on a 95% confidence interval for population standard deviation per visit. 83

List of Tables

2.1 Details of the utilized logistic functions for AD progression modeling. Note that the range of each function can be controlled by two additional parameters. 38

2.2 The utilized ρ-type M-estimators and their corresponding scale factors τ for robust regression. 40

2.3 Demographics of the obtained datasets after filtering across visits. . 44

2.4 Statistics of the visits per dataset after filtering. 44

2.5 Modeling performance as BIC (mean$\pm$SD) $\times 10^4$ for the 100-times bootstrapped ADNI training subsets using different logistic and loss functions. 49

2.6 Test modeling performance of different methods as NMAE (mean$\pm$SD) for ADNI and NACC biomarkers. Note that ADNI has 16 biomarkers while NACC has only 6 biomarkers in common between the two datasets. All the NMAEs are significantly different ($p < 0.001$). 54

2.7 Detailed information about the utilized ADNI biomarkers. 55

3.1 Test modeling performance of different methods as MAE for yearly predictions of the utilized ADNI biomarkers. All the MAEs are significantly different ($p < 0.01$). 72

3.2 Test diagnostic performance of different methods as AUC using an LDA classifier applied to the yearly estimated biomarker values. . . 72

3.3 Test modeling performance of different methods as MAE for half-yearly predictions of the utilized ADNI biomarkers. All the MAEs are significantly different ($p < 0.01$). 75
3.4 Test diagnostic performance of different methods as AUC using an LDA classifier applied to the half-yearly estimated biomarker values. 75

4.1 Test modeling performance of the two proposed methods as MAE for half-yearly predictions of the utilized ADNI biomarkers. All the MAEs are significantly different ($p < 0.01$). 77
4.2 Test diagnostic performance of the two proposed methods as AUC using an LDA classifier applied to the half-yearly estimated biomarker values. All the AUCs are significantly different ($p < 0.05$). 77
4.3 Test modeling performance of the two proposed methods as NMAE for half-yearly predictions of the utilized ADNI biomarkers. All the NMAEs are significantly different ($p < 0.01$). 78
4.4 Test diagnostic performance of the two proposed methods as AUC using an LDA classifier applied to the half-yearly estimated biomarker values. All the AUCs are significantly different ($p < 0.05$). 79
4.5 Test prediction NMAEs (mean$\pm$SD) per visit using the proposed nonparametric method. 84
4.6 Test prediction NMAEs (mean$\pm$SD) per visit using the proposed parametric method. 84

Chapter 1

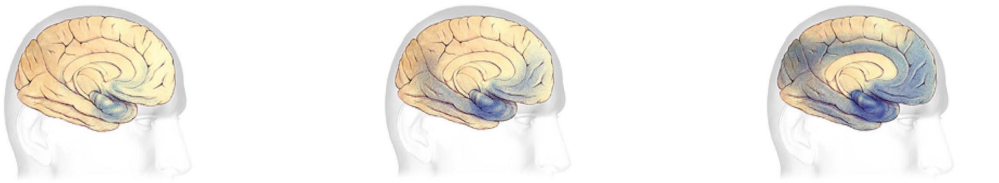
Introduction

Quantitative characterization of disease progression using longitudinal data can provide long-term predictions for the pathological stages of individuals. Accordingly, disease progression modeling (DPM) methods use longitudinal studies to develop data-driven models that can describe the evolution of the disease over time. These approaches can, therefore, provide a complete perspective of the disease by computationally exploring the available data to help with a better understanding of the disease for diagnostic, staging, monitorization, and prognostic purposes.

1.1 Alzheimer's disease

Alzheimer's disease (AD) is the most common type of dementia and leads to progressive neurodegeneration, affecting memory and behavior according to regional damage to the brain cells [5]. As shown in Figure 1.1, changes in the early stage of the disease may begin 10-20 years before diagnosis and can cause memory issues in patients [1]. In the mild and moderate AD stages that can last 2-10 years, personality changes occur and patients can have issues with recognizing objects. Finally, in the severe AD stage that may last 1-5 years, widespread cell death occurs, and patients can have the inabilities to communicate, recognize family, and care for themselves.

The hippocampus, which is the center of learning and memory, is often one of the first regions of the brain to be damaged. It has also been shown that cerebrospinal fluid (CSF) biomarkers can become abnormal in the presymptomatic phase of the disease, preceding positron emission tomography (PET) and magnetic resonance



Early AD
Learning, thinking, planning

Mild/Moderate AD
Speaking, understanding

Severe AD
Communicating, recognizing

Figure 1.1: Different stages of AD and the symptoms. Adapted from [1].

imaging (MRI) biomarkers followed by clinical markers [2, 6].

Currently, the cause of AD is not clear, and there is no cure or effective treatment to stop its progression, but early diagnosis of the disease, especially in the pre-symptomatic stages, can provide time to treat symptoms and plan for the future. Although genetic factors such as the allele e4 of the Apolipoprotein E gene (APOE $\varepsilon 4$) increase the risk of AD development [7], early diagnosis of AD is challenging mainly because elderly subjects can suffer from different age-related pathologies (e.g., cerebrovascular lesions, Lewy pathology, and TDP-43 proteinopathies) and normal aging besides AD. Therefore, methods to stage and identify at-risk individuals are critical to dementia research. These methods are applied to AD biomarkers that provide detailed measures of abnormal changes in the brain.

1.1.1 AD progression

The pathological progression of AD can be studied in three clinical phases [6]: a presymptomatic phase in which patients are still cognitively normal (CN), a prodromal phase with mild cognitive impairment (MCI), and a demented phase with AD when there are impairments in multiple cognitive domains and problems in daily living activities. Hence, biomarkers of AD can help us to detect early changes in the brains of patients at risk of AD years before showing symptoms such as memory loss. These biomarkers are used from different modalities to measure changes in the size, function, and certain protein levels of the brain.

1.1.1.1 CSF

CSF is the surrounding fluid of the brain and spinal cord that provides insulation and nutrients for keeping brain cells healthy. It is obtained in an invasive way by

a lumbar puncture and contains certain proteins such as amyloid-beta ($A\beta_{42}$) and tau that are valuable tools for the early detection of neurodegenerative diseases in presymptomatic phases [8]. Studies [6] have shown that $A\beta_{42}$ as an indicator of fibrillary amyloid deposition in plaques leaving the brain to CSF can decrease to 50% in AD patients compared with normal controls of the same age and changes may begin 20 years before symptoms, whereas tau protein abnormalities (tangles) in the CSF, can increase up to 300% in AD cases compared with normal elderly subjects where the changes may start 15 years before symptoms. It should also be noted that the aforementioned brain proteins can be measured with sensitive blood tests (proteomics) in a minimally invasive manner. However, the plasma biomarkers (proteomics, metabolomics, or transcriptomics) are less accurate than CSF biomarkers for identifying AD.

1.1.1.2 PET

PET scans are obtained by brain imaging after an injection of a radioactive substance (tracer) into the arm veins. They can reveal abnormalities in the chemical activities of the brain by measuring, for example, glucose use and protein levels in different brain regions. They can also capture the immune inflammatory responses (microglial activations) with an early and late peak in the AD course [9]. Fluorodeoxyglucose (FDG) is a widely used AD biomarker that estimates cerebral metabolic rates of glucose [10]. Studies [6] show a decline in the cerebral metabolic rates of glucose in AD patients even 15 years before symptoms, beginning in the entorhinal cortex and hippocampus and spreading to the posterior cingulate cortex, temporoparietal areas, and precuneus and prefrontal cortex. Moreover, abnormal accumulation of amyloid plaques in the brain regions be detected in the same way using the florbetapir or Amyvid (AV45) tracer [11].

1.1.1.3 MRI

MRI scans are obtained by brain imaging using magnetic fields and radio waves in a safe and painless, yet noisy manner. Depending on the type of scan, they can provide detailed information of the brain regions that can be used to measure regional changes in the size and shape of the brain functions, flows, or structures including tumors,

vascular damages, and atrophies that are related to loss of neurons and synapses and can be evidence for neurodegenerative diseases. As mentioned before, shrinkage in the medial temporal lobe including the hippocampus and entorhinal cortex is an early (presymptomatic) sign of AD neuropathology. This is followed by atrophy in the inferolateral regions of the temporal lobes in mildly or moderately impaired patients. Moreover, medial parietal lobe atrophy can be seen at all stages, with frontal lobe atrophy occurring later in the disease [12]. The early-onset AD stage may also involve changes in the precuneus/posterior cingulate gyrus [6]. Studies show that the atrophy rates in AD patients are about 3-7% per annum, whereas it accounts for less than 0.9% in elderly healthy controls.

1.1.1.4 Cognitive

Cognitive assessments through clinical and neuropsychological tests are noninvasive methods for identifying the early stages of the disease. They can measure cognitive dysfunction or decline and its severity in the mild and prodromal stages of the disease in a quantifiable manner [13] by evaluating different skills of a person such as thinking, learning, memory, and language through some questionnaires or activities. Cognitive decline is shown to be one of the latest markers becoming abnormal in the course of AD. Still, auditory verbal learning tests are found as effective markers in the early detection of AD, predicting neurodegenerative changes up to 10 years before clinical diagnosis [14, 15, 16].

1.1.2 AD subtypes

Alzheimer's disease is a heterogeneous disorder with different clinical and pathobiological subtypes that vary in age, sex distribution, cognitive status, disease duration, APOE $\varepsilon 4$ genotype, CSF biomarker levels, and clinical morphology from the early-onset AD (EOAD) to late-onset AD (LOAD). It has four major subtypes [17] of typical, limbic predominant (LP), hippocampal sparing (HcSp), and minimal atrophy (MA), according to amyloid-beta decomposition (A), distribution of tau pathology (T), and brain atrophy or neurodegeneration (N) assessed by using plasma or CSF, PET, and MRI [18, 19]. It is also shown that the risk of cognitive deterioration differs

considerably between the various subtypes (A-N- $<$ A+N- $<$ A+T-N+ $<$ A+T+N+).

Typical AD is the most frequent subtype which is characterized by tau pathology, atrophy in both hippocampus and association cortex, and greater white matter hyperintensity burden. This together with the LP subtype have higher onset and death ages, progress slowly, more involve vascular co-pathology, and are frequently seen in female patients and APOE $\varepsilon 4$ carriers. Whereas the HcSp form has a lower age, progresses more quickly with severe neurodegeneration and rapid cognitive decline, less involve Lewy co-pathology, and is frequently seen in male patients and noncarriers. In addition, in contrast to MA patients with an intermediate onset age and slow progression rate, HcSp cases have been found to have higher levels of education, suggesting that more education may help to protect the hippocampus.

Compared to the LP forms which have higher entorhinal tau load and greater amyloid-beta PET binding in frontal and parietal cortices, typical AD forms have higher amyloid-beta plaque counts in occipital regions and greater tau loads in the temporal lobe. The HcSp subtype patients show greater tau load in the frontoparietal lobe. Also, tau pathology and neurodegeneration at the molecular level can disrupt key brain networks in the medial temporal lobe of MA subtype patients, causing memory impairment comparable to LP and typical AD.

1.2 Disease progression modeling

Two types of approaches can be applied to modeling the progression of the disease, parametric and nonparametric. Parametric DPM methods describe data with a finite set of parameters independent of the number of training samples. These methods are simple, fast, and need fewer data points for training, but require temporal alignment of subjects' trajectories. Nonparametric DPM methods are flexible, make fewer assumptions for modeling, and can result in a high prediction accuracy. However, they assume that a data distribution cannot be defined in terms of a finite set of parameters, and they require substantial data for successful training to tune the parameters and avoid overfitting.

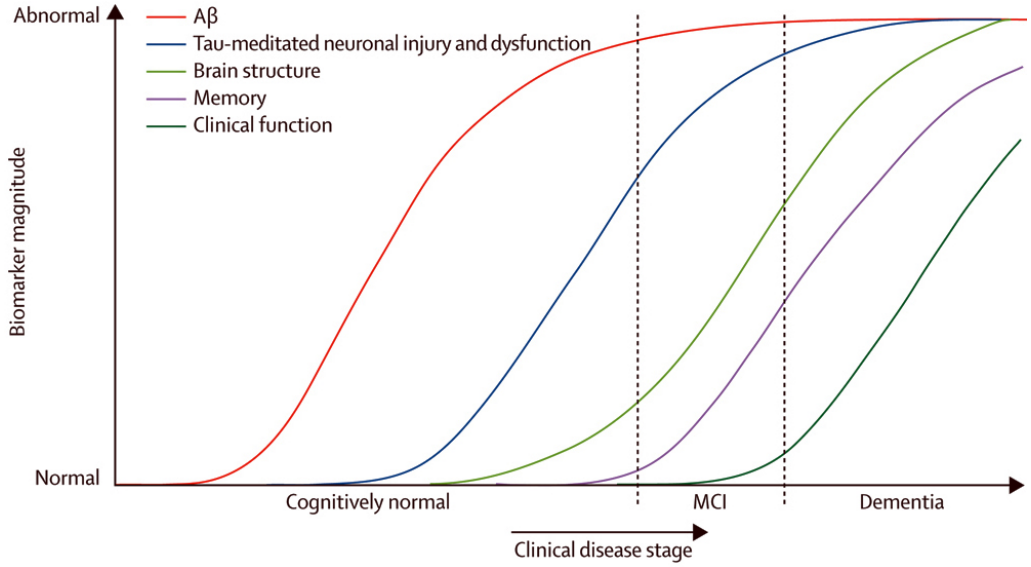


Figure 1.2: A hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. Adapted from [2].

1.2.1 Parametric DPM

Parametric disease progression modeling methods can be divided into two categories, continuous fitting for modeling the dynamics of biomarkers and discrete ordering of biomarkers for abnormality detection, both relying on unsupervised learning, e.g., by using maximum-likelihood estimation. The discrete methods focus on temporally ordering of biomarkers becoming abnormal during the disease stages by discretizing the disease progression trajectory using generative, event-based models [20, 21].

Continuous parametric methods for modeling the progression of AD have been inspired by hypothetical models, as shown in Figure 1.2, assuming a sigmoidal evolution of AD biomarkers [2, 22]. The goal of these methods is to model biomarker trajectories as a function of disease progression [3, 23]. Accordingly, a variety of approaches have been applied to fit a continuous function to the longitudinal dynamics of each biomarker using statistical models such as differential equations and mixed-effects models [24, 25, 26, 27, 28], in which one needs to align the trajectory of individuals based on some time measure, e.g., time-to-conversion. These methods are simple and require less data, but parametric assumptions on the biomarker trajectories limit the flexibility of the fits.

The parametric algorithm proposed in [3, 4] incorporates information from mul-

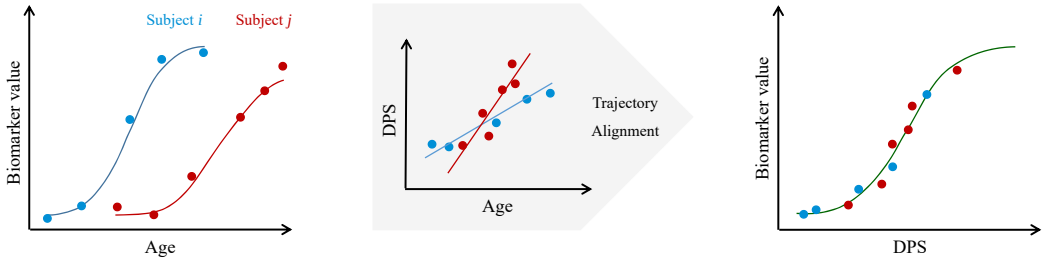


Figure 1.3: An illustration of the AD progression modeling method proposed in [3, 4]. Left: A Sigmoid function is fitted to the biomarker measurements of each subject. Middle: The biomarker trajectories are aligned by linearly transforming subject age to DPS. Right: The aligned biomarker fit is obtained for all subjects.

tiple biomarkers for modeling progression of AD over a common disease timescale. As shown in Figure 1.3, the method linearly transforms the age of the individual to a disease progression score (DPS) for the time-wise alignment of within-cohort measurements, assuming that the visit intervals in the data are short relative to the disease duration. Alternating least squares is applied to fit a sigmoid function to the longitudinal dynamics of each biomarker. In this method, biomarker trajectories are fitted independently and the biomarker dependencies are only considered when the algorithm alternates to estimate the subject-specific (age) parameters, which in turn can cause difficulties for the convergence of the alternating algorithm. Furthermore, the proposed model is not robust to outliers that can be found in more contaminated data. The first problem has been tackled in [29], but the problem with outliers remains.

1.2.1.1 Robust regression

Regression analysis is a form of statistical predictive modeling for estimating the relationships between target (dependent) variables $\boldsymbol{y}$ and predictor (independent) variables $\boldsymbol{x}$ by fitting a simple function like f to the data points as

$$\boldsymbol{y} = f(\boldsymbol{x}; \theta) + \boldsymbol{\varepsilon}\,.$$

where $\boldsymbol{\varepsilon}$ is an additive error term representing random effects and θ denotes the model parameters that can be obtained by minimizing the difference between the target and predictor variables. For example, ordinary least squares are the most

common methods for estimating the parameters of the fit by minimizing the sum of squared errors as

$$\hat{\theta} = \arg\min_{\theta} \sum_{i} (\boldsymbol{y}_i - f(\boldsymbol{x}_i; \theta))^2 ,$$

An outlier is a data point different from other observations in value and pattern they follow, which can cause serious problems, especially when learning data patterns in regression analysis by receiving more weights. To cope with outliers in data, M-estimation is introduced as a robust regression method [30] by minimizing a loss function designed to de-emphasize outliers (see Table 2.2). M-estimators are generalized types of least squares and maximum likelihood estimators for parametric models that are calculated through the minimization of an objective function with some data-dependent parameters. Assume that P is a likelihood function parameterized by θ. If the observations are independent and identically distributed, maximum likelihood estimation (MLE) can be obtained as

$$\hat{\theta} = \arg\max_{\theta} \log \left(\prod_{i} \mathrm{P}(\boldsymbol{x}_i; \theta) \right) = \arg\min_{\theta} \sum_{i} -\log(\mathrm{P}(\boldsymbol{x}_i; \theta)) ,$$

where $\boldsymbol{x}_i$'s are the observations. M-estimation employs a more generic function to estimate the parameters as

$$\hat{\theta} = \arg\min_{\theta} \sum_{i} \rho(\boldsymbol{x}_i; \theta) ,$$

where the function ρ can be chosen in such a way to reduce the effects of outliers and to provide the estimator desirable properties in terms of bias and asymptotic efficiency [31] with respect to an assumed distribution in reaching the Cramér-Rao bound on the variance of the unbiased estimators.

The model fit can further be improved by utilizing a more flexible function (see Table 2.1) and/or constraining the objective function. However, increasing the number of parameters needs to be penalized as it can increase the model complexity and result in overfitting. Figure 1.4 illustrates 1) how the use of a flexible function

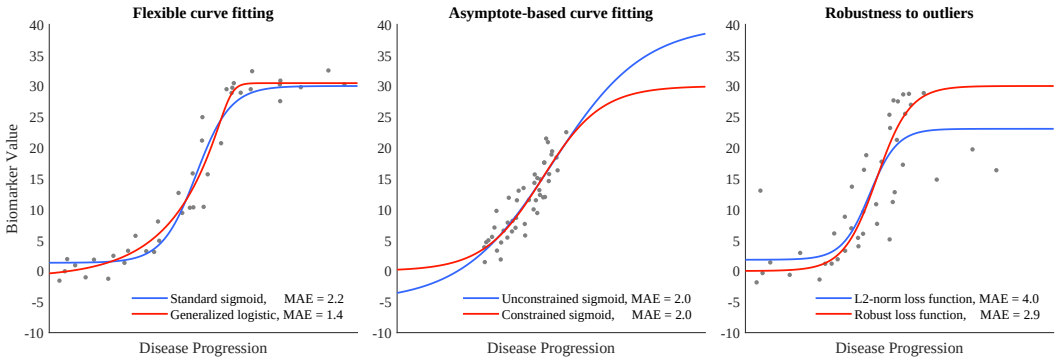


Figure 1.4: An illustration of how the proposed method (red curves) tackles the existing biomarker curve-fitting problems using simulated data generated based on logistic functions and additive white Gaussian noise. Left: A flexible function is used to fit the asymmetric shape of the simulated data points. Middle: A constrained function is utilized to estimate the exact dynamic range of the biomarker. Right: A robust estimator is applied to fit a curve to the simulated data contaminated with outliers.

improves the curve fit, 2) how the use of a constrained function moves lower and upper asymptotes to fit the exact dynamic range of the biomarker, and 3) how the use of M-estimation reduces the influence of outliers.

1.2.2 Nonparametric DPM

Nonparametric disease progression modeling methods have been introduced to model biomarkers jointly while taking temporal dependencies among measurements into account using Gaussian processes [32] or deep learning [33]. In contrast to the parametric methods, these methods do not require alignment of the trajectories of the individuals. However, a multivariate gaussian process with monotonicity constraints is computationally expensive to fit due to large covariance matrices, and deep learning methods are less interpretable and are hard to train in cases when the data is sparse or irregular. Moreover, these methods cannot easily be applied for prediction when the unseen data has fewer biomarkers than what was used for training.

1.2.2.1 Deep learning

Deep neural networks are hierarchical computing units inspired by neurobiological systems [34] that involve multiple cascaded layers with linear or nonlinear transformations, each of which learns to extract abstract features and represent specific

characteristics of the input data using several connected input-output nodes called artificial neurons [35]. The connections, which simulate the biological synapses, use learnable weights to transmit signals with different strengths between the neurons. In a fully-connected network, each layer's output node can be obtained by thresholding a weighted aggregation of the layer's input nodes using an activation function as

$$\boldsymbol{y}_i = \sigma(\sum_j w_{ij}\boldsymbol{x}_j + \boldsymbol{b}_i)\,,$$

where $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{b}$ are the input, output, and bias vectors, respectively. Also, w_{ij} is the connecting weight between the i-th input node and the j-th output node, and σ is an activation (thresholding) function such as the logistic sigmoid or hyperbolic tangent. The number of network layers and input and output nodes per layer can be assigned based on the number of available features and depending on the task, e.g., for regression, classification, and dimensionality reduction. It should be noted that although fully connected networks can be trained on multivariate data, they are not able to learn temporal dependencies among longitudinal data due to the lack of sequential nodes which limits their application to time-series prediction.

