brighton and sussex
medical school

DEFINING DISEASE PROGRESSION IN ALS:

A novel analytic approach using existing clinical and imaging datasets

Matthew Christopher Edward Gabel

A thesis submitted in partial fulfilment of the requirements of the University of Brighton and the University of Sussex for the degree of Doctor of Philosophy

August, 2017

Abstract

Background

A key aim of medical science is modelling patterns of disease progression; these patterns increase understanding of the disease, and help construct staging systems that assist diagnosis and treatment. Within Amyotrophic Lateral Sclerosis (ALS) disease progression modelling, there is a need to integrate clinical observation-based staging systems such as Roche et al. (2012), which suffer from low temporal resolution, with 'unbiased' staging of biomarkers. To this end, I have adapted and extended an Event-Based Model (EBM) for ALS from previous work in Alzheimer's disease (Fonteijn et al., 2012; Young et al., 2014). Unlike traditional models of disease progression, event-based models do not rely on *a priori* staging of patients but extract the event ordering directly from the data, thus minimising subjective bias. In MR imaging, Fractional Anisotropy (FA) derived from diffusion tensor imaging is an obvious candidate to test the hypothesis that imaging events can be staged in the ALS-adapted EBM.

Objectives

Using contemporary and historical ALS datasets comprised of diffusion MRI, clinical and neuropsychological data, I have adapted and extended a novel event-based model to analyse the likely ordering of these biomarkers in the progression of ALS.

Materials and Methods

The contemporary dataset was derived from a cross-sectional sample of 23 ALS patients and 23 matched controls (Broad et al., 2015). The two historical datasets were similarly derived from samples of i) 36 ALS patients and 22 matched controls, and ii) 28 ALS patients and 25 matched controls (Tsermentseli et al., 2015). The ALS-specific adaptations to the EBM were i) the fitting of Gaussian mixture models by constrained Expectation Maximisation, ii) the calculation of event probabilities from the cumulative distribution function to preserve the monotonicity of biomarker reading progression, and iii) accounting for the clearly delineated patient and control cohorts by performing Markov Chain Monte Carlo (MCMC) sampling on only the patient cohort.

Finally, a fully Bayesian approach to Event-Based Modelling is demonstrated.

Results

The most likely order of progression of imaging events showed that FA changes in the lower aspect of the corticospinal tracts (CSTs) occur at an early stage of disease evolution, with changes in the upper aspect occurring at a later stage. This result was found individually in all three datasets, as well as when combining them.

Discussion

This proof-of-principle study shows that data-driven models of ALS progression are feasible, as well as demonstrating a fully Bayesian approach to Event-Based Modelling. The diffusion MRI event ordering results suggest very robustly that damage to the CSTs starts in the lower aspect. Nevertheless, a general important limitation must be discussed: The small sample size may have biased our results. I have tried to address this issue by assessing how the results varied across three separate datasets, both individually and combined. While the CST results were consistent across the entire process, results for other regions such as the corpus callosum were less constant, suggesting that the biomarker ordering in the wider population may diverge from this sequence.

In order to generalise these results to the wider spectrum of ALS, future studies on larger datasets are warranted.

Conclusion

These findings provide the first solely data-driven evidence supporting a directional hypothesis of motor neurone degeneration.

Contents

1 Introduction **1**

1.1 ALS definition and overview 1

1.2 Disease modelling and staging in ALS 5

1.2.1 Existing ALS clinical staging systems 6

1.3 Diffusion Tensor Imaging (DTI) 10

1.3.1 Mechanisms of DTI 10

1.3.2 Fractional Anisotropy as an index of Neurodegeneration 13

1.3.3 Fractional Anisotropy in ALS 14

1.4 Structural MR Imaging 15

1.4.1 Volumetry 15

1.4.2 Morphometry in ALS 15

2 The Event-Based Model **16**

2.1 Theory 16

2.2 Model estimation 17

2.3 Mixture models for the likelihood of data given events 18

2.4 The MCMC process 19

2.5 Modifying the EBM for ALS 20

2.5.1 Assumptions 20

2.5.2 Mixture model fitting 21

2.5.3 Variable precision 22

2.5.4 “Patients only” vs “Patients and Controls” 24

2.5.5 CDF vs PDF 24

2.6 A Bayesian approach to event-based modelling 25

2.6.1 Bayesian inference 25

2.6.2 The Problem with Permutations and Priors 27

2.6.3 Permutations as directional data 27

3 Materials and Methods 31

3.1 MRI acquisition . . . 31

3.1.1 Set F . . . 31

3.1.2 Set K . . . 32

3.1.3 Set N . . . 33

3.2 Neuroimaging biomarker selection . . . 33

3.2.1 dMRI . . . 33

3.2.2 Structural imaging . . . 35

3.3 MRI processing and analysis . . . 37

3.3.1 Diffusion-Weighted Images . . . 37

3.3.2 Structural Images . . . 37

3.3.3 Inter-cohort dMRI heterogeneity . . . 40

3.4 Clinical and Neuropsychological data acquisition . . . 40

3.5 Clinical and Neuropsychological biomarker selection . . . 41

4 Challenging the Model *in silico* 43

4.1 Tools for comparing Permutations . . . 43

4.1.1 Terminology . . . 43

4.1.2 Definitions . . . 44

4.2 Simulations . . . 46

4.2.1 Mean separation . . . 47

4.2.2 Variance . . . 48

4.2.3 Mean separation and Variance . . . 49

4.2.4 Sample size . . . 52

4.2.5 Number of biomarkers . . . 53

4.3 Effects of model modifications . . . 56

4.3.1 CDF vs. PDF for probability calculations . . . 56

4.4 Summary . . . 57

5 Challenging the Model *in vivo* 60

5.1 Effects of model modifications . . . 60

5.1.1 Adding and removing biomarkers . . . 60

5.1.2 Too many or too few biomarkers . . . 61

5.1.3 Sample size and mixture fitting . . . 62

5.1.4 Gaussian vs Non-Gaussian . . . 64

5.1.5 Assessing Mixture types 67
5.2 Coverage of sample space 69
5.3 Summary 71

6 Model Output 72
6.1 Neuroimaging biomarkers 73
6.1.1 FA data 73
6.1.2 Structural data 79
6.2 Neuropsychological biomarkers 83

7 Discussion 86
7.1 FA results 86
7.1.1 Model output 86
7.1.2 The Uncinate Fasciculus 89
7.1.3 General observations on FA output 90
7.2 Structural results 91
7.3 Neuropsychological results 92
7.4 Aspects of the Modelling Process 92
7.4.1 Bootstrapping for mixture fitting 92
7.4.2 Bimodality of Progression 96
7.4.3 Assessing the EBM's potential as a diagnostic tool 96
7.4.4 CST pathology 100
7.4.5 The mixing proportion 103
7.4.6 Data harmonisation 108

8 Future Work 109
8.1 Validation and extension of findings 109
8.1.1 Validation 109
8.1.2 Directions for Extension 110
8.2 Model development 112
8.2.1 Mixture fitting 112
8.2.2 Disparate data modalities and collinearity 113
8.2.3 The Bayesian EBM 113
8.2.4 Fourier Transforms 115
8.2.5 MCMC methods 115
8.3 General improvements 116

Bibliography **117**

Appendix A Matlab code **125**

A.1 Fitting mixtures by EM 125

A.2 Plotting fitted mixtures 127

A.3 Calculating probabilities 129

A.3.1 Cumulative Distribution Functions 129

A.3.2 Probability Density Functions 131

A.4 Calculating ln(p(X|S)) 132

A.4.1 Double-precision implementation 132

A.4.2 Variable-precision implementation, Advanpix Toolbox 132

A.4.3 Variable-precision implementation, native Matlab 133

A.5 Comparing multiple greedy ascent runs 134

A.6 Performing MCMC iterations 136

A.7 Plotting a positional variance diagram 138

A.8 The Bayesian EBM 139

A.8.1 Embedding a permutation onto a hypersphere 139

A.8.2 Maximum likelihood estimation of parameters of von Mises-Fisher distribution 139

A.8.3 Projecting from a hypersphere 140

A.8.4 The Bayesian EBM – incorporating a prior distribution 142

A.8.5 PDF for von Mises-Fisher distribution 144

A.9 Tools for permutations 144

A.9.1 Calculating Kendall's tau distance 144

A.9.2 Calculating Kendall's tau correlation coefficient 145

List of Tables

1.1 The El Escorial criteria and its revisions 3

1.2 The King's staging system for ALS 6

1.3 The MiToS staging system for ALS 7

3.1 Numbers and demographics of subjects of Sets FKN 31

3.2 Numbers and demographics for dMRI data of Sets FKN 33

3.3 Numbers and demographics for structural T_1 MRI data of Sets KN. . . 35

3.4 Neuropsychological and functional assessments included in Set K 42

4.1 Mixture model parameters used for simulating the effects of mean separation $|\mu_a - \mu_c|$ on the EBM's output 47

4.2 Mean tau distance between S_{ML} and T, for varied $|\mu_a - \mu_c|$ with $\sigma_a - \sigma_c = 0$ 48

4.3 Mean tau correlation between S_{ML} and T when varying $|\mu_a - \mu_c|$ . . . 48

4.4 Mixture model parameters used for simulating the effects of patient biomarker variance, $\sigma_a^2 - \sigma_c^2$, on the EBM's output 48

4.5 Mean distance between S_{ML} and T, varying $\sigma_a - \sigma_c$, with $|\mu_a - \mu_c| = 0$ 49

4.6 Mean tau correlation between S_{ML} and T when varying $\sigma_a - \sigma_c$ 49

4.7 Mixture model parameters, for same-direction varying of $|\mu_a - \mu_c|$ and $\sigma_a - \sigma_c$ 50

4.8 Mean tau distance between S_{ML} and T, for same-direction varying of $|\mu_a - \mu_c|$ and $\sigma_a - \sigma_c$ 50

4.9 Mean tau correlation between S_{ML} and T, for same-direction varying of $|\mu_a - \mu_c|$ and $\sigma_a - \sigma_c$ 50

4.10 Mixture model parameters, for opposite-direction varying of $|\mu_a - \mu_c|$ and $\sigma_a - \sigma_c$ 51

4.11 Mean distance between S_{ML} and T, varying μ_a and σ_a in the opposite direction 51

4.12 Mean tau correlation between S_{ML} and T, varying μ_a and σ_a in the opposite direction 51
4.13 Mean distance between S_{ML} and T, varying μ_a 52
4.14 Simulating the effects of adding a biomarker 54
4.15 Simulating the effects of removing a biomarker 54
4.16 Mean distance between S_{ML} and T, varying $|\mu_a - \mu_c|$, using PDFs . . . 58
4.17 Mean tau correlation between S_{ML} and T, varying $|\mu_a - \mu_c|$, using PDFs 58
4.18 Mean distance between S_{ML} and T, varying $\sigma_a - \sigma_c$, using PDFs 58
4.19 Mean tau correlation between S_{ML} and T, varying $\sigma_a - \sigma_c$, using PDFs 58
4.20 Mean distance between S_{ML} and T, varying μ_a and σ_a in the same direction, using PDFs 59
4.21 Mean tau correlation between S_{ML} and T, varying μ_a and σ_a in the same direction, using PDFs 59
4.22 Mean distance between S_{ML} and T, varying μ_a and σ_a in the opposite direction, using PDFs 59
4.23 Mean tau correlation between S_{ML} and T, varying μ_a and σ_a in the opposite direction, using PDFs 59

5.1 Log-likelihood for greedy output and S_{ML}, using CDFs to calculate event probabilities 70
5.2 Log-likelihood for greedy output and S_{ML}, using PDFs to calculate event probabilities 70
5.3 Log-likelihood for greedy output and S_{ML}, using CDFs to calculate event probabilities. Outliers $> 2.5\ \sigma$ excluded during mixture fitting. 70
5.4 Log-likelihood for greedy output and S_{ML}, using PDFs to calculate event probabilities. Outliers $> 2.5\ \sigma$ excluded during mixture fitting. 70

7.1 τ correlation between S_{ML} of the population cohort groupings. EBM applied to FA data of Sets FKN 87
7.2 Temporal information parameters and mixing proportions for the outer thirds of the CSTs 102

List of Figures

1.1 Gene frequencies in ALS 5

1.2 Bar plots showing the standardised median proportion of time from onset to each disease stage for the King's and MiToS staging systems 8

1.3 The four stages of progression of abnormal pTDP-43 protein pathology in ALS, as proposed by Braak et al. 9

1.4 The diffusion tensor represented as an ellipsoid 12

1.5 An axial slice of a DTI scan of a healthy brain. The diffusion tensor of each voxel has been rendered as a 3D ellipsoid glyph 12

1.6 Single slices of a whole-brain Fractional Anisotropy (FA) image in the coronal, axial and sagittal planes 13

2.1 Illustration of the EM algorithm's sensitivity to extreme values, derived from fitting Gaussian mixtures to cortical thickness data. 23

2.2 Illustration of PDF and $(1 - \mathrm{CDF})$ values for two different readings of the same biomarker, Fractional Anisotropy 26

2.3 Permutohedrons of order 3 and 4 28

2.4 The von Mises-Fisher (vMF) density function on a 2-sphere in $\mathbb{R}^3$ for different μ (mean direction) and κ (concentration parameter) 30

3.1 White matter tracts examined in this study 34

3.2 Boxplots of mean FA biomarker values for all 157 subjects of Sets FKN 36

3.3 Cortical regions of the Desikan-Killiany atlas examined in this study . . 38

3.4 Subcortical regions examined in this study 39

4.1 Plot of $\bar{K}$ against number of simulated patients, with fitted curve. . . . 53

4.2 The effects of removing a random biomarker on EBM output 55

5.1 Positional variance diagrams showing the effects of increasing the number of biomarkers for actual patient data, Sets FKN. Constraints: mixing proportions > 0.25 . . . 61
5.2 Positional variance diagrams showing the effects of increasing the number of biomarkers for actual patient data, Sets FKN. Constraints: mixing proportions > 0.25, outliers $> 2\sigma$ removed during mixture fitting. . . . 62
5.3 Comparison of model output for Set K with outlier control subject included and excluded . . . 63
5.4 Comparison of model output for Sets FKN with outlier control subject included and excluded . . . 64
5.5 Comparison of model output for Sets FKN (outlier HC excluded) with Sets FKN (outlier HC included). Constraints: mixing proportions $>$ 0.25, outliers $> 2.5\ \sigma$ excluded during mixture fitting . . . 65
5.6 Positional variance diagrams showing the effects of transforming control and patient biomarker data to be Gaussian . . . 66
5.7 Positional variance diagrams showing the effects of refitting a single biomarker, "TROGerr", from a mixture of two Gaussians to a mixture of two Poissons . . . 67
5.8 Histograms of neuropsychological biomarker readings, with mixture model component distributions superimposed . . . 68

6.1 Positional variance diagram of the most likely event order of white matter tract neurodegeneration in ALS patients, generated from S_{ML} and the MCMC samples . . . 74
6.2 Positional variance diagrams for FA data of all possible subgroups of Sets F, K and N. Constraints: mixing proportion > 0.25 . . . 75
6.3 The fitted mixture models for Sets FKN. Constraints: mixing proportion > 0.25, no outliers removed. . . . 78
6.4 Positional variance diagrams for Sets KN subcortical volumetric data. Constraints: mixing proportion > 0.25, outliers $> 2.5\ \sigma$ removed during mixture fitting. . . . 80
6.5 Positional variance diagrams for Sets KN cortical thickness data. Constraints: mixing proportion > 0.25, outliers $> 3\ \sigma$ removed during mixture fitting. . . . 82

6.6 Positional variance diagram for Set K Clinical and Neuropsychological data 84
6.7 The fitted mixture models for Set K neuropsychological measures. . . . 85

7.1 Positional variance diagram for FA biomarkers of Sets FKN. Mix > 0.25, outliers $> 2.5\,\sigma$ removed during mixture fitting. 88
7.2 The proposed temporal order of white matter tract degeneration in ALS, as generated by the EBM 88
7.3 Positional variance diagram for FA biomarkers of Sets FKN. Mix > 0.25, outliers $> 2.5\,\sigma$ removed during mixture fitting. μ_a, σ_a swapped with μ_c, σ_c for $\mathrm{UF_L}$. 90
7.4 PVDs comparing the MCMC samples from Set K_{24} with those of Set K_{25}. Constraints: mixing proportions > 0.25 93
7.5 Meta-PVDs, summarising the exploration of bootstrapped EM mixture fitting as a tool to reduce the impact of a single outlier HC 95
7.6 Meta-PVD for Sets FKN, excluding the outlier HC 97
7.7 The distribution of the most likely stages assigned to all 156 subjects, as generated from the four CST biomarkers 99
7.8 ROC curve for the EBM as a diagnostic tool for ALS, using only the four CST biomarkers 99
7.9 The fitted mixture models for FA data of Sets FKN. Each biomarker reading is the mean FA value across the whole tract (left and right sides combined). Constraints: mixing proportion > 0.15, no outliers removed. 101
7.10 Log-likelihood calculated for all 40 320 possible event orders of an eight-biomarker model 102
7.11 PVDs for groups A and B, which form a partition of the sample space Ω of an eight-biomarker model 102
7.12 PVDs comparing S_{ML} and T_{mix}, simple implementation using CDFs . . 104
7.13 PVDs comparing S_{ML} and T_{mix}, simple implementation using PDFs . . 104
7.14 PVDs comparing S_{ML} and T_{mix}, bimodal implementation using CDFs . 106
7.15 PVDs comparing S_{ML} and T_{mix}, bimodal implementation using PDFs . 106
7.16 Estimates of the duration of each biomarker stage 107

8.1 Mapping Ω with the greedy ascent algorithm 114

Acronyms

S_{ML} most likely order of progression.

AD Alzheimer's disease.

ALS Amyotrophic Lateral Sclerosis.

ALSFRS-R ALS Functional Rating Scale-Revised.

CDF Cumulative Distribution Function.

CSF cerebrospinal fluid.

CST corticospinal tract.

DTI diffusion tensor imaging.

EBM Event-Based Model.

EM Expectation Maximisation.

ERRALS Emilia Romagna Registry for ALS.

FA Fractional Anisotropy.

GNT Graded Naming Test.

HADS Hospital Anxiety and Depression Scale.

HC healthy control.

KDT Kissing and Dancing Test.

LLH log-likelihood.

MCMC Markov Chain Monte Carlo.

MD mean diffusivity.

ML maximum likelihood.

MLOOP most likely order of progression.

MRI Magnetic Resonance Imaging.

NODDI neurite orientation dispersion and density imaging.

PDF Probability Density Function.

PLIC posterior limb of the internal capsule.

PPT Pyramids and Palm Trees.

PVD positional variance diagram.

ROI region of interest.

rsfMRI resting-state functional MRI.

TI temporal information.

TROG Test of Reception of Grammar.

vMF von Mises-Fisher.

WM white matter.

Acknowledgements

The largest thanks go to my wonderful supervisors, Mara Cercignani and Nigel Leigh. As well as their wisdom, patience and willingness to share their work, time and knowledge, it is their passion and love for their research that have made this project so much fun to work on. Somehow, they found the perfect balance between keeping an eye on me, and giving me freedom to go off and explore. Nigel, thank you for all the guidance and support you have given me; in particular, thank you for sitting and looking at my pictures of squares with genuine enthusiasm, and then trying to explain to me how on Earth they might have some biological meaning. Likewise, Mara's steady hand and complete willingness to answer my questions (no matter how ridiculous) at any time of the day have helped me learn and grow. Thank you both for making this project so enjoyable; I will always have fond memories of Mara sitting in our supervision sessions and giggling every time the name "Pippa" was mentioned. Mara and Nigel – I hope that one day, I can be a supervisor even half as fantastic as you both have been.

The person who convinced me to start down the path of medical research is Jackie Cassell. Without her encouragement, I would have missed out on the last three years' worth of happiness – and yet, if we hadn't played violin together in Musicians of All Saints many years ago, none of this would have happened. I'm still surprised that I've managed to make the jump from music to maths to research, and I owe Jackie a great deal of thanks for her part in this.

To my colleague Rebecca Broad, who has encouraged me and suffered alongside me over these three years. Your hard work in collecting Set N, and generosity in sharing the data and your time with me, have been unparalleled. You have been a joy to work with, and I wish you the very best of luck with beating your thesis into submission.

A project that covers this many different disciplines and odd corners of research requires help from collaborators, particularly when it comes to gluing all the disparate parts together. I have been so very fortunate to have (in no particular order) Ammar Al-Chalabi, Danny Alexander, Laura Goldstein and Stella Tsermentseli work with me and contribute so generously and freely to this project. Ammar – thank you for your enthusiasm and guidance, as well as your passion for using maths in this peculiar way. The tips on how to use a Viewmaster to make a VR helmet were invaluable. I want to thank Danny for his assistance with developing the model and my foolish insistence on using CDFs, and Laura and Stella for their crucial guidance with the neuropsychological data. Huge thanks must, of course, go to absolutely everybody involved in collecting the data for Sets F and K – this was a process that lasted well over thirteen years, and there are far too many people to mention by name. Without any one of them, this project would not have been possible.

I am so very grateful to Charlotte Clarke – with her sunny smile and excellent taste in hoodies – for always being in the right place at the right time. Lottie, without your prompting, there would be no Bayesian adaptations to the model, nor any bootstrapping.

How can I not mention the utterly crazy Trafford Centre Phamily? I've said this so many times before, but I'll never find a better bunch of people to work with. Thank you for all the laughter, for the company and the awful jokes (and yes, it was a papaya all along).

I sincerely thank the MND Association, who have been brave and bold enough to allow a mathematician to try his hand at medical research. Without their belief in this project, and their willingness to fund this type of research, none of this would be possible. I hope that someday I will be able to contribute something substantial to the understanding of ALS, and add to the incredible work that the Association does. Also, to all of the patients who so generously and selflessly gave up their time to be included in the original studies – thank you. We are slowly making progress in understanding ALS, and I hope you all would be pleased that your data are still being used nearly two decades later.

Finally, to my parents, without whose love and support I, and this thesis, simply would not be here right now – thank you from the bottom of my heart. Maybe – just maybe – this thesis goes some way towards making the last three years worthwhile.

In memory of Polly.

I cannot write in verse, for I am no poet. I cannot arrange the parts of speech with such art as to produce effects of light and shade, for I am no painter. Even by signs and gestures I cannot express my thoughts and feelings, for I am no dancer. But I can do so by means of sounds, for I am a musician.

—Wolfgang Amadeus Mozart

Declaration

I declare that the research contained in this thesis, unless otherwise formally indicated within the text, is the original work of the author. The thesis has not been previously submitted to this or any other University for a degree, and does not incorporate any material already submitted for a degree.

Signature:

Matthew Christopher Edward Gabel

Chapter 1

Introduction

1.1 ALS definition and overview

Amyotrophic Lateral Sclerosis (ALS) is an adult-onset neurodegenerative disease, characterised by progressive degeneration of both the upper (UMN) and lower (LMN) motor neurones in the primary motor cortex, corticospinal tracts, brain stem and spinal cord. Incidence rates range from 1.2-4.0 per 100 000/year in Caucasians, and there is a slight male preponderance (1.5:1). The condition is more common in the middle-aged and elderly, and the mean age of onset for sporadic ALS is about 60 years (Wijesekera & Leigh, 2009). Prevalence is relatively uniform in Western countries, and because of a comparatively short survival, prevalence is low (average 5.2 per 100 000). Foci of higher prevalence have appeared in the Western Pacific, with a clustering of ALS cases occurring in Guam and the Kii Peninsular in Japan in the 1950-70s.

ALS is the most common motor neurone disease (MND), and the term MND is synonymous with ALS in the UK. ALS is invariably fatal, with death usually due to respiratory complications from bulbar or respiratory muscle weakness. Median survival is 19 months from diagnosis and 30 months from symptom onset, but a small percentage of sufferers survive for 10 or more years (Gordon, 2013).

Originally thought to be a degenerative muscle disorder, Jean-Martin Charcot published clinicopathological studies in 1869 and 1874 describing ALS, emphasising the involvement of both upper and lower motor neurones. Lower motor neurones (LMNs) reside in the anterior horn of the spinal cord or motor nuclei of the brain stem, and project in peripheral nerves to make direct contact with muscle fibres. When LMNs degenerate, the muscles they activate become weak, wasted and fasciculate (noticed as involuntary muscle twitching). Upper motor neurones (UMNs) reside in the precentral gyrus of the cerebral cortex and project to the LMNs. The degeneration of UMNs results in the muscles becoming spastic, tendon reflexes becoming exaggerated and plantar responses becoming extensor.

ALS is heterogeneous in both its presentation and rate of clinical progression (Baumer et al., 2014). Approximately two-thirds of patients have a spinal form of the disease (classical Charcot's ALS, termed limb onset ALS), and they present with symptoms related to focal muscle weakness and wasting. These symptoms may start either distally or proximally in the upper and lower limbs. Bulbar onset ALS patients usually present with progressive dysarthria and dysphagia, affecting speech and swallowing for solids or liquids; the majority will go on to develop limb involvement within 1-2 years (Wijesekera & Leigh, 2009). Bulbar onset ALS and limb onset ALS have largely replaced the terms progressive bulbar palsy (PBP) and Charcot's ALS, in recognition of MND being a spectrum of diseases.

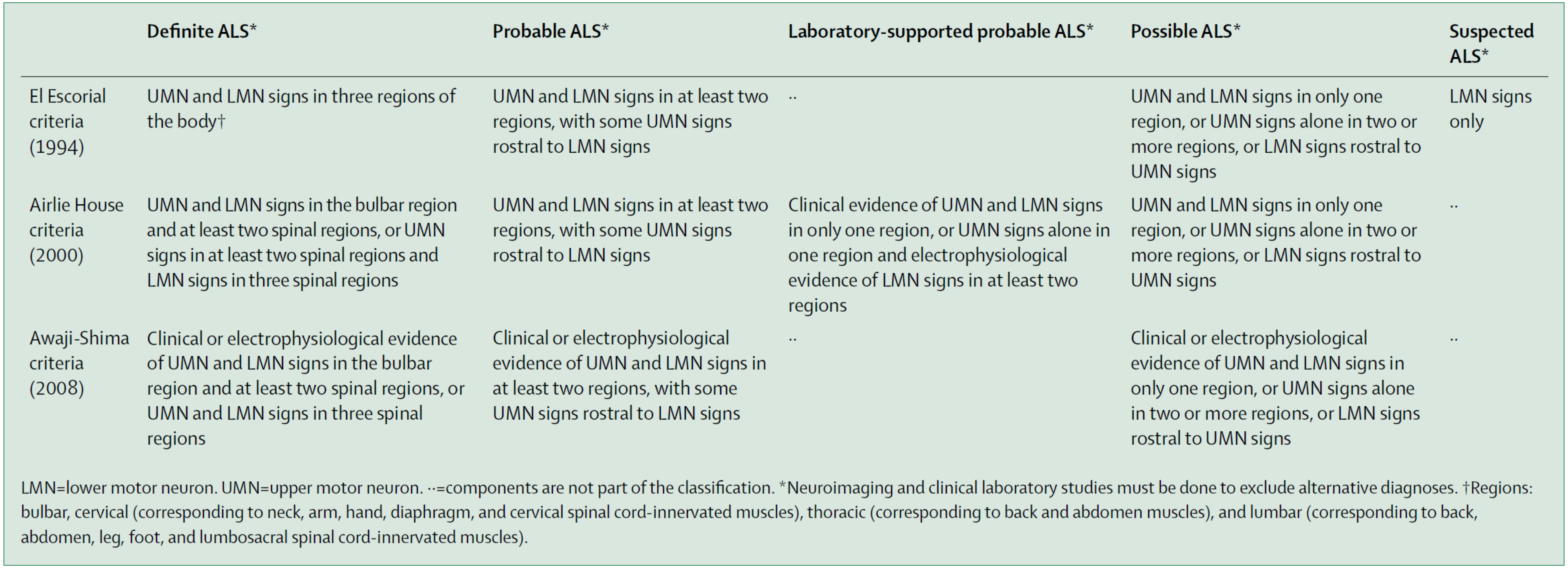
Patients who are elderly, female or have bulbar onset ALS symptoms have a worse prognosis (Fowler et al., 2003).

Other clinical phenotypes of ALS are a pure UMN variant (primary lateral sclerosis, PLS), and a pure LMN variant (progressive muscular atrophy, PMA). PLS patients have a better prognosis than full ALS sufferers, as PLS tends to follow a very slowly progressive course. PMA is uncommon amongst ALS variants, and consists of slowly progressive wasting, with possible fasciculations but no UMN signs. The wasting is usually symmetrical in both hands, and progresses to more proximal muscles and the legs over several years.

Frontotemporal dementia (FTD) shares aspects with ALS; cognitive impairment in ALS was described by Pierre Marie in 1892, but was not considered a common aspect of the disease until recently. Currently, ALS and FTD are not seen as distinct diseases, but instead are considered as extremes of a spectrum (Lattante et al., 2015).

The diagnosis of ALS is based on clinical history, examination, electromyography and exclusion of other diseases. There is a current lack of useful biomarkers, MRI or otherwise, that definitively allow for direct diagnosis of ALS (Chiò & Traynor, 2015).

The Revised El Escorial diagnostic criteria (Table 1.1) are the current gold standard criteria for diagnostic purposes, both clinically and for clinical research studies. The criteria link diagnostic certainty with the number of regions affected clinically or neurophysiologically.



	Definite ALS*	Probable ALS*	Laboratory-supported probable ALS*	Possible ALS*	Suspected ALS*
El Escorial criteria (1994)	UMN and LMN signs in three regions of the body†	UMN and LMN signs in at least two regions, with some UMN signs rostral to LMN signs	··	UMN and LMN signs in only one region, or UMN signs alone in two or more regions, or LMN signs rostral to UMN signs	LMN signs only
Airlie House criteria (2000)	UMN and LMN signs in the bulbar region and at least two spinal regions, or UMN signs in at least two spinal regions and LMN signs in three spinal regions	UMN and LMN signs in at least two regions, with some UMN signs rostral to LMN signs	Clinical evidence of UMN and LMN signs in only one region, or UMN signs alone in one region and electrophysiological evidence of LMN signs in at least two regions	UMN and LMN signs in only one region, or UMN signs alone in two or more regions, or LMN signs rostral to UMN signs	··
Awaji-Shima criteria (2008)	Clinical or electrophysiological evidence of UMN and LMN signs in the bulbar region and at least two spinal regions, or UMN and LMN signs in three spinal regions	Clinical or electrophysiological evidence of UMN and LMN signs in at least two regions, with some UMN signs rostral to LMN signs	··	Clinical or electrophysiological evidence of UMN and LMN signs in only one region, or UMN signs alone in two or more regions, or LMN signs rostral to UMN signs	··

LMN=lower motor neuron. UMN=upper motor neuron. ··=components are not part of the classification. *Neuroimaging and clinical laboratory studies must be done to exclude alternative diagnoses. †Regions: bulbar, cervical (corresponding to neck, arm, hand, diaphragm, and cervical spinal cord-innervated muscles), thoracic (corresponding to back and abdomen muscles), and lumbar (corresponding to back, abdomen, leg, foot, and lumbosacral spinal cord-innervated muscles).

Table 1.1: The El Escorial criteria and its revisions. Reproduced from Al-Chalabi et al. (2016).

Current genetic research

Approximately 10% of ALS is classed as familial, with the remaining cases considered sporadic. The discovery of genes linked to ALS has greatly increased pace in recent years; as a result, the genetic etiology of two-thirds of familial cases and approximately 10% of sporadic cases is now known (Laferriere & Polymenidou, 2015). A summary of gene frequencies in ALS is given in Figure 1.1.

In 1993, mutations in the Copper-Zinc superoxide dismutase 1 (SOD1) gene were identified as a cause of familial ALS. The first such gene to be identified, this led to the development of the transgenic SOD1 mouse. SOD1 mutations are present in 12-20% of familial ALS cases, as well as 2% of apparently sporadic patients (Chiò & Traynor, 2015).

The TARDBP gene, encoding the TAR-DNA binding protein TDP-43, was the next ALS gene to be discovered, in 2008. Mutations in TARDBP account for approximately 4% of familial ALS cases and approximately 1% of apparently sporadic cases. The discovery of TARDBP highlighted the importance of RNA processing in ALS pathology (Chiò & Traynor, 2015). This hypothesis was given further support in 2009 by the discovery that mutations of the fused in sarcoma (FUS) gene were another cause of familial ALS.

A hexanucleotide repeat expansion in C9orf72 is a major cause of ALS, accounting for 10% of all Western hemisphere ALS. The pathogenic expansion accounts appears in approximately 38% of familial ALS, and approximately 6% of apparently sporadic ALS in people of European ancestry (Alsultan et al., 2016). C9orf72 is the first large intronic repeat expansion to be implicated in ALS, and as repeat expansions are known to disrupt RNA metabolism in other neurodegenerative diseases, this further implicates altered RNA processing and protein degradation pathways as core aspects of ALS pathogenesis.

The transgenic rodent model, which overexpresses mutant SOD1, has been the primary animal model of ALS pathogenesis. More recent models include zebrafish, nematode, fruit fly and yeast.

As genetic data were not available for analysis in this study, they will play no part in the remainder of this study; this summary has been included for completeness.

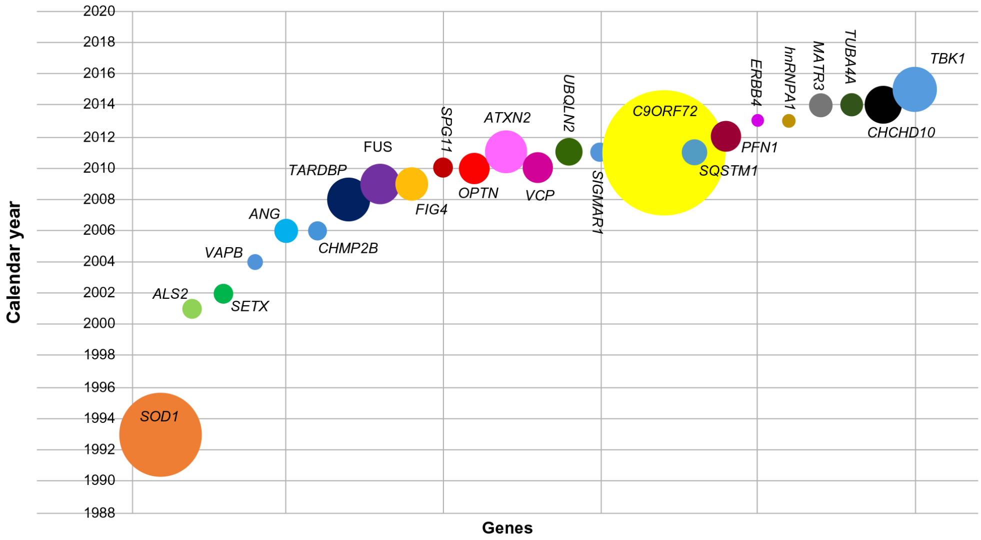


Figure 1.1: Gene frequencies in ALS, plotted against year of discovery. Circle size is proportional to the frequency of mutations, as cited in the literature. Reproduced from Alsultan et al. (2016).

1.2 Disease modelling and staging in ALS

A key aim of medical science is modelling patterns of disease progression; these patterns increase understanding of the disease, and help construct staging systems that assist diagnosis and treatment. Staging systems have multiple benefits: they enable better targeting of patient care, resource allocation, simple research classification and clinical trial design (with the aim to develop drugs that are beneficial at earlier stages, rather than later ones).

A search of the literature reveals that there is not a disease staging framework that is consistent across diseases. Instead, staging models are restricted to individual diseases, thus dependent on the clinicians and researchers specifically working on that disease, and tend to be clinically based.

These models typically use symptomatic staging to divide patients into small groups (e.g. presymptomatic, mild, moderate or severe), and assess the differences in biomarkers between the groups. Such models are relatively crude by nature; symptomatic staging relies on clinical assessment, which can be subjective and imprecise. This fundamental aspect of symptomatic staging limits the temporal resolution of any disease progression model that is constructed in this manner; lower temporal resolution reduces the effectiveness of a model to correctly stage patients, and critically, limits the ability of the model to discriminate between diseases. This last point is particularly relevant to ALS, given the heterogeneity of the disease.

1.2.1 Existing ALS clinical staging systems

The ALS Functional Rating Scale-Revised (ALSFRS-R)

The ALSFRS-R (Cedarbaum et al., 1999) is an established measure of functional decline in ALS. It is validated, correlates with survival and has been used as a primary outcome measure in clinical trials (Gordon et al., 2004). ALSFRS-R takes the form of a questionnaire-based scale, with the components of this scale grouped into four domains encompassing gross motor tasks, fine motor tasks, bulbar functions and respiratory function. Subscores over the domains are combined together to give a total score.

Although the ALSFRS-R has been extremely influential in ALS research, it now appears accepted that it is a flawed measure (Franchignoni et al., 2015). Recent evidence suggests that the scale provides greater differentiation if considering the individual subscores rather than the aggregate total score (Rooney et al., 2016); this is intuitively plausible, as using subscores avoids throwing away information.

The ALSFRS-R is a widespread measure and is routinely collected. It is currently the gold-standard measure of ALS disease progression (Menke et al., 2016).

The King's staging system

A simple clinical staging system for ALS (Table 1.2) has been developed by Roche et al. (2012). Known as the King's staging system, it is based on disease burden as measured by "clinical involvement and significant feeding or respiratory failure" (Fang et al., 2017).

Stage 1	Symptom onset (involvement of first CNS region)
Stage 2A	Diagnosis
Stage 2B	Involvement of a second CNS region
Stage 3	Involvement of a third CNS region
Stage 4A	Need for gastrostomy
Stage 4B	Need for respiratory support (non-invasive ventilation)
Stage 5	Death

Table 1.2: The King's staging system for ALS. The highest stage is taken if needed, so that a patient requiring non-invasive ventilation at diagnosis would have Stage 4B ALS, not Stage 2A. CNS = central nervous system.

Balendra et al. (2014a) devised an algorithm to estimate the King's clinical stage of ALS, based solely on a patient's ALSFRS-R score. The algorithm converts an ALSFRS-R score into one of the King's clinical stages, and can be applied retrospectively with 92% concordance (Balendra et al., 2014b).

An important aspect of this staging system is that although the clinical stage can be estimated from ALSFRS-R scores, the stages are not based on ALSFRS-R. This distinction emphasises that staging is not the same as function, thus avoiding the creation of a self-referential argument (i.e. derived ALSFRS-R scores used to create stages against which ALSFRS-R scores are correlated).

The King's staging system has limited temporal resolution, and the associated algorithm requires clinical assessment to determine the ALSFRS-R score.

The Milano-Torino (MiToS) staging system

The MiToS staging system, developed by Chiò et al. (2013), directly derives clinical stages from ALSFRS-R subscores. Commenting that the King's system stages are focused on clinical relevance rather than assessing functions directly relevant to patients, Chiò et al. define the MiToS milestones (Table 1.3) "by loss of independent function in four key domains that are included in... ALSFRS-R and that involve loss of autonomy: walking/self-care, swallowing, communicating and breathing".

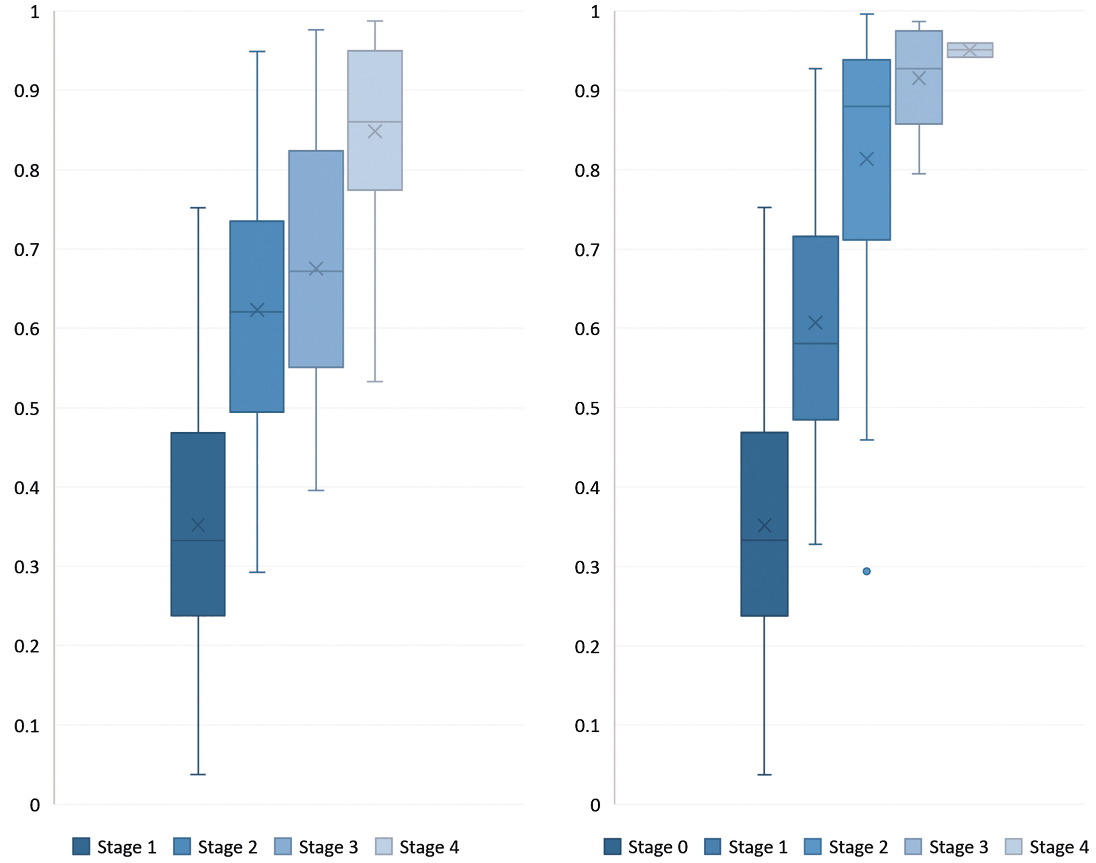
Like the King's system, MiToS suffers from limited temporal resolution. Furthermore, MiToS is wholly dependent on clinical assessment to determine the ALSFRS-R subscores.

Stage 0	Functional involvement but no loss of independence on any domain
Stages 1 to 4	Number of domains in which independence is lost
Stage 5	Death

Table 1.3: The MiToS staging system for ALS. Domains are defined on the ALSFRS-R and comprise swallowing, walking/self-care, communicating and breathing.

Comparison of the King's and MiToS systems

Fang et al. (2017) compared these two staging systems by applying them to data from the LiCALS clinical trial (Morrison et al., 2013), and concluded that there was good correspondence between the systems. The distribution of the disease stages was assessed as differing between the systems (Figure 1.2), with King's showing higher resolution



(a) King's staging system (b) MiToS staging system

Figure 1.2: Bar plots showing the standardised median proportion of time (SMT) from onset to each disease stage for the King's and MiToS staging systems, as applied to 217 ALS patients from the LiCALS clinical trial (Morrison et al., 2013). This illustrates differences in how the two systems distribute their stages throughout the progression of ALS. Reproduced from Fang et al. (2017).

through early to mid-disease and MiToS in the later stages. Because of this, the two systems were deemed to be complementary: "King's staging system summarises the clinical or anatomical spread of disease, while MiToS staging summarises the functional burden of disease". Fang et al. recommended that both systems should be used concurrently.

A second comparison of the two systems was undertaken by Ferraro et al. (2016). Both staging systems were applied to data from the Emilia Romagna Registry for ALS (ERRALS) in Italy. The authors concluded that the King's staging system showed "higher prognostic competency... especially for individual prognosis and as an outcome measure for clinical trials", but suggested that the MiToS system may be "more useful for estimating health costs and resource allocations".

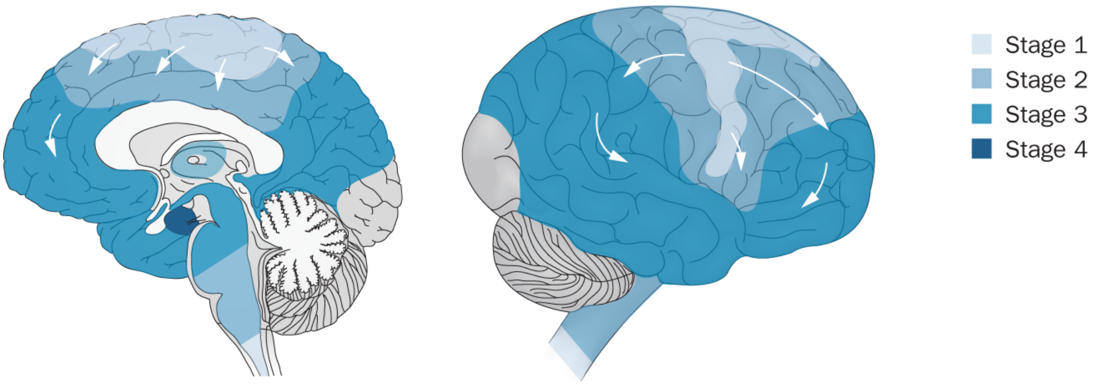


Figure 1.3: The four proposed stages of progression of abnormal pTDP-43 protein pathology in ALS. Adapted from Braak et al. (2013).

Neuropathological staging system

Employing a different approach from the aforementioned systems, Braak et al. (2013) proposed a neuropathological staging system for ALS, reporting that preliminary evidence indicated that neuronal involvement in ALS patients progressed at different rates but in a similar sequence. The Braak staging system consists of four stages, as shown in Figure 1.3, and is based on the progression of abnormal pTDP-43 protein pathology.

Braak et al. (2013) defined the four pathological stages of ALS as follows:

> Our findings indicate that the initial lesions develop in portions of the agranular frontal neocortex and in somato motor neurons of the spinal cord and lower brainstem (stage 1).
>
> In the second stage, mild pathology also develops in pre-frontal areas (such as the middle frontal gyrus) as well as in the reticular formation, precerebellar nuclei of the lower brainstem, and parvocellular portion of the red nucleus.
>
> In stage 3, the lesions progress into additional prefrontal areas (gyrus rectus, orbital frontal gyri) and then into postcentrally located sensory areas, which is accompanied by the appearance of pTDP-43 pathology in striatal medium-sized projection neurons [of the caudate nucleus and putamen].
>
> In stage 4, cortical pathology also develops in anteromedial portions of the temporal lobe, including the hippocampal formation.

It is worth noting that the data used to develop this system were obtained from staging of 22 central nervous system regions in an autopsy cohort of 76 patients. This raises the question of whether the observed variations in the disease pathology are truly indicative of ALS progression as a whole, or are instead artefacts found when extrapolating backwards from the disease end-point.

Philosophical differences between the existing staging systems and a data-driven approach to staging

The current staging systems suffer from not only low temporal resolution, but also the problem of human bias. The King's system depends on direct clinical assessment of patients to assign stages, while both the associated Balendra algorithm and the MiToS system use ALSFRS-R scores to stage patients (which again requires clinical assessment). The Braak system has human bias deeply ingrained into its design, as the data have been obtained by autopsy and the entirely subjective process of histopathological examination.

In contrast, data-driven staging attempts to reduce human bias to levels far below those of the clinical staging systems. This is achieved by designing and training a mathematical model of disease progression, and the model is then run only after declaring all relevant parameters and thresholds. Of course, bias can only be reduced in this manner, not completely eliminated. Likewise, it is critical to remember that any model will find patterns in data, and that even if these patterns exist and are not artefacts of the data, they may not have any biological meaning.

1.3 Diffusion Tensor Imaging (DTI)

1.3.1 Mechanisms of DTI

The most common technique for *in vivo* study of white matter (WM) changes in ALS is diffusion tensor imaging (DTI) (Turner & Verstraete, 2015), a type of Magnetic Resonance Imaging (MRI). DTI employs scanning sequences that are sensitive to the Brownian motion of water molecules in biological tissues. DTI analysis is based on the observation that this molecular displacement is affected by the characteristics of the medium in which the water molecules are located:

- Water in a free environment, such as cerebrospinal fluid (CSF), diffuses easily and equally in all directions.

- In biological tissue, cell walls and nerve fibres act as barriers, and the diffusion of the water molecules is hindered. In directional structures such as white matter fibre bundles, the diffusion of water molecules is greater in the direction of the fibre bodies than across them.

Assuming that the probability of water molecule displacement follows a multivariate Gaussian distribution over the observation diffusion time (Wheeler-Kingshott et al., 2003), the movement trajectories of the molecules can be described using the mathematical formalism of a second-order tensor, D:

$$D = \begin{bmatrix} D_{xx} & D_{xy} & D_{xz} \\ D_{yx} & D_{yy} & D_{yz} \\ D_{zx} & D_{zy} & D_{zz} \end{bmatrix} \tag{1.1}$$

The on-diagonal elements of D describe diffusivity along each of the cardinal axes, while the off-diagonal elements describe the covariance of the diffusion displacements for each pair of axes (Alexander et al., 2007). As the process of diffusion is symmetric for uncharged molecules such as water (Wheeler-Kingshott et al., 2003), $D_{ij} = D_{ji}$ for all $i, j \in \{x, y, z\}$, and so D is fully characterised by only six components. Combined with the physical properties of Brownian motion, this means that the diffusion tensor D is symmetric and positive definite.

From the Spectral Theorem for symmetric matrices (Axler, 1997), it follows that there exists a diagonal matrix Λ and an orthogonal matrix E such that $D = E\Lambda E^T$. The diagonal entries of Λ are the eigenvalues of D (λ_1, λ_2, λ_3), and the columns of E are the corresponding eigenvectors ($\varepsilon_1, \varepsilon_2, \varepsilon_3$):

$$D = E\Lambda E^T = \begin{bmatrix} \varepsilon_1 & \varepsilon_2 & \varepsilon_3 \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix} \tag{1.2}$$

The matrix Λ is called the diagonalised tensor, and is commonly designated $\mathbf{D}$.

Diffusion MRI scans divide the target sample into voxels, and quantify the mean diffusion within each voxel as a tensor. A second-order tensor is analogous to an ellipsoid, and the eigenvectors ($\varepsilon_1, \varepsilon_2, \varepsilon_3$) of the diffusion tensor give the three primary axes of the ellipsoid. The eigenvalues (λ_1, λ_2, λ_3) give the magnitude of water diffusion along each of the axes of the ellipsoid (Figure 1.4). The diffusion tensor can be rendered as a 3D ellipsoid glyph for every voxel of a dMRI scan, allowing for easy visual interpretation of the tensor (Figure 1.5).

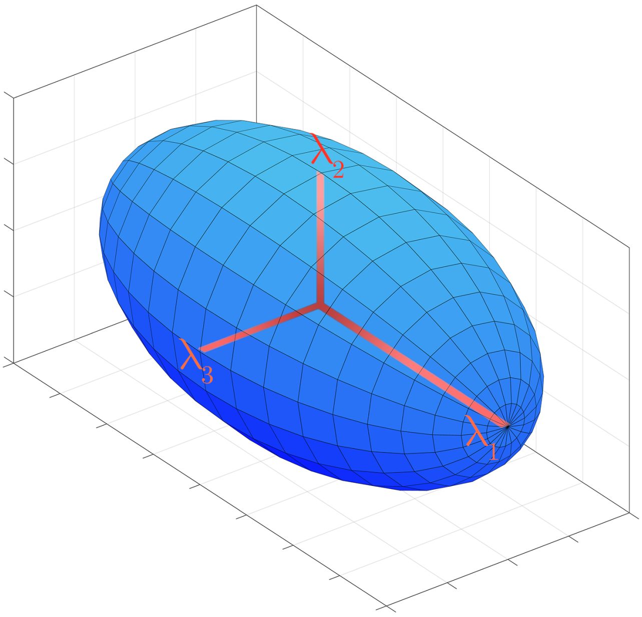


Figure 1.4: The diffusion tensor represented as an ellipsoid. The tensor eigenvectors $(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ define the ellipsoid's axes, and the tensor eigenvalues (λ_1, λ_2, λ_3) define the shape and size of the ellipsoid.

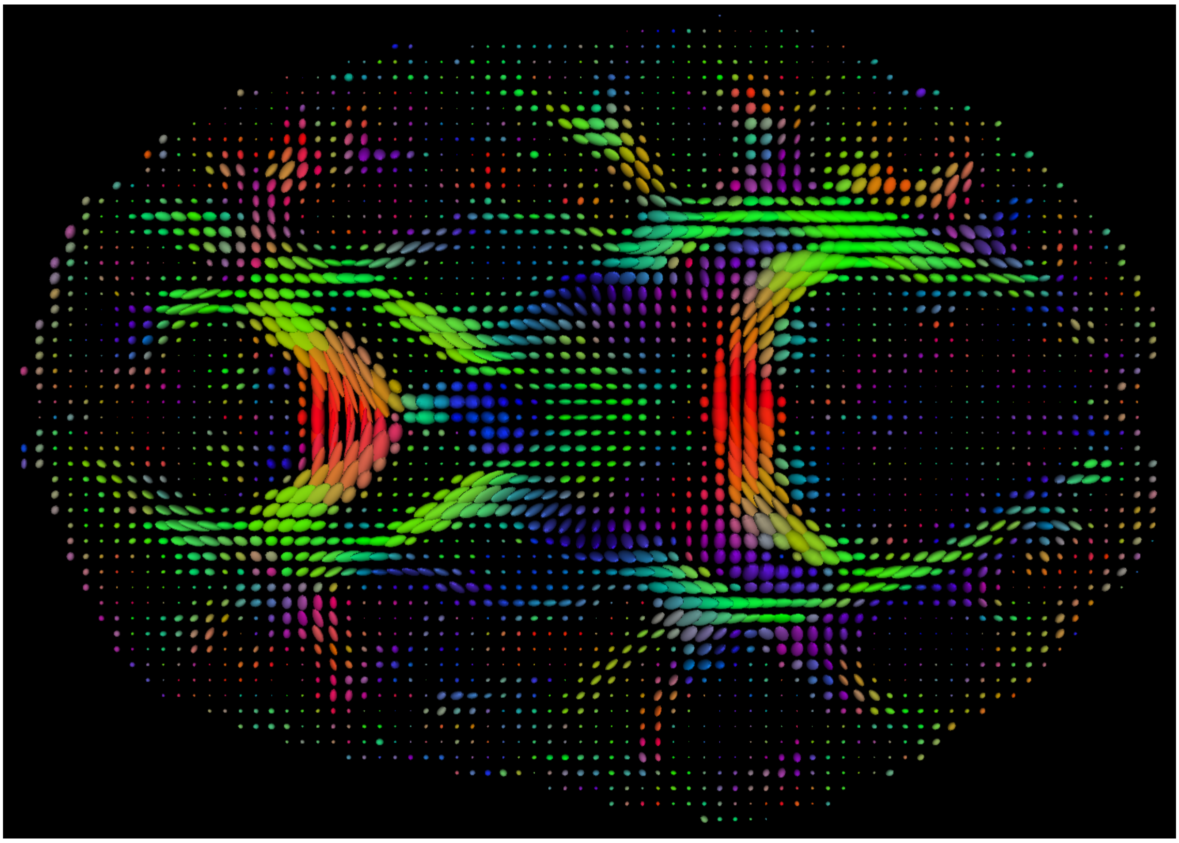
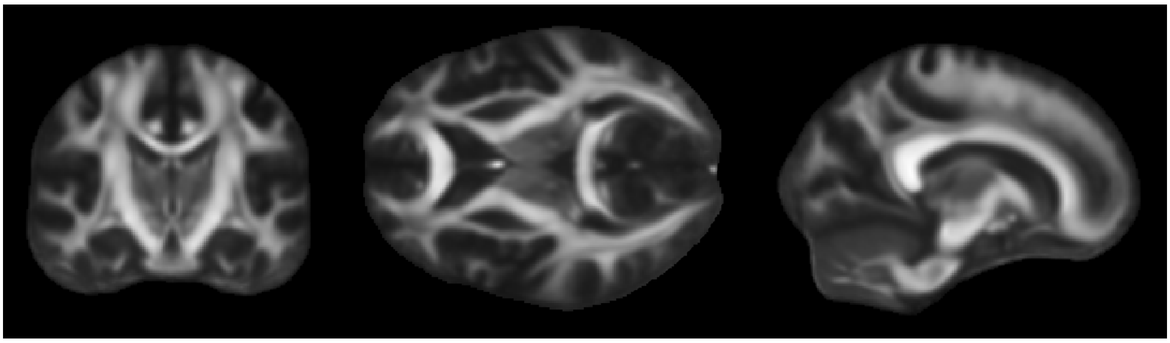


Figure 1.5: An axial slice of a DTI scan of a healthy brain. The diffusion tensor of each voxel has been rendered as a 3D ellipsoid glyph at 1.8 scale, and colour corresponds with the principle direction of diffusion. White matter structures are clearly visible.



(a) Coronal view (b) Axial view (c) Sagittal view

Figure 1.6: Single slices of a whole-brain Fractional Anisotropy (FA) image in the coronal, axial and sagittal planes. The high white-grey matter contrast of diffusion anisotropy is clearly visible, with brightness corresponding to higher FA. Compare the axial slice (b) to Figure 1.5.

The magnitude of diffusivity and the amount of anisotropy are the two scalar measurements that are most commonly used to represent the characteristic properties of the diffusion tensor.

Mean diffusivity, $\langle \mathbf{D} \rangle$, is equivalent to the mean of the eigenvalues:

$$\langle \mathbf{D} \rangle = \frac{\lambda_1 + \lambda_2 + \lambda_3}{3} \tag{1.3}$$

Several anisotropy indices have been proposed, with particular attention to their sensitivity to noise. The most commonly used is Fractional Anisotropy (FA), which is a scalar quantity, ranging from 0 to 1; i.e. isotropic to completely anisotropic diffusion (see Figure 1.6). FA is proportional to the square root of the variance of the eigenvalues divided by the square root of the sum of squares of the eigenvalues (Basser & Pierpaoli, 1996):

$$\mathrm{FA} = \sqrt{\tfrac{3}{2}} \left(\frac{\sqrt{(\lambda_1 - \langle \mathbf{D} \rangle)^2 + (\lambda_2 - \langle \mathbf{D} \rangle)^2 + (\lambda_3 - \langle \mathbf{D} \rangle)^2}}{\sqrt{{\lambda_1}^2 + {\lambda_2}^2 + {\lambda_3}^2}} \right) \tag{1.4}$$

1.3.2 Fractional Anisotropy as an index of Neurodegeneration

If the integrity of biological tissue is compromised (e.g. by neurodegeneration or trauma), then the diffusion anisotropy within the tissue will be altered. There are various different biological mechanisms by which this can occur, three of which are described here:

- *Demyelination*: the loss of the myelin sheath of a nerve cell. Myelin is an electrically insulating sleeve that is wrapped around the axon, and it acts as a diffusion barrier. The cylindrical myelin sheath is responsible for the high FA of white matter, and its loss will increase mean diffusivity (MD) and reduce FA. However, myelin alone does not fully account for anisotropy (Beaulieu & Allen, 1994).

- *Axonal loss*: nerve cell death, resulting in a reduced density of axons. It leads to a reduction in FA and an increase in MD.
- *Cell swelling*: an increase in the volumes of individual cells will reduce extracellular spaces, decreasing MD and FA.

Altered anisotropy is particularly apparent in white matter, due to its highly organised structure: "FA has the great advantage of being characterised by a high white–grey matter contrast, where FA = 1 for a cylindrically symmetric anisotropic medium with $\lambda_1 \gg \lambda_2 = \lambda_3$ and FA = 0 for complete isotropy, i.e. $\lambda_1 = \lambda_2 = \lambda_3$. Moreover, FA is very easy to interpret, having a great visual impact as white matter is white, while grey matter is grey." (Wheeler-Kingshott et al., 2003).