Network training is typically performed in two steps of feedforward and back-propagation. In the forward pass, the training data enters the network and is transmitted to its output to obtain the prediction error in the last layer. The estimated error is then propagated through the network in the backward pass until all neurons have associated error values needed to calculate the gradients and optimize the weights to achieve better performance on the training data. The gradient descent algorithm is commonly used to find the local minimum of the loss function by iteratively taking steps in the opposite direction of the gradients to reduce the overall cost as

$$W^{new} = W^{old} - \alpha\nabla_W f(W^{old})\,,$$

where W is the network weight array, $\nabla_W f(\cdot)$ is the gradient of the loss function f with respect to W, and α is a tuning parameter for faster convergence called the learning rate used to change the step size when moving toward the minimum of the

(nonconvex) problem. Depending on the problem and amount of data, the algorithm can be applied per iteration to a randomly selected data sample (stochastic), a subset of training data (mini-batch), or whole training data (batch). Also, improved variants of the gradient descent method can be used for learning such as the momentum method [36], adaptive gradient algorithm [37], and adaptive moment estimation [38].

Overfitting is an important issue in training neural networks that occurs when a trained (complex) model describes training data much better than test data. That is to say, the model tries to memorize the training data instead of learning for a reliable generalization. Several strategies have been proposed in the literature to avoid overfitting and improve generalizability to unseen test data [39]. For example, one can tackle the problem by penalizing the network loss using the regularization method, data augmentation by applying plausible modifications to the training samples, using dropout layers, using simpler network models, and applying the early-stopping method.

1.2.2.2 Recurrent neural networks

Recurrent neural networks (RNNs) are the state-of-the-art, deep learning-based methods for sequence learning that map an input sequence to an output sequence by predicting the next time steps [40]. RNN training using the backpropagation through time algorithm is challenging due to vanishing and exploding gradients where the norm of the backpropagated error gradient can increase or decrease exponentially, hindering the network in capturing long-term dependencies [41].

Three main solutions have been proposed in the literature to improve RNN training; modifications of the training algorithm, modifications of the network architecture, or different weight initialization schemes. In the first approach, advanced optimization techniques such as the Hessian-Free method [42] or regularized loss functions [43] are applied to improve the backpropagation through time algorithm for learning long sequences. The second approach is to employ nonlinear reset units in the RNN architecture to store information for a long time, for instance, using long short-term memory (LSTM) networks [44] or gated recurrent units (GRUs) [45]. The third approach is to properly initialize the RNN weight matrices, for example, to

be identity [46] or orthogonal [47], to find a solution to the long-term dependency problem.

1.2.2.3 LSTM networks and missing values

LSTM networks, the most common type of RNNs, use a gated architecture to replace the hidden unit with a memory cell to efficiently capture long-term temporal dependencies by storing and retrieving sequence information over time. The memory cell or the so-called constant error carousels (CECs) is used as feedback along with three nonlinear (multiplicative) reset units to keep the backpropagated error signal constant [48, 49]. The input and output gates of the cell learn their weights to incorporate the stored information or to control the output values. There is also a forget gate that learns to remember or forget the memory information over time by scaling the cell content.

The vanilla LSTM is the most commonly used LSTM architecture that utilizes three reset gates with full gate recurrence and can include cell-to-gates (peephole) connections [50]. Still, since longitudinal cohorts often contain missing data points due to, for instance, dropped-out patients, unsuccessful measurements, or different assessment patterns used for different subject groups – as for example seen in the Alzheimer's Disease Neuroimaging Initiative (ADNI) [51], standard RNNs including vanilla LSTMs cannot be directly applied.

Preprocessing methods such as data imputation and interpolation are the most common approaches to handling missing data in RNNs. These two-step procedures decouple missing data handling and network training, resulting in a sub-optimal performance that is heavily influenced by the choice of data preprocessing method [52]. Although RNNs themselves have been used for estimating missing data [53, 54], the lack of methods to inherently handle incomplete data in RNNs is evident [55]. Other approaches update the architecture to learn or encode the missing data patterns [55, 52]. These methods are typically biased towards specific cohort or demographic circumstances correlated with the learned missing data patterns and introduce additional parameters in the network which increases the complexity of the network.

1.2.2.4 Deep neural network initialization

Since deep neural network training is achieved by solving a nonconvex optimization problem, mostly in a stochastic way, a random weight initialization scheme is important for faster convergence and stability. Otherwise, the magnitudes of the input signal and error gradients at different layers can exponentially decrease or increase, leading to an ill-conditioned problem. Moreover, studies on the initialization, for instance, using unsupervised pre-training [56], showed its importance as a regularizer for the optimization procedure to robustly reach a local minimum and to improve generalization. Therefore, standard initialization of weights with zero-mean uniform/Gaussian distributions and heuristic variances ranging from 0.001 to 0.01 or an input layer size (N) dependent variance of $1/(3N)$ have been widely used in previous studies [57].

Training difficulties have been investigated based on the variance of the responses in each layer, when the singular values of the Jacobian are not unit, and a normalized initialization of uniform weights with a variance of $1/N$ is suggested assuming that the activation functions are identity and/or hyperbolic tangent [57]. Likewise, a scaled initialization method has been developed to train deep rectified models from scratch using zero-mean Gaussian weights whose variances are $2/N$ [58].

To resolve the long-term temporal dependencies problem in RNNs, which can be seen as deep networks when unfolded through time, the (scaled) identity matrix has been applied to initialize the hidden (recurrent) weights matrix to output the previous hidden state in the absence of the current inputs in RNNs composed of rectified linear units (ReLU) [46]. Alternatively, (nearly) orthogonal matrices [47] and scaled positive-definite weight matrices [59] have been used to address vanishing and exploding gradients in RNNs by preserving the gradient norm during backpropagation.

As can be seen, different initialization methods have been proposed to deal with the training convergence problem in deep neural networks including RNNs, assuming that LSTMs by design can handle the issue. Hence, the abovementioned

initialization methods, e.g., orthogonal recurrent weight matrices and input weight matrices, both drawn i.i.d. from zero-mean Gaussian distributions with variances of $1/N$, have also been applied to LSTMs. However, even though LSTM units by design allow gradients to flow unchanged, they can still suffer from instabilities (exploding gradient problem) when trained on long sequences [60] with improper initialization due to the stochastic nature of the optimization and using multiplicative gates and feedback signals.

1.3 Objectives

This work aims at developing machine learning and deep learning-based methods to model the progression of Alzheimer's disease using imaging biomarkers and clinical data. To achieve this goal, first, a robust extension of [3, 4] is proposed that jointly fits a constrained logistic function to the longitudinal dynamics of each biomarker using M-estimation to address the potential curve-fitting problems, e.g., outliers, in the biomarker modeling (see Figure 1.4). The estimated parameters are quantified using bootstrapping via Monte Carlo resampling, and the inflection points are used to temporally order the biomarkers in the disease course. Kernel density estimation with normal bases is applied to the estimated DPSs for clinical status classification using a Bayesian classifier. Different loss and logistic functions are considered, including a modified version of the Stannard function [61] which tends to better describe the biomarker trajectories, and they are applied to AD progression modeling of the Alzheimer's Disease Neuroimaging Initiative (ADNI) [51] and the National Alzheimer's Coordinating Center (NACC) [62] data using volumetric MRI biomarkers, CSF and PET measurements, and cognitive tests. The obtained results indicate that the modified Stannard function fitted using the logistic loss achieves the best modeling performance over different bootstraps, and it consistently outperforms the basic algorithm of [4] and state-of-the-art results of [29] and [63] in almost all experiments.

Next, we propose a generalized method for training LSTM networks that can handle missing values in both input and target. This is achieved by applying the

batch gradient descent algorithm in combination with the loss function normalized by the number of missing values in input and target. Our goal is different than the approaches that encode the missing values' patterns [55, 52]; we want to train RNNs robust to missing values to more faithfully capture the true underlying signal and to make the learned model generalizable across cohorts. The proposed LSTM algorithm is applied to AD progression modeling in the ADNI cohort, and the estimated biomarker values are used to classify the clinical status of subjects.

Finally, a simple, yet robust initialization method is proposed to tackle the training instabilities in LSTM networks. The idea is based on scaled random initialization of the network weights with the property that the input and output signals have the same variance. The proposed method is applied to the proposed LSTM training algorithm when learning from the ADNI data for multivariate disease progression modeling data.

1.4 Contributions

The main contributions based on the parametric DPM study can be listed as follows:

- A novel generalized logistic function, called modified Stannard, is proposed which better fits the AD biomarker trajectories compared to using other logistic functions.
- The utilized logistic functions are constrained to estimate the exact dynamic range of biomarkers while decreasing the number of to-be-optimized parameters.
- M-estimation is used to suppress the effect of outliers on the model fit.
- The across-cohort generalizability of the proposed model is evaluated by applying the model trained using ADNI data to the test data from the NACC cohort with fewer biomarkers.
- An end-to-end approach is introduced that performs biomarker trajectory modeling (unsupervised learning), biomarker inflection point detection (event

ordering), and clinical status classification (supervised learning). This is a holistic way to implement a system suitable for both research and clinical applications to better study, detect, and monitor AD.

The main contributions based on the nonparametric DPM study can be summarized as follows:

- A generalized formulation of the backpropagation through time algorithm for training LSTM networks is proposed to handle incomplete data, and it is shown that such built-in handling of missing values provides better modeling and prediction performance compared to using data imputation with standard LSTM networks.

- Temporal dependencies among measurements in the ADNI data are modeled using the proposed LSTM training algorithm via sequence-to-sequence learning. To the best of our knowledge, this is the first time such multidimensional sequence learning methods are applied to neurodegenerative DPM.

- An end-to-end approach, without the need for trajectory alignment, is proposed for modeling the longitudinal dynamics of biomarkers and for clinical status classification.

- A robust initialization method is proposed to address the training instabilities in LSTM networks. To the best of our knowledge, this is the first time a network initialization method is specifically introduced for training LSTM networks.

1.5 Thesis outline

The rest of this thesis is organized as follows. Chapter 2 presents the details of the parametric method proposed in [64] and provides information on how the utilized approaches address potential curve-fitting problems, e.g., outliers, in the biomarker modeling. Chapter 3 presents the deep learning-based methods proposed in [65, 33] and describes the applied techniques to LSTM network training to handle missing

values in both input and target signals and tackle the network training instabilities. Chapter 4 compares the two proposed methods for disease progression modeling from different aspects including the prediction of decline of cognitive test scores [66, 67]. Finally, Chapter 5 summarizes the thesis and provides a general discussion of the thesis content.

Chapter 2

The Proposed Parametric Method for DPM

This chapter is based on the work presented in [64] where a robust extension of [3, 4] is proposed that jointly fits a constrained logistic function to the longitudinal dynamics of each biomarker using M-estimation to address potential curve-fitting problems, e.g., outliers, in the biomarker modeling (see Figure 1.4). The proposed method makes the estimates stable and robust to outliers by minimizing the error of parametric disease progression modeling using a proposed logistic function and M-estimation.

2.1 Modeling dynamics of biomarkers robust to outliers

Two sets of parameters are estimated in the model: observed biomarker-specific parameters, which are assigned for fitting the biomarker curves, and hidden subject-specific (age-related) disease progression parameters that are defined for aligning the trajectory of subjects. Assume that $y_{i,j,k}$ is the k-th biomarker's value at the j-th visit of the i-th subject and $f(s;\boldsymbol{\theta})$ is an S-shaped logistic function of DPS s with parameters $\boldsymbol{\theta}$. Each biomarker measurement is defined as

$$y_{i,j,k} = f(s_{i,j};\boldsymbol{\theta}_k) + \sigma_k \varepsilon_{i,j,k}\,,$$

where σ_k is the standard deviation of the k-th biomarker with $\boldsymbol{\theta}_k$ parameters, $\varepsilon_{i,j,k}$ is additive white Gaussian noise (random effect) with i.i.d. assumption, and $s_{i,j}$ is the DPS for the j-th visit of the i-th subject and is obtained as

$$s_{i,j} = \alpha_i t_{i,j} + \beta_i \,,$$

where $t_{i,j}$ is the age of subject i in visit j, and $\alpha_i \in \mathbb{R}_{>0}$ and $\beta_i \in \mathbb{R}$ are the rate and onset of disease progression of subject i, respectively. Finally, the multiobjective optimization for robust nonlinear regression is defined as

$$\{\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\theta}\} = \min_{i,j,k} \sum_{i,j,k} w_i \rho \left(\frac{y_{i,j,k} - f(\alpha_i t_{i,j} + \beta_i; \boldsymbol{\theta}_k)}{\sigma_k} \right) ,$$

where $\rho(\cdot)$ is a maximum likelihood-type function and $w_i = 1/N_i$ is a weighting factor for normalizing the objective function with the number of available points per subject (N_i).

2.2 The constrained logistic functions

For fitting the longitudinal trajectories of biomarkers, four logistic functions are considered (Table 2.1). All functions have the same range $(0,1)$ and can produce the same inflection points at $c \in \mathbb{R}$, to be later used for biomarker ordering. We candidate utilization of a modified flexible logistic function based on the Stannard function [61], where the $1/\gamma$ factor is multiplied by the exponential term to create an asymmetric growth curve with an inflection point at c like other functions. This function tends to better describe asymmetrical sigmoid patterns of the biomarker trajectories with modeling both slow and rapid growths at the beginning or the end of the disease period. In the defined functions, $b \in \mathbb{R}_{>0}$ and $\gamma \in \mathbb{R}_{>0}$ denote the growth rate and symmetry parameter of the curves, respectively. The reason for restricting b to the positive real numbers is to make parameters of the estimation identifiable.

It can also be deduced from Table 2.1 that the sigmoid function first introduced by Verhulst [68] is a special (symmetric) case of both Richards' function [69] and the proposed function when $\gamma = 1$. Moreover, Gompertz's function [70] is a simplified

Table 2.1: Details of the utilized logistic functions for AD progression modeling. Note that the range of each function can be controlled by two additional parameters.

Logistic function	$g(s;\boldsymbol{\theta})$	$\boldsymbol{\theta}$	$(min,max)\,\forall b>0$	$g'(s;\boldsymbol{\theta})$	$g''(c;\boldsymbol{\theta})$
Verhulst	$\left[1+e^{-b(s-c)}\right]^{-1}$	$\{b,c\}$	$(0,1)$ at $(-\infty,+\infty)$	$be^{-b(s-c)}\left[1+e^{-b(s-c)}\right]^{-2}$	0
Gompertz	$e^{-e^{-b(s-c)}}$	$\{b,c\}$	$(0,1)$ at $(-\infty,+\infty)$	$be^{-b(s-c)}e^{-e^{-b(s-c)}}$	0
Richards	$\left[1+\gamma e^{-b(s-c)}\right]^{-1/\gamma}$	$\{b,c,\gamma\}$	$(0,1)$ at $(-\infty,+\infty)$	$be^{-b(s-c)}\left[1+\gamma e^{-b(s-c)}\right]^{-1-1/\gamma}$	0
Modified Stannard	$\left[1+\frac{1}{\gamma}e^{-\frac{b}{\gamma}(s-c)}\right]^{-\gamma}$	$\{b,c,\gamma\}$	$(0,1)$ at $(-\infty,+\infty)$	$\frac{b}{\gamma}e^{-\frac{b}{\gamma}(s-c)}\left[1+\frac{1}{\gamma}e^{-\frac{b}{\gamma}(s-c)}\right]^{-1-\gamma}$	0

form of Richards' function when γ approaches zero, i.e., $\lim_{\gamma\to 0}(1+\gamma u)^{-1/\gamma}=e^{-u}$. Finally, the upper and lower asymptotes of the curves can be adjusted by two additional parameters [71] as

$$f(s;\boldsymbol{\theta})=(a-d)g(s;\boldsymbol{\theta})+d\,.$$

The range parameters, a and d, can be set to fixed values when the exact range of biomarkers is given, which is the case with cognitive tests. This, in turn, not only reduces the number of optimization parameters but also increases the stability of the estimation. For other biomarkers, if there are, for example, sign constraints which are the cases with nonnegative CSF and PET measurements, both parameters can be constrained to lower and/or upper bounds, but otherwise remain unconstrained.

2.3 The efficient optimization algorithm for model fitting

Alternating approach, as an efficient optimization technique, is applied to solve the problem where the algorithm iteratively estimates θ using fixed values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and vice versa until the parameters converge. The proposed algorithm can be summarized as follows

Initialization: initialize $\{\boldsymbol{\alpha}^{(0)},\boldsymbol{\beta}^{(0)},\theta^{(0)}\}$ using measurements.

Optimization: iterate l until convergence.

Biomarker fitting: estimate the biomarker-specific parameters using values

of all subjects and visits.

$$\hat{\theta}^{(l)} = \min_{i,j,k} \sum_{i,j,k} w_i \rho \left(\frac{y_{i,j,k} - f(\hat{\alpha}_i^{(l-1)} t_{i,j} + \hat{\beta}_i^{(l-1)}; \boldsymbol{\theta}_k)}{\sigma_k} \right), \tag{2.1}$$

Age mapping: estimate the subject-specific parameters using values of all biomarkers and visits.

$$\{\hat{\alpha}_i^{(l)}, \hat{\beta}_i^{(l)}\} = \min_{(j,k) \in N_i} \sum_{j,k} w_i \rho \left(\frac{y_{i,j,k} - f(\alpha_i t_{i,j} + \beta_i; \hat{\boldsymbol{\theta}}_k^{(l)})}{\sigma_k} \right), \tag{2.2}$$

where N_i corresponds to the number of measurements among all biomarkers and visits available for the i-th subject. This way, in contrast to [3, 4], biomarkers are fitted jointly. The degrees of freedom of the fit is equal to $\sum_k (N_k - |\boldsymbol{\theta}_k|) - 2I$, where N_k is the number of measurements among all subjects and visits available for the k-th biomarker, $|\boldsymbol{\theta}_k|$ denotes the number of biomarker-specific parameters for the k-th biomarker, and I is the number of subjects. Therefore, the algorithm can be applied in case the data contains more than $\sum_k |\boldsymbol{\theta}_k| + 2I$ points, and if any subject has at least two distinct points considering all biomarkers and visits.

The utilized maximum likelihood-type functions for robust regression [72, 31] are described in Table 2.2. These estimators attempt to diminish the influence of the outliers while fitting curves. In general, M-estimators use a tuning parameter called τ to scale the functions as $\tau^2 \rho(r/\tau)$ in order to yield 95% asymptotic efficiency with respect to the standard normal distribution. The corresponding tuning constants for the utilized functions are also reported in Table 2.2.

Finally, the obtained DPSs are standardized with respect to the scores of the available cognitively normal visits of subjects in order to calibrate all biomarker trajectories in different experiments. This process removes the mean of the normal visits' distribution of DPSs and scales the range to give a better intuition of timing of disease progression in the course of AD. In this case, it would be necessary to

Table 2.2: The utilized ρ-type M-estimators and their corresponding scale factors τ for robust regression.

Loss function	$\rho(r)$	τ
L2	r^2	1
L1-L2	$2\left(\sqrt{1+r^2}-1\right)$	1
Logistic	$\ln(\cosh(r))$	1.205
Modified Huber	$\begin{cases} 1-\cos(\lvert r\rvert), & \lvert r\rvert \leq \pi/2 \\ \lvert r\rvert + (1-\pi/2), & \lvert r\rvert > \pi/2 \end{cases}$	1.2107
Cauchy-Lorentz	$\ln(1+r^2)$	2.3849

properly update the parameters as

$$
\begin{aligned}
s_{i,j} &= \left(s_{i,j} - \mu_{cn}\right)/\sigma_{cn}\,, \\
\alpha_i &= \alpha_i/\sigma_{cn}\,, \\
\beta_i &= \left(\beta_i - \mu_{cn}\right)/\sigma_{cn}\,, \\
b_k &= \sigma_{cn} b_k\,, \\
c_k &= \left(c_k - \mu_{cn}\right)/\sigma_{cn}\,,
\end{aligned}
$$

where μ_{cn} and σ_{cn} are the mean and standard deviation of the DPSs in the available cognitively normal visits of subjects, respectively.

2.4 Biomarker value prediction

Biomarker values can be predicted using the fitted model parameters. Age mapping part of the proposed algorithm is applied to estimate the subject-specific parameters using Equation (2.2) based on the values of those biomarkers of the test subject that have available estimated biomarker-specific parameters in the fitted model. Next, biomarker values are predicted as $f(s_{i,j};\boldsymbol{\theta}_k)$ using the estimated test DPSs where $f(\cdot)$ is the logistic function applied to model fitting.

2.5 Clinical status classification

In order to predict the clinical status of test subjects per visit, kernel density estimation (KDE) [73] is used to fit the likelihoods of cognitively normal, cognitively impaired, and AD groups in a nonparametric way. Assume that $(s_1, s_2, \ldots, s_N)$ is a set of N i.i.d. DPSs sampled from an unknown distribution with density function $p(s|c_i)$, where c_i denotes the i-th class label. KDE is expressed as

$$\hat{p}(s|c_i) = \frac{1}{Nw} \sum_{n=1}^{N} \mathcal{K}\left(\frac{s - s_n}{w}\right),$$

where $\mathcal{K}(\cdot)$ is a smooth (kernel) function with a smoothing bandwidth $w > 0$. Here, the Gaussian kernel is used as the smoothing function.

The clinical status is classified based on the DPSs with a Bayesian classifier that uses the KDE-based fitted likelihoods as

$$p(c_i|s) = \frac{p(c_i)p(s|c_i)}{\sum_i p(c_i)p(s|c_i)},$$

where $p(c_i)$ is the data-driven prior probability for the i-th class, $p(c_i|s)$ is the posterior probability for predicting the test DPS that belongs to the class c_i, and the term in the denominator specifies the evidence and can be dropped because the maximum a posteriori estimation is used for classification.

2.6 Experimental setup

2.6.1 Data

The data used in this work is obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [51] and the National Alzheimer's Coordinating Center (NACC) [62] databases.

2.6.1.1 ADNI

The ADNI was launched in 2003 as a public-private partnership, led by principal investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsycho-

logical assessment can be combined to measure the progression of mild cognitive impairment and early Alzheimer's disease. We use The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) challenge dataset [74] that includes the three ADNI phases ADNI 1, ADNI GO, and ADNI 2. This dataset contains measurements from brain MRI, PET, CSF, cognitive tests, and demographics, and genetic information.