Although FA has high sensitivity to tissue microstructure, it "lacks specificity for individual tissue microstructure features" (Pierpaoli et al., 1996), due to the many possible biological mechanisms that can effect a change in anisotropy; FA is therefore considered a general, rather than specific, index of neurodegeneration.

1.3.3 Fractional Anisotropy in ALS

DTI studies of ALS patients have consistently reported group level involvement of the corticospinal tracts (Grolez et al., 2016), typically manifesting as a bilateral decrease in FA throughout the tracts. As FA lacks specificity for individual tissue microstructure features, there are many biological mechanisms that may contribute to the observed changes in diffusion. These include loss of pyramidal motor neurons in the primary motor cortex, axonal degeneration of the CST, the proliferation of glial cells, extracellular matrix expansion and intraneuron abnormalities (Turner et al., 2012). Despite the consistent demonstration of CST involvement in ALS, the FA measurements for individual ALS patients are currently considered to have "insufficient diagnostic accuracy" (Turner & Verstraete, 2015).

The corpus callosum is the second most implicated tract in ALS DTI studies, although its involvement (a reduction in FA, greatest in the middle-posterior parts of the tract) is reported less consistently than that of the CSTs (Menke et al., 2016; Turner et al., 2012).

1.4 Structural MR Imaging

1.4.1 Volumetry

In ALS research, DTI is used for investigating white matter *in vivo*, while high-resolution T_1-weighted structural MRI is the method of choice for the investigation of grey matter (Menke et al., 2016), as T_1-weighted structural MRI produces images with good tissue contrast between grey matter, white matter and CSF. There is a subtle but important difference between these two types of MRI in terms of the information that they yield: T_1 imaging allows macroscopic volumetric quantification while DTI provides information about the tissue microstructure.

There are a number of analysis tools that will perform automated segmentation of various cortical and subcortical brain structures when applied to high-resolution T_1-weighted images. These tools produce quantitative volumetric measures, as well as cortical thickness and surface area measures.

One such tool is Surface-Based Morphometry (SBM), an image analysis technique also known as cortical thickness measures. It allows the study of regional differences in grey matter through both volumetric and cortical thickness assessment. Working from the high tissue contrast of T_1-weighted images, SBM classifies each voxel as grey matter, white matter or CSF, and then reconstructs the boundaries between the three tissue types; this allows reconstruction of the cortical surface, and the cortex can then be segmented into regions based on gyral and sulcal structure.

Another tool is Voxel-Based Morphometry (VBM), similarly derived from reconstructing the boundaries between the three tissue types of T_1-weighted images. Unlike SBM, VBM can be applied to the whole brain, including subcortical structures (although only volumetric data can be generated for these regions).

1.4.2 Morphometry in ALS

SBM studies of ALS patients have consistently demonstrated cortical thinning in the precentral gyrus or primary motor regions (Turner et al., 2012). The subcortical structures of the caudate nucleus, hippocampus, nucleus accumbens and thalamus (see Figure 3.4) are also reported as showing progressive volumetric atrophy over time (Turner & Verstraete, 2015), although these findings are less consistent than those concerning cortical thinning.

Chapter 2

The Event-Based Model

The event-based model (EBM) characterises disease progression as a series of discrete events. An event comprises a significant change in patient state; formally, an event is defined as the change of a biomarker from a healthy state to a disease state. Cut-off points between healthy and disease states are not defined *a priori*, and are instead generated through data-driven mixture fitting.

By comparing the distributions of biomarker readings in patients to those of healthy controls, the model learns the ordering of events from heterogeneous measurements over a whole patient cohort. Crucially, the EBM is capable of extracting this temporal information from purely cross-sectional data; it does not require longitudinal data, which are less readily available in ALS research. This generative model of disease progression also provides insight into the variation of progression over a cohort.

2.1 Theory

The EBM consists of a set of events $E_1, ..., E_N$ and an ordering $S = (s(1), ..., s(N))$. S is a permutation of the integers $1, ..., N$ and determines the event ordering $E_{s(1)}, ..., E_{s(N)}$. The set of events are specified *a priori* and we estimate S from a data set X.

X contains a set of measurements X_j, $j = 1, ..., J$, from each of the J patients and X_l, $l = 1, ..., L$, from each of the L controls. Thus, measurement x_{ij} is informative about the occurrence of event E_i in patient j.

Fitting the model to the data requires evaluating the likelihood $P(X|S)$ of a particular event ordering given the data. We start by fitting simple models for the likelihood function $P(x_{ij}|E_i)$ on the measurement x_{ij} given that E_i has occurred, and similarly $P(x_{ij}|\neg E_i)$ on the measurement x_{ij} given that E_i has not occurred. This process is discussed in more detail in section 2.3.

If patient j is at position k in the progression model, then events $E_{s(1)}, ..., E_{s(k)}$ have occurred, while events $E_{s(k+1)}, ..., E_{s(N)}$ have not, and we can write the likelihood of that patient's data given S as

$$P(X_j|S,k) = \prod_{i=1}^{k} P(x_{s(i)j}|E_{s(i)}) \prod_{i=k+1}^{N} P(x_{s(i)j}|\neg E_{s(i)}) \tag{2.1}$$

where we assume that individual measurements are independent. We integrate out the hidden variable k to obtain

$$P(X_j|S) = \sum_{k=0}^{N} P(k)P(X_j|S,k) \tag{2.2}$$

where $P(k)$ is the prior probability of being at position k in the ordering. We assume a uniform prior on k here, in order to impose the least information possible onto the ordering.

We sum from $k = 0$ to allow for the possibility of patients' data being recorded when they have not experienced any of the significant events. The data x_{ij} of a patient in this position closely resembles that of a control subject; however, by including $k = 0$ in the potential orderings, we account for the possibility that we have not included significant biomarkers that occur early in the disease progression.

Assuming independence of measurements between patients gives

$$P(X|S) = \prod_{j=1}^{J} P(X_j|S) \tag{2.3}$$

and the total likelihood for a dataset X given an ordering S is obtained by combining Equations 2.1 - 2.3:

$$P(X|S) = \prod_{j=1}^{J} \left[\sum_{k=0}^{N} P(k) \left(\prod_{i=1}^{k} P(x_{s(i)j}|E_{s(i)}) \prod_{i=k+1}^{N} P(x_{s(i)j}|\neg E_{s(i)}) \right) \right] \tag{2.4}$$

2.2 Model estimation

We wish to calculate the likelihood for an ordering S given a particular dataset X. We use Bayes' Theorem to obtain the posterior distribution

$$P(S|X) = \frac{P(S)P(X|S)}{P(X)} \tag{2.5}$$

The normalisation constant $P(X)$ is analytically intractable, so we use a Markov Chain Monte Carlo (MCMC) algorithm to sample from $P(S|X)$. For ease of modelling, we make the assumption that *a priori* all orderings are equally likely, and so we use a uniform prior on the ordering S. More details about the MCMC algorithm, and the greedy ascent algorithm used to initialise it, are given in section 2.4.

2.3 Mixture models for the likelihood of data given events

By the law of total probability, for each event E, the probability distribution on the associated measurement x is given by

$$P(x) = P(x|E)P(E) + P(x|\neg E)P(\neg E) \tag{2.6}$$

and so the evaluation of Equation 2.1 requires separate models for both $P(x|E)$ and $P(x|\neg E)$.

To obtain these models for a particular event E_i, we need to fit a mixture of distributions to all the measurements associated with that event (i.e. for all patients and controls). The components of the mixture can then be extracted and used as models for $P(x|E)$ and $P(x|\neg E)$.

There are many different options for the choice of mixture distributions:

- a mixture of Gaussians with more than two components, as in Fonteijn et al. (2011). When applying the EBM to Alzheimer's disease (AD), there is reasonable justification for fitting three-component Gaussian mixtures, given the natural division of an AD cohort into i) cognitively normal, ii) mild cognitive impairment and iii) AD. However, in ALS populations, there is less uncertainty in the diagnosis than with AD, and the progression of the disease is fast enough to assume a sharp separation between patients and controls. Thus, this family of mixture distributions was discarded due to the risk of over-fitting.

- a Gaussian distribution mixed with a uniform distribution, as in Fonteijn et al. (2012). In general this is a reasonable alternative; it was not employed in this work simply because it was not appropriate for any of the selected biomarkers.

- a Gaussian distribution combined with a single fixed value for controls. ALSFRS-R scores could be modelled in this manner, as all controls retain the maximum score of 48 while patients' scores decrease throughout the course of the disease. However, for the ALSFRS-R biomarker, we have invoked Ockham's Razor and chosen a simpler approach: to discount the control data and fit a single Gaussian distribution to the patients' ALSFRS-R scores. This is possible only because the main EBM process is run on just patients, as opposed to the mixture model fitting which is performed on patients and controls.

Biomarkers that are binary events, such as a patient's Riluzole status (0 = not taking Riluzole, 1 = taking Riluzole), require different modelling from events characterised by continuous variables. Previous implementations of the EBM (Fonteijn et al., 2011) assigned fixed high and low likelihood values to $P(x_{ij}|E_i)$ and $P(x_{ij}|\neg E_i)$ respectively. In this study, I have not used binary variables as biomarkers, as I feel that they do not contain enough temporal information to have their progression order modelled meaningfully.

For this implementation of the EBM, we fit a mixture of two Gaussian distributions to each biomarker. Each mixture has five parameters: mean and standard deviation for healthy controls (μ_c, σ_c), mean and standard deviation for patients (μ_a, σ_a), and mixing proportion. Following the methods of Fonteijn et al. (2012) and Young et al. (2014), the mixing proportion is discarded; we then designate our models for $P(x|E)$ and $P(x|\neg E)$ as $X|E \sim \mathcal{N}(\mu_a, \sigma_a^2)$ and $X|\neg E \sim \mathcal{N}(\mu_c, \sigma_c^2)$, respectively.

2.4 The MCMC process

The MCMC process is a classic Metropolis algorithm (Metropolis et al., 1953). For each iteration t, it proceeds as follows:

- $P(X|S_t)$ is calculated for the current model order S_t
- a perturbation S' of the current model order S_t swaps the positions of two randomly chosen events
- $P(X|S')$ is calculated for the order S'
- We set $S_{t+1} = S'$ with probability $\min(a, 1)$, where $a = P(X|S')/P(X|S_t)$
- otherwise $S_{t+1} = S_t$

MCMC algorithms depend on appropriate initialisation in order to achieve good mixing properties within a practical time span, particularly if the chain is slow-mixing.

We therefore initialise the MCMC algorithm with an initial sequence close to the maximum likelihood solution S_{ML} for $P(X|S)$. This initial sequence is found using a greedy ascent algorithm, and is thus designated as S_g.

The greedy ascent algorithm uses the same perturbation rule as the MCMC algorithm, but sets $a = 1$ if $P(X|S') > P(X|S_t)$ or $a = 0$ otherwise.

We run the greedy ascent algorithm for 5000 iterations, which is generally enough to reach a local maximum. We repeat this from 100 different initialisation points, to increase the likelihood of the algorithm reaching the global maximum.

The MCMC algorithm is then run for 1 100 000 iterations, of which the first 100 000 are discarded as burn-in iterations.

While Young et al. (2014) use the MCMC algorithm to extract the maximum likelihood ordering S_{ML}, Fonteijn et al. (2012) suggest taking the characteristic ordering $\bar{S}$, which is the event order with the modal likelihood in the MCMC samples. Taking the characteristic ordering potentially aids in generalising the event order to the disease population as a whole, rather than restricting it to the specific dataset from which it has been obtained.

The ultimate output of the EBM is either the maximum likelihood ordering S_{ML} or a characteristic ordering $\bar{S}$, and a positional variance diagram illustrating the uncertainty in the ordering over the set of MCMC samples.

2.5 Modifying the EBM for ALS

2.5.1 Assumptions

Mathematical modelling, by definition, involves simplifications and making assumptions about real-world processes. The model assumptions made by Fonteijn et al. (2012) are appropriate for the EBM's original purpose:

> Two key assumptions of the event-based model are: first, measurements decrease monotonically as the disease progresses; second, the event sequence is consistent over all patients. The first assumption is in line with existing models of disease progression... thus, patients for whom event E_i has occurred cannot revert to an earlier state where E_i has not occurred. This assumption is essential because it ensures that snapshots are informative about the full event ordering. The second assumption is essential to enable

> pooling of snapshots from individual patients to inform the complete event sequence for the whole cohort.

The first assumption may hold for ALS progression, although ALSFRS-R is likely to be an exception as its score - although ultimately decreasing over time - has been observed to undergo local increases (Mandrioli et al., 2015). Nonetheless, this assumption seems reasonable for biomarkers directly assessing neurodegeneration, such as changes in regional FA or cortical thickness.

Given the heterogeneity of ALS, the second assumption is extremely unlikely to hold true. Consequences of this include greater positional variance (i.e. an increase in the uncertainty of results). The converse of this is that any clear and consistent results found, in spite of the heterogeneity of progression, are likely to be common to all phenotypes of ALS.

2.5.2 Mixture model fitting

There is an extensive literature on the subject of mixture model fitting (see McLachlan & Peel (2000)); consequently there are a great many methods for performing this task. To fit mixture models to data, Young et al. (2014) employ constrained non-linear minimisation in the form of Matlab's *fmincon* function; this method is extremely flexible and powerful, as well as allowing for the imposition of constraints on the values for the mixture model components. Recalling that ALS differs from AD in having a clearly defined control population, we potentially may require fewer constraints when fitting mixtures to ALS data.

An alternative and widely used method of fitting mixture models is the Expectation Maximisation (EM) algorithm, which is a class of optimising functions designed for maximum likelihood problems. The EM algorithm also functions well when presented with missing data values, a common problem in ALS research.

> ...the fitting of finite mixture distributions by maximum likelihood is a classic example of a problem that is simplified considerably by the EM's conceptual unification of maximum likelihood (ML) estimation from data that can be viewed as being incomplete. (McLachlan & Peel, 2000)

Matlab's *fitgmdist* function is specifically designed to fit mixtures of Gaussian distributions by EM. As the control population group is clearly defined in ALS, we can perform EM initialised by status (i.e. telling the algorithm how the data points are classified, rather than needing to infer the classifications). Lastly, the 'RegularizationValue' option of *fitgmdist* is used to place weak constraints upon the Gaussian mixture proportions. This is achieved by adding $\min(\sigma_c^2, \sigma_a^2)/100$ to the diagonal of the mixture parameter covariance matrix and refitting; this process is repeated until the mixing proportions meet the required constraints.

A final modification was made to allow for the removal of outliers during the mixture fitting process. In practice, the EM algorithm appears to be sensitive to extreme values (see Figure 2.1); thus, the option to exclude outliers during the fitting (and only during the fitting) has been included. This is achieved by separately fitting single Gaussian distributions to a biomarker's control data and patient data; outliers can then be defined by distance in standard deviations from the fitted mean. These outlier biomarker readings are still included when running the EBM, as they potentially represent patients at a late stage of progression.

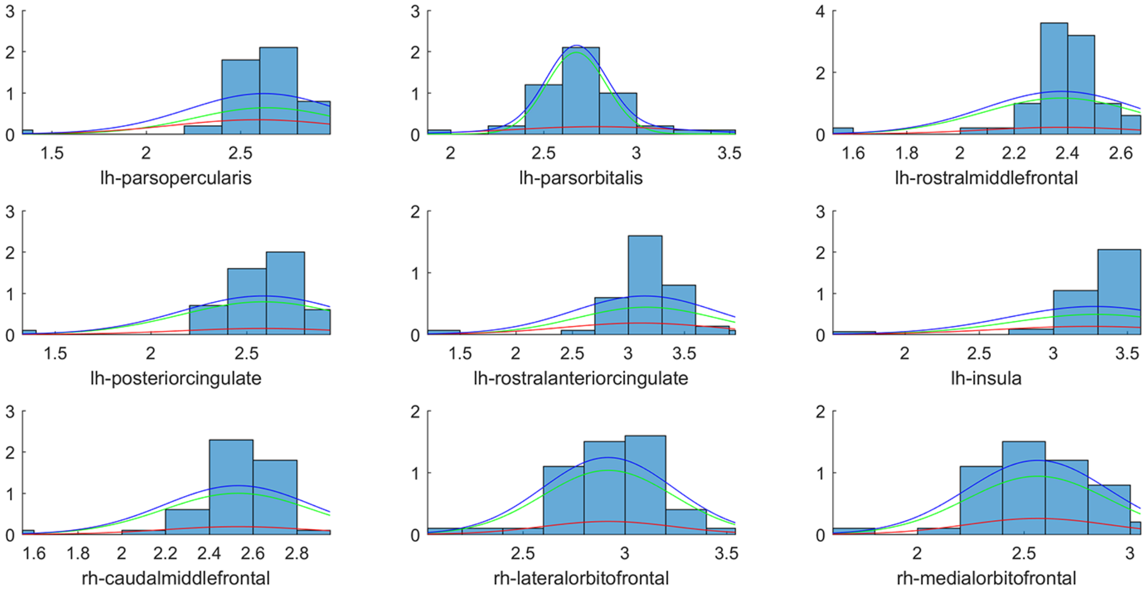
Unsurprisingly, the effects of outlier removal are felt most strongly in smaller cohorts. A word of caution: this is a blunt tool, and its use depends on the biomarker data being normally distributed.

2.5.3 Variable precision

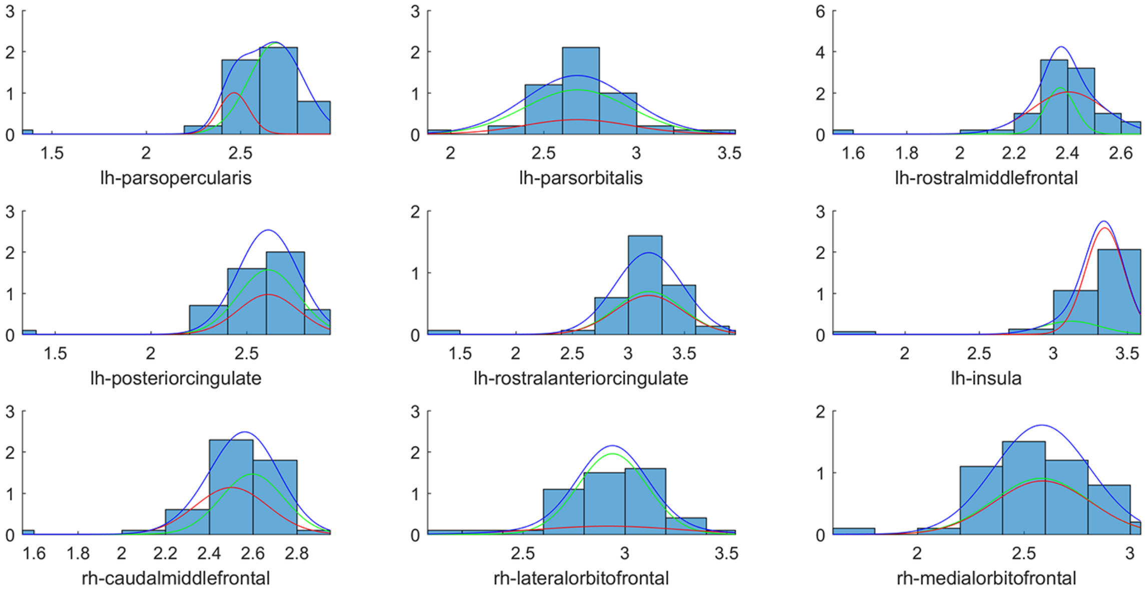
Step-by-step testing of the EBM revealed that for when modelling (in the region of) 45 biomarkers for 100 ALS subjects, Equation 2.4 was returning a probability of zero, regardless of proposed event order. Further investigation showed that this was a direct result of multiplying such a large number of probabilities together (or the summing of the logarithms of probabilities, which is equivalent). The large number of biomarkers meant that although $P(X|S)$ was not actually zero, it was extremely small; in this case, smaller than the precision with which Matlab is capable of working. By default, Matlab works in double-precision floating-point arithmetic, which has 16 digits of precision.

As a result of this, the main functions of this implementation of the EBM have been adapted to work in variable precision; quadruple precision has proved sufficient so far, although arbitrary levels of precision are possible. Two different methods for running in variable precision are supported: i) natively through Matlab's Symbolic Toolbox, and ii) through Advanpix's Multiprecision Computing Toolbox (Holoborodko, 2017), an external plugin for Matlab.

This enables the EBM to be used with larger datasets and more biomarkers than was previously possible.



(a) Fitted mixtures for cortical thicknesses, Set K. Outliers not excluded during fitting.



(b) Fitted mixtures for cortical thicknesses, Set K. Outliers $> 3\sigma$ excluded during fitting.

Figure 2.1: Illustration of the EM algorithm's sensitivity to extreme values, derived from fitting Gaussian mixtures to cortical thickness data.

2.5.4 "Patients only" vs "Patients and Controls"

As it was originally developed for use in Alzheimer's disease, previous implementations of the EBM have classified subjects into three categories: cognitively normal (CN), mild cognitive impairment (MCI) and AD. The mixture models were fitted to CN and AD data, but as "a significant proportion of the cognitively normal control group may have had presymptomatic AD" (Young et al., 2014), the EBM itself was run on all subjects from all three categories.

Unlike in Alzheimer's disease, the control population available in ALS research is clearly defined. We can make the assumption that ALS patients and controls will have very different progressions; if, indeed, controls can be said to progress in the same manner at all. This suggests that including control biomarker data in the progression modelling would greatly increase noise and the uncertainty of the results. Thus, once the mixture models have been fitted, we elect to run the EBM on only ALS patients' data. This has the added benefit of reducing the required processing time, as the EBM has substantially fewer data to manipulate.

2.5.5 CDF vs PDF

Probability Density Functions (PDFs) are often used when working with probability distributions. In this section, I argue that Cumulative Distribution Functions (CDFs) are better suited to the event-based modelling process in ALS. This argument is based on Fonteijn's first assumption that the progression of biomarker readings is monotonic, and relies on taking this assumption to its logical conclusion; this results in a stronger assumption than monotonicity of event order, but one that is still reasonable for a variety of biomarkers.

Having assigned models for $P(x|E)$ and $P(x|\neg E)$, we then use these model distributions to calculate specific probability values for individual patients' biomarker readings (see section 2.3). Recalling our first modelling assumption ("biomarker measurements decrease monotonically as the disease progresses"), subjects with a lower biomarker reading are therefore more likely to be in a disease state for that biomarker. This appears reasonable for the biomarker of FA, which is used as an index of neurodegeneration in white matter tracts (i.e. lower FA implies greater neurodegeneration of the tract), as dMRI studies have consistently shown reduced FA in the CSTs of ALS patients compared with healthy controls (Turner & Verstraete, 2015).

Using parameters approximately derived from the historical datasets, we can model mean FA tract values as $\mathcal{N}(0.9, 0.03^2)$. For two subjects with biomarker readings of 0.86 and 0.89, we can calculate PDF or CDF values to determine which subject has a higher probability of being in a disease state (see also Figure 2.2):

	FA = 0.86	$>$	FA = 0.89
PDF	5.47	$<$	12.58
$1 -$ CDF	0.91	$>$	0.63

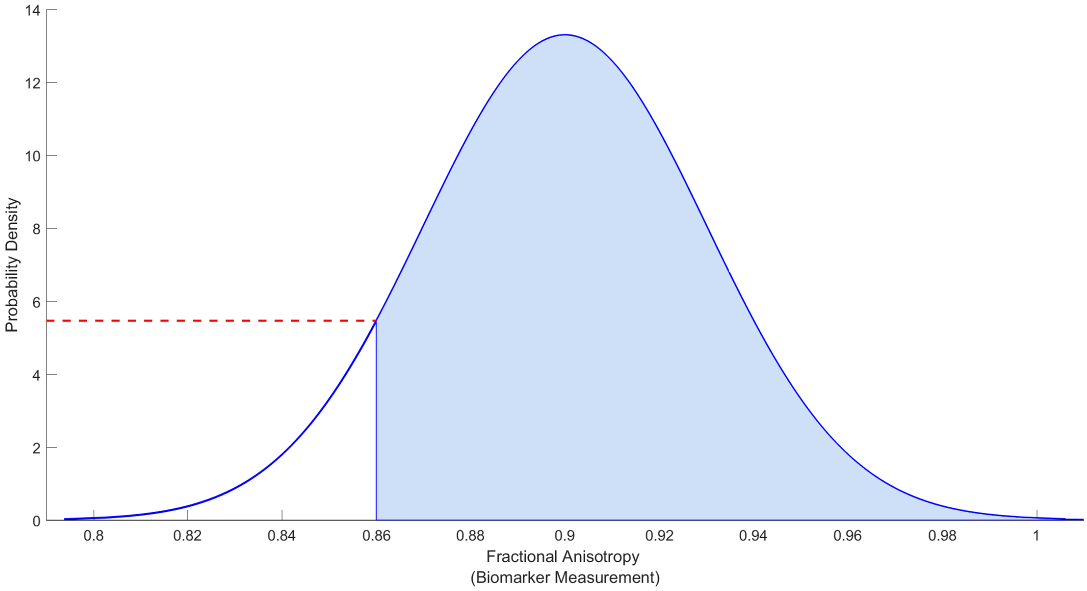
Putting aside the fact that a PDF is a *density* rather than a probability – hence able to be greater than 1 – we find that the PDF value is greater for the subject with the higher biomarker reading of 0.89; this is because the biomarker reading is close to the mean of the fitted Gaussian distribution. The lower biomarker reading of 0.86 is further away from the mean and so has a lower PDF value, even though it is likely to have come from a subject with greater neurodegeneration. Note that the reason that we use ($1 -$ CDF) is precisely because FA decreases with neurodegeneration (if the opposite were true and FA increased with neurodegeneration, we would instead use the CDF value to calculate the probability of being in a disease state).

To summarise: PDFs place emphasis on readings near the centre of the distribution, meaning that far-progressed patients will be considered less likely to be in a disease state than patients mid-way through progression. We can employ CDFs to reframe the problem; by using them, we are effectively asking "What is the maximum probability of a patient being in a healthy state, for a biomarker reading this extreme or worse?". See subsection 4.3.1 for testing and exploration of the effect that this has on the model output.

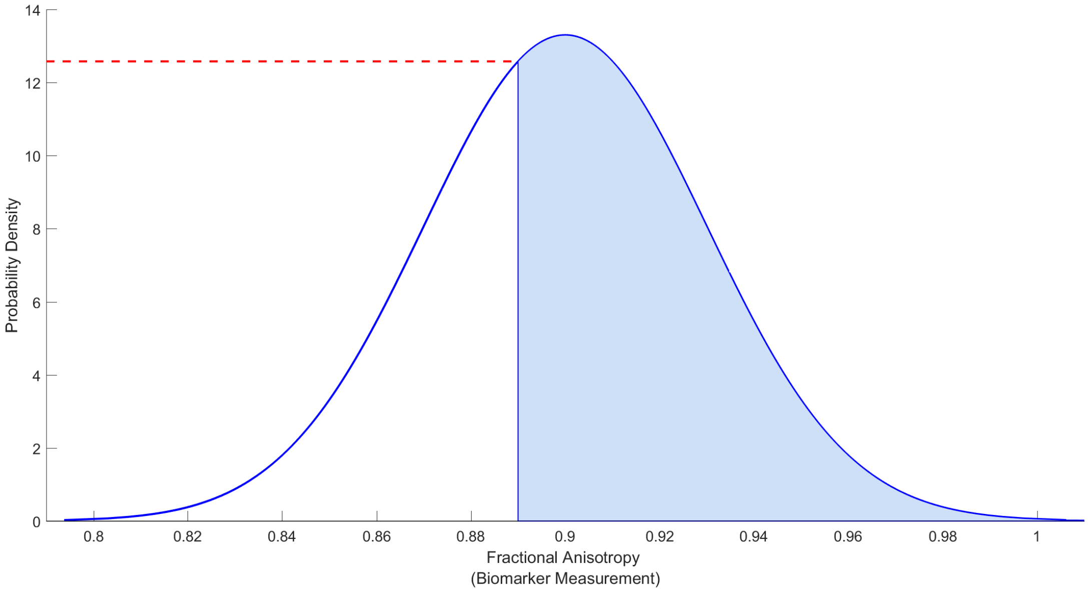
2.6 A Bayesian approach to event-based modelling

2.6.1 Bayesian inference

Probabilities are a numerical representation of a set of rational beliefs. For example, a fair coin toss will have $P(\text{Heads}) = 0.5$, which is equivalent to a belief that the coin is equally likely to land either side up when tossed. If, however, we toss the coin ten times and get all Tails, then we might reconsider our belief in the coin's fairness - but how should we change our belief? Although it seems possible that $P(\text{Heads}) = 0.5$ is too high, ten Tails in a row does not rule out ever getting a Heads. Bayes' Theorem (Equation 2.5) provides a rational method for updating beliefs in light



(a) Plot of $P(x_i|E_i)$. FA = 0.86 is marked.



(b) Plot of $P(x_i|E_i)$. FA = 0.89 is marked.

Figure 2.2: Illustration of PDF and $(1 - \text{CDF})$ values for two different readings of the same biomarker, Fractional Anisotropy. Decreasing FA values are associated with increasing neurodegeneration, so we can assume the subject in (a) is more likely to be in a disease stage than the subject in (b). As shown by the shaded areas, $(1 - \text{ CDF}(0.86)) > (1 - \text{CDF}(0.89))$. This would mean that subject (a) is more likely than subject (b) to be in a disease state, as expected. However, as indicated by the dotted red lines, $\text{PDF}(0.86) < \text{PDF}(0.89)$. Thus the PDF values imply that subject (b) is more likely to be in a disease state than subject (a), which is the converse of what the actual biomarker reading tells us.

of new information, and Bayesian inference is the process of inductive learning through the use of Bayes' Theorem (Gelman et al., 2013; Hoff, 2010).

The process of Bayesian inference involves passing from a prior distribution, $p(S)$, to a posterior distribution, $p(S|X)$, updating our beliefs via a sampling distribution, $p(X|S)$.

2.6.2 The Problem with Permutations and Priors

As mentioned in section 2.2, the current approach to event-based modelling uses a uniform prior on the ordering S. In Bayesian terms, we are stating that we have no prior knowledge about the order of the events. This imposes the least possible information on the model output, which is an accepted method in Bayesian inference (Hoff, 2010). However, the motivation here is primarily ease of modelling, as fitting a distribution to a sample space Ω consisting of all possible orderings turns out to be non-trivial. This means that currently we cannot fit a distribution to our prior or posterior probabilities; thus the output of the EBM is limited to only reporting the most likely event order, and we are unable to describe or characterise Ω in any detail.

Not being able to fit distributions to the prior or posterior means that we cannot update our prior knowledge (i.e. use our posterior probability as a prior probability), and so we are not taking advantage of one of the main strengths of Bayesian inference.

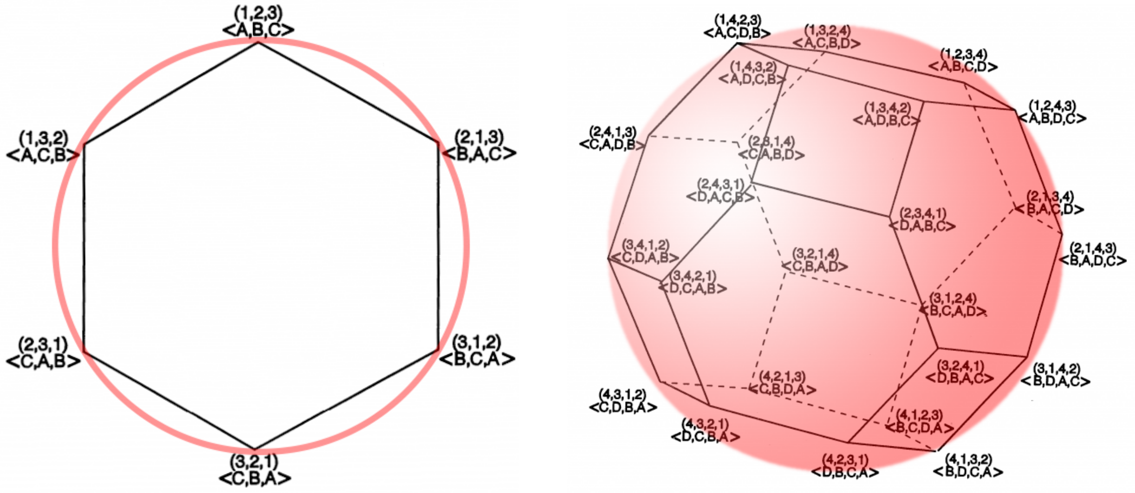
Below, I give a method for fitting a distribution to a sample space Ω consisting of permutations. This method is taken from Plis et al. (2010), and its incorporation into the EBM represents a substantial improvement to the model.

2.6.3 Permutations as directional data

The ultimate output of the EBM is the most likely event order, S_{ML}. During the application of the EBM, an arbitrary order is assigned to biomarkers (the order in which they are entered into the model), and so S_{ML} can be considered a permutation of this order. The core idea from Plis et al. (2010) is that permutations, a form of *ranked* data, can be transformed into *directional* data.

Background definitions

As a geometrical object, a sphere is defined as the set of points equidistant from a given point in three-dimensional space. More intuitively, a sphere is the surface of a ball, and is thus a two-dimensional object embedded in three-dimensional Euclidean



(a) (b)

Figure 2.3: (a) For permutations of order $n = 3$, the permutohedron is a hexagon inside a circle (a circle is a 1-sphere embedded in $\mathbb{R}^2$). (b) For $n = 4$, the permutohedron is a truncated octahedron inside an ordinary sphere (i.e. a 2-sphere embedded in $\mathbb{R}^3$). Adapted from Thompson (1993).

space. This concept can be generalised to higher-dimensional spaces: a *hypersphere* is an n-sphere embedded in an $(n + 1)$-dimensional Euclidean space.

A *permutohedron* of order n is a convex polyhedron whose vertices are in one-to-one correspondence with the permutations of order n (Thompson, 1993). Permutohedrons of order 3 and 4 are illustrated in Figure 2.3.

Finally, QR decomposition (also called a QR factorisation) of a matrix is a decomposition of a matrix A into a product A = QR of an orthogonal matrix Q and an upper triangular matrix R.

Transforming permutations into directional data

Plis et al. (2010) show how to embed a permutation set consisting of $n!$ elements onto the surface of a hypersphere embedded in $\mathbb{R}^{n-1}$ space. This is done by proving that the extreme points of a permutohedron of order n are located on the surface of a hypersphere of radius $r_s = \sqrt{\frac{1}{12}(n^3 - n)}$ with origin at the centre of mass of all $n!$ permutation vectors. The embedding process is as follows:

- Form a permutohedron from Ω. This permutohedron is located on a hyperplane in $\mathbb{R}^n$, and each vertex of the permutohedron is a point in $\mathbb{R}^n$ space whose coordinates are the permutations of n distinct numbers (Thompson, 1993).

- The extreme points (vertices) of this permutohedron are located on the surface of a d-hypersphere in $\mathbb{R}^{n-1}$ space, where $d = n - 2$.
- Transform elements of the discrete $n!$ permutation space into the continuous space of the d-hypersphere by:
  - a translation (move the centre of mass of the permutation vectors to the origin).
  - a change of basis (by projecting into $\mathbb{R}^{n-1}$). The basis is obtained by a QR factorisation of the centre of mass vector.
  - rescaling the hypersphere radius to give a unit length vector.

We can then use techniques from the field of directional statistics to fit a probability distribution on the hypersphere.

Fitting a von Mises-Fisher distribution to MCMC samples

The vMF distribution $M_p(\mu, \kappa)$ is a basic probability distribution from the field of directional statistics (Mardia & Jupp, 2000). The $(p-1)$-dimensional vMF density is parametrised by the mean direction μ and the concentration parameter κ, named because it characterises how strongly the unit vectors are concentrated around the mean direction (as shown in Figure 2.4). Note that $p = n$ in our permutohedron terminology.

An efficient method for maximum likelihood estimation for the vMF distribution is given by Sra (2012). This method requires a set of points sampled from the vMF distribution, in order to estimate μ and κ; we use the MCMC samples produced by the EBM for this purpose.

Updating the prior distribution

Returning to Equation 2.5, we are no longer restricted to a uniform prior on the ordering S, as we can now designate our model for $P(S)$ as $S \sim M_p(\mu, \kappa)$.

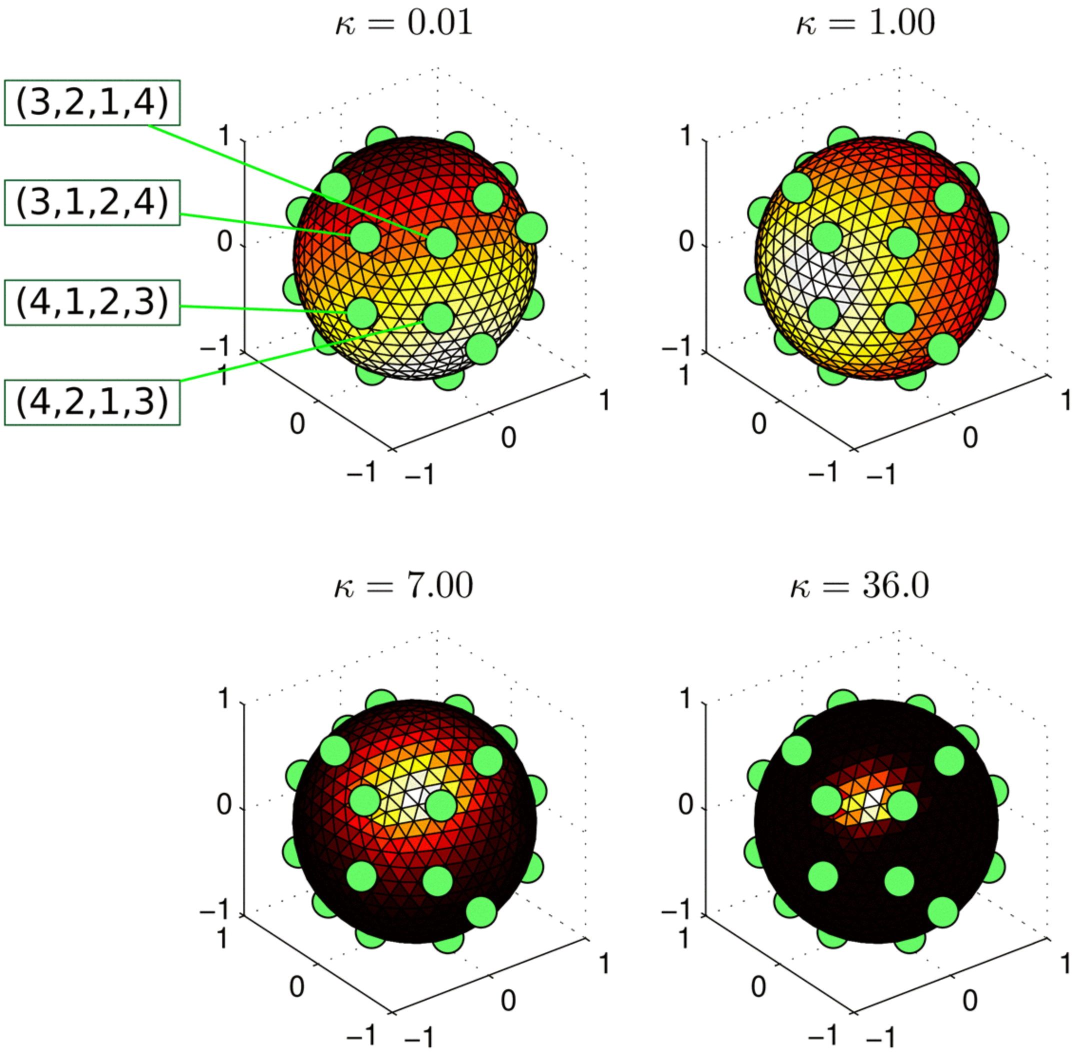


Figure 2.4: The vMF density function on a 2-sphere in $\mathbb{R}^3$ for different μ (mean direction) and κ (concentration parameter). Lighter colour indicate more probable areas. Green circles indicate points corresponding to vertices of the relevant permutohedron, the truncated octahedron shown in Figure 2.3 (b). Adapted from Plis et al. (2010).

Chapter 3

Materials and Methods

There are two legacy data sets available for use in this study, Sets F and K. The data sets were collected between 1999 and 2011 at the Centre for Neuroimaging Sciences, King's College London. Both sets are pre-existing, and comprise both clinical and MRI data for patients and healthy controls. The MRI data were collected on two different scanners, one per data set.

A contemporary data set, Set N, has also been made available. This data set was collected between 2014 and 2016 at the Clinical Imaging Sciences Centre, Brighton and Sussex Medical School. Table 3.1 lists the number of participants available with each data set, and more details about data acquisition are provided in section 3.1.

Data set	ALS (M/F)	Age ± SD	Controls (M/F)	Age ± SD	ALSFRS-R	Duration
Set F	36 (20/16)	53.9 ± 11.9	22 (14/8)	49.8 ± 15.6	37.5 ± 6.5*	35.5 ± 33.4
Set K	28 (25/3)	52.6 ± 11.8	25 (19/6)	48.1 ± 8.9	40.6 ± 4.1	25.6 ± 15.2
Set N	23 (16/7)	64.4 ± 8.0	23 (14/9)	61.5 ± 9.3	40.0 ± 5.2	19.3 ± 8.9

Table 3.1: Numbers and demographics of subjects of Sets FKN. Age in years, Duration in months. ± indicates standard deviation. * data available for 28 patients.

3.1 MRI acquisition

3.1.1 Set F

Dataset F was obtained on a 1.5T MRI system between 1999 and 2008. The protocol included dual echo fast spin echo to yield proton density and T_2-weighted data, and diffusion-weighted EPIs, with diffusion gradients applied along 32 directions and maximum b-value of 1300 s mm^{-2}. T_1-weighted data were not acquired.

Data were acquired using a GE Signa HDx system (General Electric, Waukesha, WI, USA), with actively shielded magnetic field gradients (maximum amplitude 40 mT m^{-1}). The protocol was based on that described by Jones et al. (2002). The body coil was used for RF transmission, and an 8 channel head coil for signal reception. Each volume was acquired using a multi-slice peripherally-gated doubly refocused spin echo EPI sequence, optimised for precise measurement of the diffusion tensor in parenchyma, from 60 contiguous near-axial slice locations with anisotropic ($1.875 \times 1.875 \times 2.5$ mm) voxels. The echo time was 101.3 ms while the effective repetition time varied between subjects in the range 12 and 20 RR intervals, depending on individual participants' heart rates. The maximum diffusion weighting was 1300 s mm^{-2}, and at each slice location, 7 images were acquired with no diffusion gradients applied, together with 64 diffusion-weighted images in which gradient directions were uniformly distributed in space.

3.1.2 Set K

All imaging for dataset K was obtained on a 3T MRI system between 2008 and 2011. The protocol included T_2-weighted fast spin echo, T_1-weighted volumes using inversion-recovery fast spoiled gradient recalled echo (FSPGR), diffusion-weighted echo-planar imaging (EPI), with diffusion gradients applied along 32 non-collinear directions and maximum b-value of 1300 s mm^{-2}, and arterial spin labelling data for cerebral blood flow quantification.

Data were acquired using a GE Signa HDx system (General Electric, Waukesha, WI, USA), with actively shielded magnetic field gradients (maximum amplitude 40 mT m^{-1}). The body coil was used for RF transmission, and an 8 channel head/neck/spine coil for signal reception, allowing a parallel imaging (ASSET) speed up factor of two. Each volume was acquired using a multi-slice peripherally-gated doubly refocused spin echo EPI sequence, optimised for precise measurement of the diffusion tensor in parenchyma, from 60 contiguous near-axial slice locations with isotropic ($2.4 \times 2.4 \times 2.4$ mm) voxels. The echo time was 104.5 ms while the effective repetition time varied between subjects in the range 12 and 20 RR intervals, depending on individual participants' heart rates. The maximum diffusion weighting was 1300 s mm^{-2}, and at each slice location, 4 images were acquired with no diffusion gradients applied, together with 32 diffusion-weighted images in which gradient directions were uniformly distributed in space.

3.1.3 Set N

Dataset N was obtained on a 1.5T MRI system between 2014 and 2016. The protocol included T_2-weighted dual echo turbo spin echo, fast fluid-attenuated inversion recovery (FLAIR), and volumetric MPRAGE with isotropic ($1 \times 1 \times 1$ mm) voxels . Multi-shell diffusion-weighted images were acquired with single-shot, twice-refocused pulse-gradient spin-echo EPI, using 3 b-values (9 directions for b=300 s mm^{-2}, 30 directions with b=800 s mm^{-2}, and 60 diffusion directions with b=2400 s mm^{-2}), optimised for neurite orientation dispersion and density imaging (NODDI) as detailed in Zhang et al. (2012). Ten non-diffusion weighted (b=0) volumes were acquired. A parallel imaging (GRAPPA) speed up factor of two was used.

Data were acquired using a Siemens Avanto system (Siemens AG Medical Solutions, Erlangen, Germany), with actively shielded magnetic field gradients (maximum amplitude 40 mT m^{-1}).

3.2 Neuroimaging biomarker selection

3.2.1 dMRI

Data set	ALS (M/F)	Age ± SD	Controls (M/F)	Age ± SD	ALSFRS-R ± SD
Set F	36 (20/16)	53.9 ± 11.9	22 (14/8)	49.8 ± 15.6	37.5 ± 6.5
Set K	28 (25/3)	52.6 ± 11.8	24 (19/5)	47.3 ± 8.2	40.6 ± 4.1
Set N	23 (16/7)	64.4 ± 8.0	23 (14/9)	61.5 ± 9.3	40.0 ± 5.2

Table 3.2: Numbers and demographics for dMRI data of Sets FKN, after the exclusion of one control subject from Set K.

In consultation with colleagues and collaborators, neuroimaging biomarkers were selected as the mean FA values of WM tracts previously studied and implicated in ALS. The tracts employed in this study and their locations are given in Figure 3.1.

Mean FA values were calculated for the left and right sides of each tract (with the exception of the corpus callosum which was divided into the genu, body and splenium). The left and right sides of the tracts were considered as separate regions in order to explore the possibility of bilateral asymmetry in ALS progression.

To allow the investigation of a directional progression of ALS WM neurodegeneration, the CSTs were split in the inferior to superior direction. The boundary was

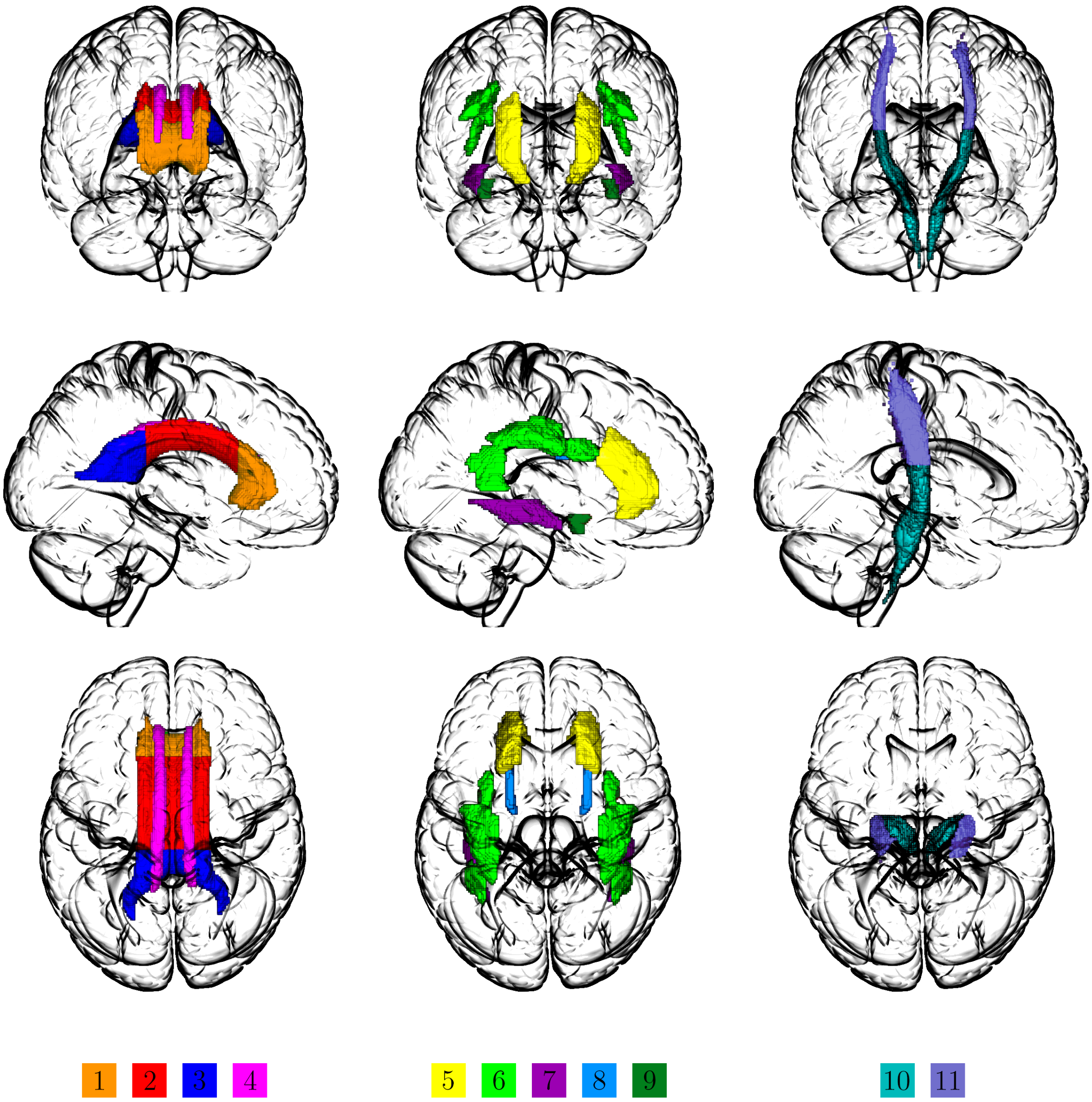


Figure 3.1: White matter tracts examined in this study. 1/2/3 = Corpus callosum Genu/Body/Splenium, 4 = Cingulum (cingulate gyrus), 5 = Anterior corona radiata, 6 = Superior longitudinal fasciculus, 7 = Sagittal stratum, 8 = Superior fronto-occipital fasciculus, 9 = Uncinate fasciculus, 10/11 = Inferior/Superior corticospinal tracts. Tracts and nomenclature are derived from the JHU DTI-based white-matter atlases (Mori et al., 2005).

chosen to be in the posterior limb of the internal capsule (PLIC), at the junction of the lateral ventricles and third ventricle (MNI coordinates -24 -23 15). The CSTs were selected for this subdivision due to their central and indeed pathognomonic role in the pathophysiology of ALS, as delineated in the original descriptions of the disease , and because these tracts have been consistently identified as abnormal in ALS DTI studies (Menke et al., 2016). In addition, their large size and coherent organisation renders these tracts especially suitable for studies using DTI and related MRI modalities.

After calculation of mean FA values, it was necessary to exclude one control subject from Set K, as their biomarker readings were classed as outliers for 6 of the 19 biomarkers. Here, outliers were defined in terms of sample quartiles, as described in Frigge et al. (1989): taking Q_1 and Q_3 to be the lower quartile and upper quartile respectively, a data point x is classed as an outlier if $x < Q_1 - k(Q_3 - Q_1)$ or $x > Q_3 + k(Q_3 - Q_1)$, where $k = 1.5$. Boxplots for the mean FA values of Sets FKN are given in Figure 3.2.

3.2.2 Structural imaging

Data set	ALS (M/F)	Age ± SD	Controls (M/F)	Age ± SD
Set K	28 (25/3)	52.6 ± 11.8	25 (19/6)	48.1 ± 8.9
Set N	23 (16/7)	64.4 ± 8.0	23 (14/9)	61.5 ± 9.3

Table 3.3: Numbers and demographics for structural T_1 MRI data of Sets KN.

There were no T_1 structural data available for Set F. All regions were selected from imaging atlases currently in widespread use, and were chosen in consultation with colleagues and collaborators.

Cortical regions

Due to the complexity inherent in volumetric analyses, biomarker selection was restricted to regions predefined in the Desikan-Killiany cortical atlas (Desikan et al., 2006) rather than defining them manually. The regions selected are given in Figure 3.3.

Subcortical regions

The automatic subcortical segmentation of a brain volume in FreeSurfer 5.3.0 is based on the Fischl et al. (2002) probabilistic atlas. Regions were selected from those defined in this atlas, which contains probabilistic information on the location of subcortical structures. The regions selected are given in Figure 3.4.

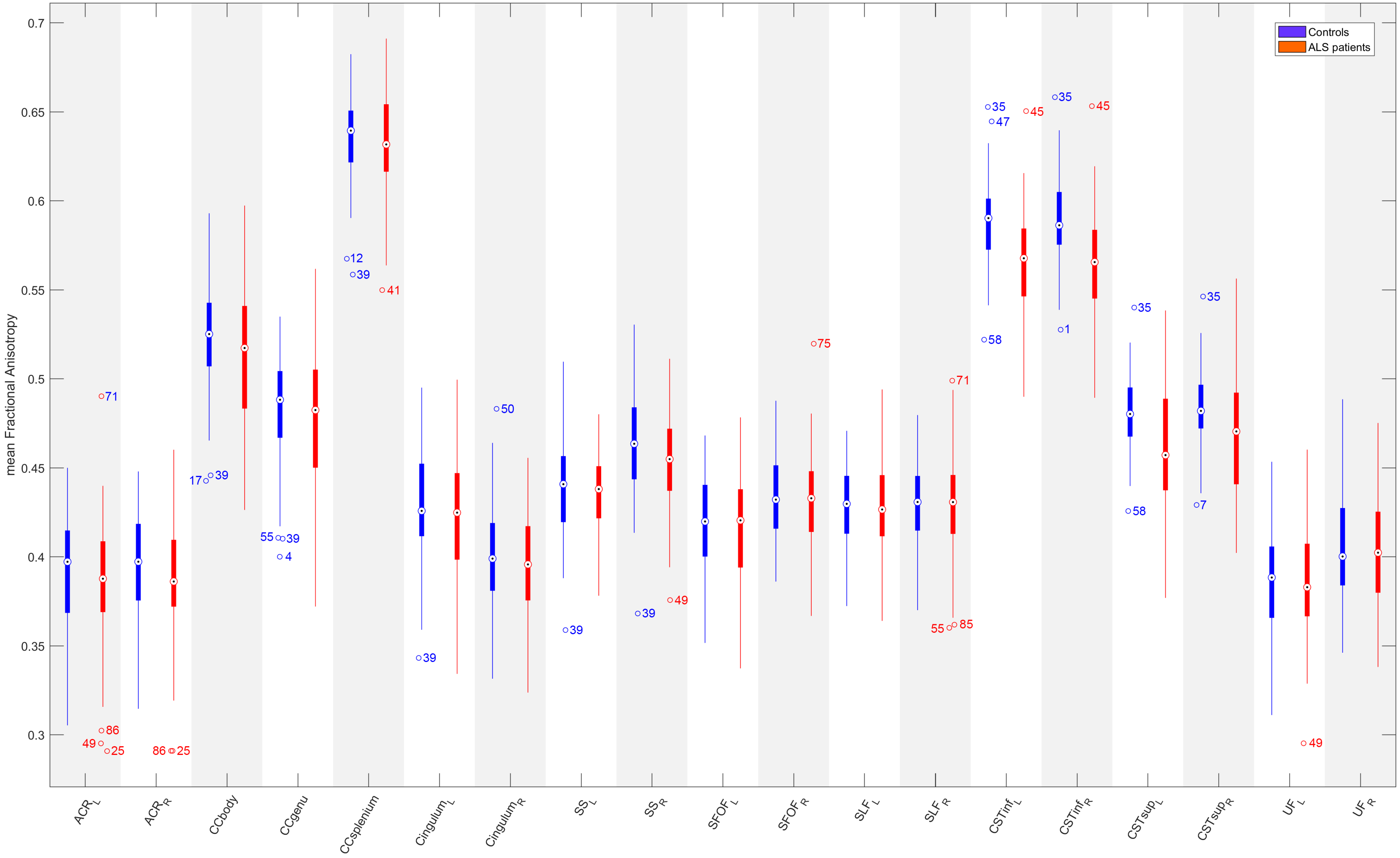


Figure 3.2: Boxplots of mean FA biomarker values for all 157 subjects of Sets FKN. Outliers are labelled with subject number.

3.3 MRI processing and analysis

3.3.1 Diffusion-Weighted Images

All diffusion-weighted images were corrected for involuntary motion and eddy current distortions using affine registration and the FMRIB's Linear Registration Tool (FLIRT). The images were skull-stripped using FMRIB's Brain Extraction Tool (BET).

Manual quality control was performed on all datasets, comprising of inspection for low signal to noise ratio and movement artefacts. The latter are a common feature of ALS imaging data, due to the nature of the disease. Due to its age, Set F required more extensive manual checking than Sets KN, and a greater proportion of data was deemed unusable.

For all diffusion imaging analysis, the single tensor (ST) model was used to derive FA. The ST model was fitted with weighted least squares, using FMRIB's *dtifit*.

Normalisation into MNI space was performed using ANTs 2.1.0. For each of the three datasets, a study-specific template was created using *antsMultiTemplateConstruction.sh*. Each subject's FA map was warped to the template, and all templates were warped to MNI space. These warps were combined to produce a single warp for each subject, which was then applied to their FA map.

For each tract, region of interest (ROI) masks were created from the two JHU DTI-based white-matter tractography atlases (Mori et al., 2005) included in FSL 5.0.7 (Jenkinson et al., 2012). The ROI masks for the corticospinal tracts were created from the "JHU white-matter tractography" atlas, thresholded at 25%; all other ROI masks were created from the "ICBM-DTI-81" atlas. FMRIB's *fslstats* was used to calculate mean FA values for each ROI.

3.3.2 Structural Images

The structural data of Set N were automatically processed and analysed using the FreeSurfer 5.3.0 *recon-all* pipeline. Manual pial surface editing was then performed by a colleague on Set N, and the pipeline re-run. The structural data of Set K were likewise processed and analysed, but did not receive the manual pial surface editing. These processes generated i) volumetric and cortical thickness data for the cortical regions given in Figure 3.3, and ii) volumetric data for the subcortical anatomical structures given in Figure 3.4.