The labels cognitively normal (CN), significant memory concern (SMC), and normal (NL) are merged under CN; mild cognitive impairment (MCI), early MCI (EMCI), and late MCI (LMCI) under MCI; and AD and dementia under AD. In addition, subjects converting from one clinical status to another, e.g., MCI-to-AD, are assigned the latter label (AD in this example). The utilized ADNI data includes T1-weighted brain MRI volumes of ventricles, hippocampus, whole brain, fusiform, and entorhinal cortex precomputed by ADNI using FreeSurfer tools [75], PET scan measures of florbetapir (AV45-PET) and fludeoxyglucose (FDG-PET) averaged across multiple areas of the cerebrum including temporal and parietal lobes and cingulate regions [76], CSF measures of Amyloid beta, total tau, and phosphorylated tau, as well as the cognitive tests of clinical dementia rating sum of boxes (CDR-SB), Alzheimer's disease assessment scale 13 items (ADAS-13), mini-mental state examination (MMSE), functional activities questionnaire (FAQ), Montreal cognitive assessment (MOCA), and Rey auditory verbal learning test of immediate recall (RAVLT-IR). Detailed information about the utilized biomarkers can be found in Table 2.7.

2.6.1.2 NACC

The NACC, established by the National Institute on Aging of the National Institutes of Health in 1999, has been developing a large database of standardized clinical and neuropathological data from both exploratory and explanatory Alzheimer's disease research [62]. The data has been collected from different Alzheimer's disease centers across the United States and among others contains measurements from different modalities such as MRI, PET, and cognitive tests.

Labels with numerical cognitive status of one (normal cognition) and two

(impaired-not-MCI) are merged under CN, and cognitive status of three (MCI) and four (Dementia) are set to MCI and AD, respectively. It should be noted that we only keep subjects with primary etiologic diagnosis of normal, AD, or missing. This is to exclude subjects diagnosed with other types of dementia, non-neurodegenerative disease, or a non-neurological condition. The used NACC data includes T1-weighted brain MRI volumes of hippocampus and whole brain precomputed by NACC using IDeA Lab's software [77], and the cognitive tests of MMSE, MOCA, FAQ (sum of the 10-item scores), and CDR-SB using the CDR® Dementia Staging Instrument.

2.6.1.3 Data filtering

For our analysis, in each of the ADNI and NACC datasets, measurements outside known biomarker ranges, e.g., RAVLT-IR < 0, are rejected and assumed as missing values. The volumetric MRI outliers observed in the ADNI dataset are removed by assuming intracranial volume (ICV) estimates that are proportionally smaller than estimated corresponding MRI measurements, i.e., MRI / ICV > 1, as missing values.

Clinical follow-up visits with reverting clinical diagnoses are removed per subject considering the neighboring visits. In the ADNI dataset, clinical follow-up visits with wrongly ordered dates are discarded per subject. Also, MRI, PET, and CSF measurements that are already matched to the cognitive visits with any extreme time gaps are excluded. The acceptable time gap is obtained based on the data statistics and is set to three months. In the NACC dataset, we perform the matching of MRI and clinical visits. However, due to the relatively smaller sample size in NACC compared with ADNI, matches more than three months apart are kept and treated as two distinct visits. In this analysis, we assign a missing clinical status for any MRI visits that do not fall within the 3-month window.

In order to be able to apply the proposed method, measurements and clinical diagnoses with missing date information per visit are set to missing values, and subjects with less than two distinct visits are omitted. This results in 219 ADNI subjects and 151 NACC subjects being excluded.

Table 2.3: Demographics of the obtained datasets after filtering across visits.

	clinical status	age, year (mean±SD)	education, year (mean±SD)	MMSE (mean±SD)
ADNI	CN	76.93±6.03	15.76±2.92	29.05±1.20
	MCI	75.07±7.67	15.80±2.90	27.43±2.26
	AD	76.47±7.51	15.80±2.90	21.61±4.61
	Missing	74.44±7.87	16.10±2.64	27.34±3.07
NACC	CN	79.06±7.34	13.76±4.00	28.46±1.71
	MCI	80.83±8.57	13.79±4.03	25.32±3.03
	AD	81.09±8.14	13.73±4.08	19.60±5.11
	Missing	78.88±11.69	13.56±4.69	28.29±2.36

Note that missing clinical status after filtering is indicated as 'Missing'.

Table 2.4: Statistics of the visits per dataset after filtering.

	# visits per clinical status				# visits per subject	visit interval, year	# measurements per subject
	CN	MCI	AD	Missing	(mean±SD)	(mean±SD)	(mean±SD)
ADNI	2,285	3,850	2,064	899	5.99±2.37	0.74±0.43	58.60±23.38
NACC	1,140	205	318	9	7.00±2.91	1.15±0.37	21.61±9.03

2.6.1.4 The obtained study population

After filtering the data, the utilized 16 ADNI biomarkers are acquired from 1,518 subjects (854 males and 664 females) in 9,098 visits between August 2005 and May 2017, and the six NACC biomarkers are acquired from 239 subjects (75 males and 164 females) in 1,672 visits between October 2005 and July 2018. All subjects in both datasets have at least one cognitive test. In the NACC data, 203 subjects underwent MRI imaging while in the ADNI data, 1,515 and 1,220 subjects underwent MRI and PET imaging, respectively, and 1,088 subjects have CSF measures. Table 2.3 and Table 2.4 summarize statistics of the demographics and measurements in the two datasets after data filtering. Note that both datasets include missing values and missing clinical status, the latter indicated as 'Missing'.

2.6.1.5 Data preprocessing

In the ADNI dataset, the volumetric measurements were obtained using two different versions of the FreeSurfer software, and in the NACC dataset, they were calculated using IDeA Lab's software following ADNI protocols. Therefore, the MRI measurements need to be corrected for software version [78], software package, and hence for different cohorts (ADNI-NACC). In addition, biological difference in brain size hinders direct utilization of MRI measurements for disease progression estimation. Total intracranial volume (TIV) or ICV is a commonly used measure for normalization to correct for head size. To overcome both aforementioned problems of difference in cohort/software (version) and head size, we employ the residual approach [79] based on the analysis of covariance, which takes data from control groups and linearly regresses MRI volumes on their corresponding ICV as a covariate of interest. The corrected measurements can thus be calculated as the estimated residuals $\hat{R}$ of the volumes using the regression parameters obtained from the control data as

$$\hat{R}_{i,j,k,v} = ROI_{i,j,k,v} - \left[\hat{\beta}^{cn}_{k,v} + \hat{\alpha}^{cn}_{k,v} ICV_{i,j,k,v}\right],$$

where $ROI_{i,j,k,v}$ is the k-th MRI volume for subject i at visit j calculated (observed) using software or software version v, ICV is the corresponding intracranial volume, and $\hat{\beta}^{cn}$ and $\hat{\alpha}^{cn}$ are the intercept and slope of the regression estimated from the CN group. Finally, the estimated residuals are standardized per cohort/software (version) so that all variables have zero mean and unit variance.

2.6.1.6 Data partitioning and bootstrapping

To evaluate the algorithms, each of the ADNI and NACC datasets is partitioned into two non-overlapping sets for training and testing. To be more specific, based on the first and last available diagnoses of subjects, i.e., CN-CN, CN-MCI, ..., AD-AD, we divide each of these types of pairs into two groups including few and many visits using the median number of visits as threshold and randomly select 20% of the subjects from each group for testing.

Additionally, bootstrapping is used in the experiments for quantification of the estimation, and in each bootstrap, a subset of the training subjects is randomly sampled with replacement based on the first and last available pair of diagnoses and the number of available visits per subject, to make sure each bootstrap sampling contains data from any diagnostic status and sequence lengths. The unused subjects are assigned for validation and account for $1/e \approx 0.37$ of the subjects where e is the base of the natural logarithm. This also means that the estimated variance using the bootstrapped model will account for approximately 63% of the total variance.

To facilitate future research in AD progression modeling and comparison with the current study, all the data splits, including each bootstrap split, are available online (`https://arxiv.org/src/1908.05338v3/anc`) as supplementary material [64].

2.6.2 Evaluation metrics

Robust Bayesian information criterion (BIC) is used as a criterion for model selection among the robust models [80]. The criterion is penalized with the number of parameters to avoid overfitting, where the model with the lowest BIC is preferred, and it is defined as

$$\mathrm{BIC} = 2E_{train}^{(L_{opt})} + Q\ln(N),$$

where $E_{train}^{(L_{opt})}$ is the training loss at the optimum iteration number L_{opt} obtained through biomarker fitting using Equations (2.1) and (2.2), N is the total number of measurements, and Q is the total number of parameters which is equal to $\sum_k |\boldsymbol{\theta}_k| + 2I$.

The mean absolute error (MAE) is used to assess the modeling performance as a measure less sensitive to outliers [81]. It is calculated based on the absolute differences between actual and estimated values as follows

$$\mathrm{MAE} = \frac{1}{N_k} \sum_{(i,j)\in N_k} \left| y_{i,j,k} - f(s_{i,j}; \boldsymbol{\theta}_k) \right|,$$

where N_k is the number of measurements among all subjects and visits available

for the k-th biomarker, and $y_{i,j,k}$ and $f(s_{i,j};\boldsymbol{\theta}_k)$ are the ground-truth and estimated values of the k-th biomarker for the i-th subject at the j-th visit. Absolute errors of different biomarkers can be normalized with the corresponding standard deviation of the biomarkers and averaged across all normalized biomarkers to obtain a single performance measure called normalized MAE (NMAE). The modeling performance of two different methods is statistically compared using the paired, two-sided Wilcoxon signed-rank test [82] applied to the NMAEs obtained from different bootstraps.

Additionally, multiclass area under the receiver operating characteristic (ROC) curve (AUC) [83] is used to measure the diagnostic performance in a multiclass test set and is calculated using the posterior probabilities as

$$\mathrm{AUC} = \frac{1}{(n_c(n_c-1))}\sum_{i=1}^{n_c-1}\sum_{k=i+1}^{n_c}\frac{1}{n_i n_k}\left[\mathrm{SR}_i - \frac{n_i(n_i+1)}{2} + \mathrm{SR}_k - \frac{n_k(n_k+1)}{2}\right],$$

where n_c is the number of distinct classes, n_i denotes the number of available observations belonging to the i-th class, and SR_i is the sum of the ranks of posteriors $p(c_i|\boldsymbol{s}_i)$ after sorting all concatenated posteriors $\{p(c_i|\boldsymbol{s}_i), p(c_i|\boldsymbol{s}_k)\}$ in an ascending order, where $\boldsymbol{s}_i$ and $\boldsymbol{s}_k$ are vectors of DPSs belonging to the true classes c_i and c_k, respectively. Likewise, SR_k is the sum of the ranks of posteriors $p(c_k|\boldsymbol{s}_k)$ after sorting all concatenated posteriors $\{p(c_k|\boldsymbol{s}_k), p(c_k|\boldsymbol{s}_i)\}$ in an ascending order.

2.6.3 Initialization of the optimization algorithm

Since the fitting algorithm is iteratively performed using an alternating approach starting from values optimized in the previous step, initialization is an important step for efficiently reaching the optimum. We initially set $\boldsymbol{\alpha}^{(0)}$ and $\boldsymbol{\beta}^{(0)}$ to $\mathbf{1}$ and $\mathbf{0}$, respectively. Moreover, we initialize the slope of the trajectories ($\boldsymbol{\lambda}$) to either -1 or 1 depending on the diagnoses. A positive slope is considered when the average of the k-th biomarker's values for cognitively normal visits is less than that for AD visits and vice versa.

Next, the parameters of the logistic functions are initially estimated as $\gamma_k = 1$, $c_k = 0$, and $b_k = 4\lambda_k/(a_k - d_k)$, where d_k and a_k are the minimum and maximum of the k-th biomarker's values, respectively, provided that the slope λ_k is positive,

and vice versa if the slope is negative. Finally, we repeat the alternating procedure using the logistic functions and the trust-region algorithm [84] considering robust estimators for at most 50 iterations.

2.6.4 Stopping criteria

To avoid overfitting, the optimal parameter values are selected according to the optimum generalization loss obtained using the following criteria [85]

$$\{\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\theta}\} = \min_{L_{min} \leq l \leq L_{max}} E_{valid}^{(l)},$$

where $E_{valid}^{(l)}$ is the validation loss at the l-th iteration obtained through biomarker fitting using Equations (2.1) and (2.2). The minimum number of iterations, L_{min}, is set to 10 to allow for enough training progress. The maximum number of iterations, L_{max}, is set to 50. This avoids unnecessary computations since it was empirically observed that E_{valid} attained a minimum well within this iteration range in all cases.

2.7 Results and discussion

2.7.1 Biomarker modeling

First, the proposed method is applied to model the dynamics of the ADNI biomarkers. Table 2.5 illustrates the modeling performance (BIC) for ADNI training subsets obtained from 100 bootstraps using different logistic and loss functions. The combination of the modified Stannard function for biomarker fitting and the logistic loss achieve the best modeling performance with both the lowest average BIC and the smallest standard deviation and a validation NMAE of 0.985 ± 0.029. This configuration will be used in all the remaining experiments.

To further investigate the stability and robustness of the model with the chosen configuration of logistic and loss functions, we visualize the fitted trajectories for each of the 100-bootstrap runs together with their average per biomarker in Figures 2.1 and 2.2. As it can be seen, the bootstrap curves follow almost the same logistic growth pattern per biomarker. Moreover, although the confidence intervals are relatively narrow in CSF biomarkers, RAVLT-IR, AV45-PET, and Entorhinal measurements,

Table 2.5: Modeling performance as BIC (mean±SD) $\times 10^4$ for the 100-times bootstrapped ADNI training subsets using different logistic and loss functions.

Logistic function \ Loss function	L2	L1-L2	Logistic	Modified Huber	Cauchy-Lorentz
Verhulst	2.090±0.039	1.901±0.028	1.830±0.027	1.836±0.027	1.925±0.029
Gompertz	2.101±0.042	1.902±0.028	1.831±0.027	1.836±0.027	1.927±0.029
Richards	2.077±0.038	1.899±0.028	1.829±0.027	1.835±0.027	1.924±0.029
Modified Stannard	2.077±0.038	1.898±0.028	**1.828±0.026**	1.834±0.027	1.924±0.028

The best result is shown in boldface and its corresponding configuration is selected for the remaining experiments.

we can simply separate bootstrap curves of the other biomarkers into different clusters, including CDR-SB, ADAS-13, MMSE, MOCA, FDG-PET, Ventricles, Whole Brain, and Fusiform measurements. Since we have applied robust regression to reduce the effects of possible outliers on the bootstrapped data curve fits, there should be other explanations for the seen differences. We hypothesize that some of the observed outliers and differently distributed bootstrapped data curve fits may represent AD subtypes discussed in Section 1.1.2. In other words, the proposed method has attempted to robustly fit a logistic function to the trajectory of each biomarker considering ages of bootstrapped subjects but disregarding some other subject-specific factors such as gender, cognitive status, APOE $\varepsilon 4$ genotype, CSF changes, and brain atrophy which may affect the regression performance. In addition, some of the biomarkers and AD subtypes do not necessarily follow an S-shaped trajectory pattern [9] and may progress with different rates in the disease course.

2.7.2 Temporal ordering of biomarkers

To indicate the timing and the dynamics of the different biomarkers relative to each other, Figure 2.3 shows the average curves scaled to $[0, 1]$ using the estimated upper and lower asymptotes per biomarker and superimposed in the same figure. The distribution of the inflection points of the biomarkers, quantified through bootstrapping, can be used to see how biomarkers proceed in the course of AD with respect to each other. The inflection point is considered a turning point at which the direction of biomarker curvature changes. Figure 2.4 displays the temporal ordering of the ADNI biomarkers based on the estimated inflection points. As can be seen, CSF

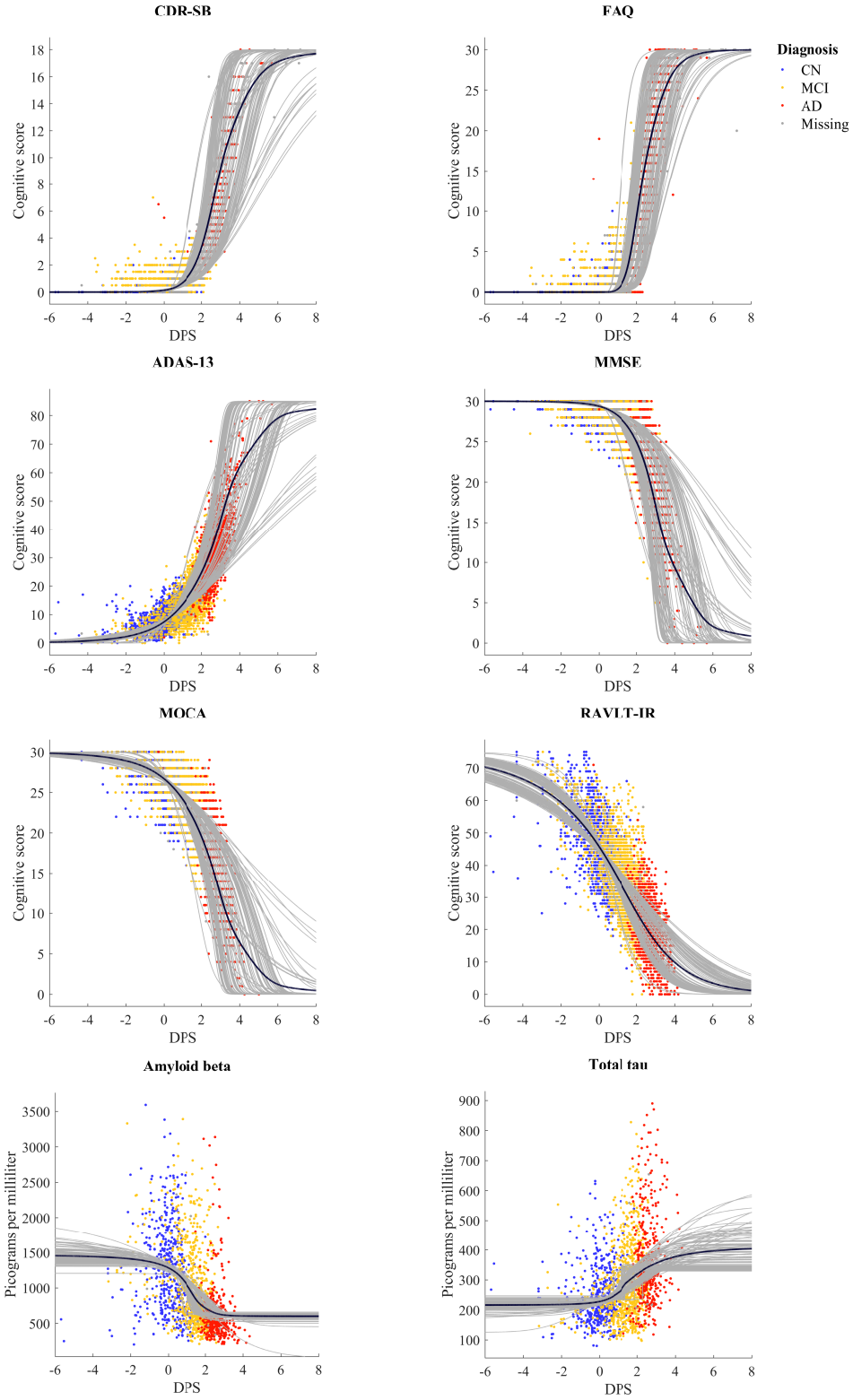


Figure 2.1: Estimated curves per bootstrap (in gray) for the ADNI biomarkers using the modified Stannard function and the logistic loss. The average of the bootstrapped curves per biomarker is shown as the black curve.

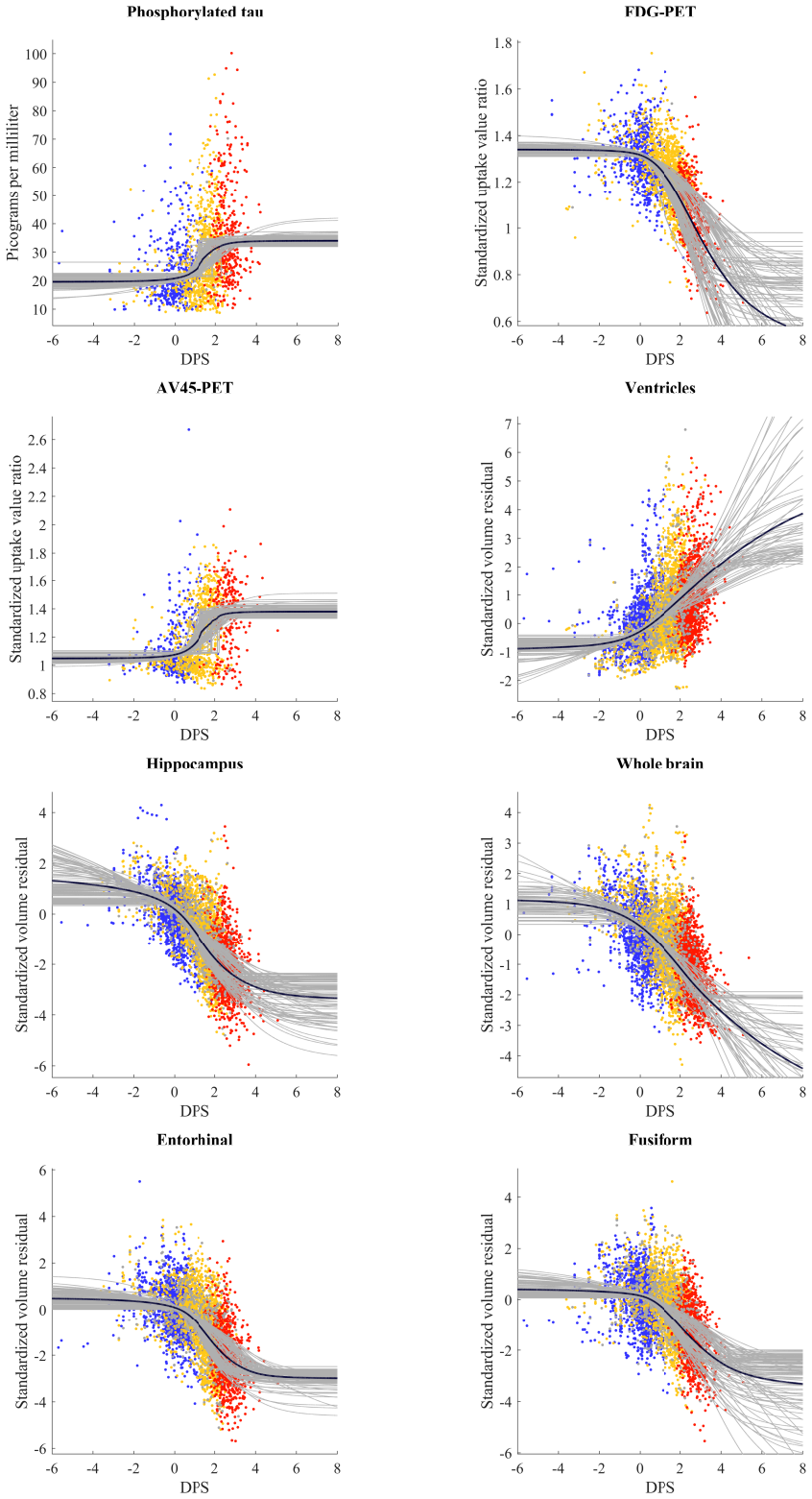
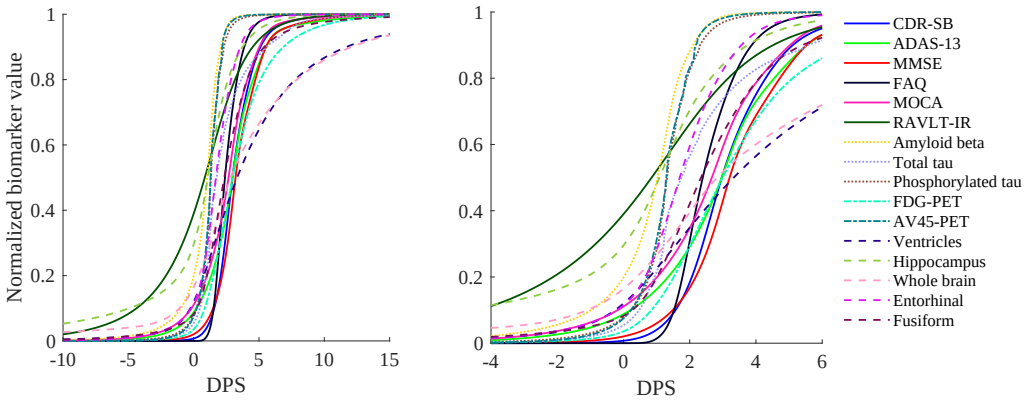


Figure 2.2: Estimated curves per bootstrap (in gray) for the ADNI biomarkers using the modified Stannard function and the logistic loss. The average of the bootstrapped curves per biomarker is shown as the black curve.