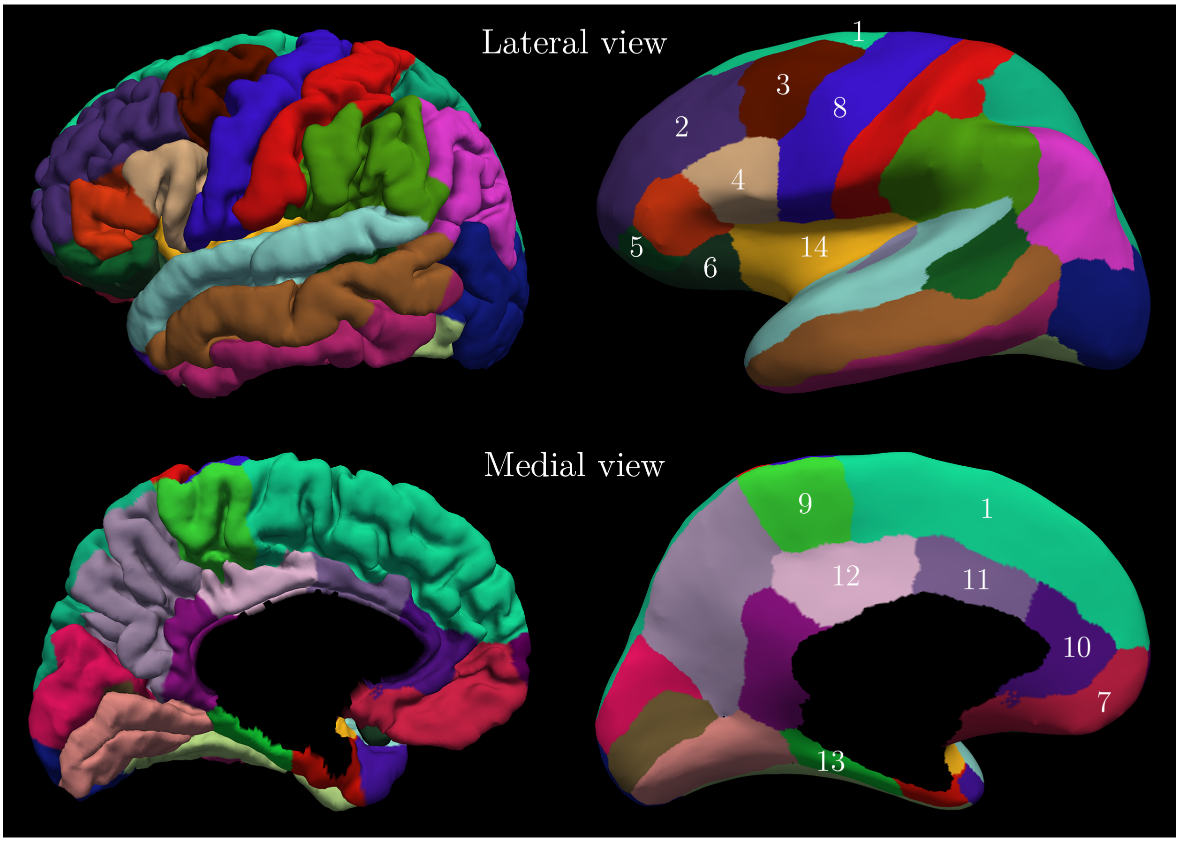


Figure 3.3: Cortical regions of the Desikan-Killiany atlas examined in this study. The right side of the figure shows the inflated cortical surface. 1 = superior frontal gyrus, 2 = rostral middle frontal gyrus, 3 = caudal middle frontal gyrus, 4 = pars opercularis, 5 = pars orbitalis, 6 = lateral orbitofrontal cortex, 7 = medial orbitofrontal cortex, 8 = precentral gyrus, 9 = paracentral lobule, 10 = rostral anterior cingulate cortex, 11 = caudal anterior cingulate cortex, 12 = posterior cingulate cortex, 13 = parahippocampal gyrus, 14 = insular cortex. Figure created using Freeview 1.0, included in FreeSurfer 5.3.0.

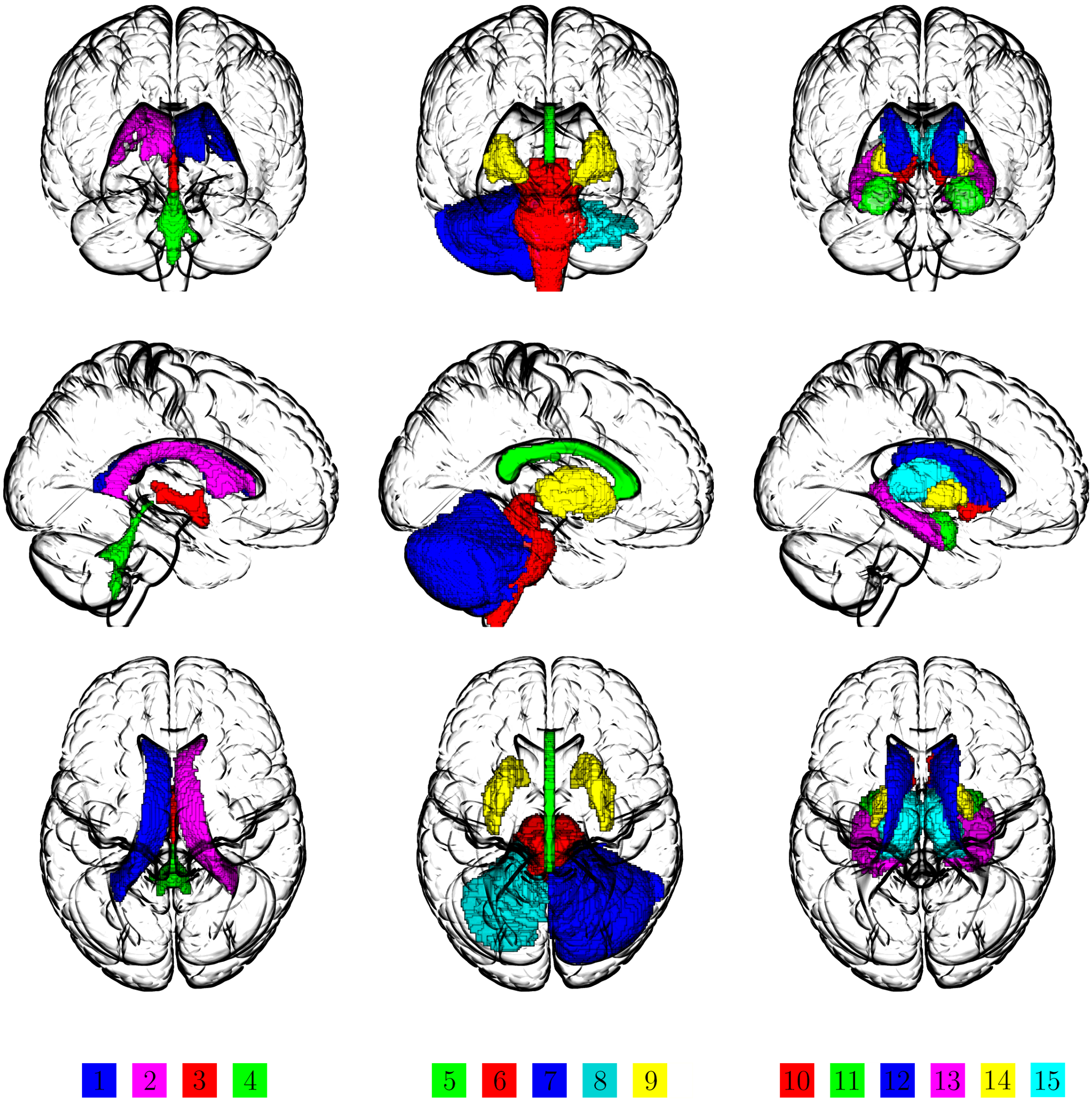


Figure 3.4: Subcortical regions examined in this study. 1/2 = Left/Right lateral ventricles, 3 = Third ventricle, 4 = Fourth ventricle, 5 = Corpus callosum, 6 = Brain stem, 7 = Cerebellum cortex (right side shown), 8 = Cerebellum white matter (left side shown), 9 = Putamen, 10 = Accumbens area, 11 = Amygdala, 12 = Caudate nucleus, 13 = Pallidum, 14 = Hippocampus, 15 = Thalamus proper.

3.3.3 Inter-cohort dMRI heterogeneity

In order to impose the minimum of information on the data, we chose to assess inter-cohort heterogeneity in as simple a manner as possible. This was accomplished by testing for significant centre differences in whole brain mean FA values, rather than comparing regional mean FA values. SPSS 23.0 was used to perform one-way ANOVA tests with the following manual contrasts:

- Between sets: F/K, F/N, K/N
- Between CTLs only: F/K, F/N, K/N
- Between ALS only: F/K, F/N, K/N

There were no significant differences at group level for any contrast, and *post hoc* tests (Bonferroni and Tukey) also showed no evidence of differing means.

3.4 Clinical and Neuropsychological data acquisition

There were a wide variety of clinical and neuropsychological data available in Set K, the vast majority of which were not available in Sets FN. We elected to model from only Set K to ensure a sufficiently large number of clinical and neuropsychological biomarkers.

Details of the assessments are reproduced here from Tsermentseli et al. (2015):

> Patients' functional abilities were assessed using the ALSFRS-R. Symptoms of anxiety and depression were measured using the Hospital Anxiety and Depression Scale (HADS).
>
> Formal standardised language testing included measurements of confrontation naming, semantic access, and single word and syntactic comprehension. Confrontation naming was measured with the Graded Naming Test (GNT). Semantic access of nouns/objects and verbs/action was measured by the Pyramids and Palm Trees (PPT) Test and the Kissing and Dancing Test (KDT), respectively. Syntactic and single word comprehension was assessed with the Test of Reception of Grammar (TROG) and the modified Token Test.

The connected speech sample was taken from the Cookie Theft picture description component of the Boston Diagnostic Aphasia Examination. Quantitative production analysis (QPA) was used to analyse the transcribed language samples following the approach used by Wilson et al. Transcripts were coded by two graduates trained in analysis of recorded discourse, blind to participants' status.

Distortions were defined as articulatory speech errors that did not involve frank phonemic substitutions. For patients with dysarthria, every word could potentially be classified as a distortion, so for these patients only words that were distorted above a patient's most accurate speech were coded as distortions. Semantic errors were recorded when participants produced sentences that were syntactically well-formed, but were either nonsensical or were semantically inappropriate for the context. The most common type of semantic errors involved substitutions of semantically related items. Discrepancies in coding were resolved following advice from a speech and language therapist.

The linguistic profile of connected speech production assessed measures under four main categories: 1) speech production (numbers of words, duration of narratives, speech rate, distortions and phonological paraphasias); 2) disruptions to fluency (false starts, filled pauses, repaired sequences and incomplete sentences); 3) lexical content (proportional frequencies of closed class words, pronouns and verbs); and 4) syntactic structure and complexity (mean length of utterances, proportional frequencies of words in sentences, number of embeddings, and semantic errors).

3.5 Clinical and Neuropsychological biomarker selection

Biomarkers were chosen in consultation with collaborators, and details of the biomarkers are given in Table 3.4. The main selection criterion was if a biomarker had been shown to be significantly different between patients and controls in previous analysis (Tsermentseli et al., 2015); biomarkers without significant differences were excluded, as were those with data distributions that led to poor Gaussian mixture fitting.

Biomarker name	**Details of Assessment**
log delta ALSFRS	(Log of) rate of change of ALSFRS-R score = log ((48 - ALSFRS score)/(duration in months))
HADSANX rev	Hospital Anxiety and Depression Scale – anxiety (revised)
CVLTSDR	California Verbal Learning Test short-delay Free Recall Correct; a list learning task
log CVB	(Log of) 4-letter C Word Fluency Index
Hayling Latency	Latency of response in the Hayling Sentence Completion Test. The difference in times between two conditions: one where the subject has to suppress an automatic response and complete a sentence with a word that renders the sentence meaningless, and one where they have to complete sentences in a meaningful way
Telephone SC	Telephone Search While Counting Scaled-Score. This is a measure of divided attention, and can be thought of as executive dysfunction
WCST err	Wisconsin Card Sorting Test; total errors
Distortions	Distortions per hundred words; linguistic errors during spoken speech describing a picture
Incomplete Sentences	Incomplete sentences per hundred words
Utterance	Mean length of utterance during spoken speech describing a picture
GNT	Graded Naming Test total score; expressive language, naming of line drawings of objects
Token	Modified Token test score; language comprehension test
TROG err	Test for the Reception of Grammar; number of errors on a sentence comprehension test

Table 3.4: Neuropsychological and functional assessments included in Set K. The biomarkers have been grouped into function and clinical, memory, executive function and speech and language.

Chapter 4

Challenging the Model *in silico*

The purpose of the EBM is to extract temporal information (TI) from cross-sectional data. In this chapter, I use simulations to examine some of the sources of TI, as well as exploring the effects of the main modifications that have been made to the model.

4.1 Tools for comparing Permutations

Altering parameters of the EBM results in different event orders, and to assess the impact of these parameter alterations, we require methods of quantifying differences between permutations. This section gives details of two metrics that will be used for this purpose: the normalised Kendall tau distance, and Kendall's rank correlation coefficient.

4.1.1 Terminology

We shall define and explain by example. Let S_1 and S_2 be two different event orders (also called permutations or rankings), such as

	A	B	C	D	E
S_1	1	2	3	4	5
S_2	2	1	5	4	3

For each event order S, we then calculate the sign of each possible pair of events. The sign or "score" (Kendall, 1970) of each event pair (i, j) is equal to $+1$ if $i < j$ and -1 if $i > j$. We declare an event pair to be *concordant* if the pair's sign is the same in both orders (and *discordant* if the sign differs):

Pair	S_1	S_2	Concordant
(A,B)	$1 < 2$	$2 > 1$	No
(A,C)	$1 < 3$	$2 < 5$	Yes
(A,D)	$1 < 4$	$2 < 4$	Yes
(A,E)	$1 < 5$	$2 < 3$	Yes
(B,C)	$2 < 3$	$1 < 5$	Yes
(B,D)	$2 < 4$	$1 < 4$	Yes
(B,E)	$2 < 5$	$1 < 3$	Yes
(C,D)	$3 < 4$	$5 > 4$	No
(C,E)	$3 < 5$	$5 > 3$	No
(D,E)	$4 < 5$	$4 > 3$	No

Thus, when comparing these two permutations, there are six concordant pairs and four discordant pairs.

For two rankings, S_1 and S_2, we formally define a concordant pair as a pair of two-variable observations i and j such that

$$\text{sgn}(S_{1j} - S_{1i}) = \text{sgn}(S_{2j} - S_{2i}) \tag{4.1}$$

and a discordant pair as

$$\text{sgn}(S_{1j} - S_{1i}) = -\text{sgn}(S_{2j} - S_{2i}) \tag{4.2}$$

4.1.2 Definitions

Kendall's tau distance is simply defined as the number of discordant pairs, with the **normalised Kendall tau distance**, $\boldsymbol{K}$, given by

$$K = \frac{\text{number of discordant pairs}}{\text{number of possible pairs}} \tag{4.3}$$

We recall that the binomial theorem tells us that for n events, there are $\binom{n}{2} = n(n-1)/2$ possible pairs. Returning to the example given in the previous section, the normalised Kendall tau distance is therefore

$$K(S_1, S_2) = \frac{4}{5(5-1)/2} = 0.4 \tag{4.4}$$

Note that K lies in the interval $[0, 1]$, where 0 indicates no distance (i.e. identical permutations) and 1 indicates maximum distance (i.e. the permutations are "diametrically opposite" when embedded on a hypersphere).

The **Kendall rank correlation coefficient**, $\boldsymbol{\tau}$, explicitly quantifies the degree of similarity between two permutations:

$$\tau(S_1, S_2) = \frac{\text{no. of concordant pairs} - \text{no. of discordant pairs}}{\text{no. of possible pairs}} \tag{4.5}$$

τ lies in the interval $[-1, 1]$ where 1 indicates maximum correlation, 0 indicates no correlation and -1 indicates maximum inverse correlation (i.e. the same permutation reversed).

The correlation coefficient allows us to perform tests of significance to establish whether two permutations are significantly similar. Kendall (1970) proves that for $n > 10$, the sampling distribution of τ approximates to the normal distribution, while for $n \leq 10$, the distribution can be calculated exactly. The significance tests for τ are available in Matlab, through the *corr* function. If the EBM is functioning better than blind chance, then S_1 and S_2 should be positively correlated, and thus we can perform a one-tailed test.

Both of the above metrics (K and τ) provide useful information for comparing two permutations; even though they are closely related to each other ($\tau = 1 - 2K$), they can be viewed in conceptually different ways. For the Bayesian adaption of the EBM, we have transformed permutations into directional data by embedding them on to the surface of a hypersphere; the nature of this embedding is such that for any permutation, its nearest neighbours on the hypersphere surface will have one discordant pair. This is a direct consequence of the nature of the permutohedron, which is constructed so that vertices are adjacent and connected if their orderings differ by a pairwise adjacent transposition (Thompson, 1993). Thus, the non-normalised tau distance represents an actual physical quantity: the smallest possible number of edges linking two permutations.

However, the correlation coefficient is a measure of similarity, and as such does not have an equivalent physical representation.

Finally, we acknowledge that notation for these metrics is not consistent throughout the literature, with τ being used to represent both Kendall's tau distance and the Kendall rank correlation coefficient. For the sake of clarity and consistency, the following notation will be used in this work:

$$K(S_1, S_2) = \text{normalised Kendall tau distance between event orders } S_1, S_2 \tag{4.6}$$

$$\tau(S_1, S_2) = \text{Kendall rank correlation coefficient between event orders } S_1, S_2 \tag{4.7}$$

4.2 Simulations

As one of the building blocks of the modelling process, it is clear that the fitted mixture models will be responsible for a large amount of TI. In this study, each Gaussian mixture model has five parameters (μ_c, μ_a, σ_c, σ_a and mixing proportion); we note that the current implementation of the EBM makes use of the first four parameters, discarding mixing proportion.

For the purposes of challenging the model, I have chosen to simulate data for a theoretical disease consisting of ten biomarkers. We begin by selecting a "true" event order (for simplicity, we let this be Events 1 to 10 in order), and impose the appropriate mixture component parameters on the model. Then, we vary one or both of the TI-rich mixture parameters for ALS patients (μ_a, σ_a) , while keeping the control parameters (μ_c, σ_c) fixed. Data are randomly generated from the component distributions, and the EBM is run on these data. The output of the model, the most likely event order S_{ML}, can then be compared to our theoretical "true" order T. (Essentially, this is a reversal of the mixture fitting process; instead of fitting distributions to the data, we are choosing the mixture distribution parameters in order to create a "true" event order, and then generating data from those distributions.)

For simplicity, we choose $X|\neg E \sim \mathcal{N}(0.9, 0.03^2)$ for all ten biomarkers, giving the same distribution for all healthy controls. Matlab's *normrnd* was used to generate a 50 by 10 matrix of values from this distribution, to act as our healthy control data. This matrix was kept the same for every simulation; this means that there was a certain amount of noise within the simulated control data, but that noise remained constant during all iterations of the model.

The process of testing the model (varying one TI parameter while keeping controls the same, simulating patient data, running the EBM) was repeated 1000 times for each TI parameter. This required upwards of 150 hours of cluster processing time, dependent on the number of patients that were being simulated; for each TI parameter that was investigated, we performed four separate runs of the 1000 repetition process (one each for 25, 50, 100 and 200 simulated ALS patients). In total, each results table in this section required 4.4 billion MCMC iterations.

4.2.1 Mean separation

An immediate source of TI is the difference between the mixture component means, $|\mu_a - \mu_c|$. Intuitively, if an event occurs early on in the disease, then we would expect a large separation between the component means. Conversely, an event from late on in the disease progression should have only a small separation between the component means. This is dependent on the modelling assumption that biomarker progression is monotonic, and thus the differences between ALS patients and healthy controls cannot decrease over time. The TI parameters for this set of simulations are given in Table 4.1.

	Controls	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10
μ	0.9	0.81	0.82	0.83	0.84	0.85	0.86	0.87	0.88	0.89	0.9
σ	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03

Table 4.1: Mixture model parameters used for simulating the effects of mean separation $|\mu_a - \mu_c|$ on the EBM's output. B1 = Biomarker 1, B2 = Biomarker 2, and so on.

For each run of the EBM process, we calculate the tau distance K between the model output, S_{ML}, and the theoretical order obtained from sorting the biomarker means, T_μ. We then take the mean K over all 1000 runs, $\bar{K}(S_{ML}, T_\mu)$. If mean separation does affect the output of the EBM in the manner we hypothesise, then we expect $\bar{K}$ to be low. This implies that the mean tau correlation coefficient $\bar{\tau}$ should be positive and high.

As a sanity check, we repeat this process for the tau distance K between the model output, S_{ML}, and the theoretical order obtained from reverse-sorting (see subsection 4.2.2) the biomarker standard deviations, T_σ. We again take the mean K across all 1000 runs, giving us $\bar{K}(S_{ML}, T_\sigma)$. As we have chosen $\sigma_a = \sigma_c$, we expect the variance separation to have no effect on the model output, suggesting $\bar{K}(S_{ML}, T_\sigma) \simeq 0.5$.

Finally, four series of 1000 EBM runs are performed, for different numbers of simulated ALS patients. The results of these four series are given in Table 4.2, and appear to confirm that mean separation is a strong source of TI. We use the mean tau correlation coefficient to perform a test of significance (Table 4.3), which shows that regardless of number of simulated patients, there is strong evidence to reject the null hypothesis of independence of S_{ML} and T_μ. As expected, there is no evidence to reject the hypothesis of independence of S_{ML} and T_σ.

Patients	$\bar{K}(S_{ML}, T_\mu)$	$\bar{K}(S_{ML}, T_\sigma)$
25	0.1801	0.5006
50	0.1352	0.5041
100	0.1005	0.5074
200	0.0750	0.4995

Table 4.2: Mean tau distance between S_{ML} and T, for varied $|\mu_a - \mu_c|$ with $\sigma_a - \sigma_c = 0$

Patients	$\bar{\tau}(S_{ML}, T_\mu)$	$\bar{\tau}(S_{ML}, T_\sigma)$
25	0.6398*	-0.0012
50	0.7296*	-0.0082
100	0.7990**	0.0148
200	0.8500**	0.0011

Table 4.3: Mean tau correlation between S_{ML} and T when varying $|\mu_a - \mu_c|$.
*$p < 0.005$, **$p < 0.001$

4.2.2 Variance

The effect that the component variances (σ_c^2, σ_a^2) may have on TI is less conceptually clear than for mean separation. Biologically, it seems reasonable to expect σ_a^2 to be greater than σ_c^2, due to the varying rates of ALS progression. As with mean separation, we intuitively would expect $\sigma_a^2 - \sigma_c^2$ to increase over time (i.e. earlier event = greater variance separation).

The TI parameters for this set of simulations are given in Table 4.4. Note that Biomarkers 7 - 10 have been assigned standard deviations such that $\sigma_a < \sigma_c$. While this may be biologically unlikely, these parameter values were dictated by wanting $|\sigma_{B_i} - \sigma_{B_{i+1}}| \geq 0.005$ in order to ensure that $\sigma_{B_i} \neq \sigma_{B_{i+1}}$ after simulating small numbers of patients.

	Controls	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10
μ	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9
σ	0.03	0.055	0.05	0.045	0.04	0.035	0.03	0.025	0.02	0.015	0.01

Table 4.4: Mixture model parameters used for simulating the effects of patient biomarker variance, $\sigma_a^2 - \sigma_c^2$, on the EBM's output. B1 = Biomarker 1, and so on.

Again, we calculate mean K after 1000 EBM runs. If variance separation does affect the output of the EBM in the manner we hypothesise, then we expect $\bar{K}(S_{ML}, T_\sigma)$ to be low, and $\tau(S_{ML}, T_\sigma)$ to be positive and high. Our sanity check is that we expect $\bar{K}(S_{ML}, T_\mu) \simeq 0.5$ and $\bar{\tau}(S_{ML}, T_\mu) \simeq 0$.

The results of the four series of 1000 EBM runs are given in Table 4.5, and appear to confirm that variance is a source of TI. We use the mean tau correlation coefficient to perform a test of significance (Table 4.6): for fifty or more simulated patients, there is evidence at the 5% level to reject the null hypothesis of independence of S_{ML} and T_σ. These significance levels are lower than those of Table 4.3, suggesting that for these parameters, variance is a weaker source of TI than mean separation. As expected, there is no evidence to reject the hypothesis of independence of S_{ML} and T_μ.

Patients	$\bar{K}(S_{ML}, T_\mu)$	$\bar{K}(S_{ML}, T_\sigma)$
25	0.5325	0.3174
50	0.5298	0.2626
100	0.5318	0.2092
200	0.5283	0.1709

Table 4.5: Mean distance between S_{ML} and T, varying $\sigma_a - \sigma_c$, with $|\mu_a - \mu_c| = 0$

Patients	$\bar{\tau}(S_{ML}, T_\mu)$	$\bar{\tau}(S_{ML}, T_\sigma)$
25	-0.0650	$0.3652^{\dagger}$
50	-0.0596	0.4747*
100	-0.0637	0.5816**
200	-0.0567	0.6581***

Table 4.6: Mean tau correlation between S_{ML} and T when varying $\sigma_a - \sigma_c$.
$^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.005$

4.2.3 Mean separation and Variance

We have seen that i) mean separation is proportional to progression through disease (i.e. greater difference between biomarker means implies an early event), and ii) variance is also proportional to progression through disease (i.e. greater variance of patient readings implies an early event). The next question to ask is how these sources of TI are related: what happens when varying both $|\mu_a - \mu_c|$ and $\sigma_a - \sigma_c$?

Firstly, we vary these two parameters in the **same direction** (high mean separation with high variance), as given in Table 4.7.

	Controls	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10
μ	0.9	0.81	0.82	0.83	0.84	0.85	0.86	0.87	0.88	0.89	0.9
σ	0.03	0.055	0.05	0.045	0.04	0.035	0.03	0.025	0.02	0.015	0.01

Table 4.7: Mixture model parameters, for same-direction varying of $|\mu_a - \mu_c|$ and $\sigma_a - \sigma_c$. B1 = Biomarker 1, B2 = Biomarker 2, and so on.

After performing four series of 1000 EBM runs, we calculate mean K and find that both $\bar{K}(S_{ML}, T_\mu)$ and $\bar{K}(S_{ML}, T_\sigma)$ are low (Table 4.8), while both $\bar{\tau}(S_{ML}, T_\mu)$ and $\bar{\tau}(S_{ML}, T_\sigma)$ are significantly high and positive (Table 4.9). In other words, both TI parameters are acting in the same manner, and so we cannot distinguish which parameter is having the greater effect on the model.

Patients	$\bar{K}(S_{ML}, T_\mu)$	$\bar{K}(S_{ML}, T_\sigma)$
25	0.2053	0.2260
50	0.1631	0.1803
100	0.1357	0.1465
200	0.1037	0.1090

Table 4.8: Mean tau distance between S_{ML} and T, for same-direction varying of $|\mu_a - \mu_c|$ and $\sigma_a - \sigma_c$

Patients	$\bar{\tau}(S_{ML}, T_\mu)$	$\bar{\tau}(S_{ML}, T_\sigma)$
25	0.5893**	0.5481*
50	0.6738**	0.6394**
100	0.7287***	0.7071***
200	0.7926****	0.7821****

Table 4.9: Mean tau correlation between S_{ML} and T, for same-direction varying of $|\mu_a - \mu_c|$ and $\sigma_a - \sigma_c$. $*p < 0.05$, $**p < 0.01$, $***p < 0.005$, $****p < 0.001$

Secondly, we vary our TI parameters in **opposite directions** (i.e. high mean separation with low variance), as given in Table 4.10.

	Controls	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10
μ	0.9	0.81	0.82	0.83	0.84	0.85	0.86	0.87	0.88	0.89	0.9
σ	0.03	0.01	0.015	0.02	0.025	0.03	0.035	0.04	0.045	0.05	0.055

Table 4.10: Mixture model parameters, for opposite-direction varying of $|\mu_a - \mu_c|$ and $\sigma_a - \sigma_c$. B1 = Biomarker 1, B2 = Biomarker 2, and so on.

Following another four series of 1000 EBM runs, we again calculate mean K and τ. This time, we find that $\bar{K}(S_{ML}, T_\mu)$ is low, while $\bar{K}(S_{ML}, T_\sigma)$ is high (Table 4.11). Equivalently, $\bar{\tau}(S_{ML}, T_\mu)$ is significantly strong and positive, while $\bar{\tau}(S_{ML}, T_\sigma)$ is significantly strong but negative (Table 4.12). This means that the model output is likely to be similar to the mean separation order T_μ and the reverse of the variance order T_σ.

Here, we have shown that for these particular parameter values, $|\mu_a - \mu_c|$ outweighs $\sigma_a - \sigma_c$ as a TI parameter. In order to quantify the relative importance of these TI parameters, further simulations with more widely varying values of σ_a are needed.

Patients	$\bar{K}(S_{ML}, T_\mu)$	$\bar{K}(S_{ML}, T_\sigma)$
25	0.1727	0.8165
50	0.1244	0.8675
100	0.0879	0.9066
200	0.0592	0.9388

Table 4.11: Mean distance between S_{ML} and T, varying μ_a and σ_a in the opposite direction

Patients	$\bar{\tau}(S_{ML}, T_\mu)$	$\bar{\tau}(S_{ML}, T_\sigma)$
25	0.6545*	-0.6330*
50	0.7512**	-0.7350**
100	0.8242**	-0.8132**
200	0.8816***	-0.8776***

Table 4.12: Mean tau correlation between S_{ML} and T, varying μ_a and σ_a in the opposite direction. $*p < 0.005$, $**p < 0.0005$, $***p < 0.00005$

4.2.4 Sample size

Small cohorts are a long-standing problem in ALS research, due in part to the difficulties in diagnosing the disease and its rapid progression. Also, the cost of collecting large data sets is high, and so there is a great deal of ongoing research into overcoming the difficulties of combining smaller datasets; consequently, there are many ideas on how best to harmonise data sets from different sources.

Large sample sizes enable us to study small effect sizes. For the EBM, it is not clear how the concept of effect size relates to TI, but it is possible to use the simulated data to quantify the relationship between sample size N and mean tau distance $\bar{K}$. We focus on varied $|\mu_a - \mu_c|$, and as we currently have only four sizes of cohort (Table 4.2), we run four more series of simulations for 75, 150, 350 and 500 patients (Table 4.13), and then plot $\bar{K}$ against number of simulated patients (Figure 4.1).

Patients	$\bar{K}(S_{ML}, T_\mu)$
25	0.1801
50	0.1352
75	0.1159
100	0.1005
150	0.0840
200	0.0750
350	0.0567
500	0.0458

Table 4.13: Mean distance between S_{ML} and T, varying μ_a

We can clearly see that there is a strong relationship here. We use Matlab's Curve Fitting Toolbox to fit a power distribution of the form $y = ax^b + c$ to the data, giving distribution coefficients (95% confidence bounds) as $a = 0.6250$ $(0.5780, 0.6721)$, $b = -0.3414$ $(-0.3880, -0.2949)$ and $c = -0.02838$ $(-0.04617, -0.01059)$. Thus, the relationship between mean tau distance and simulated patients is

$$\bar{K} = (0.625 \times N^{-0.341}) - 0.0284 \tag{4.8}$$

Recalling that $\tau = 1 - 2K$, this means that the relationship between simulated patients and mean tau correlation is given by

$$\bar{\tau} = 1.057 - (1.250 \times N^{-0.341}) \tag{4.9}$$

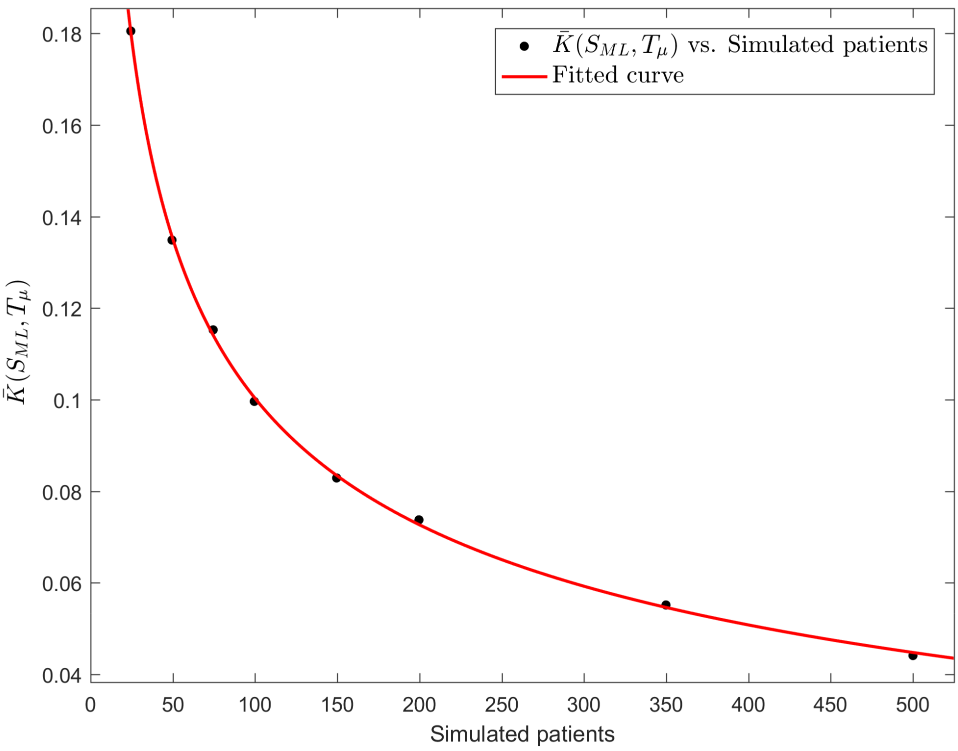


Figure 4.1: Plot of $\bar{K}$ against number of simulated patients, with fitted curve.

Clearly, increased patient numbers will result in greater accuracy (defining accuracy as reduction in distance between S_{ML} and our theoretical "true" order T), and indeed, all the simulations of the previous sections support this idea.

It is impossible to say how accurate the EBM will be for non-simulated data, as we cannot know what the "true" order will be. However, we suggest that increasing the number of actual patients from 200 to 500 would not result in a worthwhile increase in accuracy. A sample size in the region of 100 patients appears be sufficient to obtain $\bar{K} \leq 0.1$ (although we note that for these TI parameters, even a sample size of 25 simulated patients results in a significant correlation between S_{ML} and T, $p < 0.005$).

4.2.5 Number of biomarkers

In this section, I investigate how the EBM output, S_{ML}, is affected when biomarkers are added to or removed from the data set being modelled.

Adding biomarkers

For each iteration of the comparison process, the EBM is run on a theoretical disease consisting of ten biomarkers, generated as described in section 4.2. This gives us the most likely event order for ten biomarkers, $S_{ML_{10}}$. Mixture parameters for an eleventh biomarker are chosen such that $\mu_c = 0.9$, $\sigma_c = 0.03$, $\mu_a \in (0.8, 0.9)$ and $\sigma_a = 0.03$. Data are then generated from these parameters, in the same manner as the other biomarkers. A random μ_a is selected in order to simulate the effects of adding biomarkers that occur at different points in the disease progression. The EBM is then re-run on the original ten plus one new biomarkers, which gives us the most likely event order for eleven biomarkers, $S_{ML_{11}}$. We then remove the eleventh biomarker from $S_{ML_{11}}$. This gives us the most likely position of biomarkers 1 to 10, calculated from a model run on eleven biomarkers, which we designate as $S^*_{ML_{11}}$. We can now directly compare $S_{ML_{10}}$ and $S^*_{ML_{11}}$, by examining mean tau distance and the tau correlation coefficient:

Patients	$\bar{K}(S_{ML_{10}}, S^*_{ML_{11}})$	$\bar{\tau}(S_{ML_{10}}, S^*_{ML_{11}})$
25	0.0782	0.8436*

Table 4.14: Simulating the effects of adding a biomarker. $^*p < 0.0001$

Multiplying $\bar{K}$ by the number of possible pairs tells us that the mean tau distance is approximately 3.5. This means that on average, the effect of adding an eleventh biomarker is to increase the number of discordant pairs by 3-4, a remarkably small amount given that we have simulated only 25 patients. As an aside, 474 of the 1000 iterations had $\bar{K} = 0$ (i.e. adding a biomarker had no effect on the order of the original ten biomarkers).

Removing biomarkers

As before, we run the EBM on a theoretical disease consisting of ten biomarkers, generating $S_{ML_{10}}$. This time, however, we remove a biomarker at random, and then re-run the EBM on the remaining nine biomarkers, generating S_{ML_9}. The removed biomarker is then deleted from $S_{ML_{10}}$, giving us the most likely position of the nine persisting biomarkers, as calculated from a model run on ten biomarkers; we designate this as $S^*_{ML_{10}}$. We can now directly compare S_{ML_9} and $S^*_{ML_{10}}$, by examining mean tau distance and the tau correlation coefficient:

Patients	$\bar{K}(S_{ML_9}, S^*_{ML_{10}})$	$\bar{\tau}(S_{ML_9}, S^*_{ML_{10}})$
25	0.0842	0.8316*

Table 4.15: Simulating the effects of removing a biomarker. $^*p < 0.0005$

Table 4.15 shows that, on average, there is low distance and high correlation between S_{ML_9} and $S^*_{ML_{10}}$, even though we have simulated only 25 patients. We find that 497 of the 1000 iterations had $\bar{K} = 0$ (i.e. removing a biomarker had no effect on the order of the nine persisting biomarkers).

One more question arises: does the position of the removed biomarker have an effect on the output of the EBM? To answer this, we take all 1000 samples of S_{ML_9} and $S^*_{ML_{10}}$, group them by position of the removed biomarker, and then plot histograms of $K(S_{ML_9}, S^*_{ML_{10}})$ for each of the ten positions (Figure 4.2). We find that there appears to be a relationship between the position of the removed biomarker and the tau distance (e.g. removing biomarkers 1, 2 or 3 gives higher probabilities of $K = 0$ than removing biomarkers 6, 7 or 8). The implications of this are not immediately apparent, but as some form of this relationship may well occur when applying the EBM to actual patient data, this is an area that requires further investigation.

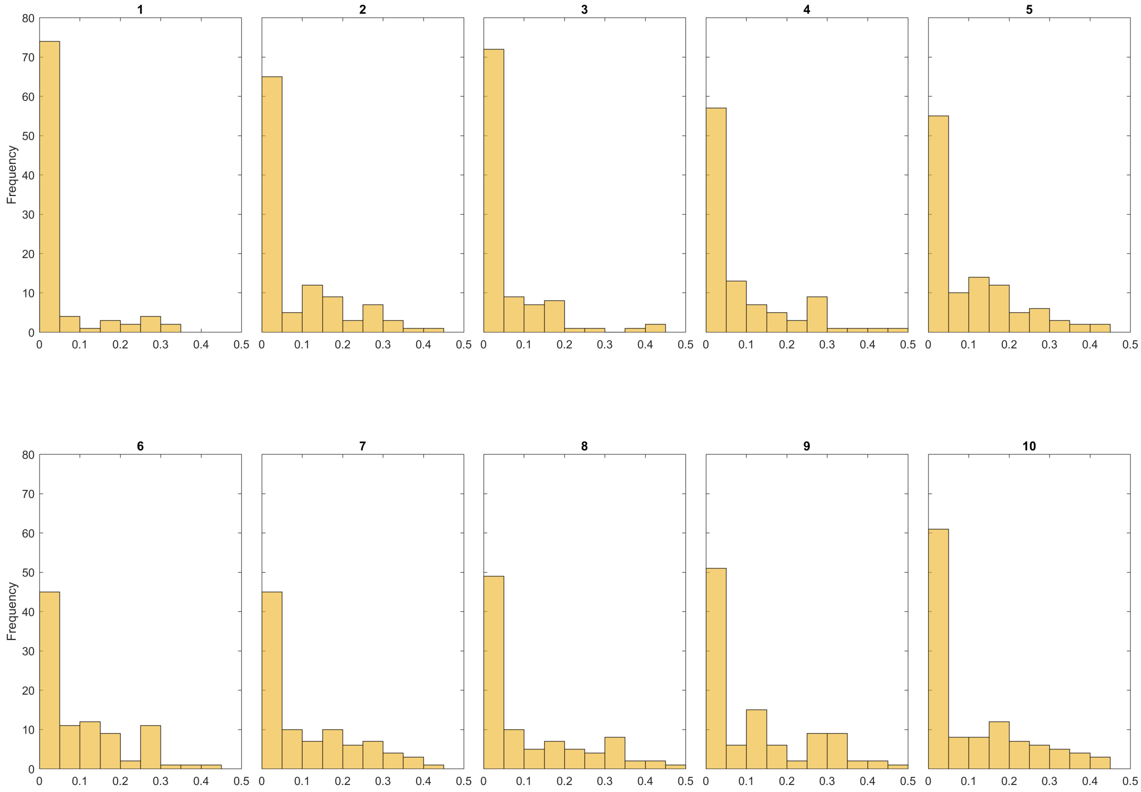


Figure 4.2: The effects of removing a random biomarker on EBM output, as shown by histograms of $K(S_{ML_9}, S^*_{ML_{10}})$, grouped by the position of the removed biomarker.

4.3 Effects of model modifications

4.3.1 CDF vs. PDF for probability calculations

Repetition of Simulations

In order to compare the effects of using Cumulative Distribution Functions (CDFs) and Probability Density Functions (PDFs) upon the EBM's output, each series of simulations in sections 4.2.1 - 4.2.3 was re-run using PDFs for the probability calculations. Each re-run used the same random number generator seed as its corresponding CDF series. Thus, the results in this section were produced from simulated patient data identical to those used in sections 4.2.1 - 4.2.3, meaning that any differences in the model output are purely due to differences between using CDFs and PDFs.

The detailed results are given in Tables 4.16 - 4.23. To summarise, PDFs appear to give more accurate results than CDFs when i) varying $|\mu_a - \mu_c|$ and ii) varying $|\mu_a - \mu_c|$ and $\sigma_a - \sigma_c$ in the opposite direction. However, PDF results are less accurate when i) varying $\sigma_a - \sigma_c$ and ii) varying $|\mu_a - \mu_c|$ and $\sigma_a - \sigma_c$ in the same direction.

These simulations do not necessarily show that PDFs are less appropriate than CDFs for calculating probabilities in the EBM, only that CDFs are more accurate with respect to sample variance - something which is known to be high in ALS. Recall that we defined accuracy as a reduction in distance between S_{ML} and our theoretical "true" order T; this relies on the assumption that we have accounted for all other sources of TI and that the theoretical "true" order is therefore actually correct. The safest conclusion from these simulations is that CDFs allow us to impose TI on the model output to a much higher degree than PDFs.

Biomarker directionality

Unlike PDFs, the use of CDFs requires that we take account of the directionality of biomarker change. If biomarker readings decrease as the disease progresses, as in the case of FA, then we take $p(X|E) = 1 - \mathrm{CDF}_a$ and $p(X|\neg E) = \mathrm{CDF}_c$. If, however, biomarker readings increase as the disease progresses, then instead we take $p(X|E) = \mathrm{CDF}_a$ and $p(X|\neg E) = 1 - \mathrm{CDF}_c$.

Although I consider taking account of biomarker directionality to be a strength, as it reduces the amount of information that is being discarded, it does require a decision to be made: how to decide which direction the biomarker progresses in? Three options are: i) comparing the parameters from the fitted mixtures $(\hat{\mu}_c, \hat{\mu}_a)$, ii) comparing the means of the patients and controls ($\bar{\mu}_c, \bar{\mu}_a$), and iii) separately fitting single Gaussians to the patients and the controls, and comparing the single Gaussian parameters ($\bar{\mu}_1, \bar{\mu}_2$).

It is possible that some or all of these options risk imposing too much information on the model. Using the fitted mixture parameters ($\hat{\mu}_c, \hat{\mu}_a$) would seem to impose a minimum of extra information, but could increase the effects of poor modelling if the parameters are not representative of the data; it is possible for the mixture means to "cross over" during the mixture fitting.

In the current implementation of the EBM, I have chosen to use the cohort means ($\bar{\mu}_c, \bar{\mu}_a$) to assess biomarker directionality, as I feel it provides a good balance between imposing minimal extra information and ensuring that the direction matches the actual data. However, this is an area that requires further development; if possible, it should be an integral component when developing EM methods for fitting mixtures other than Gaussians.

4.4 Summary

In this chapter, different sources of temporal information (TI) have been examined through simulations. Mean separation (the difference between a mixture's component means, $|\mu_a - \mu_c|$) was found to correspond with event position, such that earlier events have greater mean separation than later events. Variance separation (the difference between a mixture's component variances, $\sigma_a^2 - \sigma_c^2$) also influenced event position, such that earlier-progressing biomarkers were found to have greater variance separation than later ones.

The effect of sample size on the EM mixture fitting algorithm was examined in the context of simulated data (i.e. data for which the target distributions are known). This was assessed by calculating the τ correlation between the predicted event order and the actual order generated by the model, for varying numbers of simulated patients. Sample size was found to be strongly related to τ correlation by an inverse power law.

Finally, the effects of using two different methods for calculating individual probabilities (per subject, per biomarker) were examined. The accuracy when using CDFs was compared with using PDFs, defining accuracy as a reduction in tau distance between the predicted and actual event orders. The accuracy of both methods varied across different conditions, with neither method shown to conclusively outperform the other. However, these results do show that the use of CDFs permits a greater degree of TI imposition on the model output.

Patients	$\bar{K}(S_{ML}, T_\mu)$	$\bar{K}(S_{ML}, T_\sigma)$
25	0.11764	0.5034
50	0.0900	0.5009
100	0.0648	0.5086
200	0.0489	0.4968

Table 4.16: Mean distance between S_{ML} and T, varying $|\mu_a - \mu_c|$, using PDFs

Patients	$\bar{\tau}(S_{ML}, T_\mu)$	$\bar{\tau}(S_{ML}, T_\sigma)$
25	0.7671*	-0.0068
50	0.8200*	-0.0018
100	0.8704**	-0.0172
200	0.9023***	0.0064

Table 4.17: Mean tau correlation between S_{ML} and T, varying $|\mu_a - \mu_c|$, using PDFs. The correlations are stronger than when using CDFs (see Table 4.3). $^*p < 0.0005$, $^{**}p < 0.0001$, $^{***}p < 0.00005$

Patients	$\bar{K}(S_{ML}, T_\mu)$	$\bar{K}(S_{ML}, T_\sigma)$
25	0.5579	0.5818
50	0.5640	0.5742
100	0.5824	0.5586
200	0.5786	0.5494

Table 4.18: Mean distance between S_{ML} and T, varying $\sigma_a - \sigma_c$, using PDFs

Patients	$\bar{\tau}(S_{ML}, T_\mu)$	$\bar{\tau}(S_{ML}, T_\sigma)$
25	-0.1158	-0.1636
50	-0.1280	-0.1485
100	-0.1648	-0.1171
200	-0.1572	-0.0988

Table 4.19: Mean tau correlation between S_{ML} and T, varying $\sigma_a - \sigma_c$, using PDFs. No correlations are significant.

Patients	$\bar{K}(S_{ML}, T_\mu)$	$\bar{K}(S_{ML}, T_\sigma)$
25	0.3529	0.3599
50	0.3192	0.3261
100	0.2877	0.2932
200	0.2608	0.2626

Table 4.20: Mean distance between S_{ML} and T, varying μ_a and σ_a in the same direction, using PDFs

Patients	$\bar{\tau}(S_{ML}, T_\mu)$	$\bar{\tau}(S_{ML}, T_\sigma)$
25	0.2943	0.2801
50	0.3617	0.3478
100	0.4246	0.4135
200	0.4784*	0.4748*

Table 4.21: Mean tau correlation between S_{ML} and T, varying μ_a and σ_a in the same direction, using PDFs. $^*p < 0.05$

Patients	$\bar{K}(S_{ML}, T_\mu)$	$\bar{K}(S_{ML}, T_\sigma)$
25	0.1322	0.8534
50	0.1078	0.8860
100	0.0819	0.9128
200	0.0681	0.9299

Table 4.22: Mean distance between S_{ML} and T, varying μ_a and σ_a in the opposite direction, using PDFs

Patients	$\bar{\tau}(S_{ML}, T_\mu)$	$\bar{\tau}(S_{ML}, T_\sigma)$
25	0.7356*	-0.7069*
50	0.7843**	-0.7719**
100	0.8362***	-0.8257***
200	0.8637***	-0.8599***

Table 4.23: Mean tau correlation between S_{ML} and T, varying μ_a and σ_a in the opposite direction, using PDFs. $^*p < 0.005$, $^{**}p < 0.0005$, $^{***}p < 0.0001$

Chapter 5

Challenging the Model *in vivo*

In this chapter, I apply the EBM to actual patient data in order to explore the effects of the main modifications that have been made to the model. As simulations are designed to be simplified and possess reduced sources of noise, it is necessary to explore the effects of applying the model to real-world data.

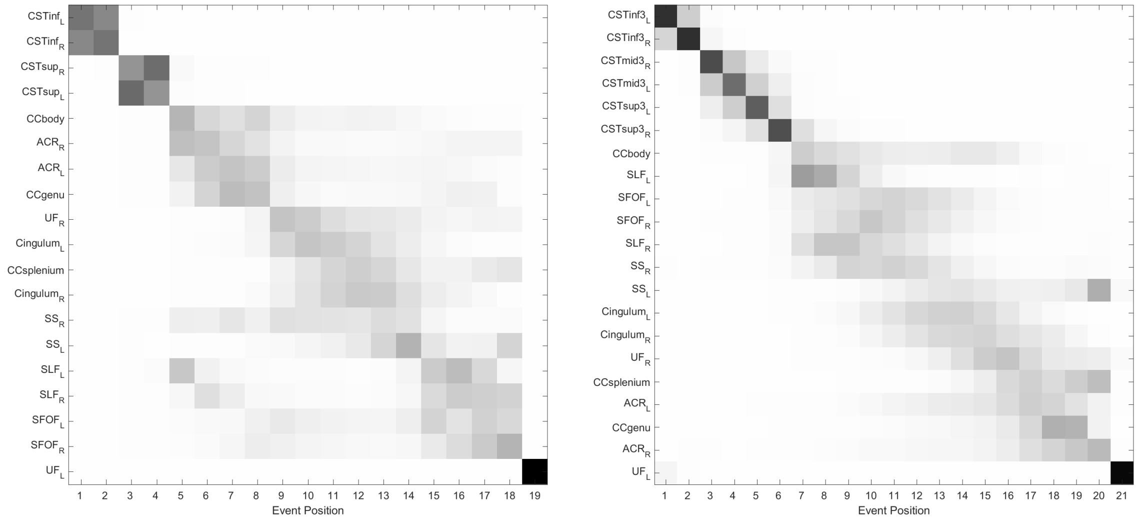
5.1 Effects of model modifications

5.1.1 Adding and removing biomarkers

This section examines what happens if the number of biomarkers is altered when using actual patient data. Here, we apply the EBM to our complete DTI data set, obtained by combining Sets F, K and N. To examine the effects of adding and removing biomarkers based on FA, we split each side of the CST into halves, run the EBM, and compare the results when split into thirds. Thus the change being compared is the removal of four biomarkers (CST inferior/superior, L/R) and replacement with six related biomarkers (CST inferior/middle/superior, L/R), giving an overall increase of two biomarkers.

Figure 5.1 shows the model output for this comparison, and it can be seen that the CST progression retains the inferior to superior directionality. However, the other biomarkers substantially change positions; taking only the non-CST tracts into consideration, the Kendall tau distance between the two most likely orders of progression is 0.5810, meaning that almost 60% of possible event pairs are different. We note that events 8 to 20 of the six-region-CST progression order are approximately a reversal of events 6 to 18 of the four-region-CST progression order.

If the comparison is re-run after removing outliers $> 2\sigma$ during mixture fitting (Figure 5.2), we then find that the results remain almost completely stable: the only



(a) 19 events: CSTs divided into four regions.

(b) 21 events: CSTs divided into six regions.

Figure 5.1: Positional variance diagrams showing the effects of increasing the number of biomarkers for actual patient data, Sets FKN. Constraints: mixing proportions > 0.25

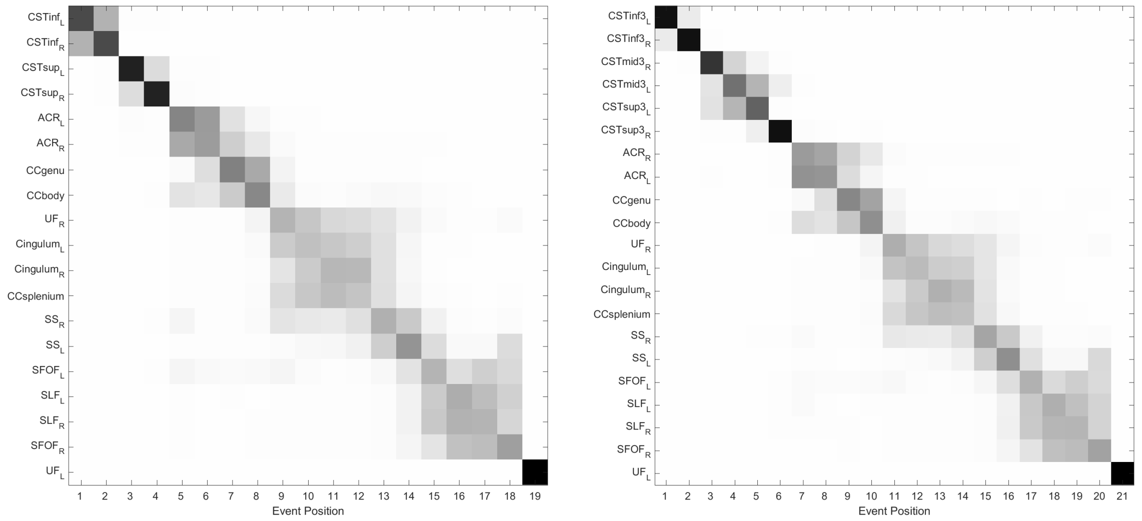
difference between the progression orders for four-region-CSTs and six-region-CSTs is that ACR L/R swap positions.

This reinforces how crucial the mixture fitting process is to the EBM. Further discussion of the impact of the mixture models is given in subsection 5.1.3.

5.1.2 Too many or too few biomarkers

Too many biomarkers results in Matlab encountering floating-point precision problems, hence the adaptations for variable precision (see subsection 2.5.3). The number of subjects will also affect Matlab's precision. Quantifying "too many" biomarkers is not straightforward, as the magnitude of probabilities will depend on the data being modelled, the fitted mixture model parameters and the distribution of patients' biomarker readings; anecdotal evidence suggests that this problem can arise when working with upwards of 45 biomarkers for 100 ALS patients.

Too few biomarkers defeats the point of the EBM, as we can then exhaustively calculate $P(S|X)$ for all Ω (see subsection 7.4.4) and definitively find the most likely event order. Quantifying "too few" is also difficult, as the processing time required will depend not only on the number of biomarkers, but also on the number of subjects. However, experience during the development of the model suggests that at least 14 biomarkers are needed to make it worthwhile using the EBM.



(a) 19 events: CSTs divided into four regions.

(b) 21 events: CSTs divided into six regions.

Figure 5.2: Positional variance diagrams showing the effects of increasing the number of biomarkers for actual patient data, Sets FKN. Constraints: mixing proportions > 0.25, outliers $> 2\sigma$ removed during mixture fitting.

5.1.3 Sample size and mixture fitting

The importance of sample size to the EBM should not be underestimated, as the EM algorithm appears to be sensitive to outliers. Recalling that one control subject was excluded during data preparation (as 6 of 19 biomarker readings were outliers; see subsection 3.2.1), this section will examine the effects of reintroducing this subject into our data sample.

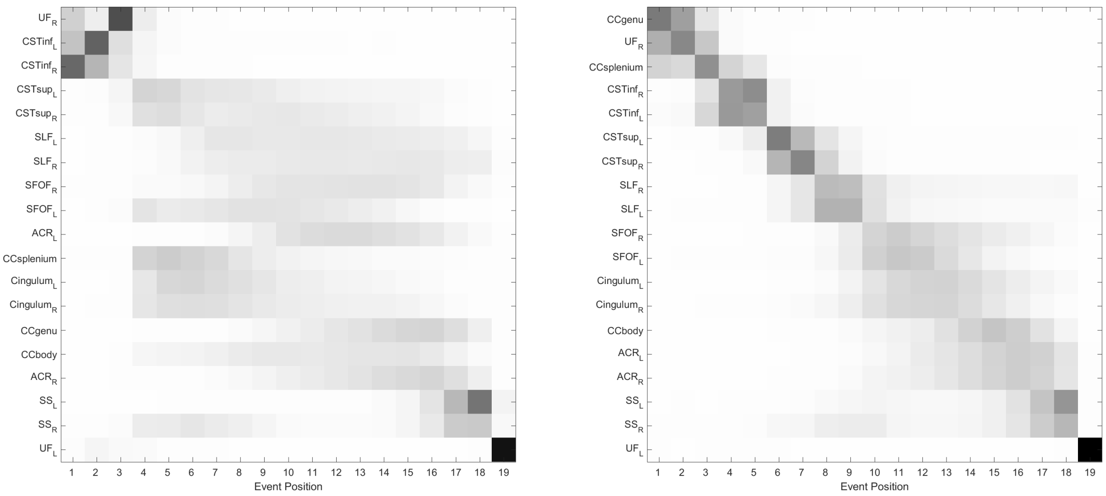
As the outlier healthy control (HC) was a member of Set K, we begin by running the EBM on Set K. We shall refer to Set K with the outlier HC excluded as K_{24}, since the sample contains 24 controls. Set K with the outlier HC included is similarly designated K_{25}.

As can be seen in Figure 5.3, the impact of the outlier HC is substantial. The majority of the events for K_{24} have extremely high positional variance, suggesting that the subject sample size may be too small; we should certainly be hesitant to draw any conclusions from this model output. Conversely, the MCMC samples for K_{25} appear to have an acceptable level of positional variance. Extrapolating from this result would not seem unreasonable; without having run the EBM on K_{24}, we would have a potentially false level of confidence in the validity of the model output for K_{25}.

The size of Set K is limited (28 ALS patients, 24 or 25 HCs), which could provide an explanation for the disproportionate impact of a single outlier HC. To test this, we repeat the above comparison process, applying the model to all available data (87 ALS patients, 69 or 70 HCs). As expected, the larger sample size reduces the effect of the outlier HC (Figure 5.4), although it is not completely eliminated. This suggests that the outlier HC is affecting all biomarkers, including those for which their readings are not outliers.

However, the desired similarity of EBM output can be recovered by removal of outliers $> 2.5\ \sigma$ during mixture fitting. Indeed, this seems to nullify the effects of the outlier HC, as the results now appear identical (Figure 5.5 (a) - (b)). It is only by applying a custom colourmap to the PVDs that differences in the MCMC iterations become visible (Figure 5.5 (c) - (d)).

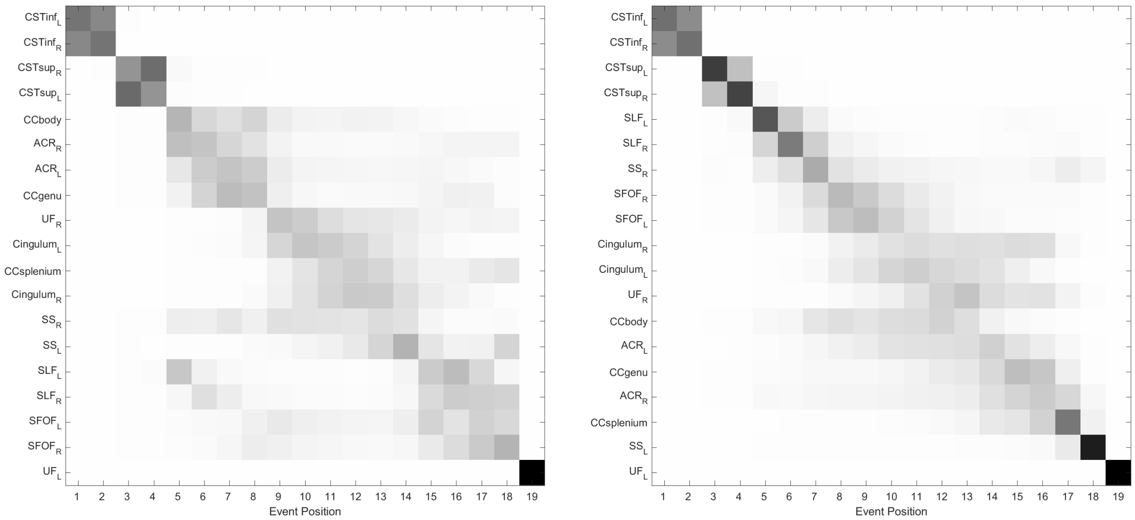
The EM algorithm's apparent sensitivity to outliers suggests that further exploration of the resilience of EM Gaussian mixture fitting is needed. These effects can be reduced by the removal of outliers during the mixture fitting (as described in subsection 2.5.2), although this method lacks finesse. An alternative method of reducing sensitivity to outliers is to fit the EM mixture models by bootstrapping ("resampling with replacement"). This process is discussed in greater detail in subsection 7.4.1.