(a) The entire trajectory of all biomarkers. **(b)** A zoom on the DPS axis showing the most dynamic area.

Figure 2.3: The average of the normalized curves of the ADNI biomarkers across 100 bootstraps.

and PET biomarkers, as well as RAVLT-IR, precede all other biomarkers followed by MRI biomarkers and cognitive tests. These findings are in line with the results of [3, 4, 86, 29]. More interestingly, RAVLT-IR starts becoming abnormal early in the disease course which is consistent with several clinical studies concluding that some cognitive tests including RAVLT are significant predictors that can predict neurodegenerative changes up to 10 years before clinical diagnosis [14, 15, 16]. However, some of the MRI biomarkers such as the ventricles and whole-brain are noisy measurements for modeling the progression of AD in this dataset, as also seen in Figure 2.4. It is important to note that the inflection points are utilized to order the biomarkers in the disease course. These points do not measure when the biomarkers start becoming abnormal and hence, cannot be used for early abnormality detection.

2.7.3 DPS distribution versus biomarker timing

Figure 2.5 shows the variance of the estimated inflection points per biomarker alongside the estimated class-conditional likelihoods of the obtained DPSs from 100 bootstraps. As can be seen, there are moderate overlaps between the DPS distributions of CN-MCI and MCI-AD while the CN and AD groups can be discriminated easily. Moreover, the estimated inflection points per biomarker are almost in line with those of the hypothetical model by [2] that illustrates when biomarkers are dynamic versus disease stages. Especially, inflection points of the MRI biomarkers

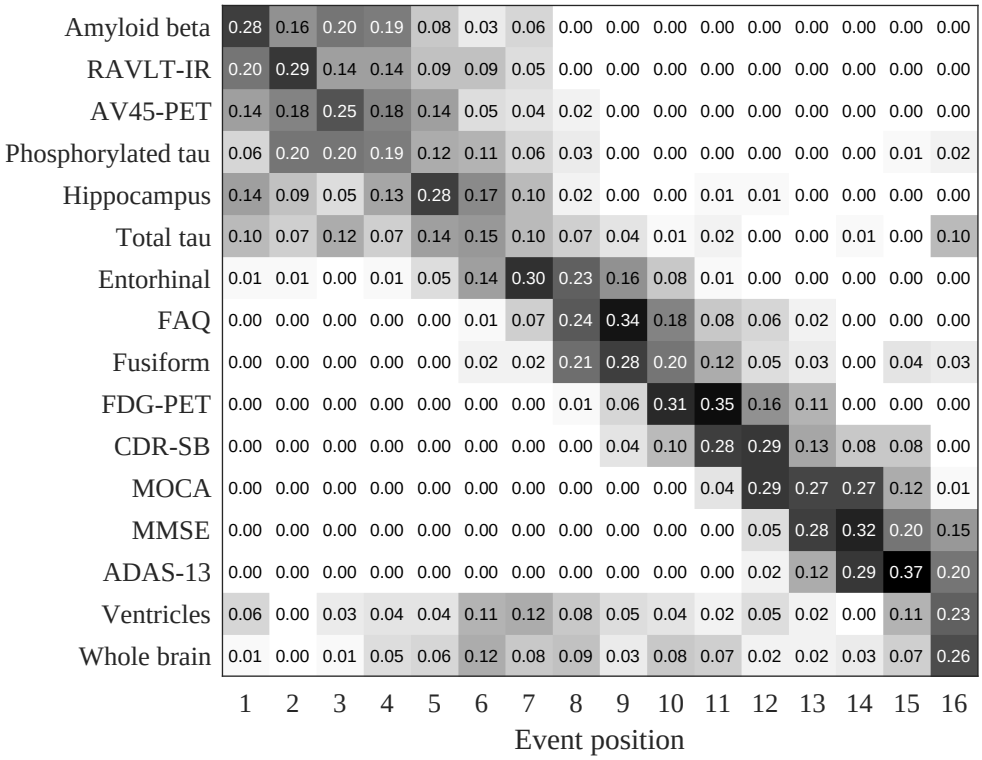


Figure 2.4: Temporal ordering of the ADNI biomarkers in the disease course obtained using inflection points and quantified through 100 bootstraps. The values in the matrix represent the frequency of occurrences (probabilities) and the units in the x-axis indicate the relative ordering of the biomarkers.

(brain structure) are mainly located in the MCI stage while those of the cognitive tests (memory), except for RAVLT-IR, lie on the AD stage.

2.7.4 Predicting biomarker values

The biomarker-specific parameters estimated using the bootstrapped training set are applied to map the ages of test individuals to DPSs using Equation (2.2). The obtained DPSs are then fed to the estimated biomarker functions in each bootstrap. Table 2.6 shows the test NMAEs of the 100-times bootstrapped ADNI dataset for the proposed model and the analogous model by [4] that independently fits the basic sigmoid function using an unconstrained, L2-norm loss function. The proposed model significantly ($p < 0.001$) outperforms the analogous model with an average NMAE of 0.991 vs. 1.552 and an average BIC of 1.828×10^4 vs. 3.303×10^4. Table 2.7 shows the average test MAE per biomarker across 100 bootstraps.

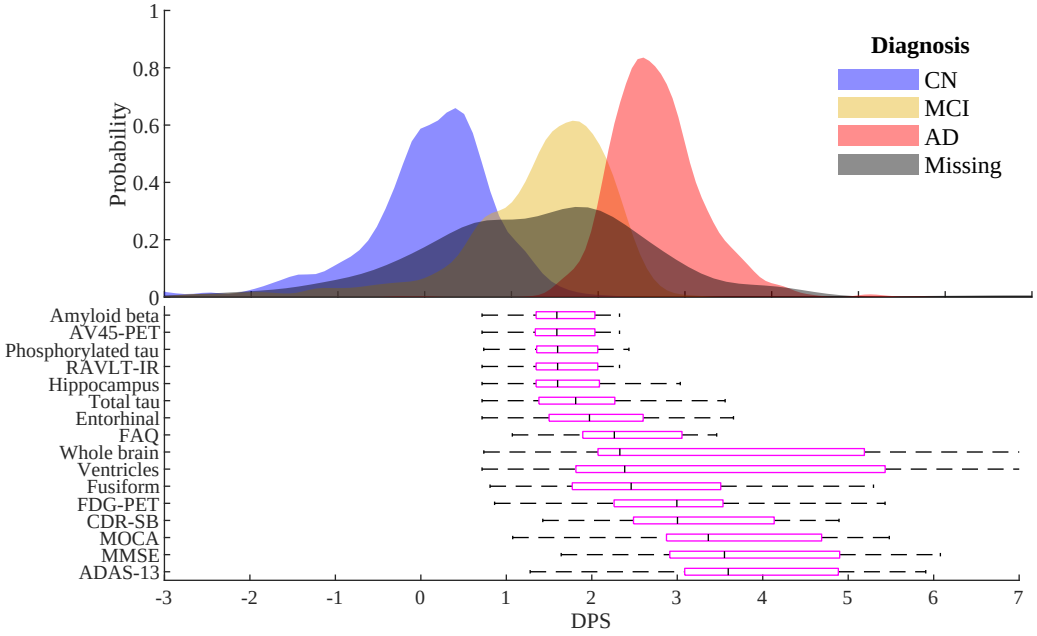


Figure 2.5: Estimated class-conditional likelihoods using the DPSs obtained from 100 ADNI-trained bootstraps. The box plots indicate the 25th to 75th percentiles of the estimated inflection points per biomarker, centrally marked with the median, and they are extended to the most extreme non-outlier inflection points using dashed lines.

2.7.5 Classifying clinical status

To evaluate the diagnostic predictive performance, the obtained training DPSs are used to estimate class-conditional likelihood functions per bootstrap using KDE and fed to a three-class Bayesian classifier with prior probabilities proportional to the number of training observations in each class. The classifiers, one for each bootstrap, are applied to the test DPSs estimated as described in Section 2.7.4 to compute the posterior probabilities of the clinical labels. The proposed model achieves an AUC

Table 2.6: Test modeling performance of different methods as NMAE (mean±SD) for ADNI and NACC biomarkers. Note that ADNI has 16 biomarkers while NACC has only 6 biomarkers in common between the two datasets. All the NMAEs are significantly different ($p < 0.001$).

	Within cohort		Across cohort
Method	ADNI	NACC	ADNI to NACC
Regression-L2 [4]	1.552±0.069	1.040±0.210	2.665±0.311
Regression-proposed [64]	0.991±0.023	0.833±0.061	1.182±0.087

Table 2.7: Detailed information about the utilized ADNI biomarkers.

Biomarker	Interpretation	Unit	Range	Inflection point (median)	Test MAE (mean)
CDR-SB	The sum of scores of six sets of questions. Lower values indicate less cognitive dysfunction.	Cognitive score	[0, 18]	3.003	0.562
ADAS-13	The sum of scores of 13 itemized tasks. Lower values indicate less cognitive dysfunction.	Cognitive score	[0, 85]	3.596	4.236
MMSE	The sum of scores of a set of questions. Lower values indicate more cognitive dysfunction.	Cognitive score	[0, 30]	3.552	1.506
FAQ	The sum of scores of 10 sets of questions. Lower values indicate less cognitive dysfunction.	Cognitive score	[0, 30]	2.264	1.415
MOCA	The sum of scores of 30 questions. Lower values indicate more cognitive dysfunction.	Cognitive score	[0, 30]	3.363	2.154
RAVLT-IR	The sum of scores from five trials in remembering a list of 15 words immediately after each trial. Lower values indicate more cognitive dysfunction.	Cognitive score	[0, 75]	1.600	5.983
CSF amyloid-beta	The concentration level of brain beta-amyloid protein. Lower values indicate more concentration.	Picograms per milliliter	$(0, \infty)$	1.591	374.4
CSF total tau and phosphorylated tau	The concentration level of neurofibrillary tangles of brain tau protein. Lower values indicate less concentration.	Picograms per milliliter	$(0, \infty)$	1.811 1.600	95.19 10.10
FDG-PET	The regional cerebral metabolic rate of glucose. Lower values indicate less activity.	Standardized uptake value ratio	$(0, \infty)$	2.995	0.104
AV45-PET	The cerebral amyloid deposition. Lower values indicate less deposition.	Standardized uptake value ratio	$(0, \infty)$	1.591	0.151
Adjusted T1-weighted brain MRI volumes of ventricles, hippocampus, whole brain, fusiform, and entorhinal cortex	The regional brain atrophies. Except in the case of ventricles, lower values indicate more atrophy.	Standardized volume residual	$(-\infty, \infty)$	2.385 1.600 2.328 1.973 2.461	0.899 0.791 0.716 0.883 0.789

of 0.931 ± 0.004 in classifying the clinical status of the test ADNI subjects per visit, which reveals the effect of modeling on classification performance.

The obtained posterior probabilities from the different classifiers can be combined using ensemble learning techniques to potentially improve prediction performance and robustness [87]. For example, by fusing the posteriors based on taking the average of the within-class posteriors over an ensemble of models from different bootstraps (bagging), the AUC of the proposed method increases to 0.934.

2.7.6 Comparison with state-of-the-art results

In order to fairly compare our results with those of state-of-the-art methods, we apply the proposed method to the TADPOLE training and test subsets of D1 and D2 using the same 16 ADNI biomarkers. The proposed model achieves an average AUC of 0.937 which is on a par with the best performance of TADPOLE with an average AUC of 0.931 [63]. Besides, our obtained average MAE of 3.93 for ADAS-13 outperforms the best reported result with an average MAE of 4.70. However, the proposed model does not perform well on the normalized ventricles compared to the best reported result with an average MAE of 0.0086 vs.0.0041.

Next, we employ the same ADNI data splits and biomarkers as used by [29] and make a head-to-head comparison with the results reported in the aforementioned study. This also enables a head-to-head comparison with both [28] and [32] based on their results reported by [29]. To do so, biomarker trajectories need to be described as a function of time-from-AD-conversion. Hence, inspired by [29], we select any subjects converting to AD and calculate the time from AD conversion using the difference between the visiting age and the age at which the first AD status is diagnosed. The corresponding DPSs are then mapped to the obtained times from the AD conversion of the selected subjects using a linear regression model. These estimates can later be used to calculate the time-from-AD-conversion for any subject's visits using the estimated DPSs. Since the time-from-AD-conversion is a linear function of DPS, i.e., $\hat{m}_0 + \hat{m}_1 s_{i,j}$, we can adjust the biomarker parameters as $b_k = b_k/\hat{m}_1$ and $c_k = \hat{m}_0 + \hat{m}_1 c_k$ to obtain biomarker trajectories as a function of time-from-AD-conversion. The obtained results indicate that the proposed model

outperforms [29] with a root-mean-square-error of 0.68 vs. 1.48; yet it has a larger maximum absolute error (4.20 vs. 3.79).

2.7.7 Generalizability across cohorts

As the final set of experiments, the generalizability of the proposed model to an independent cohort is assessed using the NACC data. First, the same configuration of logistic function and M-estimator, i.e., the modified Stannard and logistic loss is applied to model the progression of AD within NACC. Figure 2.6 depicts the modeled NACC biomarkers for 100 bootstraps. Second, the optimal model previously trained on ADNI is utilized to predict the NACC test measurements using the estimates of the common ADNI-NACC biomarkers, i.e., CDR-SB, MMSE, FAQ, MOCA, hippocampus, and whole brain. Table 2.6 compares the modeling performance of the ADNI-trained and NACC-trained models applied to the NACC test set. As it can be noticed from the obtained results, the previously selected configuration for training ADNI data is also a good choice when applied to NACC data. Moreover, the proposed model significantly ($p < 0.001$) outperforms the analogous model of [4] in all cases. Additionally, modeling performance of the proposed method degrades less than that of the analogous model of [4] when applying the ADNI-trained model to the NACC test set, which indicates the generalizability of the proposed method across cohorts. It should also be noted that the utilized NACC subset have fewer biomarkers and measurements than the used ADNI subset, which likely is the reason why it results in a smaller within-cohort modeling error.

We also apply the ADNI and NACC trained classifiers to the estimated test NACC DPSs to classify the clinical status per subject per visit. The proposed method achieves AUCs of 0.929 ± 0.012 and 0.928 ± 0.016, respectively. This reveals that diagnostic performance improves when applying the ADNI-trained model to the NACC test set.

2.8 Conclusions

In this chapter, a robust parametric model of Alzheimer's disease progression was proposed based on alternating M-estimation using the logistic loss to address potential

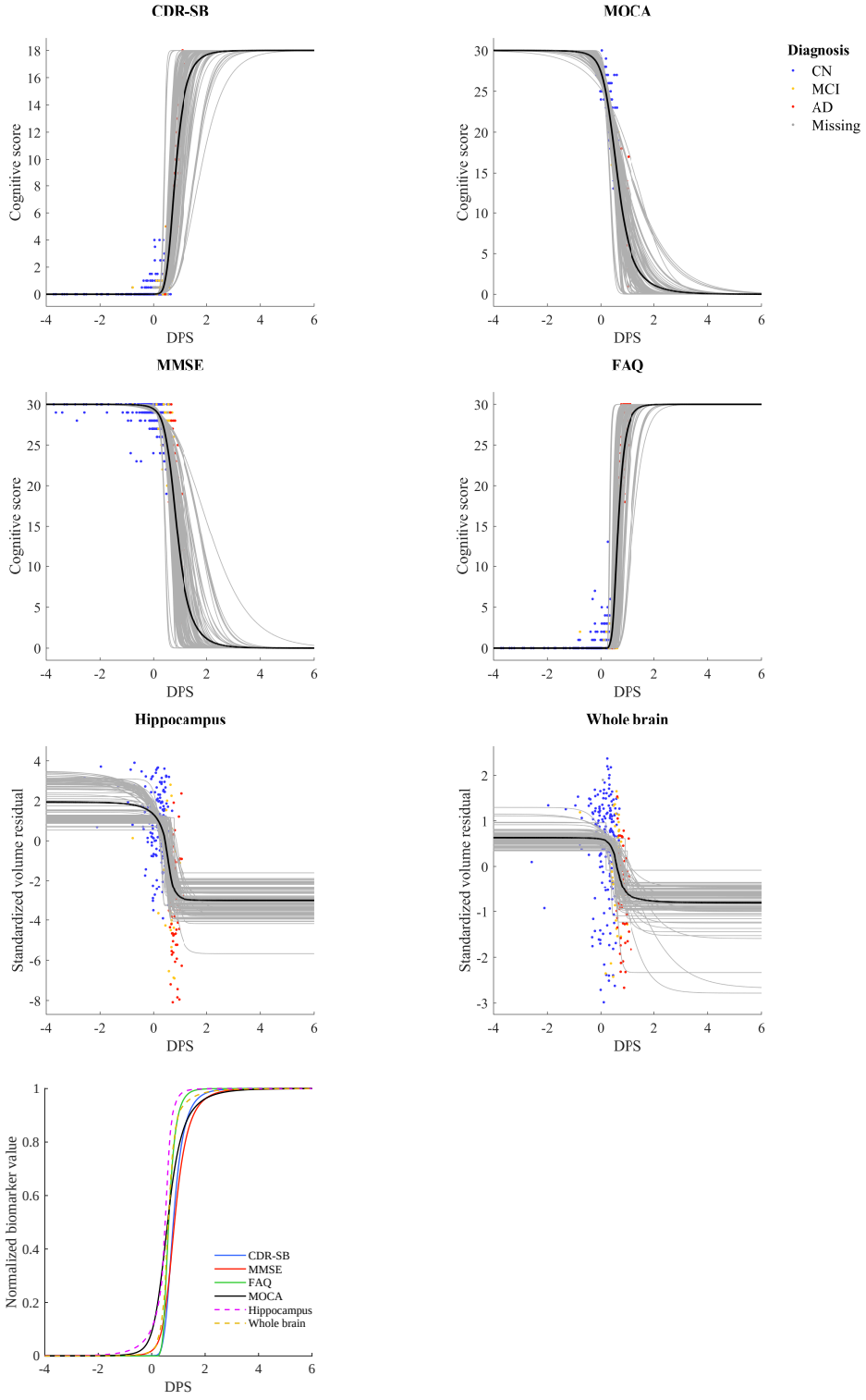


Figure 2.6: Estimated curves per bootstrap (in gray) for the NACC biomarkers using the modified Stannard function and the logistic loss. The average of the bootstrapped curves per biomarker is shown as the black curve. The last subfigure shows the average of the normalized curves of the NACC biomarkers across 100 bootstraps.

curve-fitting problems such as outliers. The proposed method linearly transformed individuals' ages to disease progression scores and jointly fitted modified Stannard functions to the longitudinal dynamics of biomarkers. The estimated parameters were then used to temporally order the biomarkers in the disease course and to predict biomarker values as well as to classify the clinical status per subject visit in an independent test set. The obtained results showed the superiority of the proposed method over the state-of-the-art results in terms of prediction performance, and this method generalized well across cohorts.

The proposed approach can be applied to different time-series data including missing data points and labels, or to biomarkers with other characteristics than the monotonic behavior that one typically encounters in, for example, neurodegenerative disease progression modeling using MRI/PET biomarkers, as long as suitable functions are used for biomarker modeling. Moreover, as an alternative to using M-estimators, resistant estimators such as the least trimmed sum of squares and least median of squares [88] with higher breakdown points can be used to fit biomarker trajectories. Though, this will result in an additional parameter to be optimized for the coverage (range) needed for trimming the residuals.

Chapter 3

The Proposed Nonparametric Method for DPM

This chapter is based on the work presented in [65, 33], where a generalized method for training LSTM networks is proposed that can handle missing values in both input and target signals, and the work presented in [89], where a robust initialization method is proposed to tackle the training instabilities in LSTM networks. The proposed training method uses the batch gradient descent algorithm in combination with the weighted input and loss function to regularize the network according to the number of available data points. Moreover, a normalized random initialization of the network weights is applied to preserve the variance of the network input and output in the same range.