(a) Outlier HC excluded

(b) Outlier HC included

Figure 5.3: Comparison of model output for (a) Set K with outlier HC excluded, and (b) Set K with outlier HC included. Constraints: mixing proportions > 0.25



(a) Sets FKN, outlier control subject excluded (b) Sets FKN, outlier control subject included

Figure 5.4: Comparison of model output for Sets FKN (outlier HC excluded) with Sets FKN (outlier HC included). Constraints: mixing proportions > 0.25

5.1.4 Gaussian vs Non-Gaussian

Inappropriate Gaussians

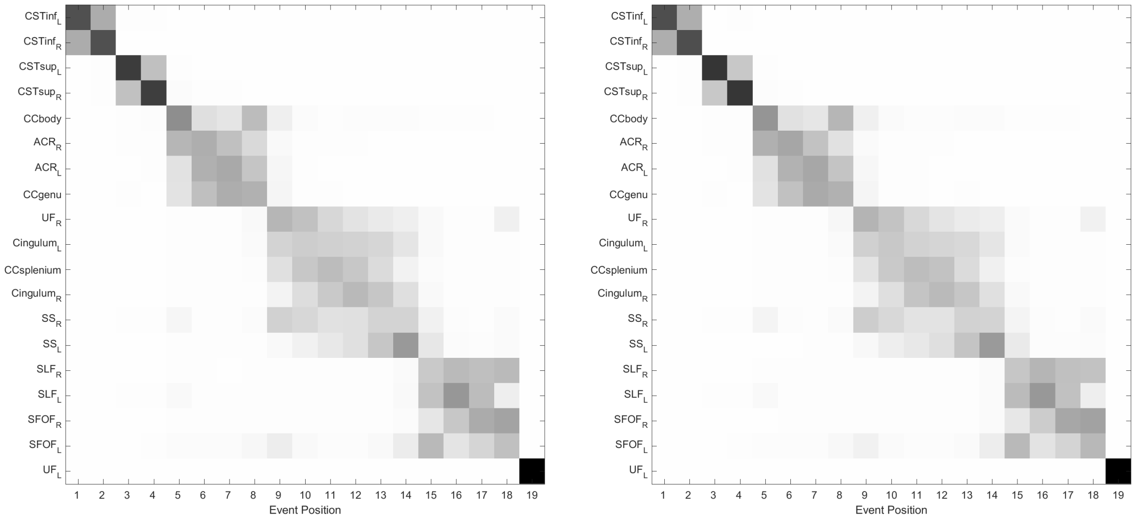
Applying the EBM to cortical thickness data derived from structural MRI scans allows us to compare what happens when data are mostly Gaussian or not Gaussian.

For 28 cortical thickness biomarkers, we consider controls and patients separately, giving 56 groups that could be Gaussian. The Lilliefors test (the *lillietest* function in Matlab) was used to test for normality, for a significance level of $p < 0.05$. This revealed that 10 of the possible 56 groups were unlikely to be normally distributed. All 28 biomarkers were then transformed by dividing by the mean thickness of the corresponding hemisphere on a subject-by-subject basis, and the Lilliefors test was performed as before. This time, only 5 of the possible 56 groups were unlikely to be normally distributed. The EBM was then run on the original data and the transformed data.

A visual comparison of the EBM output (Figure 5.6) confirms that the EBM appears to perform better if the data are modelled with more appropriate mixture distributions. Because of this, support for other mixture types is a desirable addition to the model.

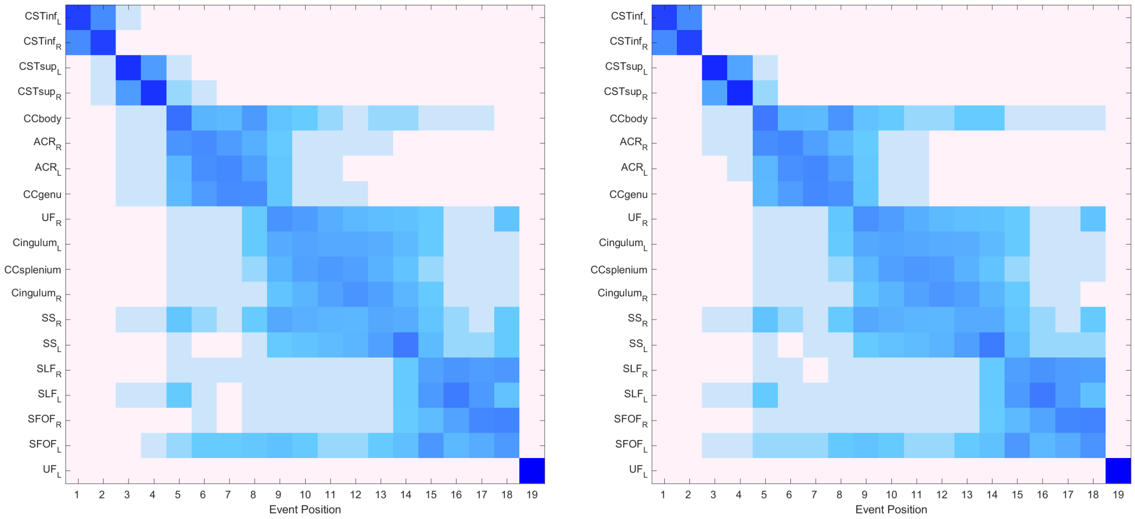
Refitting a single biomarker

For a test of model stability, we change the mixture modelling of the neuropsychological biomarker TROGerr from two Gaussians to two Poissons. This biomarker was selected



(a) Sets FKN, outlier HC excluded

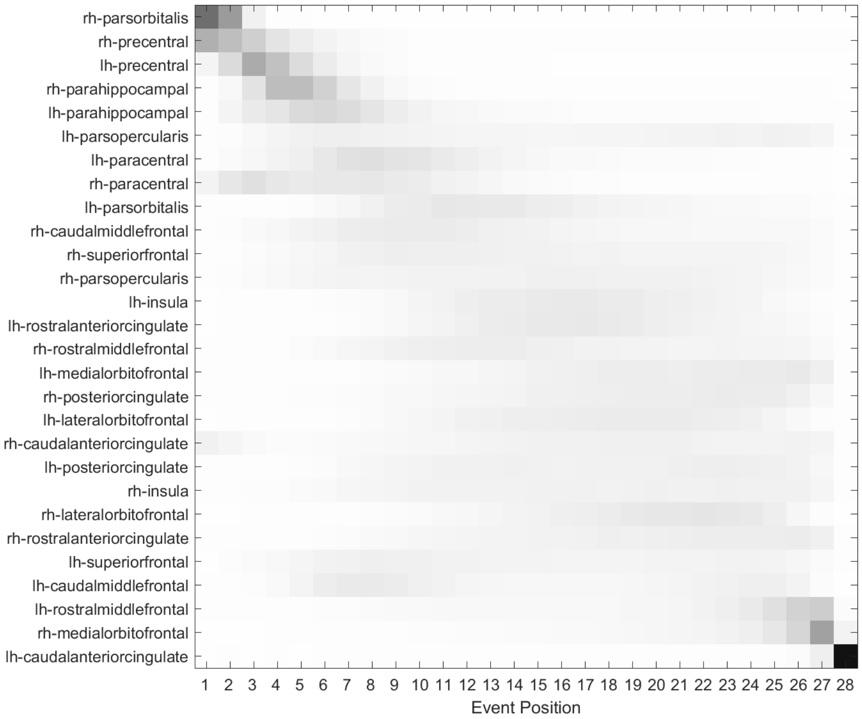
(b) Sets FKN, outlier HC included



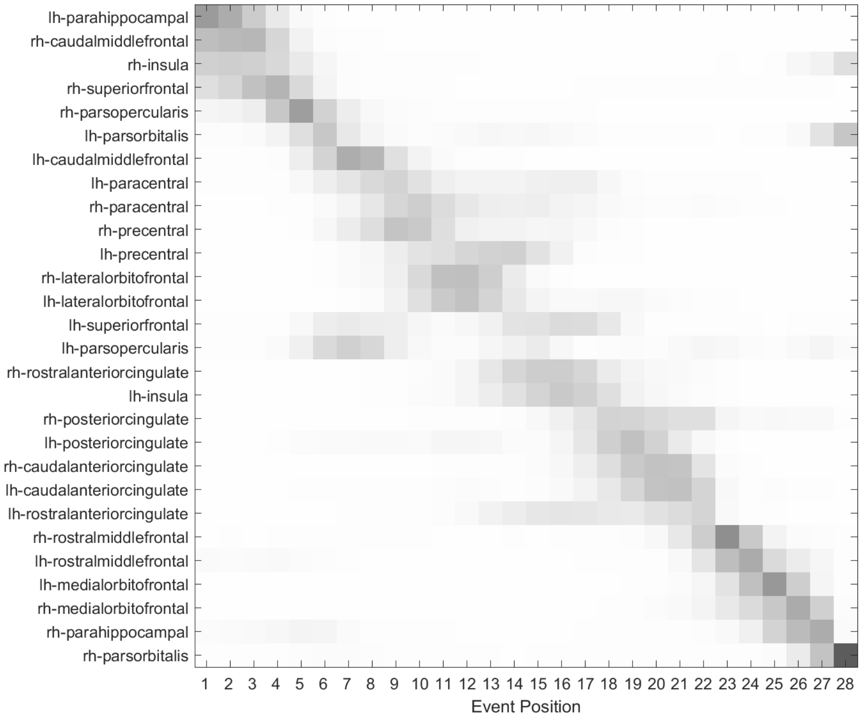
(c) Sets FKN, outlier HC excluded*

(d) Sets FKN, outlier HC included*

Figure 5.5: Comparison of model output for Sets FKN (outlier HC excluded) with Sets FKN (outlier HC included). Constraints: mixing proportions > 0.25, outliers $> 2.5\ \sigma$ excluded during mixture fitting. * = Custom colourmap applied to increase contrast.



(a) Cortical thickness data from Sets KN, with no transformations applied. 10 of the 56 groups are not Gaussian.

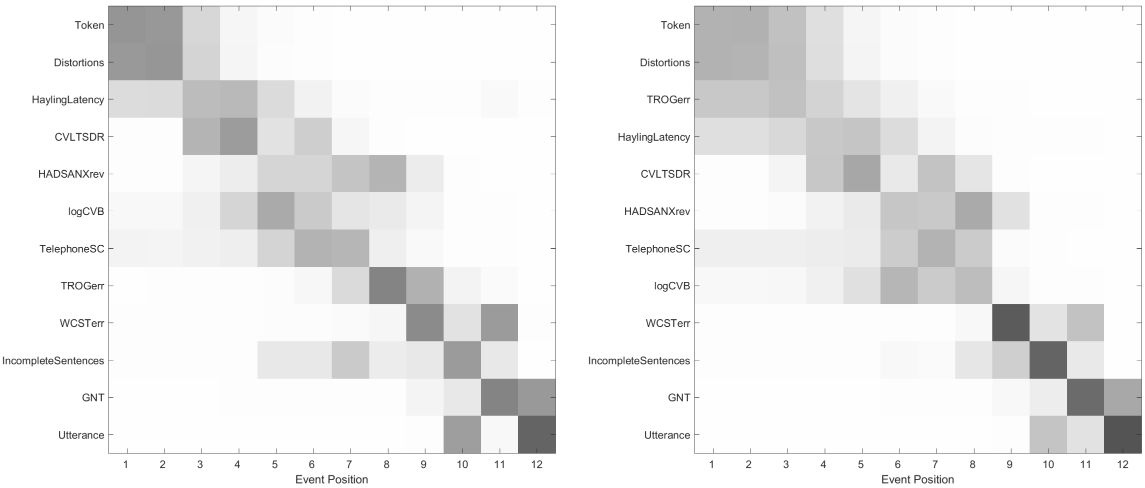


(b) Cortical thickness data from Sets KN, transformed by dividing by mean hemisphere cortical thickness. 5 of the 56 groups are not Gaussian.

Figure 5.6: Positional variance diagrams showing the effects of transforming control and patient biomarker data to be Gaussian. Controls and patients were separately assessed for normality, giving 56 groups that could be Gaussian. The positional variance of (a) is considerably higher than (b). Constraint: mixing proportion > 0.15.

for mixture refitting as its data can plausibly be modelled as a Poisson distribution; “TROGerr” is a test of verbal comprehension of syntax, and is measured as the number of errors made during a fixed number of tests. This can be interpreted as a rate of discrete events (i.e. errors per fixed period), which in turn is appropriate for modelling as a Poisson distribution.

The results of this change of mixture distribution are shown in Figure 5.7, and the model output remains remarkably stable. Aside from “TROGerr” moving to an earlier position, there is only one other change: a pair of neighbouring events swap places. As the majority of the event order appears to be preserved, we can only speculate whether a Poisson mixture is actually more appropriate for the biomarker in question, although we note that the positional variance of the new event order, Figure 5.7 (b), appears to be more symmetrical around the diagonal.



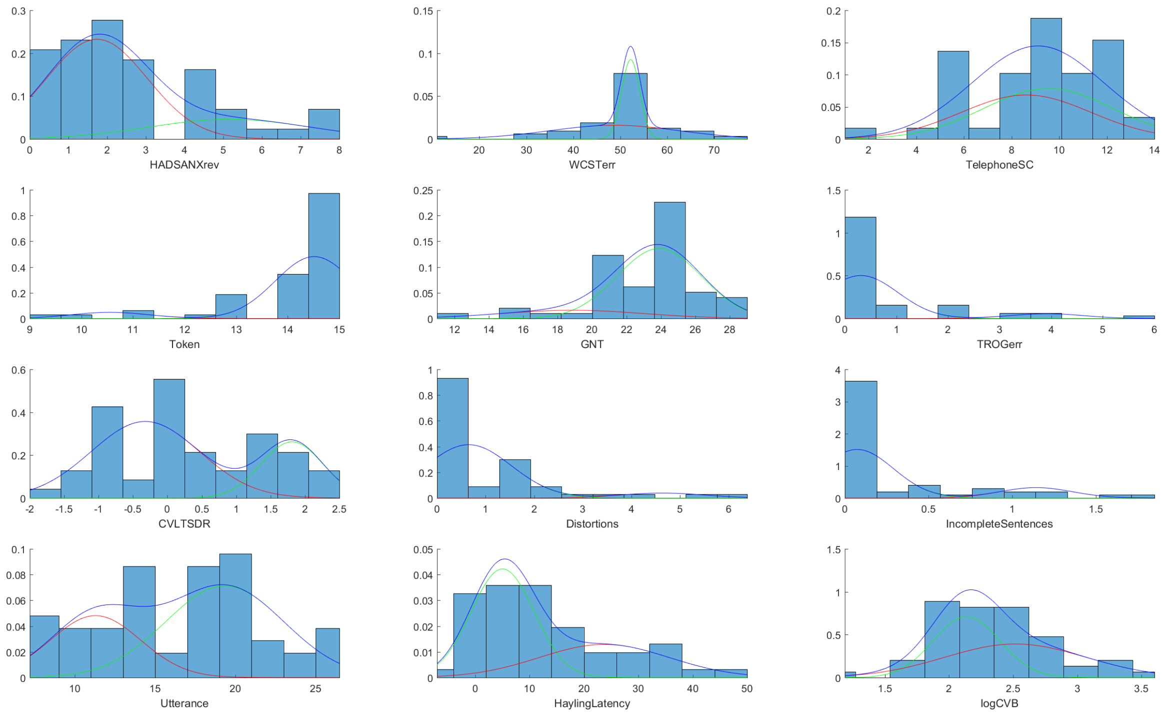
(a) All fitted as mixtures of two Gaussians

(b) “TROGerr” as a mixture of two Poissons

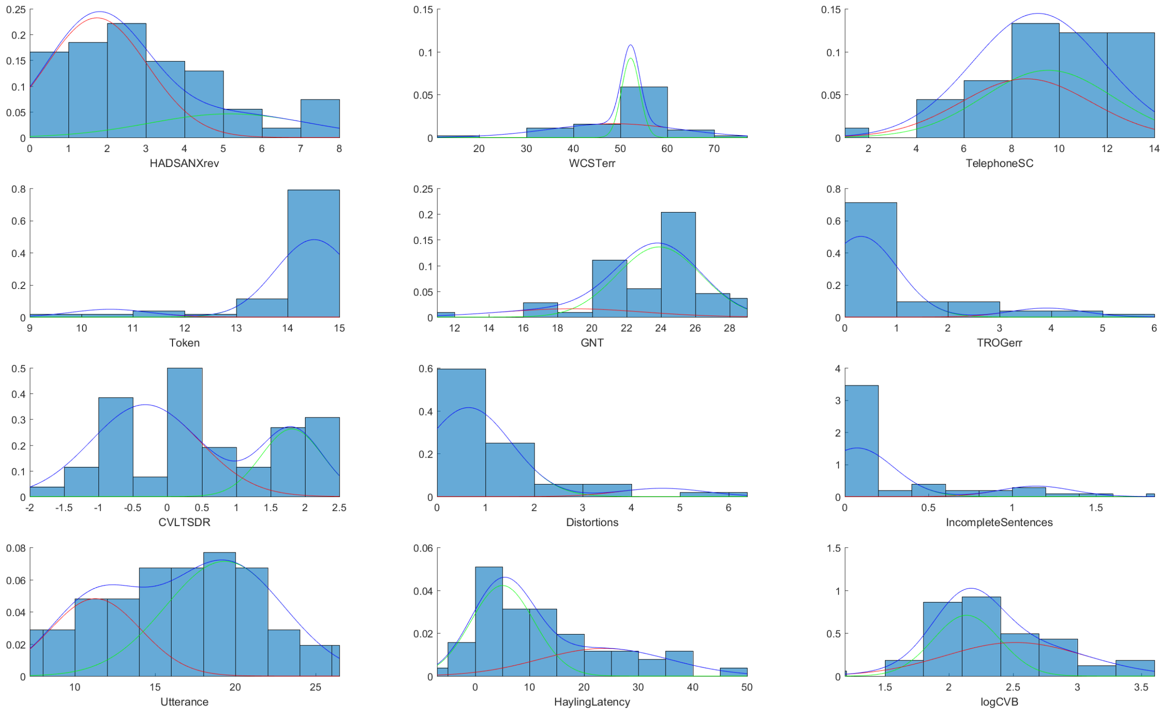
Figure 5.7: Positional variance diagrams showing the effects of refitting a single biomarker, “TROGerr”, from a mixture of two Gaussians to a mixture of two Poissons. The neuropsychological biomarkers are modelled on data from Set K.

5.1.5 Assessing Mixture types

It is important to note that the appropriateness of mixture fitting should not be judged by eye from the mixture plots. The selection of the bin method for plotting the histograms can have a tremendous effect on the appearance of the histograms, as can be seen in Figure 5.8.



(a) Histograms using ten equal width bins.



(b) Histograms using the "square root" method to size the bins.

Figure 5.8: Histograms of neuropsychological biomarker readings, with mixture model component distributions superimposed. The fitted mixture model parameters are identical in (a) and (b); the only difference is the method used for choosing the limits of the histogram bins. "Utterance" appears particularly affected.

5.2 Coverage of sample space

Here, for each of the seven possible dataset combinations, we compare the output of the greedy ascent algorithm, S_g, to the EBM output, S_{ML}. This acts as a simple method of checking the coverage of the sample space of all possible event orders, Ω.

In justification, we recall that the greedy ascent algorithm is a crude method of finding an event order S_g with (possibly local) maximum likelihood, and that we use this greedy output to initialise the more sophisticated MCMC process. If S_g is the global maximum, then the MCMC process should remain broadly within the nearby area of Ω. If S_g is only a local maximum, then the MCMC process should move to the area of Ω containing the global maximum. This leads us to the following conclusions:

- If S_g is the global maximum, then $S_{ML} = S_g$, and so $p(S_g|X) = p(S_{ML}|X)$
- If S_g is a local maximum, then $S_{ML} \neq S_g$, and so $p(S_g|X) < p(S_{ML}|X)$
- If $p(S_g|X) > p(S_{ML}|X)$, then some part of the modelling is not appropriate

The third situation will arise if the MCMC process has moved out of the area of Ω containing the global maximum and not returned. This means that the MCMC output will not, in fact, be the event order with maximum likelihood. There are many reasons why this could occur, such as inaccurate mixture fitting, too few MCMC iterations or an inappropriate choice of MCMC algorithm; irrespective of the causes, it is an indication of problems within the modelling process.

Tables 5.1 - 5.4 give the log-likelihoods for S_g and S_{ML}, calculated from patient datasets F, K, N and all possible combinations thereof. We find that when using CDFs to calculate event probabilities, then $p(S_g|X) > p(S_{ML}|X)$ for two out of the seven groups (Table 5.1). The two groups affected, Sets K and N, have the smallest sample sizes of the subgroups, which would be expected to increase the chances of poor modelling. Conversely, when using PDFs to calculate event probabilities, $p(S_g|X) > p(S_{ML}|X)$ for all seven groups (Table 5.2), indicating much more widespread inappropriate coverage of the sample space.

As the EM algorithm appears to be sensitive to outliers, we re-run the EBM excluding outliers $> 2.5\,\sigma$ during the mixture fitting process, as described in subsection 2.5.2. As shown in Tables 5.3 - 5.4, this does not appear to affect the coverage of the sample space for either CDFs or PDFs, as the cohorts with $S_g > S_{ML}$ remain unchanged.

The results of this section serve as a reminder that it is always worthwhile performing MCMC diagnostic tests. In general, Brooks et al. (2011) recommend simple diagnostics such as examining a time series plot or an autocorrelation plot of the MCMC iterations. For the EBM, I suggest that when using CDFs, a comparison of $p(S_g|X)$ and $p(S_{ML}|X)$ should also be included.

	F	K	N	FK	FN	KN	FKN
S_g	-582.66	-429.37	-324.63	-876.77	-828.52	-678.87	-1212.98
S_{ML}	-582.66	-429.71	-324.74	-876.77	-828.52	-678.87	-1212.69

Table 5.1: Log-likelihood for greedy output and S_{ML}, using CDFs to calculate event probabilities. $S_g > S_{ML}$ for groups K and N.

	F	K	N	FK	FN	KN	FKN
S_g	1523.31	1198.43	932.97	2661.02	2440.91	2046.32	3546.73
S_{ML}	1522.72	1197.63	932.79	2660.55	2440.35	2045.80	3546.40

Table 5.2: Log-likelihood for greedy output and S_{ML}, using PDFs to calculate event probabilities. $S_g > S_{ML}$ for all groups.

	F	K	N	FK	FN	KN	FKN
S_g	-628.19	-442.98	-355.53	-924.33	-853.34	-782.82	-1283.16
S_{ML}	-627.63	-443.08	-355.87	-924.33	-853.34	-782.82	-1280.56

Table 5.3: Log-likelihood for greedy output and S_{ML}, using CDFs to calculate event probabilities. Outliers $> 2.5\ \sigma$ were excluded during mixture fitting. $S_g > S_{ML}$ for groups K and N.

	F	K	N	FK	FN	KN	FKN
S_g	1496.52	1195.55	934.93	2660.10	2434.57	2066.65	3553.47
S_{ML}	1496.27	1194.76	934.52	2659.85	2433.99	2066.62	3553.14

Table 5.4: Log-likelihood for greedy output S_g and S_{ML}, using PDFs to calculate event probabilities. Outliers $> 2.5\ \sigma$ were excluded during mixture fitting. $S_g > S_{ML}$ for all groups.

5.3 Summary

In this chapter, the effects of the main modifications to the EBM were explored by applying the model to actual patient data. It was concluded that the model output generally remains stable, although changing the biomarkers being modelled (whether by adding biomarkers, removing biomarkers or changing mixture parameters) has knock-on effects that were not always predictable. It was established that the event position of an individual biomarker is dependent on the other events being modelled.

The model was applied to non-Gaussian data, and then to the same data after applying a transformation to make the data Gaussian in nature. As expected, the EBM output for the non-Gaussian data had consistently greater positional variance than for the transformed data.

The effects of sample size and outliers on the EBM output were examined in the context of actual patient data. Sample size was found to have the expected impact on the accuracy of the EM mixture fitting algorithm (i.e. larger sample sizes permit greater accuracy when fitting mixtures). The EM algorithm was shown to be highly sensitive to the effects of outlier data points, and it was concluded that further research is required into methods to account for this sensitivity.

Finally, the effects of CDFs and PDFs on coverage of the sample space Ω were investigated. By comparing the likelihood for the output of the greedy ascent algorithm S_g to that of the EBM output S_{ML}, it was found that PDFs appear to consistently generate output such that $p(S_g|X) > p(S_{ML}|X)$. The implications of the greedy ascent algorithm seeming more effective than the MCMC process are not entirely clear, particularly as CDFs do not appear to suffer from the same problem. It is possible that this may indicate inappropriate coverage of the sample space while using PDFs, and further investigation is warranted.

Chapter 6

Model Output

As with any modelling process, the parameters of the mixture fitting and EBM algorithms have been adjusted over the course of this study. These alterations have effects of varying sizes on the model output; this means that the model has generated many possible most likely orders of progression, all of which are related but not identical. As the actual course of ALS progression is unknown, we cannot know which are the ideal model parameters that will lead to the ALS-generalised most likely order of progression. As it is poor practice to run many different parameter variants and then select the results that appear interesting, we must instead declare the model parameters beforehand:

- Mixture proportions (“mix”) constrained such that mix > 0.25
- Outliers to be removed during mixture fitting if:
    - sample size is “small” (provisionally defined as fewer than 30 patients or controls)
    - positional variance is unacceptably high (as in Figure 5.3 (a))
    - the MCMC trace plot shows evidence of poor mixing
- Outlier removal should be as lenient as possible:
    - Removal of outliers $> 3\sigma$, then assess the changes to the mixture fitting
    - If changes are not significant, then remove outliers $> 2.5\,\sigma$, or $> 2\sigma$
    - Each outlier threshold is equivalent to removing 0.0027, 0.0124 and 0.0455 respectively of the data from the Gaussian distribution

In this chapter I present the results obtained by applying the EBM to three different modalities of data: DTI metrics, structural T_1-weighted MRI data and neuropsychological readings. The results are given in the form of positional variance diagrams, which show the most likely event order for the modelled cohort, as well as each event's variation across the MCMC samples. This variation may be considered to represent the uncertainty of an event's ordering. To aid in interpreting and understanding the model output, the fitted mixture models are also given and examined.

6.1 Neuroimaging biomarkers

6.1.1 FA data

Changes in FA are used as an index of neurodegeneration, particularly in white matter. Thus, I have applied the EBM to this metric in order to examine the progression of white matter tract degeneration in ALS.

Figure 6.1 gives the results obtained by applying the model to the entirety of the data available, a single data set comprised of all three cohorts: Sets F, K and N. Under this model, the CSTs are the first tracts to be affected, with neurodegeneration taking place in the inferior CST (positions 1 to 2) immediately followed by the superior CST (positions 3 to 4). The CSTs have low positional variance, which is almost entirely restricted to switching between the left and right halves of each biomarker. The rest of the white matter tracts have higher positional variance, and appear to be grouped in three "clusters". Each cluster roughly outlines a square, indicating a group of tracts that progress at similar times. $\mathrm{UF_L}$ is not included in these clusters, and instead appears rigidly in last position.

Figure 6.2 gives the results of applying the model to i) individual cohorts and ii) all pairs of cohorts. It can be seen that the most likely event order is not consistent across the cohort subgroups, and frequently differs from the event order seen in Figure 6.1; this is not unexpected, given the heterogeneity of ALS and the datasets being examined. As diffusion MRI studies of ALS patients have consistently shown involvement of the CSTs and the corpus callosum (Grolez et al., 2016), we instead examine the relative positions of these biomarkers.