3.1 The basic LSTM architecture

Figure 3.1 shows a typical schematic of a vanilla LSTM architecture. As can be seen, the topology includes a memory cell, an input modulation gate, and three nonlinear reset gates, namely input gate, forget gate, and output gate, each of which accepting current and recurrent inputs. The memory cell learns to maintain its state over time while the multiplicative gates learn to open and close access to the constant error, to prevent exploding or vanishing gradients. The input gate protects the memory contents from perturbation by irrelevant inputs, and the output gate protects other units from perturbation by currently irrelevant memory contents. The forget gate

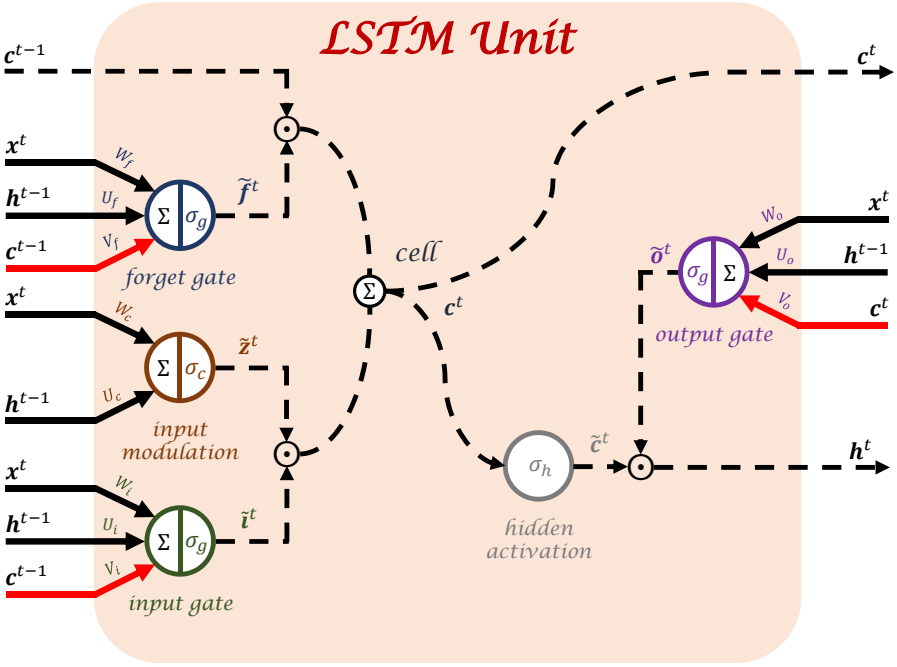


Figure 3.1: An illustration of a vanilla LSTM unit with peephole connections in red. The solid and dashed lines show weighted and unweighted connections, respectively.

deals with continual or very long input sequences, and finally, cell-to-gate (peephole) connections allow the gates to inspect the current cell state even if the output gate is closed, and consequently help to improve the performance, especially when the task involves a precise duration of intervals [90].

3.2 The proposed training algorithm for handling missing values

The proposed algorithm sets input missing values to zero, passes the input data normalized with the number of available input data points, and backpropagates zero errors corresponding to the target missing data points, where the residuals are weighted according to the number of available target data points. Figure 3.2 illustrates how the normalization factors are related to the input and output of an unfolded RNN. Note that the use of batch gradient descend ensures the availability of at least one data point per biomarker node that can proportionally contribute to the weight update rule.

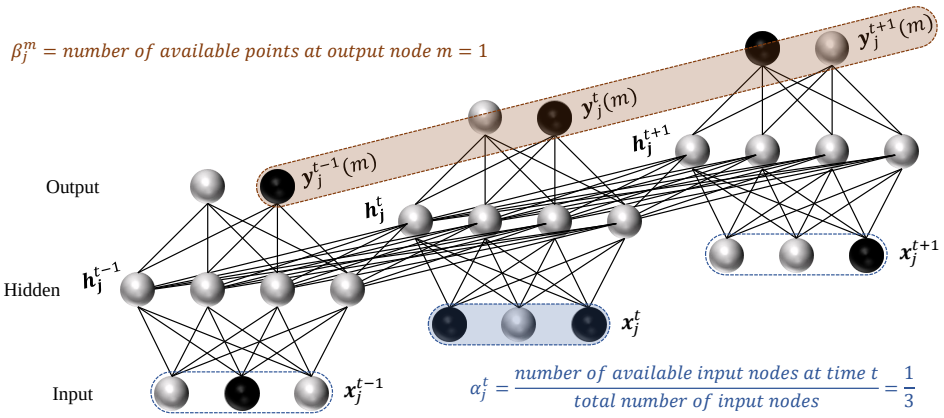


Figure 3.2: Illustration of how the normalization factors are related to the input and output of an unfolded RNN. Assume an RNN with three consecutive time points $\{t-1, t, t+1\}$, three input nodes, four hidden nodes, and two output nodes. Missing data for an instance subject j is illustrated as black nodes. We wish to weight the input vector and loss function according to the number of available data points in the input and output nodes. In this example, the subject j has only one input measurement at time t and one data point in the m-th output node. Hence, the input signal and loss function are weighted by 1/3 and 1, respectively.

3.3 Feedforward

Assume $\boldsymbol{x}_j^t \in \mathbb{R}^{N\times 1}$ is the j-th subject's sample of an N-dimensional input vector at current time t. Feedforward calculations of a peephole LSTM can be summarized as

$$\begin{aligned}
\boldsymbol{f}_j^t &= W_f[\alpha_j^t \boldsymbol{x}_j^t] + U_f \boldsymbol{h}_j^{t-1} + \boldsymbol{V}_f \odot \boldsymbol{c}_j^{t-1} + \boldsymbol{b}_f, \\
\tilde{\boldsymbol{f}}_j^t &= \sigma_g(\boldsymbol{f}_j^t), \\
\boldsymbol{i}_j^t &= W_i[\alpha_j^t \boldsymbol{x}_j^t] + U_i \boldsymbol{h}_j^{t-1} + \boldsymbol{V}_i \odot \boldsymbol{c}_j^{t-1} + \boldsymbol{b}_i, \\
\tilde{\boldsymbol{i}}_j^t &= \sigma_g(\boldsymbol{i}_j^t), \\
\boldsymbol{z}_j^t &= W_c[\alpha_j^t \boldsymbol{x}_j^t] + U_c \boldsymbol{h}_j^{t-1} + \boldsymbol{b}_c, \\
\tilde{\boldsymbol{z}}_j^t &= \sigma_c(\boldsymbol{z}_j^t), \\
\boldsymbol{c}_j^t &= \tilde{\boldsymbol{f}}_j^t \odot \boldsymbol{c}_j^{t-1} + \tilde{\boldsymbol{i}}_j^t \odot \tilde{\boldsymbol{z}}_j^t, \\
\tilde{\boldsymbol{c}}_j^t &= \sigma_h(\boldsymbol{c}_j^t), \\
\boldsymbol{o}_j^t &= W_o[\alpha_j^t \boldsymbol{x}_j^t] + U_o \boldsymbol{h}_j^{t-1} + \boldsymbol{V}_o \odot \boldsymbol{c}_j^t + \boldsymbol{b}_o, \\
\tilde{\boldsymbol{o}}_j^t &= \sigma_g(\boldsymbol{o}_j^t), \\
\boldsymbol{h}_j^t &= \tilde{\boldsymbol{o}}_j^t \odot \tilde{\boldsymbol{c}}_j^t,
\end{aligned}$$

where $\{\boldsymbol{f}^t_j, \boldsymbol{i}^t_j, \boldsymbol{z}^t_j, \boldsymbol{c}^t_j, \boldsymbol{o}^t_j, \boldsymbol{h}^t_j\} \in \mathbb{R}^{M\times 1}$ and $\{\tilde{\boldsymbol{f}}^t_j, \tilde{\boldsymbol{i}}^t_j, \tilde{\boldsymbol{z}}^t_j, \tilde{\boldsymbol{c}}^t_j, \tilde{\boldsymbol{o}}^t_j\} \in \mathbb{R}^{M\times 1}$ are the j-th sample of forget gate, input gate, modulation gate, cell state, output gate, and hidden output at time t before and after activation, respectively, and M is the number of output nodes of the LSTM unit, which is equal to N in this regression problem. The normalization factor $\alpha^t_j = |\boldsymbol{x}^t_j|/N$ is responsible for the input missing values, where $|\boldsymbol{x}^t_j|$ denotes the number of available data points (nodes) of the j-th subject at time t. Moreover, $\{W_f, W_i, W_o, W_c\} \in \mathbb{R}^{M\times N}$ and $\{U_f, U_i, U_o, U_c\} \in \mathbb{R}^{M\times M}$ are sets of connecting weights from current and recurrent inputs to the gates and cell, respectively, $\{\boldsymbol{V}_f, \boldsymbol{V}_i, \boldsymbol{V}_o\} \in \mathbb{R}^{M\times 1}$ is the set of peephole connections from the cell to the gates, $\{\boldsymbol{b}_f, \boldsymbol{b}_i, \boldsymbol{b}_o, \boldsymbol{b}_c\} \in \mathbb{R}^{M\times 1}$ represents corresponding biases of the nodes, and $\odot$ is the Hadamard product. Finally, σ_g, σ_c, and σ_h are nonlinear activation functions assigned for the gates, input modulation, and hidden output, respectively.

3.4 Backpropagation through time

Let $\mathscr{L} \in \mathbb{R}$ be the loss function defined based on the actual target $\boldsymbol{s}$ and network output $\boldsymbol{y}$. Here, we assume an L2-norm loss function and one layer of LSTM units for sequence learning which means that the network output is the hidden output. The main idea is to calculate the partial derivatives of the normalized loss function (δ) with respect to the weights using the chain rule.

$$\mathscr{L} = \frac{1}{2JM} \sum_{j,t,m} \frac{1}{\beta^m_j} \left[\boldsymbol{y}^t_j(m) - \boldsymbol{s}^t_j(m)\right]^2,$$
$$\delta \boldsymbol{y}^t_j(m) = \frac{1}{JM\beta^m_j} \left[\boldsymbol{y}^t_j(m) - \boldsymbol{s}^t_j(m)\right],$$

where $\beta^m_j = |\boldsymbol{y}_j(m)|$ is the normalization factor to handle the target missing values of the j-th subject with batch size J, and $|\boldsymbol{y}_j(m)|$ denotes the number of available data points of the m-th target node of the j-th subject. The backpropagation calculations through time using full gradients can be obtained as

$$\delta \boldsymbol{h}^t_j = U^T_f \delta \boldsymbol{f}^{t+1}_j + U^T_i \delta \boldsymbol{i}^{t+1}_j + U^T_c \delta \boldsymbol{z}^{t+1}_j + U^T_o \delta \boldsymbol{o}^{t+1}_j + \delta \boldsymbol{y}^t_j,$$

$$\delta\tilde{\boldsymbol{o}}_j^t = \delta\boldsymbol{h}_j^t \odot \tilde{\boldsymbol{c}}_j^t,$$
$$\delta\boldsymbol{o}_j^t = \delta\tilde{\boldsymbol{o}}_j^t \odot \sigma_g'(\boldsymbol{o}_j^t),$$
$$\delta\tilde{\boldsymbol{c}}_j^t = \delta\boldsymbol{h}_j^t \odot \tilde{\boldsymbol{o}}_j^t,$$
$$\delta\boldsymbol{c}_j^t = \boldsymbol{V}_f \odot \delta\boldsymbol{f}_j^{t+1} + \boldsymbol{V}_i \odot \delta\boldsymbol{i}_j^{t+1} + \boldsymbol{V}_o \odot \delta\boldsymbol{o}_j^t + \delta\tilde{\boldsymbol{c}}_j^t \odot \sigma_h'(\boldsymbol{c}_j^t) + \delta\boldsymbol{c}_j^{t+1} \odot \tilde{\boldsymbol{f}}_j^{t+1},$$
$$\delta\tilde{\boldsymbol{z}}_j^t = \delta\boldsymbol{c}_j^t \odot \tilde{\boldsymbol{i}}_j^t,$$
$$\delta\boldsymbol{z}_j^t = \delta\tilde{\boldsymbol{z}}_j^t \odot \sigma_c'(\boldsymbol{z}_j^t),$$
$$\delta\tilde{\boldsymbol{i}}_j^t = \delta\boldsymbol{c}_j^t \odot \tilde{\boldsymbol{z}}_j^t,$$
$$\delta\boldsymbol{i}_j^t = \delta\tilde{\boldsymbol{i}}_j^t \odot \sigma_g'(\boldsymbol{i}_j^t),$$
$$\delta\tilde{\boldsymbol{f}}_j^t = \delta\boldsymbol{c}_j^t \odot \boldsymbol{c}_j^{t-1},$$
$$\delta\boldsymbol{f}_j^t = \delta\tilde{\boldsymbol{f}}_j^t \odot \sigma_g'(\boldsymbol{f}_j^t),$$
$$\delta\boldsymbol{x}_j^t = W_f^T\,\delta\boldsymbol{f}_j^t + W_i^T\,\delta\boldsymbol{i}_j^t + W_c^T\,\delta\boldsymbol{z}_j^t + W_o^T\,\delta\boldsymbol{o}_j^t,$$

where T is the transpose operator, and $\sigma'(\cdot)$ is the derivative of the activation function $\sigma(\cdot)$. Finally, if $\theta \in \{f,i,z,o\}$ and $\phi \in \{f,i\}$, the gradients of the loss function with respect to the weights are calculated as

$$\delta W_\theta = \sum_{j,t} \delta\boldsymbol{\theta}_j^t [\alpha_j^t \boldsymbol{x}_j^t]^T,$$
$$\delta U_\theta = \sum_{j,t} \delta\boldsymbol{\theta}_j^{t+1} [\boldsymbol{h}_j^t]^T,$$
$$\delta\boldsymbol{V}_\phi = \sum_{j,t} \delta\boldsymbol{\phi}_j^{t+1} \odot \boldsymbol{c}_j^t,$$
$$\delta\boldsymbol{V}_o = \sum_{j,t} \delta\boldsymbol{o}_j^t \odot \boldsymbol{c}_j^t,$$
$$\delta\boldsymbol{b}_\theta = \sum_{j,t} \delta\boldsymbol{\theta}_j^t,$$

3.5 Momentum batch gradient descent

As an efficient iterative algorithm, momentum batch gradient descent is applied to find the local minimum of the loss function calculated over a batch while speeding

up the convergence. The update rule using L2 regularization can be written as

$$\vartheta^{new} = \mu\,\vartheta^{old} - \alpha(\delta\omega + \gamma\omega^{old}),$$
$$\omega^{new} = \omega^{old} + \vartheta^{new},$$

where ϑ is the weight update initialized to zero, ω is the to-be-updated weight array, $\delta\omega$ is the gradient of the loss function with respect to ω, and α, γ, and μ are the learning rate, weight decay or regularization factor, and momentum weight, respectively.

3.6 The proposed initialization for efficient training of LSTMs

To address training instability and slow convergence in LSTMs, we propose a scaled random weights initialization method that aims to keep the variance of the network input and output in the same range. Here we consider a regression problem, where $M = N$ and $\boldsymbol{h}_j^{t-1}$ is an estimation of $\boldsymbol{x}_j^t$. The regression assumptions can still be applied to sequence-to-sequence or sequence-to-label learning problems by adding a fully-connected layer with N input nodes and the desired number of output units.

Assume that the study data is complete and all of the weight matrices are independently initialized with zero-mean i.i.d. random values obtained from symmetric distributions. The goal is to derive the condition(s) on the initialization of the weights to achieve $\mathrm{Var}(\boldsymbol{h}_j^t) = \mathrm{Var}(\boldsymbol{x}_j^t)$. Since the weights are independent from the input, assuming an exact estimation for the recurrent value, i.e., $\boldsymbol{h}_j^{t-1} = \boldsymbol{x}_j^t$, and mutually independent zero-mean input features – sharing the same distribution, the variance of the forget gate can be calculated as

$$\begin{aligned}
\mathrm{Var}(\tilde{\boldsymbol{f}}_j^t) &= \mathrm{Var}(\sigma_g(W_f\boldsymbol{x}_j^t + U_f\boldsymbol{h}_j^{t-1} + \boldsymbol{V}_f \odot \boldsymbol{c}_j^{t-1} + \boldsymbol{b}_f)),\\
&= \mathrm{Var}(W_f\boldsymbol{x}_j^t + U_f\boldsymbol{h}_j^{t-1} + \boldsymbol{V}_f \odot \boldsymbol{c}_j^{t-1} + \boldsymbol{b}_f),\\
&= \mathrm{Var}(\left(W_f + U_f\right)\boldsymbol{x}_j^t) + \mathrm{Var}(\boldsymbol{V}_f \odot \boldsymbol{c}_j^{t-1}),\\
&= N\left(\mathrm{Var}(w_f) + \mathrm{Var}(u_f)\right)\mathrm{Var}(\boldsymbol{x}_j^t) + \mathrm{Var}(v_f)\mathrm{Var}(\boldsymbol{c}_j^{t-1}),
\end{aligned} \tag{3.1}$$

where w_f, u_f, and v_f are the elements of W_f, U_f, and $\boldsymbol{V}_f$, respectively. The bias in the variance calculation is canceled out as it is an independent constant initialized to zero. Moreover, the second equality holds under the assumption that σ_g is an identity function. We will discuss other commonly used functions in LSTM units in Section 3.6.1.

Variance calculations for the input, modulation, and output gates can be performed in a similar way to the forget gate. That is to say,

$$\mathrm{Var}(\tilde{\boldsymbol{i}}_j^t) = N\left(\mathrm{Var}(w_i) + \mathrm{Var}(u_i)\right)\mathrm{Var}(\boldsymbol{x}_j^t) + \mathrm{Var}(v_i)\mathrm{Var}(\boldsymbol{c}_j^{t-1}), \tag{3.2}$$

$$\mathrm{Var}(\tilde{\boldsymbol{z}}_j^t) = N\left(\mathrm{Var}(w_c) + \mathrm{Var}(u_c)\right)\mathrm{Var}(\boldsymbol{x}_j^t), \tag{3.3}$$

$$\mathrm{Var}(\tilde{\boldsymbol{o}}_j^t) = N\left(\mathrm{Var}(w_o) + \mathrm{Var}(u_o)\right)\mathrm{Var}(\boldsymbol{x}_j^t) + \mathrm{Var}(v_o)\mathrm{Var}(\boldsymbol{c}_j^t), \tag{3.4}$$

where w_i, u_i, w_c, u_c, w_o, u_o, v_i, and v_o are the elements of W_i, U_i, W_c, U_c, W_o, U_o, $\boldsymbol{V}_i$, and $\boldsymbol{V}_o$, respectively.

The cell state formula is of a form of the stochastic recurrence equation [91], also known as growing perpetuity, in which the moments of the cell state are time-varying. Therefore, one tractable way to stabilize the network training is to set $\mathrm{Var}(\boldsymbol{c}_j^t) = \mathrm{Var}(\boldsymbol{c}_j^{t-1})$. Accordingly,

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{c}_j^t) &= \mathrm{Var}(\tilde{\boldsymbol{f}}_j^t \odot \boldsymbol{c}_j^{t-1} + \tilde{\boldsymbol{i}}_j^t \odot \tilde{\boldsymbol{z}}_j^t), \\
&= \mathrm{Var}(\tilde{\boldsymbol{f}}_j^t)\mathrm{Var}(\boldsymbol{c}_j^{t-1}) + \mathrm{Var}(\tilde{\boldsymbol{i}}_j^t)\mathrm{Var}(\tilde{\boldsymbol{z}}_j^t), \\
&= \mathrm{Var}(\tilde{\boldsymbol{i}}_j^t)\mathrm{Var}(\tilde{\boldsymbol{z}}_j^t)/(1 - \mathrm{Var}(\tilde{\boldsymbol{f}}_j^t)),
\end{aligned} \tag{3.5}$$

where the above equation is obtained based on the zero-mean assumption and independence assumption between all of the gates and the cell state to avoid terms containing covariance matrices in the last expression. Note that $0 < \mathrm{Var}(\tilde{\boldsymbol{f}}_j^t) < 1$.

Finally, the variance of the network output is computed as

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{h}_j^t) &= \mathrm{Var}(\tilde{\boldsymbol{o}}_j^t \odot \sigma_h(\boldsymbol{c}_j^t)), \\
&= \mathrm{Var}(\tilde{\boldsymbol{o}}_j^t)\mathrm{Var}(\boldsymbol{c}_j^t),
\end{aligned} \tag{3.6}$$

where the last equality is obtained assuming an identity activation function and independence between the output gate and the cell state. Merging Equations (3.4) and (3.6) under the assumption that $\mathrm{Var}(\boldsymbol{h}_j^t) = \mathrm{Var}(\boldsymbol{x}_j^t) = 1$ results in a quadratic equation that can be expressed as

$$\beta_{01} + \beta_{11}\mathrm{Var}(\boldsymbol{c}_j^t) + \beta_{21}\mathrm{Var}^2(\boldsymbol{c}_j^t) = 0\,, \tag{3.7}$$

where $\beta_{01} = -1$, $\beta_{11} = N\left(\mathrm{Var}(w_o) + \mathrm{Var}(u_o)\right)$, and $\beta_{21} = \mathrm{Var}(v_o)$. Since the discriminant $\Delta_1 = \beta_{11}^2 - 4\beta_{21}\beta_{01}$ is always positive considering nonzero variances, there are two possible solutions for Equation (3.7): $\mathrm{Var}(\boldsymbol{c}_j^t) = (-\beta_{11} \pm \sqrt{\Delta_1})/(2\beta_{21})$. However, since $\beta_{21} > 0$ and $\beta_{01} < 0$, with a positive discriminant and based on the sign of the product of the roots (β_{01}/β_{21}), one of the real solutions would be negative, which cannot be accepted as $\mathrm{Var}(\boldsymbol{c}_j^t) > 0$. Therefore, the desired solution to Equation (3.7) will be obtained as

$$\mathrm{Var}(\boldsymbol{c}_j^t) = (-\beta_{11} + \sqrt{\Delta_1})/(2\beta_{21})\,. \tag{3.8}$$

Likewise, combining Equations (3.1) to (3.3) and (3.5) using the same assumptions leads to another quadratic equation that can be written as

$$\beta_{02} + \beta_{12}\mathrm{Var}(\boldsymbol{c}_j^t) + \beta_{22}\mathrm{Var}^2(\boldsymbol{c}_j^t) = 0\,, \tag{3.9}$$

where $\beta_{02} = N^2\left(\mathrm{Var}(w_i) + \mathrm{Var}(u_i)\right)\left(\mathrm{Var}(w_c) + \mathrm{Var}(u_c)\right)$, $\beta_{22} = \mathrm{Var}(v_f)$, and $\beta_{12} = N\mathrm{Var}(v_i)\left(\mathrm{Var}(w_c) + \mathrm{Var}(u_c)\right) + N\left(\mathrm{Var}(w_f) + \mathrm{Var}(u_f)\right) - 1$. The two possible solutions for Equation (3.9) will be obtained as $\mathrm{Var}(\boldsymbol{c}_j^t) = (-\beta_{12} \pm \sqrt{\Delta_2})/(2\beta_{22})$, where $\Delta_2 = \beta_{12}^2 - 4\beta_{22}\beta_{02}$ is the discriminant of the equation. Here, since $\beta_{02}, \beta_{22} > 0$, assuming a nonnegative discriminant and based on the sign of the sum and product of the roots ($-\beta_{12}/\beta_{22}$ and β_{02}/β_{22}), both real solutions could be positive and acceptable provided that $\beta_{12} < 0$. However, to achieve a simple solution for initialization, one can set $\Delta_2 = 0$ and $\beta_{12} < 0$ which produces repeated real positive roots for the

problem. Therefore, the real solution to Equation (3.9) can be obtained as

$$\mathrm{Var}(\boldsymbol{c}_j^t) = (-\beta_{12})/(2\beta_{22})\,. \tag{3.10}$$

Finally, conditions for the existence of a common solution to Equations (3.7) and (3.9) can be obtained using Equations (3.8) and (3.10) as follows

$$\begin{aligned}
&0 < \mathrm{Var}(v_i)\left(\mathrm{Var}(w_c)+\mathrm{Var}(u_c)\right) + \left(\mathrm{Var}(w_f)+\mathrm{Var}(u_f)\right) < 1/N\,,\\
&\frac{\mathrm{Var}(v_o)}{\mathrm{Var}(v_f)}\sqrt{4N^2\mathrm{Var}(v_f)\left(\mathrm{Var}(w_i)+\mathrm{Var}(u_i)\right)\left(\mathrm{Var}(w_c)+\mathrm{Var}(u_c)\right)} =\\
&\sqrt{N^2\left(\mathrm{Var}(w_o)+\mathrm{Var}(u_o)\right)^2 + 4\mathrm{Var}(v_o)} - N\left(\mathrm{Var}(w_o)+\mathrm{Var}(u_o)\right)\,.
\end{aligned} \tag{3.11}$$

Similar to the forward pass, some initialization conditions can be derived to ensure that the variance of the backpropagated gradient remains unchanged, i.e., $\mathrm{Var}(\delta \boldsymbol{h}_j^t) = \mathrm{Var}(\delta \boldsymbol{x}_j^t)$. However, as shown in [57] and [58], initialization with properly scaling the forward signal is equivalent to initialization with properly scaling the backward signal, and since the number of units in the input and output of the LSTM network are the same, similar conditions for weight initialization using backpropagation will be obtained.

3.6.1 The nonlinear activation functions in LSTM training

All the abovementioned equations are obtained based on the assumption that the activation functions are identity functions. In general, symmetric functions with zero intercepts such as the identity and hyperbolic tangent are suggested for σ_h and σ_c, respectively, and logistic sigmoid is suggested for σ_g [90]. Both the hyperbolic tangent and logistic sigmoid are nonlinear symmetric functions that can be linearly approximated using a Taylor series expansion. The former has a zero intercept and its expansion about zero leads to an identity function ($\sigma_c(x) \approx x$). The latter, however, has a nonzero intercept and its Taylor series about zero is approximated as $\sigma_g(x) \approx 0.5 + 0.25x$. Therefore, the sigmoid function approximately increases the input signal mean by $1/2$ and scales its variance by $1/16$. Note that the nonzero mean value of the sigmoid can induce important singular values in the Hessian

matrix, resulting in saturation of the top layers and prohibition of gradients to flow backward to learn useful features in the lower layers [57]. Using the suggested activation functions in the gates, the variance calculations for the peephole LSTM network are updated as follows based on the aforementioned Taylor series expansion

$$\begin{aligned}
\mathrm{Var}(\tilde{\boldsymbol{f}}_j^t) &= N\left(\mathrm{Var}(w_f)+\mathrm{Var}(u_f)\right)\mathrm{Var}(\boldsymbol{x}_j^t)/16+\mathrm{Var}(v_f)\mathrm{Var}(\boldsymbol{c}_j^{t-1})/16\,,\\
\mathrm{Var}(\tilde{\boldsymbol{i}}_j^t) &= N\left(\mathrm{Var}(w_i)+\mathrm{Var}(u_i)\right)\mathrm{Var}(\boldsymbol{x}_j^t)/16+\mathrm{Var}(v_i)\mathrm{Var}(\boldsymbol{c}_j^{t-1})/16\,,\\
\mathrm{Var}(\tilde{\boldsymbol{z}}_j^t) &= N\left(\mathrm{Var}(w_c)+\mathrm{Var}(u_c)\right)\mathrm{Var}(\boldsymbol{x}_j^t)\,,\\
\mathrm{Var}(\tilde{\boldsymbol{o}}_j^t) &= N\left(\mathrm{Var}(w_o)+\mathrm{Var}(u_o)\right)\mathrm{Var}(\boldsymbol{x}_j^t)/16+\mathrm{Var}(v_o)\mathrm{Var}(\boldsymbol{c}_j^t)/16\,,\\
\mathrm{Var}(\boldsymbol{c}_j^t) &= \mathrm{Var}(\boldsymbol{c}_j^{t-1}) = (\mathrm{Var}(\tilde{\boldsymbol{i}}_j^t)+0.25)\mathrm{Var}(\tilde{\boldsymbol{z}}_j^t)/(0.75-\mathrm{Var}(\tilde{\boldsymbol{f}}_j^t))\,,
\end{aligned}$$

where the last equation is obtained bearing in mind that $\mathrm{Var}(xy)=\mathrm{Var}(x)\mathrm{Var}(y)+\mathbb{E}^2(x)\mathrm{Var}(y)+\mathbb{E}^2(y)\mathrm{Var}(x)$ for two independent random variables x and y, and considering $\mathbb{E}(\tilde{\boldsymbol{z}}_j^t)=0$, $\mathbb{E}(\tilde{\boldsymbol{f}}_j^t)=\mathbb{E}(\tilde{\boldsymbol{i}}_j^t)=0.5$, and, hence, $\mathbb{E}(\boldsymbol{c}_j^t)=\mathbb{E}(\boldsymbol{c}_j^{t-1})=0$. Here also using Equation (3.6), two quadratic equations can be obtained similar to Equations (3.7) and (3.9), where $\beta_{01}=-16$, $\beta_{11}=N\left(\mathrm{Var}(w_o)+\mathrm{Var}(u_o)\right)$, $\beta_{21}=\mathrm{Var}(v_o)$, $\beta_{02}=N\left(\mathrm{Var}(w_c)+\mathrm{Var}(u_c)\right)\left(N\left(\mathrm{Var}(w_i)+\mathrm{Var}(u_i)\right)+4\right)$, $\beta_{22}=\mathrm{Var}(v_f)$, and $\beta_{12}=N\mathrm{Var}(v_i)\left(\mathrm{Var}(w_c)+\mathrm{Var}(u_c)\right)+N\left(\mathrm{Var}(w_f)+\mathrm{Var}(u_f)\right)-12$. Likewise, conditions for the existence of a common solution to Equations (3.7) and (3.9) can be obtained using Equations (3.8) and (3.10) as follows

$$\begin{aligned}
&0<\mathrm{Var}(v_i)\left(\mathrm{Var}(w_c)+\mathrm{Var}(u_c)\right)+\left(\mathrm{Var}(w_f)+\mathrm{Var}(u_f)\right)<12/N\,,\\
&\frac{\mathrm{Var}(v_o)}{\mathrm{Var}(v_f)}\sqrt{4N\mathrm{Var}(v_f)\left(\mathrm{Var}(w_c)+\mathrm{Var}(u_c)\right)\left(N\left(\mathrm{Var}(w_i)+\mathrm{Var}(u_i)\right)+4\right)}=\\
&\sqrt{N^2\left(\mathrm{Var}(w_o)+\mathrm{Var}(u_o)\right)^2+64\mathrm{Var}(v_o)}-N\left(\mathrm{Var}(w_o)+\mathrm{Var}(u_o)\right)\,.
\end{aligned}\tag{3.12}$$

3.7 Experiments and results

3.7.1 Data

The data used in this study is obtained from the ADNI-based datasets of the TADPOLE challenge after performing the same data filtering and preprocessing steps

mentioned in Section 2.6.1. The third visiting month is excluded from the obtained data to confine the matched time points to half-yearly regular follow-ups including baseline. Finally, subjects with less than two consecutive visits are removed to ensure the possibility of sequence learning through the feedforward and backpropagation steps. This results in 16 ADNI biomarkers acquired from 1,400 subjects (789 males and 611 females) in 8,133 visits, where 82% of the actual data is missing.

3.7.2 Experimental setup

For evaluation purposes, since a large amount of data is missing (82%), we initially confine the matched time points to yearly regular visits by excluding the sixth and 18th visiting months and use a subgroup of subjects that have at least one available measurement per biomarker during all visits. This results in 16 ADNI biomarkers acquired from 582 subjects (322 males and 260 females) in 3,031 visits, where 68% of the actual data is missing. The entire dataset is partitioned into three non-overlapping subsets for training, validation, and testing. More specifically, based on the first and last available diagnoses of subjects, i.e., CN-CN, CN-MCI, ..., AD-AD, we divide each of these types of pairs into two groups including few and many visits using the median number of visits as threshold and randomly select 20% of the subjects from each group for testing and the same amount for validation.

A one-layer peephole LSTM is used with an identity function, hyperbolic tangent, and logistic sigmoid as activation functions for σ_h, σ_c, and σ_g, respectively. The hidden output is used as the network output with the same number of input nodes for the regression of 16 biomarker values over time. The network biases are initialized to zero, and values of the weight matrices are drawn from the zero-mean i.i.d. Gaussian distributions using Equation (3.12) with the following variances suggested in [89]: $\mathrm{Var}(v_f) = \mathrm{Var}(v_i) = \mathrm{Var}(v_o) = 1$, $\mathrm{Var}(w_f) = \mathrm{Var}(w_c) = \mathrm{Var}(u_c) = 1/(4N)$, $\mathrm{Var}(u_f) = 3/(4N)$, $\mathrm{Var}(w_i) = 1/N$, $\mathrm{Var}(u_i) = 3/N$, $\mathrm{Var}(w_o) = 2/N$, and $\mathrm{Var}(u_o) = 4/N$.

The input data is standardized to have a zero mean and unit variance per feature dimension, and the first to penultimate time points are utilized to estimate the second to last visits using the following methods:

- LSTM-Proposed: an LSTM network trained based on the proposed training algorithm [33] by setting input missing values to zero and backpropagating zero errors corresponding to the target missing points while training.
- LSTM-Mean: an LSTM network trained using the standard backpropagation through time algorithm with missing values imputed based on the mean imputation method before training [92].
- LSTM-Forward: an LSTM network trained using the standard backpropagation through time algorithm with missing values imputed based on the forward imputation method before training [52].

The batch size is set to the number of available training subjects and the validation set is used to tune all the networks' optimization parameters, each time by adjusting one of the parameters while keeping the rest at fixed values to achieve the lowest validation set error. Based on these strategies, the optimal parameters are obtained as $\alpha = 0.5$, $\mu = 0.9$, and $\gamma = 0.001$ with 500 epochs.

3.8 Results and discussion

3.8.1 Modeling biomarkers

Table 3.1 compares the test modeling performance (MAE) using the utilized methods. Even though the performance is reported per biomarker, the models are jointly fitted to all biomarkers. As can be deduced from Table 3.1, LSTM-Proposed and LSTM-Forward significantly outperform LSTM-Mean in all cases with $p < 0.01$ using the paired, two-sided Wilcoxon signed-rank test. In general, the proposed method performs better than the LSTM-Forward approach in modeling MRI biomarkers and cognitive tests while the LSTM-Forward is superior in predicting CSF and PET measurements.

3.8.2 Classifying clinical status

To assess the ability of the estimated measurements in classifying the clinical status, we train a linear discriminant analysis (LDA) classifier using the estimated training measurements and apply it to the estimated test data to compute the posterior

Table 3.1: Test modeling performance of different methods as MAE for yearly predictions of the utilized ADNI biomarkers. All the MAEs are significantly different ($p < 0.01$).

	LSTM-Proposed [33]	LSTM-Forward [52]	LSTM-Mean [92]
CDR-SB	0.920	0.843	1.086
ADAS-13	4.098	3.734	4.971
MMSE	1.412	1.441	1.728
FAQ	2.265	2.385	3.028
MOCA	1.900	2.008	2.648
RAVLT-IR	5.379	5.505	6.924
Amyloid-beta	270.8	217.5	525.6
Total tau	53.82	38.64	95.10
Phosphorylated tau	5.860	3.866	10.05
FDG-PET	0.075	0.069	0.125
AV45-PET	0.097	0.082	0.185
Ventricles	0.338	0.353	0.708
Hippocampus	0.465	0.414	0.689
Whole brain	0.372	0.411	0.671
Fusiform	0.461	0.474	0.782
Entorhinal cortex	0.551	0.573	0.885

Table 3.2: Test diagnostic performance of different methods as AUC using an LDA classifier applied to the yearly estimated biomarker values.

	LSTM-Proposed [33]	LSTM-Forward [52]	LSTM-Mean [92]
AUC	0.723	0.734	0.725

probabilities. The obtained scores are then used to calculate diagnostic AUCs. The diagnostic performances on the utilized test set are shown in Table 3.2. As can be seen, although the obtained results are almost in the same range, LSTM-Forward outperforms all other methods in classifying the clinical status of the subjects per visit with a multiclass AUC of 0.734. One could of course use other classifiers or train the LSTM network directly for classification based on sequence-to-label learning to potentially improve the diagnostic AUCs. However, the focus of this work is on DPM based on sequence-to-sequence learning. Besides, sequence-to-label learning would only be able to utilize the part of the training data which has available clinical status.

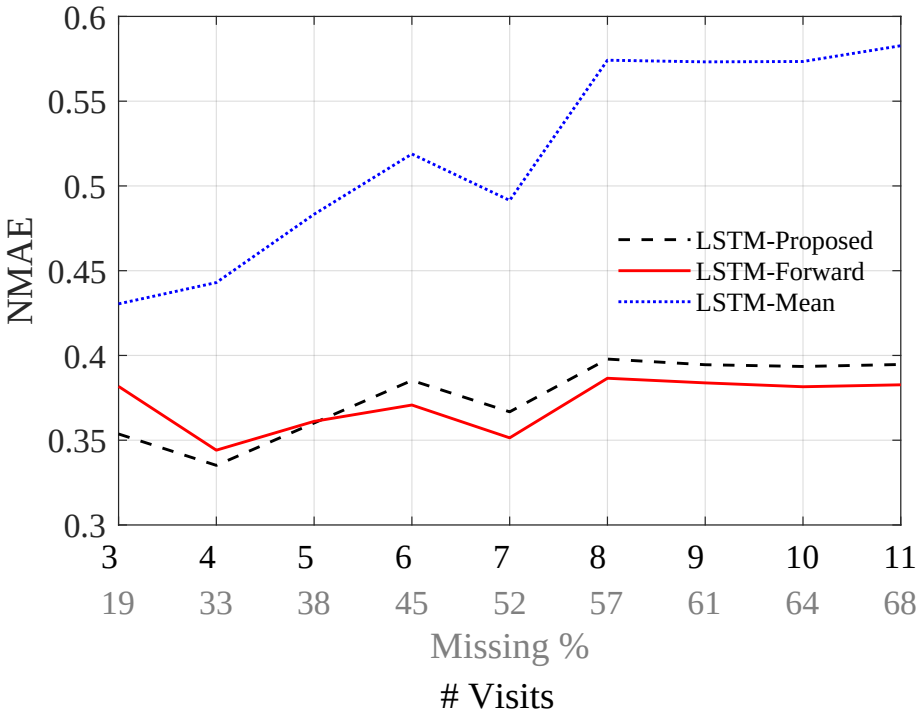


Figure 3.3: Test modeling performance of different methods for various amounts of data.

3.8.3 Robustness to missing values

To evaluate the modeling robustness of the proposed method compared to the alternatives with different amounts of missing data and number of visits, we construct subsamples of the training dataset by respectively removing later follow-ups of the subjects which include more missing values and train the methods on the smaller datasets. Figures 3.3 and 3.4 illustrates the modeling and diagnostic performances of the utilized methods on various amounts of missing measurements, and accordingly, the number of visits. As can be seen, both modeling and diagnostic performances of the proposed method are superior to those of the benchmarks in lower rates of missing data (up to 40%), and LSTM-Forward performs slightly better in higher missing rates.

For higher rates of missing data, we utilize the larger subset obtained from 1,400 subjects, where 82% of the actual data is missing, for half-yearly predictions. Tables 3.3 and 3.4 compare the modeling performance and diagnostic performance of the utilized methods on the test ADNI subsets, respectively. As can be seen,

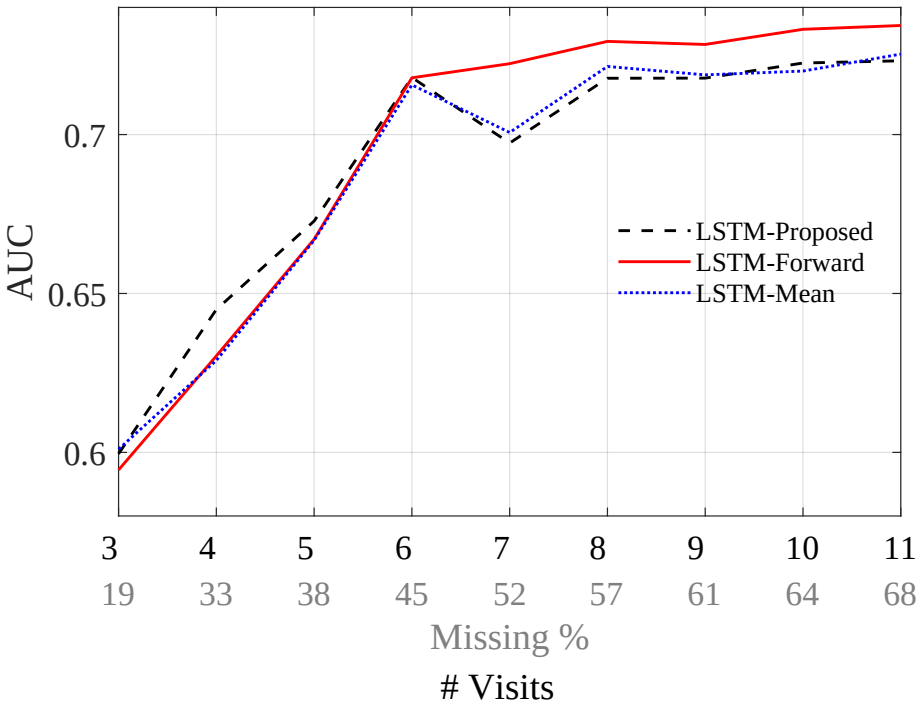


Figure 3.4: Test diagnostic performance of different methods for various amounts of data.

the standard LSTM with the forward imputation outperforms the other methods in all cases with $p < 0.01$ except for the FAQ, MOCA, and entorhinal cortex where the LSTM-Proposed works better. One reason why LSTM-Forward is robust to the higher rates of missing data could be due to providing more information for training by filling the missing months (30, 42, 54, 66, 78, 90, 102, and 114) using the forward imputation while replacing the missing values placed at the beginning of the biomarker sequences with the median of the available data.

3.9 Conclusions

In this chapter, a training algorithm was proposed for LSTM networks aiming to improve robustness against missing data. Moreover, a robust initialization method was proposed for LSTM networks to address training instability and slow convergence. The trained LSTM network was applied to AD progression modeling using longitudinal measurements of 16 ADNI biomarkers. This was the first time RNNs have been studied and applied to DPM within a neurodegenerative disease. Besides, since RNNs are nonparametric learning methods, the proposed approach can be applied

Table 3.3: Test modeling performance of different methods as MAE for half-yearly predictions of the utilized ADNI biomarkers. All the MAEs are significantly different ($p < 0.01$).

	LSTM-Proposed [33]	LSTM-Forward [52]	LSTM-Mean [92]
CDR-SB	1.039	0.970	1.645
ADAS-13	4.311	3.946	6.847
MMSE	1.647	1.614	2.458
FAQ	2.692	2.709	5.039
MOCA	2.153	2.239	3.252
RAVLT-IR	5.481	5.421	8.610
Amyloid-beta	275.4	170.9	503.9
Total tau	47.55	24.27	84.69
Phosphorylated tau	5.235	2.871	9.556
FDG-PET	0.081	0.061	0.117
AV45-PET	0.110	0.073	0.195
Ventricles	0.335	0.310	0.707
Hippocampus	0.463	0.407	0.937
Whole brain	0.367	0.360	0.714
Fusiform	0.474	0.469	0.892
Entorhinal cortex	0.614	0.617	1.024

Table 3.4: Test diagnostic performance of different methods as AUC using an LDA classifier applied to the half-yearly estimated biomarker values.

	LSTM-Proposed [33]	LSTM-Forward [52]	LSTM-Mean [92]
AUC	0.722	0.741	0.737

to different time-series data and characteristics than the monotonic behavior that one typically encounters in neurodegenerative disease progression modeling. The proposed training method demonstrated better performance than using imputation prior to standard LSTM network training in terms of biomarker value prediction, especially when trained on data with lower rates of missing values. The proposed methods are applicable for other types of RNNs such as gated recurrent units (GRUs) [45]. This study highlights the potential of RNNs for modeling the progression of AD using longitudinal measurements, provided that proper care is taken to handle missing values and time intervals.

Chapter 4

Comparison of the Two Proposed Methods for DPM

This chapter is based on the work presented in [66, 67], where the decline prediction of cognitive test scores in stable and converting MCI subjects is investigated using both nonparametric [33] and parametric [64] AD progression modeling methods. Moreover, a comprehensive study is done to compare the two proposed methods for disease progression modeling from different aspects in terms of biomarker value prediction and clinical status classification.

4.1 Biomarker value prediction

The same data subset obtained in Section 3.7.1 is used to evaluate the ability of the two proposed methods in predicting half-yearly matched 16 ADNI biomarker values. Table 4.1 compares the modeling performance of the two proposed methods on the test ADNI subsets. As can be seen, the nonparametric method outperforms the parametric method in all cases with $p < 0.01$ except for the CDR-SB and FAQ where the proposed parametric method results in lower prediction errors.

4.2 Clinical status classification

To assess the ability of the estimated measurements in classifying the clinical status, we train an LDA classifier using the estimated training measurements and apply it to the estimated test data to compute the posterior probabilities. The obtained scores

Table 4.1: Test modeling performance of the two proposed methods as MAE for half-yearly predictions of the utilized ADNI biomarkers. All the MAEs are significantly different ($p < 0.01$).

	LSTM-Proposed [33]	Regression-Proposed [64]
CDR-SB	1.039	0.570
ADAS-13	4.311	4.432
MMSE	1.647	1.657
FAQ	2.692	1.506
MOCA	2.153	2.229
RAVLT-IR	5.481	6.428
Amyloid-beta	275.4	419.5
Total tau	47.55	85.09
Phosphorylated tau	5.235	9.285
FDG-PET	0.081	0.092
AV45-PET	0.110	0.156
Ventricles	0.335	0.813
Hippocampus	0.463	0.865
Whole brain	0.367	0.795
Fusiform	0.474	0.870
Entorhinal cortex	0.614	0.926

Table 4.2: Test diagnostic performance of the two proposed methods as AUC using an LDA classifier applied to the half-yearly estimated biomarker values. All the AUCs are significantly different ($p < 0.05$).

	LSTM-Proposed [33]	Regression-Proposed [64]
AUC	0.722	0.926

are then used to calculate diagnostic AUCs. The diagnostic performances on the test set are shown in Table 4.2. As can be seen, the proposed parametric method significantly outperforms the nonparametric method in classifying the clinical status of the subjects per visit with $p < 0.05$. Note that the proposed methods are compared with each other using McNemar's test [93] applied to the hard classification results (clinical status) obtained from the LDA classifier. One reason why the regression-based method is performing significantly better in the classification task could be because of using age information while modeling the biomarker trajectories and estimating DPSs. Moreover, the biomarker modeling results shown in both Figure

2.1 and Table 4.1 reveal that the parametric method is performing well in fitting and estimating the cognitive scores, which are very important features for clinical status classification, compared to other biomarkers.

4.3 The effects of different modalities on the performance

To see how utilizing different modalities can affect the prediction and classification performances, we repeat the previously conducted experiments each time using a subset of biomarkers from MRI volumetric measures (ventricles, hippocampus, whole brain, fusiform, and entorhinal cortex), PET scan measures (FDG-PET and AV45-PET), CSF measures (Amyloid beta, total tau, and phosphorylated tau), and cognitive tests (CDR-SB, ADAS-13, MMSE, FAQ, MOCA, and RAVLT-IR). Tables 4.3 and 4.4 illustrate the modeling and diagnostic performances of the two methods on the test set using different biomarkers.

As can be seen in Table 4.3, the proposed nonparametric method significantly outperforms the parametric method in predicting the biomarker values in all cases with $p < 0.01$ except for the cognitive tests where the parametric method achieves the best result. Also, compared to the parametric method, the nonparametric method is more successful in modeling and prediction of MRI biomarkers. The ability of the proposed nonparametric method in modeling six volumetric MRI biomarkers was already seen in [33]. It can also be deduced from Table 4.4 that a combination of

Table 4.3: Test modeling performance of the two proposed methods as NMAE for half-yearly predictions of the utilized ADNI biomarkers. All the NMAEs are significantly different ($p < 0.01$).

	LSTM-Proposed [33]	Regression-Proposed [64]
Cognitive	0.352	0.335
MRI	0.348	0.464
Cognitive & MRI	0.361	0.462
Cognitive & CSF	0.367	0.371
Cognitive & PET	0.377	0.385
Cognitive & MRI & CSF & PET	0.379	0.487

Table 4.4: Test diagnostic performance of the two proposed methods as AUC using an LDA classifier applied to the half-yearly estimated biomarker values. All the AUCs are significantly different ($p < 0.05$).

	LSTM-Proposed [33]	Regression-Proposed [64]
Cognitive	0.735	0.940
MRI	0.603	0.659
Cognitive & MRI	0.737	0.944
Cognitive & CSF	0.734	0.932
Cognitive & PET	0.727	0.934
Cognitive & MRI & CSF & PET	0.722	0.926

cognitive tests and MRI biomarkers results in the best diagnostic performance in both modeling methods. The proposed parametric method significantly outperforms the nonparametric method in all cases with $p < 0.05$. Other modalities, i.e., PET and CSF, have fewer biomarkers and available measurements, which could be a reason for declining the overall performance.

4.4 Robustness to missing values

To evaluate the modeling robustness of the two proposed methods compared to each other with different amounts of missing data, we utilize the same smaller data subsets used in Section 3.8.3 and repeat the same experiments with the parametric method. Figures 4.1 and 4.2 display the modeling and diagnostic performances of the utilized methods on various amounts of missing measurements, and accordingly, the number of visits.

The obtained results indicate that both methods are robust to the various amounts of missing data and the number of visits. However, the nonparametric method performs significantly better in predicting the biomarker values while the parametric method achieves remarkably better diagnostic performance.

4.5 Cognitive decline prediction using few visits

The objective is to investigate the decline prediction of cognitive test scores in stable MCI (sMCI) and converting MCI (cMCI) subjects using both nonparametric and parametric Alzheimer's disease progression modeling methods trained on data from

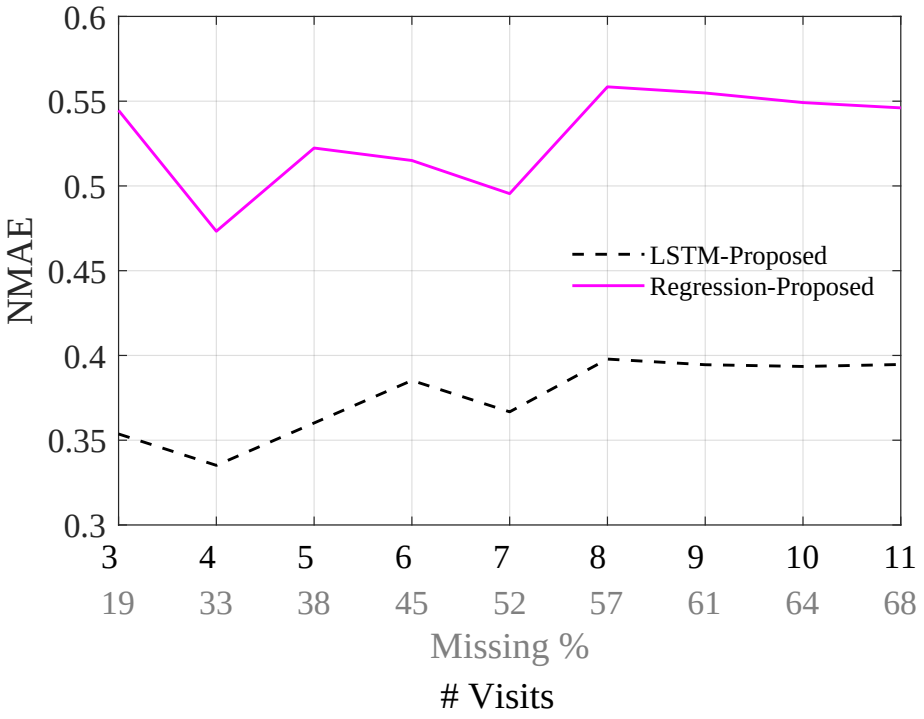


Figure 4.1: Test modeling performance of different methods for various amounts of data.

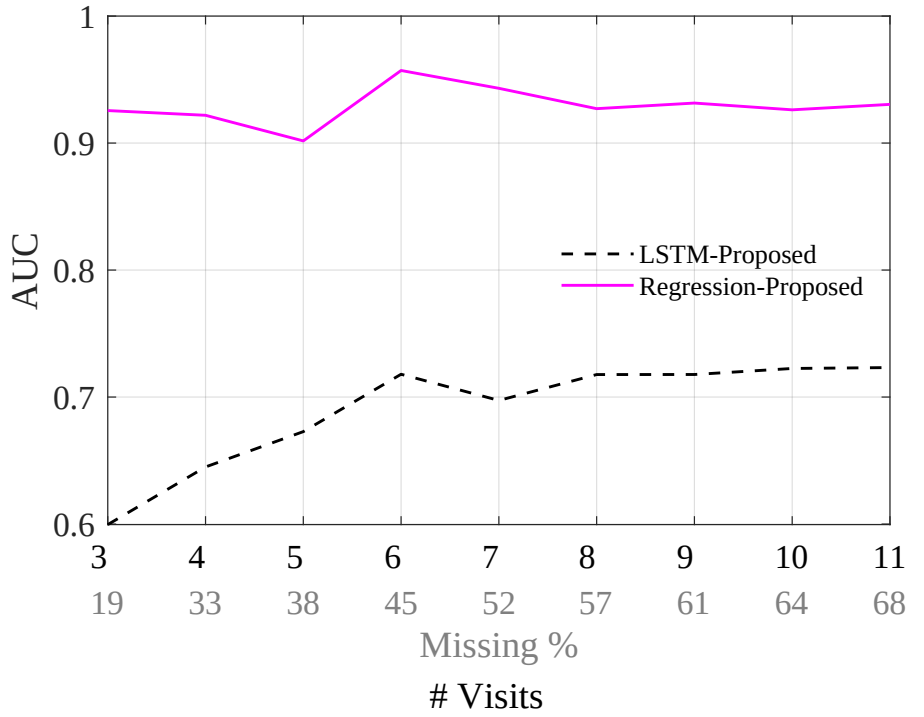


Figure 4.2: Test diagnostic performance of different methods for various amounts of data.

multiple modalities. To do so, the proposed methods were trained on the same previously utilized data and subsequently applied to predict month 18 to 60 cognitive scores of the test subjects using at most their baseline, month 6, and month 12 data. Figures 4.3 and 4.4 display the cognitive test prediction results for the test subjects per visit using the proposed nonparametric and parametric methods.

Moreover, the corresponding test prediction NMAEs per visit are reported in Tables 4.5 and 4.6. As can be seen, in almost all cases, the nonparametric method outperforms the parametric model in predicting the cognitive scores. Moreover, predictions from both nonparametric and parametric methods can significantly discriminate between sMCI and cMCI groups using a two-sample t-test with $p < 0.01$ and $p < 0.001$, respectively. Though, the discrimination capability of the nonparametric method is superior in the long-term prediction of cognitive decline.

4.6 Conclusions

In this chapter, a comprehensive study was performed to compare the two proposed methods for AD progression modeling on the ADNI dataset. It was shown that the nonparametric method outperformed the parametric method in predicting biomarker values and cognitive decline, even when using a few time points of the test subjects. On the other hand, the parametric method performed significantly better in classifying clinical status.

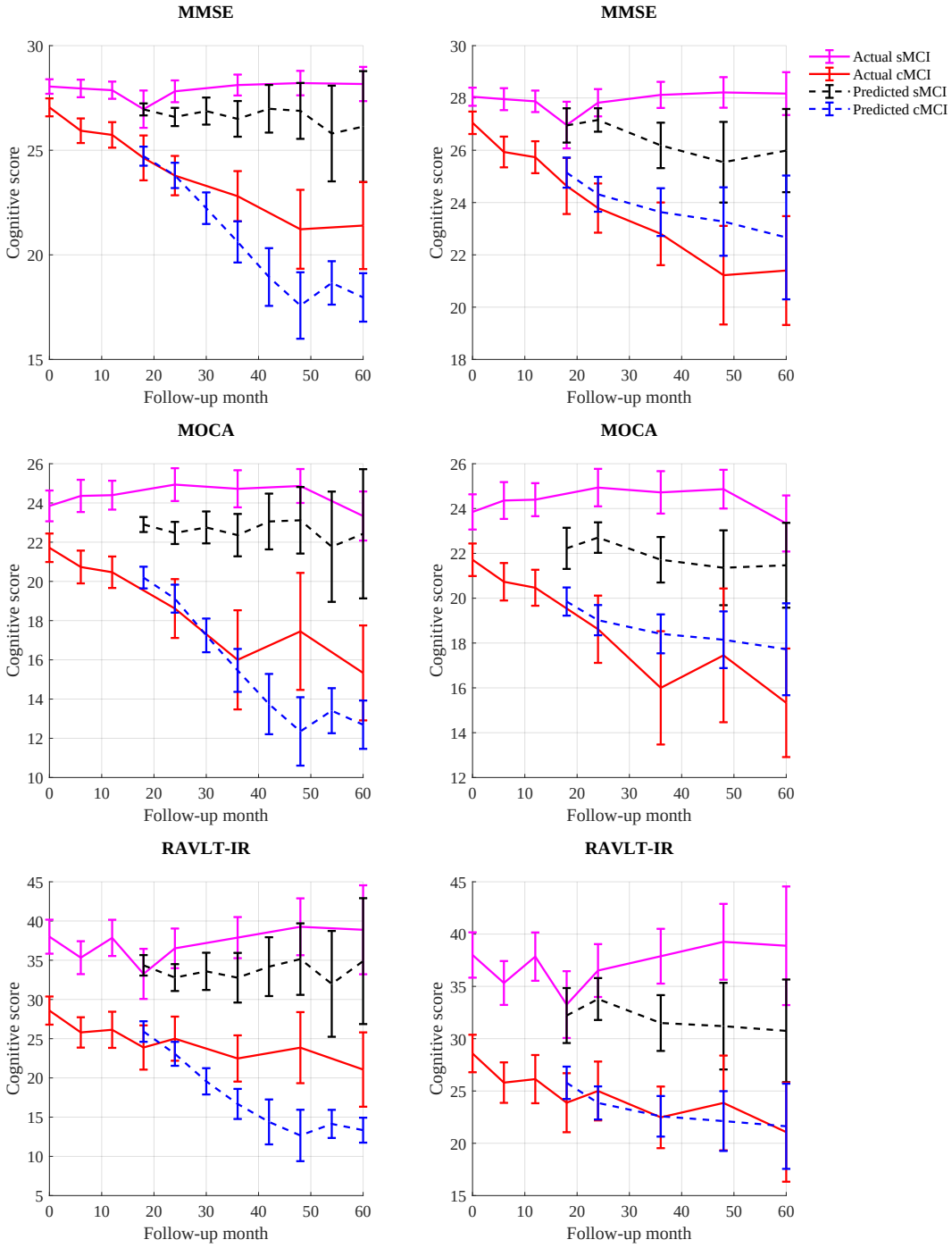


Figure 4.3: Cognitive test prediction results for the test subjects per visit using the nonparametric (left) and parametric (right) methods. The error bars are calculated based on a 95% confidence interval for population standard deviation per visit.

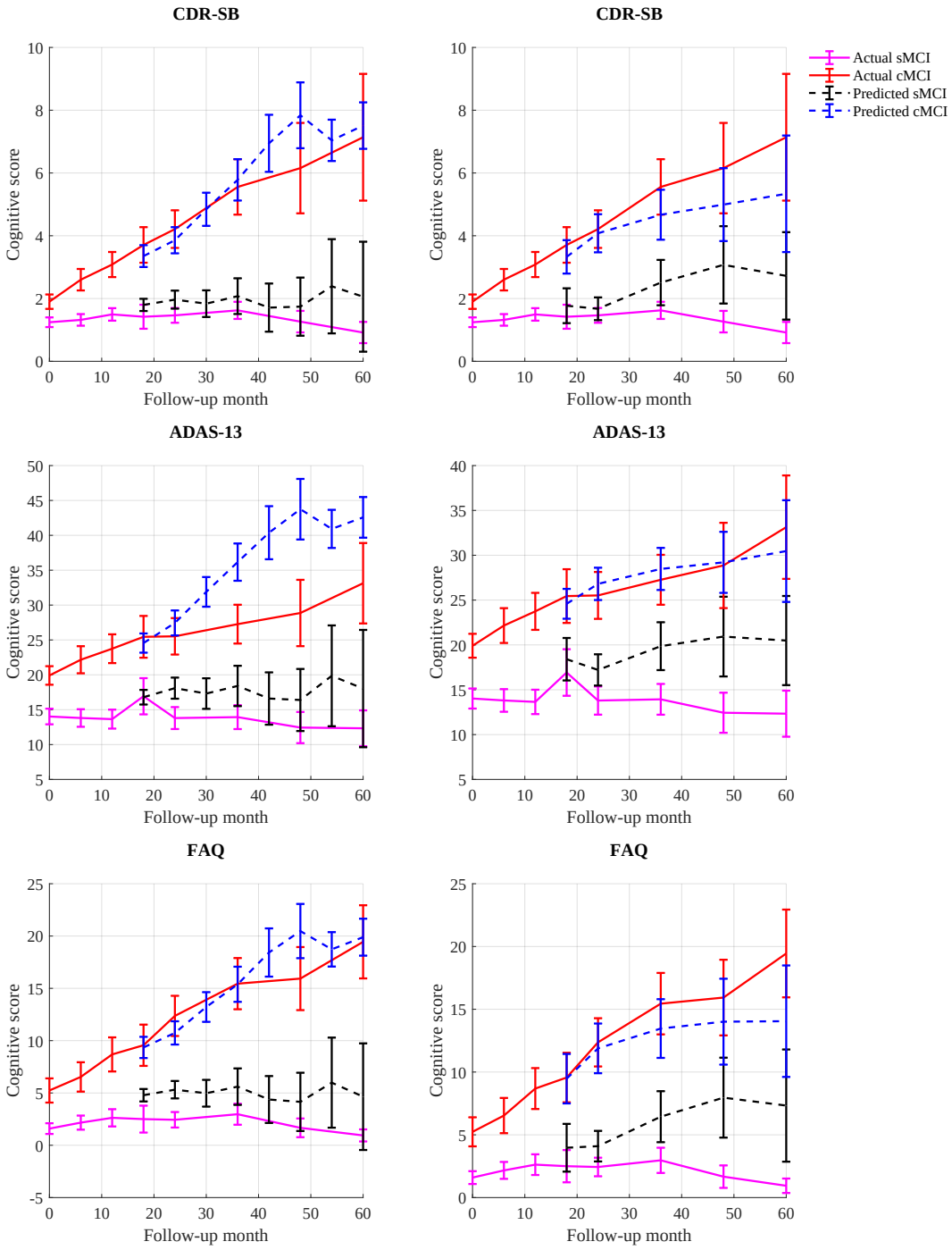


Figure 4.4: Cognitive test prediction results for the test subjects per visit using the nonparametric (left) and parametric (right) methods. The error bars are calculated based on a 95% confidence interval for population standard deviation per visit.

Table 4.5: Test prediction NMAEs (mean±SD) per visit using the proposed nonparametric method.

	Month	18	24	36	48	60
cMCI	CDR-SB	0.988±0.893	1.354±1.080	1.928±1.575	3.694±2.414	3.359±2.867
	ADAS-13	4.447±3.967	4.919±3.987	9.670±6.551	16.62±11.65	12.35±10.29
	MMSE	1.842±1.656	2.509±1.999	3.323±2.874	5.991±4.089	4.504±3.749
	FAQ	3.510±2.775	5.166±3.167	5.850±4.698	8.467±5.798	5.784±6.065
	MOCA		2.310±1.897	3.097±2.131	6.320±3.869	5.172±2.775
	RAVLT-IR	5.360±4.289	6.826±4.615	8.196±5.795	12.99±9.418	9.935±7.721
sMCI	CDR-SB	0.662±0.481	0.938±0.664	1.800±1.219	2.710±1.368	3.057±1.704
	ADAS-13	3.876±3.422	5.909±3.511	10.39±6.295	12.04±7.137	13.73±8.552
	MMSE	1.729±0.952	2.046±1.343	3.124±1.852	3.957±2.333	4.402±3.002
	FAQ	2.891±1.723	3.685±1.957	6.049±3.232	8.484±4.042	8.871±4.398
	MOCA		2.846±1.807	3.601±2.183	4.432±2.996	5.483±2.976
	RAVLT-IR	5.079±3.646	7.097±5.504	12.03±7.873	10.97±9.039	13.53±7.414

The blanks indicate that the subjects have no data points in the visits.

Table 4.6: Test prediction NMAEs (mean±SD) per visit using the proposed parametric method.

	Month	18	24	36	48	60
cMCI	CDR-SB	0.956±1.043	1.272±1.068	2.376±1.779	3.173±2.759	3.589±2.914
	ADAS-13	6.424±4.746	6.255±4.231	8.245±6.293	10.37±8.321	9.372±7.852
	MMSE	2.156±2.104	2.820±2.364	3.418±3.016	4.881±3.825	3.831±2.969
	FAQ	3.085±3.507	4.037±3.557	6.018±5.209	7.076±6.365	7.852±5.756
	MOCA		2.427±1.337	3.587±2.977	4.876±4.206	4.347±3.306
	RAVLT-IR	7.266±4.596	8.126±5.331	8.835±5.972	11.55±7.156	8.669±4.698
sMCI	CDR-SB	0.600±1.097	0.860±1.017	1.614±2.200	2.627±3.438	2.450±2.548
	ADAS-13	5.251±4.024	5.741±5.266	8.586±9.112	10.01±11.87	9.151±7.956
	MMSE	1.814±1.317	1.854±1.647	2.725±3.083	3.546±4.117	2.771±3.057
	FAQ	1.894±2.435	3.057±4.245	5.247±6.461	7.991±9.520	7.362±9.190
	MOCA		2.591±2.130	3.721±3.194	4.446±4.402	4.222±3.161
	RAVLT-IR	5.592±4.494	6.903±5.180	9.377±9.032	10.64±10.54	12.17±9.006

The blanks indicate that the subjects have no data points in the visits.

Chapter 5

Conclusion

In this work, two different methods were proposed for modeling the progression of Alzheimer's disease using longitudinal measurements of ADNI and NACC biomarkers based on a parametric method robust to outliers and missing data and a nonparametric method robust to missing values and training instabilities.

5.1 Summary

The proposed parametric method linearly mapped the individual's age to a disease progression score (DPS) and jointly fitted constrained modified Stannard functions to the longitudinal dynamics of biomarkers as functions of the DPS based on alternating M-estimation using the logistic loss. The estimated parameters were then used to temporally order the biomarkers in the disease course and to predict biomarker values as well as to classify the clinical status per subject visit in an independent test set. The obtained results showed the superiority of the proposed method over the state-of-the-art methods in terms of prediction and classification performances, and this method generalized well across cohorts.

The proposed nonparametric method applied a generalized training rule based on normalization of the input and loss to the number of available data points to LSTMs to handle missing input and target values. Moreover, a robust initialization method was developed to address the training instability and slow convergence in LSTM networks based on a scaled random initialization of the network weights, aiming at preserving the variance of the network input and output in the same

range. This was the first time LSTMs were studied and applied to DPM within a neurodegenerative disease. The results showed that the proposed deep learning-based training algorithm achieves superior results to standard LSTMs with data imputation before training, especially when applied to data with lower rates of missing values.

A thorough comparison of the two proposed methods for neurodegenerative AD progression modeling on the ADNI dataset revealed that the deep learning methods outperformed the parametric method in predicting biomarker values and cognitive decline, even when using a few time points of the training and/or test subjects. On the other hand, the parametric method performed significantly better in classifying clinical status.

5.2 Discussion

In general, the nonparametric method was good in modeling MRI biomarkers, while the parametric method performed well in modeling cognitive tests. In both cases, MRI measurements helped to improve cognitive-based diagnostic performance. Also, the parametric method in contrast to the nonparametric method can be applied to the prediction of test data with fewer biomarkers than what was used for training.

Both methods can be applied to different time-series data including missing data points and labels, or to biomarkers with other characteristics than the monotonic behavior that one typically encounters in, for example, neurodegenerative disease progression modeling using MRI/PET biomarkers. However, suitable functions need to be used in the parametric method for biomarker modeling, and proper care needs to be taken to handle missing labels and time intervals in LSTMs.

The proposed training and initialization methods are applicable for other types of RNNs such as GRUs and various activation functions. In addition, the same robust loss functions used for the parametric method can be applied to the LSTM to make the nonparametric predictions robust to outliers. LSTM networks can be directly applied to classification based on sequence-to-label learning to potentially improve diagnostic performance. However, the focus of this work was on DPM based on sequence-to-sequence learning. Besides, sequence-to-label learning would only be

able to utilize the part of the training data which has available clinical status.

The methods presented in this work have the potential to be used in clinical environments for a better understanding of AD for diagnostic, staging, monitorization, and prognostic purposes. The proposed robust tools can automatically analyze the complete perspective of the disease using longitudinal data in an end-to-end fashion. This is also a holistic way to implement a system suitable for both (academic) research and (industrial) clinical applications to better study, detect, and monitor AD. Finally, the proposed methods developed to deal with heterogeneous patterns, missing data, and outliers can be applied to longitudinal studies other than AD.

5.3 Future work

The proposed deep learning method performed better than the regression-based approach in almost all tasks except for classification. As discussed before, two possible reasons for this problem could be the absence of age information and fair performance in modeling the cognitive biomarkers. Therefore, as a potential direction for future work, we can investigate the effects of adding age information and clinical status to the model, e.g., as extra input feature dimensions, on classification performance.

Moreover, the proposed parametric method is not very flexible in modeling non-monotonic biomarkers, such as activation markers with several peaks, and different AD subtypes with different progression rates. However, we would like to benefit from some useful features of the parametric model such as the robust regression scheme in the proposed nonparametric method. The required modification can be applied to the output layer of the LSTM unit by calculating the network loss using M-estimation.

Although the availability of large datasets provides the opportunity for deep learning, learning temporal patterns from longitudinal healthcare data is challenging due to the irregularity and asynchronicity of the data points. Hence, we plan to extend our work based on using a combination of RNNs and continuous-time models for modeling multiple temporal features in sporadic data.

Last but not least, we expect to develop the proposed deep learning tool for clustering longitudinal data into homogeneous subgroups sharing similar trajectories or future outcomes for efficient phenotyping of patients and designing treatment plans for AD. This will pave the way for understanding or interpreting the latent space representations concerning different subtypes of AD.

Bibliography

[1] Alzheimer's association, alzheimer's & dementia. `www.alz.org`.

[2] Clifford R Jack Jr, David S Knopman, William J Jagust, Leslie M Shaw, Paul S Aisen, Michael W Weiner, Ronald C Petersen, and John Q Trojanowski. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *The Lancet Neurology*, 9(1):119–128, 2010.

[3] Bruno M. Jedynak, Andrew Lang, Bo Liu, Elyse Katz, Yanwei Zhang, Bradley T. Wyman, David Raunig, C. Pierre Jedynak, Brian Caffo, and Jerry L Prince. A computational neurodegenerative disease progression score: method and results with the Alzheimer's Disease Neuroimaging Initiative cohort. *NeuroImage*, 63(3):1478–1486, 2012.

[4] Bruno M Jedynak, Bo Liu, Andrew Lang, Yulia Gel, and Jerry L Prince. A computational method for computing an Alzheimer's disease progression score; experiments and validation with the ADNI data set. *Neurobiology of Aging*, 36:S178–S184, 2015.

[5] Michael Ewers, Reisa A Sperling, William E Klunk, Michael W Weiner, and Harald Hampel. Neuroimaging markers for the prediction and early diagnosis of Alzheimer's disease dementia. *Trends in Neurosciences*, 34(8):430–442, 2011.

[6] Milton C Biagioni and James E Galvin. Using biomarkers to improve detection of Alzheimer's disease. *Neurodegenerative Disease Management*, 1(2):127–139, 2011.

[7] Robert W Mahley, Karl H Weisgraber, and Yadong Huang. Apolipoprotein E4: a causative factor and therapeutic target in neuropathology, including Alzheimer's disease. *Proceedings of the National Academy of Sciences*, 103(15):5644–5651, 2006.

[8] Tero Tapiola, Irina Alafuzoff, Sanna-Kaisa Herukka, Laura Parkkinen, Päivi Hartikainen, Hilkka Soininen, and Tuula Pirttilä. Cerebrospinal fluid β-amyloid 42 and tau proteins as biomarkers of Alzheimer-type pathologic changes in the brain. *Archives of Neurology*, 66(3):382–389, 2009.

[9] Zhen Fan, David J Brooks, Aren Okello, and Paul Edison. An early and late peak in microglial activation in Alzheimer's disease trajectory. *Brain*, 140(3):792–803, 2017.

[10] Lisa Mosconi, Valentina Berti, Lidia Glodzik, Alberto Pupi, Susan De Santi, and Mony J de Leon. Pre-clinical detection of Alzheimer's disease using FDG-PET, with or without amyloid imaging. *Journal of Alzheimer's Disease*, 20(3):843–854, 2010.

[11] Seok Rye Choi, Geoff Golding, Zhiping Zhuang, Wei Zhang, Nathaniel Lim, Franz Hefti, Tyler E Benedum, Michael R Kilbourn, Daniel Skovronsky, and Hank F Kung. Preclinical properties of 18F-AV-45: a PET agent for Aβ plaques in the brain. *Journal of Nuclear Medicine*, 50(11):1887–1894, 2009.

[12] Rachael I Scahill, Jonathan M Schott, John M Stevens, Martin N Rossor, and Nick C Fox. Mapping the evolution of regional atrophy in Alzheimer's disease: unbiased analysis of fluid-registered serial MRI. *Proceedings of the National Academy of Sciences*, 99(7):4703–4707, 2002.

[13] Steve Balsis, Jared F Benge, Deborah A Lowe, Lisa Geraci, and Rachelle S Doody. How do scores on the ADAS-Cog, MMSE, and CDR-SOB correspond? *The Clinical Neuropsychologist*, 29(7):1002–1009, 2015.

[14] Mary C Tierney, Christie Yao, Alex Kiss, and Ian McDowell. Neuropsychological tests accurately predict incident Alzheimer disease after 5 and 10 years. *Neurology*, 64(11):1853–1859, 2005.

[15] SM Landau, D Harvey, CM Madison, EM Reiman, NL Foster, PS Aisen, Ronald Carl Petersen, LM Shaw, JQ Trojanowski, CR Jack, et al. Comparing predictors of conversion and decline in mild cognitive impairment. *Neurology*, 75(3):230–238, 2010.

[16] Azar Zandifar, Vladimir Fonov, Simon Ducharme, Sylvie Belleville, and D Louis Collins. MRI and cognitive scores complement each other to accurately predict Alzheimer's dementia 2 to 7 years before clinical onset. *bioRxiv*, page 567867, 2019.

[17] Kurt A Jellinger. Pathobiological subtypes of Alzheimer disease. *Dementia and Geriatric Cognitive Disorders*, pages 1–13, 2020.

[18] Daniel Ferreira, Chloë Verhagen, Juan Andrés Hernández-Cabrera, Lena Cavallin, Chun-Jie Guo, Urban Ekman, J-Sebastian Muehlboeck, Andrew Simmons, José Barroso, Lars-Olof Wahlund, and Eric Westman. Distinct subtypes of Alzheimer's disease based on patterns of brain atrophy: longitudinal trajectories and clinical applications. *Scientific reports*, 7(1):1–13, 2017.

[19] Daniel Ferreira, Agneta Nordberg, and Eric Westman. Biological subtypes of Alzheimer disease: A systematic review and meta-analysis. *Neurology*, 94(10):436–448, 2020.

[20] Hubert M Fonteijn, Marc Modat, Matthew J Clarkson, Josephine Barnes, Manja Lehmann, Nicola Z Hobbs, Rachael I Scahill, Sarah J Tabrizi, Sebastien Ourselin, Nick C Fox, et al. An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. *NeuroImage*, 60(3):1880–1889, 2012.

[21] Vikram Venkatraghavan, Esther E Bron, Wiro J Niessen, and Stefan Klein. Disease progression timeline estimation for Alzheimer's disease using discriminative event based modeling. *NeuroImage*, 186:518–532, 2019.

[22] Clifford R Jack Jr, David S Knopman, William J Jagust, Ronald C Petersen, Michael W Weiner, Paul S Aisen, Leslie M Shaw, Prashanthi Vemuri, Heather J Wiste, Stephen D Weigand, et al. Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *The Lancet Neurology*, 12(2):207–216, 2013.

[23] Neil P. Oxtoby and Daniel C. Alexander. Imaging plus X: multimodal models of neurodegenerative disease. *Current Opinion in Neurology*, 30(4):371, 2017.

[24] Neil P. Oxtoby, Alexandra L. Young, Nick C. Fox, Pankaj Daga, David M. Cash, Sebastien Ourselin, Jonathan M. Schott, and Daniel C. Alexander. Learning imaging biomarker trajectories from noisy Alzheimer's disease data using a Bayesian multilevel model. In *Bayesian and Graphical Models for Biomedical Imaging*, pages 85–94. 2014.

[25] Wai-Ying Wendy Yau, Dana L. Tudorascu, Eric M. McDade, Snezana Ikonomovic, Jeffrey A. James, Davneet Minhas, Wenzhu Mowrey, Lei K. Sheu, Beth E. Snitz, Lisa Weissfeld, et al. Longitudinal assessment of neuroimaging and clinical markers in autosomal dominant Alzheimer's disease: a prospective cohort study. *The Lancet Neurology*, 14(8):804–813, 2015.

[26] Ricardo Guerrero, Alexander Schmidt-Richberg, Christian Ledig, Tong Tong, Robin Wolz, and Daniel Rueckert. Instantiated mixed effects modeling of Alzheimer's disease markers. *NeuroImage*, 142:113–125, 2016.

[27] Murat Bilgel, Jerry L Prince, Dean F Wong, Susan M Resnick, and Bruno M Jedynak. A multivariate nonlinear mixed effects model for longitudinal image analysis: application to amyloid imaging. *NeuroImage*, 134:658–670, 2016.

[28] Dan Li, Samuel Iddi, Wesley K Thompson, and Michael C Donohue. Bayesian latent time joint mixed effect models for multicohort longitudinal data. *Statistical Methods in Medical Research*, 28(3):835–845, 2019.

[29] Murat Bilgel and Bruno M Jedynak. Predicting time to dementia using a quantitative template of disease progression. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 11(C):205–215, 2019.

[30] Peter J Huber. *Robust Statistics*, volume 523. John Wiley & Sons, 2004.

[31] Paolo Pennacchi. Robust estimate of excitations in mechanical systems using M-estimators–theoretical background and numerical applications. *Journal of Sound and Vibration*, 310(4-5):923–946, 2008.

[32] Marco Lorenzi, Maurizio Filippone, Giovanni B Frisoni, Daniel C Alexander, and Sébastien Ourselin. Probabilistic disease progression modeling to characterize diagnostic uncertainty: application to staging and prediction in Alzheimer's disease. *NeuroImage*, 190:56–68, 2017.

[33] Mostafa Mehdipour Ghazi, Mads Nielsen, Akshay Pai, M. Jorge Cardoso, Marc Modat, Sébastien Ourselin, and Lauge Sørensen. Training recurrent neural networks robust to incomplete data: application to Alzheimer's disease progression modeling. *Medical Image Analysis*, 53:39–46, 2019.

[34] Dario Floreano and Claudio Mattiussi. *Bio-inspired artificial intelligence: theories, methods, and technologies*. MIT Press, 2008.

[35] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386, 1958.

[36] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[37] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7), 2011.

[38] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[39] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press Cambridge, 2016.

[40] Barak A. Pearlmutter. Learning state space trajectories in recurrent neural networks. *Neural Computation*, 1(2):263–269, 1989.

[41] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In *A Field Guide to Dynamical Recurrent Neural Networks*. IEEE Press, 2001.

[42] James Martens and Ilya Sutskever. Learning recurrent neural networks with Hessian-free optimization. In *Proceedings of the International Conference on Machine Learning*, pages 1033–1040, 2011.

[43] Trieu H. Trinh, Andrew M. Dai, Minh-Thang Luong, and Quoc V. Le. Learning longer-term dependencies in RNNs with auxiliary losses. *CoRR*, abs/1803.00144, 2018.

[44] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[45] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734, 2014.

[46] Quoc V. Le, Navdeep Jaitly, and Geoffrey E. Hinton. A simple way to initialize recurrent networks of rectified linear units. *CoRR*, abs/1504.00941, 2015.

[47] Eugene Vorontsov, Chiheb Trabelsi, Samuel Kadoury, and Chris Pal. On orthogonality and learning recurrent networks with long term dependencies. *CoRR*, abs/1702.00071, 2017.

[48] Felix A. Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with LSTM. In *Proceedings of the 9th International Conference on Artificial Neural Networks (ICANN 99)*, volume 2, pages 850–855, 1999.

[49] Felix A. Gers and Jürgen Schmidhuber. LSTM recurrent networks learn simple context-free and context-sensitive languages. *IEEE Transactions on Neural Networks*, 12(6):1333–1340, 2001.

[50] Klaus Greff, Rupesh K. Srivastava, Jan Koutník, Bas R. Steunebrink, and Jürgen Schmidhuber. LSTM: a search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10):2222–2232, 2017.

[51] Ronald Carl 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. *Neurology*, 74(3):201–209, 2010.

[52] Zachary C. Lipton, David C. Kale, and Randall Wetzel. Modeling missing data in clinical time series with RNNs. In *Proceedings of Machine Learning for Healthcare*, 2016.

[53] Shahla Parveen and Phil Green. Speech recognition with missing data using recurrent neural nets. In *Advances in Neural Information Processing Systems*, pages 1189–1195, 2002.

[54] Jinsung Yoon, William R Zame, and Mihaela van der Schaar. Estimating missing data in temporal data streams using multi-directional recurrent neural networks. *IEEE Transactions on Biomedical Engineering*, 2018.

[55] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1):6085, 2018.

[56] Dumitru Erhan, Pierre-Antoine Manzagol, Yoshua Bengio, Samy Bengio, and Pascal Vincent. The difficulty of training deep architectures and the effect of unsupervised pre-training. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 153–160, 2009.

[57] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010.

[58] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the 2015 IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

[59] Sachin S. Talathi and Aniket Vartak. Improving performance of recurrent neural network with ReLU nonlinearity. *CoRR*, abs/1511.03771, 2015.

[60] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[61] CJ Stannard, AP Williams, and PA Gibbs. Temperature/growth relationships for psychrotrophic food-spoilage bacteria. *Food Microbiology*, 2(2):115–122, 1985.

[62] Duane L Beekly, Erin M Ramos, William W Lee, Woodrow D Deitrich, Mary E Jacka, Joylee Wu, Janene L Hubbard, Thomas D Koepsell, John C Morris, Walter A Kukull, et al. The National Alzheimer's Coordinating Center (NACC) database: the uniform data set. *Alzheimer Disease & Associated Disorders*, 21(3):249–258, 2007.

[63] Razvan V Marinescu, Neil P Oxtoby, Alexandra L Young, Esther E Bron, Arthur W Toga, Michael W Weiner, Frederik Barkhof, Nick C Fox, Arman

Eshaghi, Tina Toni, et al. The Alzheimer's disease prediction of longitudinal evolution (TADPOLE) challenge: results after 1 year follow-up. *CoRR*, abs/2002.03419, 2020.

[64] Mostafa Mehdipour Ghazi, Mads Nielsen, Akshay Pai, Marc Modat, M. Jorge Cardoso, Sébastien Ourselin, and Lauge Sørensen. Robust parametric modeling of Alzheimer's disease progression. *CoRR*, abs/1908.05338, 2019.

[65] Mostafa Mehdipour Ghazi, Mads Nielsen, Akshay Pai, M. Jorge Cardoso, Marc Modat, Sébastien Ourselin, and Lauge Sørensen. Robust training of recurrent neural networks to handle missing data for disease progression modeling. In *International Conference on Medical Imaging with Deep Learning*, 2018.

[66] Mostafa Mehdipour Ghazi, Mads Nielsen, Akshay Pai, Marc Modat, M. Jorge Cardoso, Sébastien Ourselin, and Lauge Sørensen. MRI biomarkers improve disease progression modeling-based prediction of cognitive decline. In *Radiological Society of North America – Scientific Assembly and Annual Meeting*, 2019.

[67] Mostafa Mehdipour Ghazi, Lauge Sørensen, Akshay Pai, M. Jorge Cardoso, Marc Modat, Sébastien Ourselin, and Mads Nielsen. Disease progression modeling-based prediction of cognitive decline. In *Alzheimer's Association International Conference*, 2020.

[68] PF Verhulst. La loi d'accroissement de la population. *Nouv. Mém. de l'Academie Royale des Sci. et Belles-Lettres de Bruxelles*, 18(1):1–41, 1845.

[69] FJ Richards. A flexible growth function for empirical use. *Journal of Experimental Botany*, 10(2):290–301, 1959.

[70] Benjamin Gompertz. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philosophical Transactions of the Royal Society of London*, (115):513–583, 1825.

[71] MH Zwietering, Il Jongenburger, FM Rombouts, and K Van't Riet. Modeling of the bacterial growth curve. *Applied and Environmental Microbiology*, 56(6):1875–1881, 1990.

[72] Paul W Holland and Roy E Welsch. Robust regression using iteratively reweighted least-squares. *Communications in Statistics - Theory and Methods*, 6(9):813–827, 1977.

[73] Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

[74] Razvan V. Marinescu, Neil P. Oxtoby, Alexandra L. Young, Esther E. Bron, Arthur W. Toga, Michael W. Weiner, Frederik Barkhof, Nick C. Fox, Stefan Klein, and Daniel C. Alexander. TADPOLE challenge: prediction of longitudinal evolution in Alzheimer's disease. *CoRR*, abs/1805.03909, 2018.

[75] Bruce Fischl, David H Salat, Evelina Busa, Marilyn Albert, Megan Dieterich, Christian Haselgrove, Andre Van Der Kouwe, Ron Killiany, David Kennedy, Shuna Klaveness, Albert Montillo, Nikos Makris, Bruce Rosen, and Anders M Dale. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3):341–355, 2002.

[76] William J Jagust, Susan M Landau, Robert A Koeppe, Eric M Reiman, Kewei Chen, Chester A Mathis, Julie C Price, Norman L Foster, and Angela Y Wang. The Alzheimer's disease neuroimaging initiative 2 PET core: 2015. *Alzheimer's & Dementia*, 11(7):757–771, 2015.

[77] Peter A Freeborough and Nick C Fox. The boundary shift integral: an accurate and robust measure of cerebral volume changes from registered repeat MRI. *IEEE Transactions on Medical Imaging*, 16(5):623–629, 1997.

[78] Ed HBM Gronenschild, Petra Habets, Heidi IL Jacobs, Ron Mengelers, Nico Rozendaal, Jim Van Os, and Machteld Marcelis. The effects of FreeSurfer version, workstation type, and Macintosh operating system version on anatomical volume and cortical thickness measurements. *PloS One*, 7(6):e38234, 2012.

[79] Liam M O'Brien, David A Ziegler, Curtis K Deutsch, Jean A Frazier, Martha R Herbert, and Joseph J Locascio. Statistical adjustments for brain size in volumetric neuroimaging studies: some practical implications in methods. *Psychiatry Research: Neuroimaging*, 193(2):113–122, 2011.

[80] Jose AF Machado. Robust model selection and M-estimation. *Econometric Theory*, 9(3):478–493, 1993.

[81] Tianfeng Chai and Roland R Draxler. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3):1247–1250, 2014.

[82] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[83] David J. Hand and Robert J. Till. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, 45(2):171–186, 2001.

[84] Thomas F Coleman and Yuying Li. An interior trust region approach for nonlinear minimization subject to bounds. *SIAM Journal on Optimization*, 6(2):418–445, 1996.

[85] Lutz Prechelt. Early stopping–but when? In *Neural Networks: Tricks of the Trade*, pages 55–69. 1998.

[86] Alexandra L Young, Neil P Oxtoby, Jonathan Huang, Razvan V Marinescu, Pankaj Daga, David M Cash, Nick C Fox, Sebastien Ourselin, Jonathan M Schott, and Daniel C Alexander. Multiple orderings of events in disease progression. In *International Conference on Information Processing in Medical Imaging*, pages 711–722, 2015.

[87] Ludmila I Kuncheva. *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons, 2014.

[88] Peter J Rousseeuw and Annick M Leroy. *Robust regression and outlier detection*, volume 1. Wiley Online Library, 1987.

[89] Mostafa Mehdipour Ghazi, Mads Nielsen, Akshay Pai, Marc Modat, M. Jorge Cardoso, Sébastien Ourselin, and Lauge Sørensen. On the initialization of long short-term memory networks. In *International Conference on Neural Information Processing*, pages 275–286. Springer, 2019.

[90] Felix A Gers, Nicol N Schraudolph, and Jürgen Schmidhuber. Learning precise timing with LSTM recurrent networks. *Journal of Machine Learning Research*, 3:115–143, 2002.

[91] Dariusz Buraczewski, Ewa Damek, and Thomas Mikosch. *Stochastic models with power-law tails*. Springer, 2016.

[92] Pedro J García-Laencina, José-Luis Sancho-Gómez, and Aníbal R Figueiras-Vidal. Pattern classification with missing data: a review. *Neural Computing and Applications*, 19(2):263–282, 2010.

[93] Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.