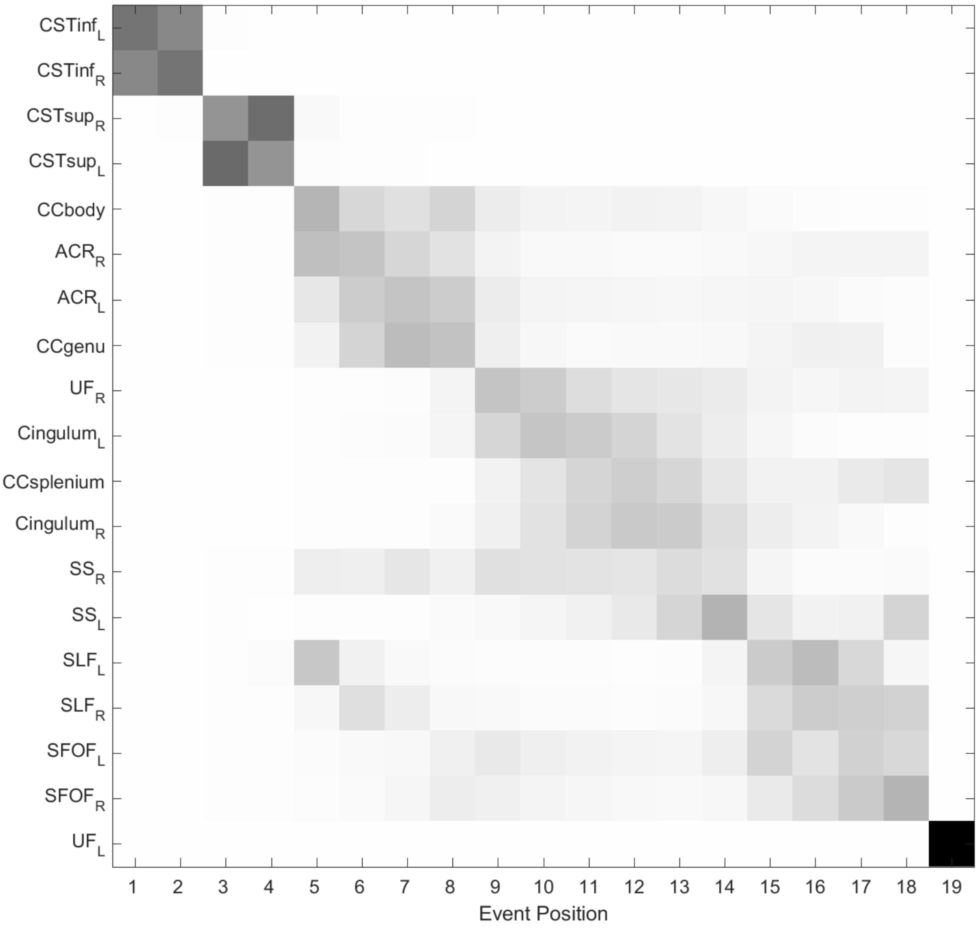
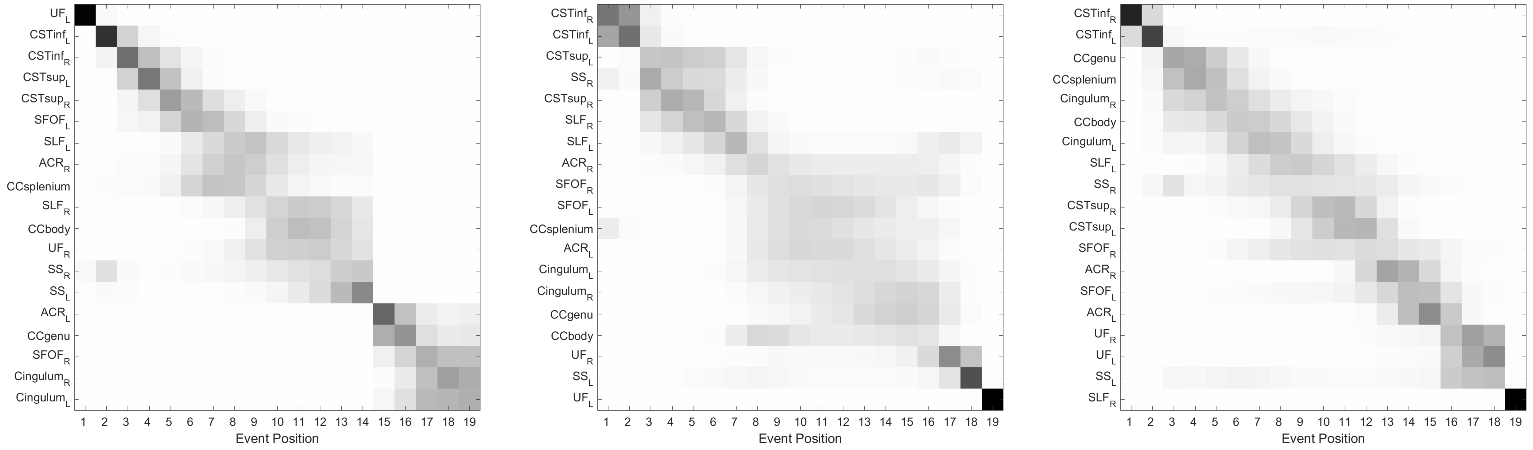


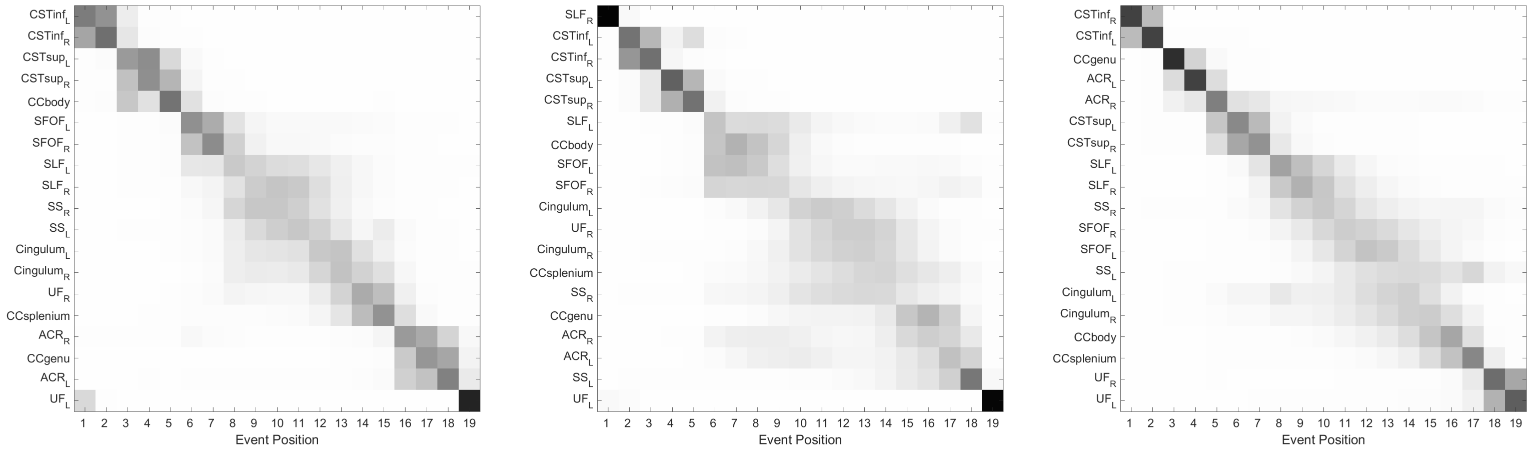
Figure 6.1: Positional variance diagram of the most likely event order of white matter tract neurodegeneration in ALS patients, generated from S_{ML} and the MCMC samples. The EBM was run on the FA data of all ALS patients of the combined data set (Sets FKN). Constraints: mixing proportion > 0.25, no outliers removed.



(a) Set F*

(b) Set K*

(c) Set N*



(d) Sets FK

(e) Sets FN

(f) Sets KN

Figure 6.2: Positional variance diagrams for FA data of all possible subgroups of Sets F, K and N. Constraints: mixing proportion > 0.25. * = outliers $> 2\sigma$ removed during mixture fitting.

Corticospinal Tracts

From the positional variance diagrams of the combined data and all six possible subgroups (Figures 6.1 - 6.2), it is apparent that the CSTs consistently occur early in disease progression. The degeneration of the CSTs is shown to be strongly directional, as the inferior aspect occurs before the superior aspect for all possible groupings. We also find that the positional variance for the CSTs is low, regardless of cohort. These features suggest that the directional degeneration of the CSTs is an important and early feature of white matter damage progression in ALS.

As external validation of the directionality observation is not possible, we have performed a sanity check by splitting the CST into thirds and re-running the EBM. As shown and discussed in Figure 5.1, we find the directionality of CST progression is preserved, moving from inferior through middle to superior.

Further discussion of the importance of the CSTs to overall event progression in ALS can be found in subsection 7.4.4.

Corpus Callosum

Examining all positional variance diagrams, we find that both the relative and absolute positions for CC genu, CC body and CC splenium show no consistency. This is in contrast with the regularity of the CSTs' positional appearance in the subgroups. However, the EBM output for the combined data, as shown in Figure 6.1, suggests that CC degeneration is nonetheless an early feature of ALS white matter disease progression (beginning centrally and then spreading outwards to the genu and then the splenium). We note that the CC biomarkers occur after the CSTs in 5 out of the 7 groupings, with the exceptions being Sets N and the related KN.

This pattern of CC degeneration could be characterised as secondary to and less certain than that of the CSTs, and this relationship appears to be mirrored in the wider literature; the CST is almost invariably reported as being strongly involved in DTI studies of ALS, while there is less consensus as to the extent of CC involvement (Grolez et al., 2016).

We conclude that the relationship between CST and CC degeneration in ALS requires further examination in future studies.

Fitted Mixture Models

Visual inspection of the plots of the fitted mixture models for Sets FKN (mix > 0.25, no outliers removed), given in Figure 6.3, reveals further details about the modelling process:

- All the CST and CC biomarkers have clear mean separation, suggesting that they are responsible for a good proportion of the temporal information extracted by the model.
- We see from Figure 6.1 that the left and right sides of each tract progress in close proximity to each other. However, Figure 6.3 shows that a majority of the tracts have been fitted with dissimilar mixtures for both sides. This is particularly noticeable for the ACR, which is assigned to neighbouring events despite its differing mixtures.
- The UF has been modelled such that its left and right sides do not progress in close proximity, and we observe that $\mathrm{UF_L}$ is the only biomarker that has been fitted with $\hat{\mu_a} > \hat{\mu_c}$.

Conclusions from FA data

I conclude from the modelling of these WM tracts that:

- the corticospinal tracts are the earliest to be affected
- the neurodegeneration of the corticospinal tracts is directional, proceeding from inferior to superior aspect
- We can propose a 5-stage hypothesis of WM degeneration in ALS:
    1. Inferior corticospinal tract
    2. Superior corticospinal tract
    3. CC Body and Genu, ACR
    4. CC Splenium, Cingulum and SS
    5. SLF and SFOF

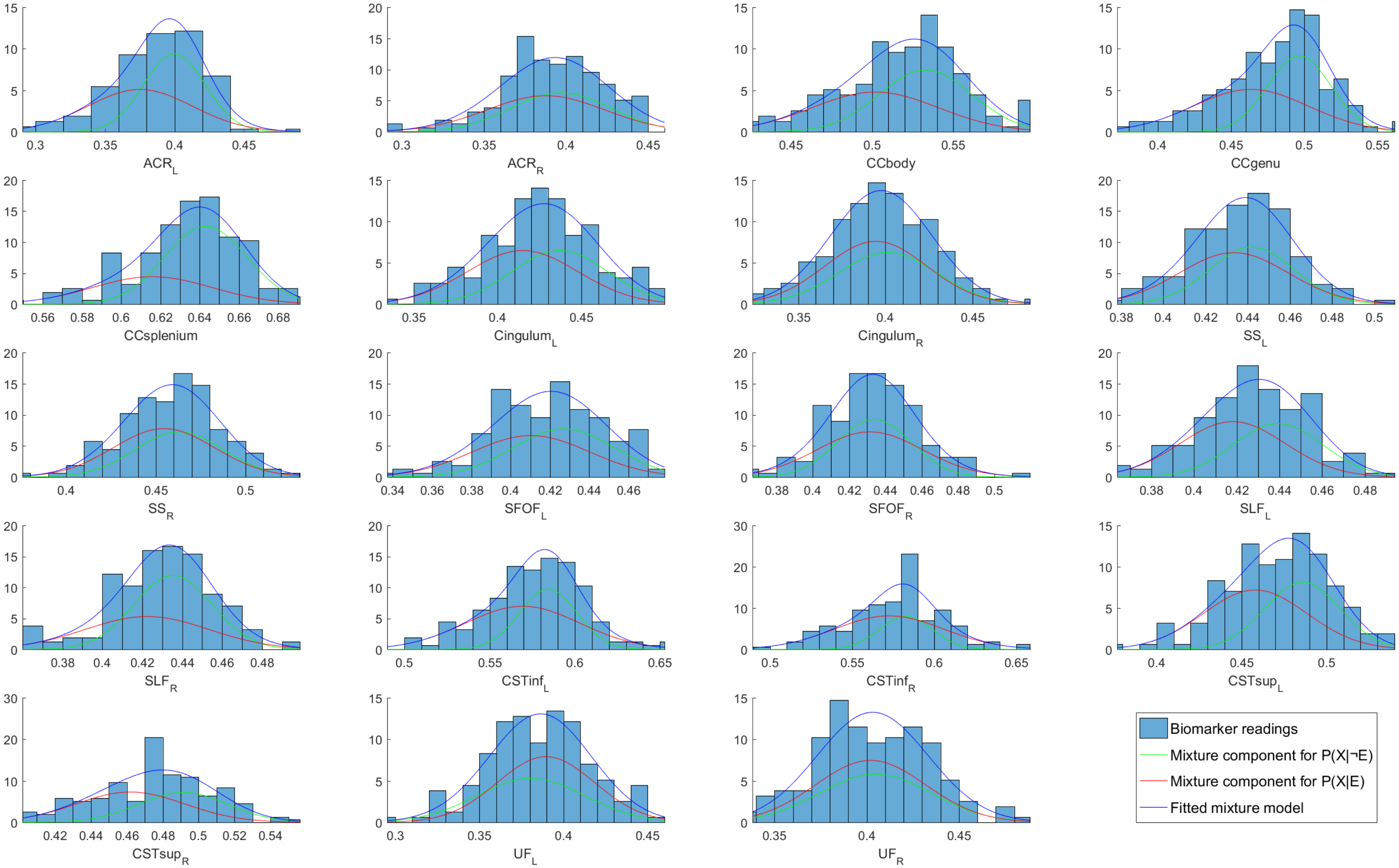


Figure 6.3: The fitted mixture models for Sets FKN. Constraints: mixing proportion > 0.25, no outliers removed.

6.1.2 Structural data

Some general comments on the difficulties of modelling the progression of structural biomarkers, as opposed to FA-based metrics:

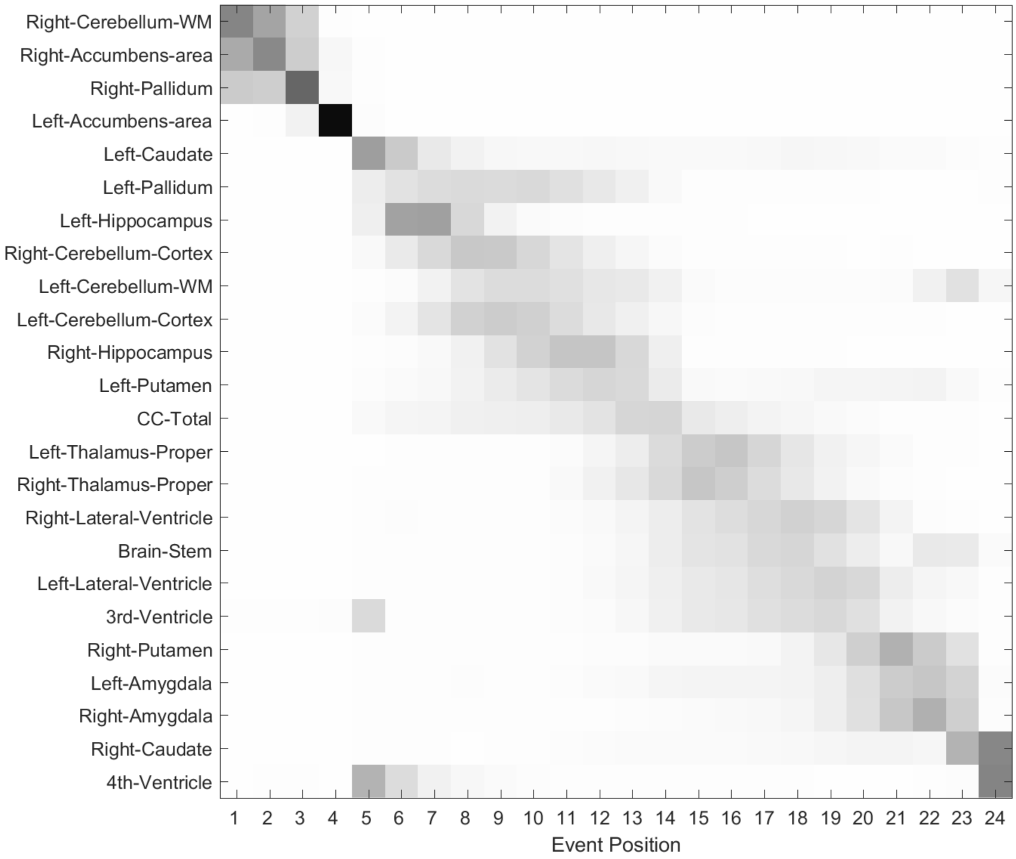
- Patients' structural biomarker data were less often normally distributed; this had an impact on the mixture fitting process, although we note that the EM algorithm did reach convergence for all biomarkers.
- Matters were improved by normalising region volumes by total intracranial volume, and region cortical thicknesses by mean thickness of the corresponding hemisphere.
- When compared to the results for FA, overall positional variance of structural biomarkers was increased. As would be expected, positional variance was higher when modelling each dataset (Set K and Set N) individually than when combined into Sets KN.

These difficulties are probably related to the smaller sample size, as well as the extra parametric assumptions made during the FreeSurfer segmentation and parcellation processes.

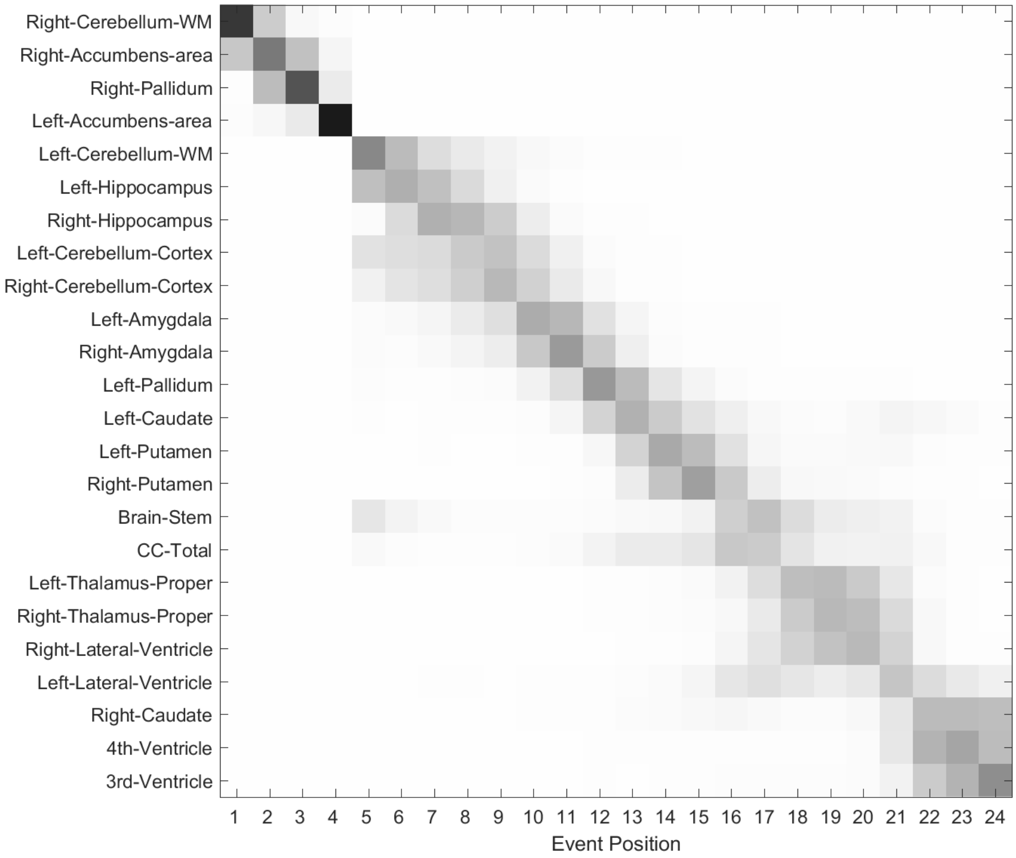
Subcortical Segmentation

These data are comprised of volumes in mm^3 for subcortical structures, as generated by FreeSurfer's automatic subcortical segmentation process. The regions were taken from the atlas by Fischl et al. (2002), included in FreeSurfer 5.3.0. In order to account for inter-subject variability, region volumes were normalised by the subject's total intracranial volume. The EBM parameters were set as mix > 0.25 and outliers $> 2.5\,\sigma$ excluded during the mixture fitting process. The model was run separately on ALS patients and controls; we expect volume changes due to healthy aging, and this allows a comparison.

Under this model, we compare the most likely order of progression for patients with that of controls (Figure 6.4). We find that ventricles progress earlier in patients; this is most apparent for the lateral ventricles, but the third and fourth ventricles also appear to be affected (and have increased positional variance). The corpus callosum also progresses slightly earlier in patients; although the change is not as dramatic as might be expected, the corpus callosum's positional variance is more heavily weighted towards the latter half for controls, suggesting that the change is noteworthy.



(a) Normalised volumes of subcortical structures, EBM run on ALS patients



(b) Normalised volumes of subcortical structures, EBM run on healthy controls

Figure 6.4: Positional variance diagrams for Sets KN subcortical volumetric data. Constraints: mixing proportion > 0.25, outliers $> 2.5\,\sigma$ removed during mixture fitting.

The positions of the thalamus and brain stem “cleanly” swap (i.e. the location of their most frequent positions are not equal), with the thalamus progressing first for patients. We also observe that the amygdala is affected much later in patients than in controls; it seems unlikely that ALS confers a neuroprotective effect on the structure, so this may instead suggest that the absolute progression of the intervening biomarkers has been accelerated.

Cortical Parcellation

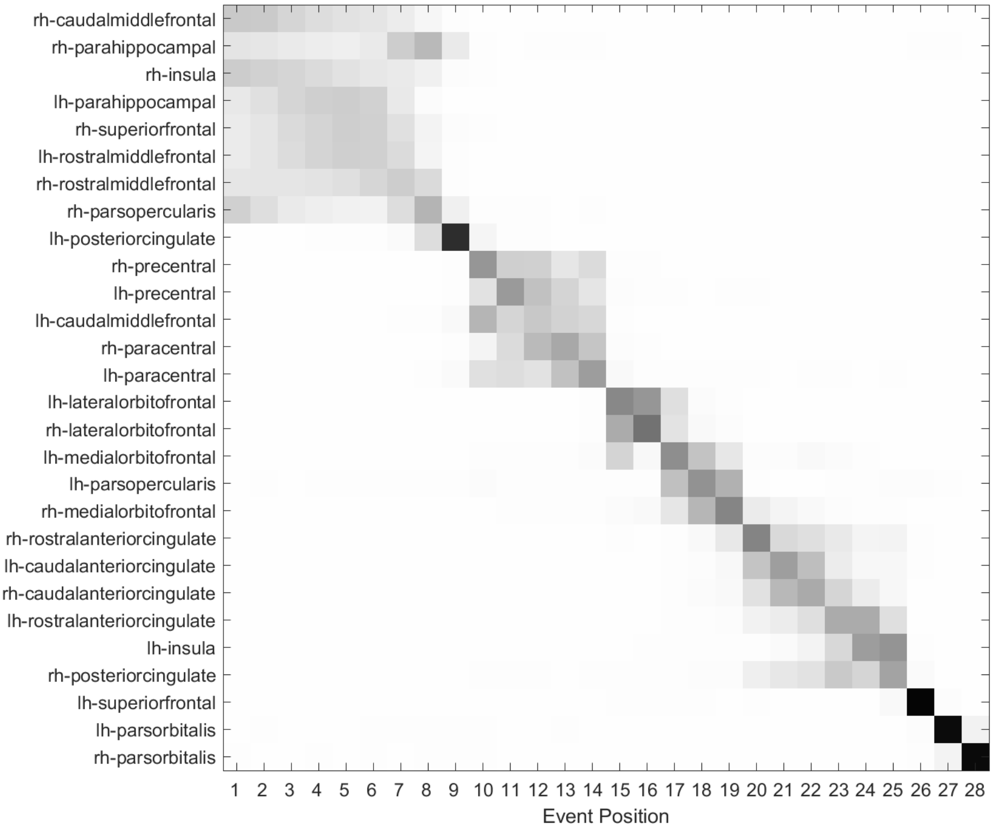
These data are comprised of mean thicknesses in mm for cortical structures, as generated by FreeSurfer’s automatic cortical parcellation process. The regions were taken from the Desikan-Killiany atlas (Desikan et al., 2006), included in FreeSurfer 5.3.0. In order to account for inter-subject variability, and to transform the data to be Gaussian, region thicknesses were normalised by mean cortical thickness of the corresponding brain hemisphere. The EBM parameters were set as mix > 0.25 and outliers $> 3\,\sigma$ excluded during the mixture fitting process. The model was run separately on ALS patients and controls; we expect cortical thinning due to healthy aging, and this allows a comparison.

The outlier removal threshold is different from that used in the previous section; we have previously stated that we wish to use the least strict constraints possible, and there did not appear to be substantial differences to the mixture fitting when using a threshold of $2.5\,\sigma$ or 3σ.

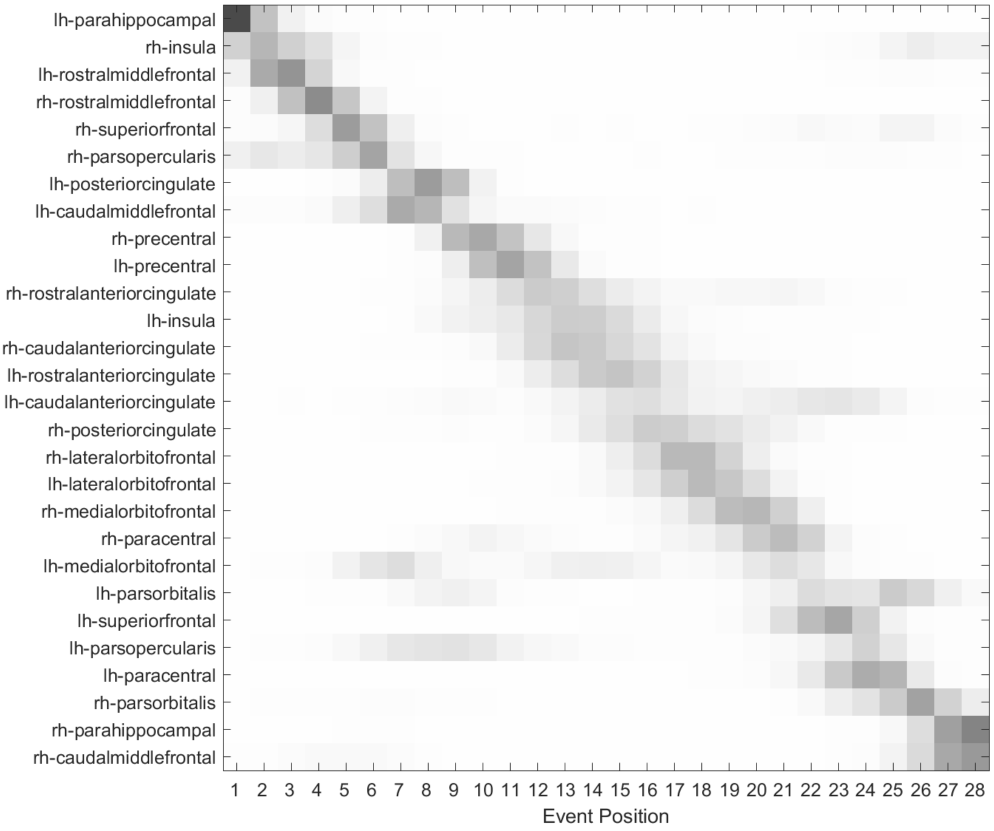
Again, under this model, we compare the most likely order of progression for patients with that of controls (Figure 6.5). It is immediately apparent that Events 1 to 8 and 10 to 14 are strongly clustered for ALS patients, unlike for controls, which display a more consistent positional variance across the entire progression.

The precentral and paracentral gyri events form a cluster in ALS patients; these regions are the primary motor cortex and supplementary motor area, respectively, so it appears noteworthy that they progress in such close proximity. Examining these same regions in controls reveals although the precentral gyrus also progresses early on, the paracentral gyrus is affected considerably later in overall progression. However, we note that the right paracentral gyrus event has greater positional variance in controls than in ALS patients.

In ALS patients, the anterior cingulate cortex regions are clustered together approximately three-quarters through the progression order; the same regions occur slightly



(a) Normalised thicknesses of cortical structures, EBM run on ALS patients



(b) Normalised thicknesses of cortical structures, EBM run on healthy controls

Figure 6.5: Positional variance diagrams for Sets KN cortical thickness data. Constraints: mixing proportion > 0.25, outliers $> 3\ \sigma$ removed during mixture fitting.

less than halfway through progression for controls. Again, we recall that the positional variance diagrams show only relative event progression, and that the overall time scales may not be directly comparable between patients and controls.

Finally, some regions appear to progress at a similar point for both patients and controls: the rostral middle frontal gyrus occurs early on in progression, the orbitofrontal cortex regions occur in the middle of progression, and the pars orbitalis occur towards the end of progression.

6.2 Neuropsychological biomarkers

The results in this section are included as proof of concept that the EBM can be applied to non-MRI data.

Set K

Here, the EBM was applied to patients and controls combined (as in the AD implementations of the model), as the neuropsychological tests measure variables that have age-related changes. The small size of cohort was also a factor in this decision, as the full range of neuropsychological data were only available in Set K.

Within this cohort and under this model, the most likely order of progression (Figure 6.6) suggests that changes in executive function occurred at an early stage of disease evolution, while language impairment occurred after executive dysfunction. Changes in functional ability as assessed by rate of change of ALSFRS-R occurred during the executive function changes and before impairments of speech and language.

Small subject numbers strongly affect mixture fitting, and modelling different types of biomarkers (some integer, some scale-based and some continuous) will also have had an impact. While I continue to caution against judging the appropriateness of the mixture fitting solely by eye, Figure 6.7 reveals that biomarkers such as Distortions, TROGerr and IncompleteSentences contain limited variation within their readings; as such, fitting a mixture of two Gaussians may not be entirely appropriate. Indeed, it proved necessary to remove all constraints from the EM algorithm in order for the mixture fitting to reach convergence for all biomarkers.

Because of these concerns, I suggest that this outcome is treated as proposing a hypothesis of neuropsychological changes in ALS.

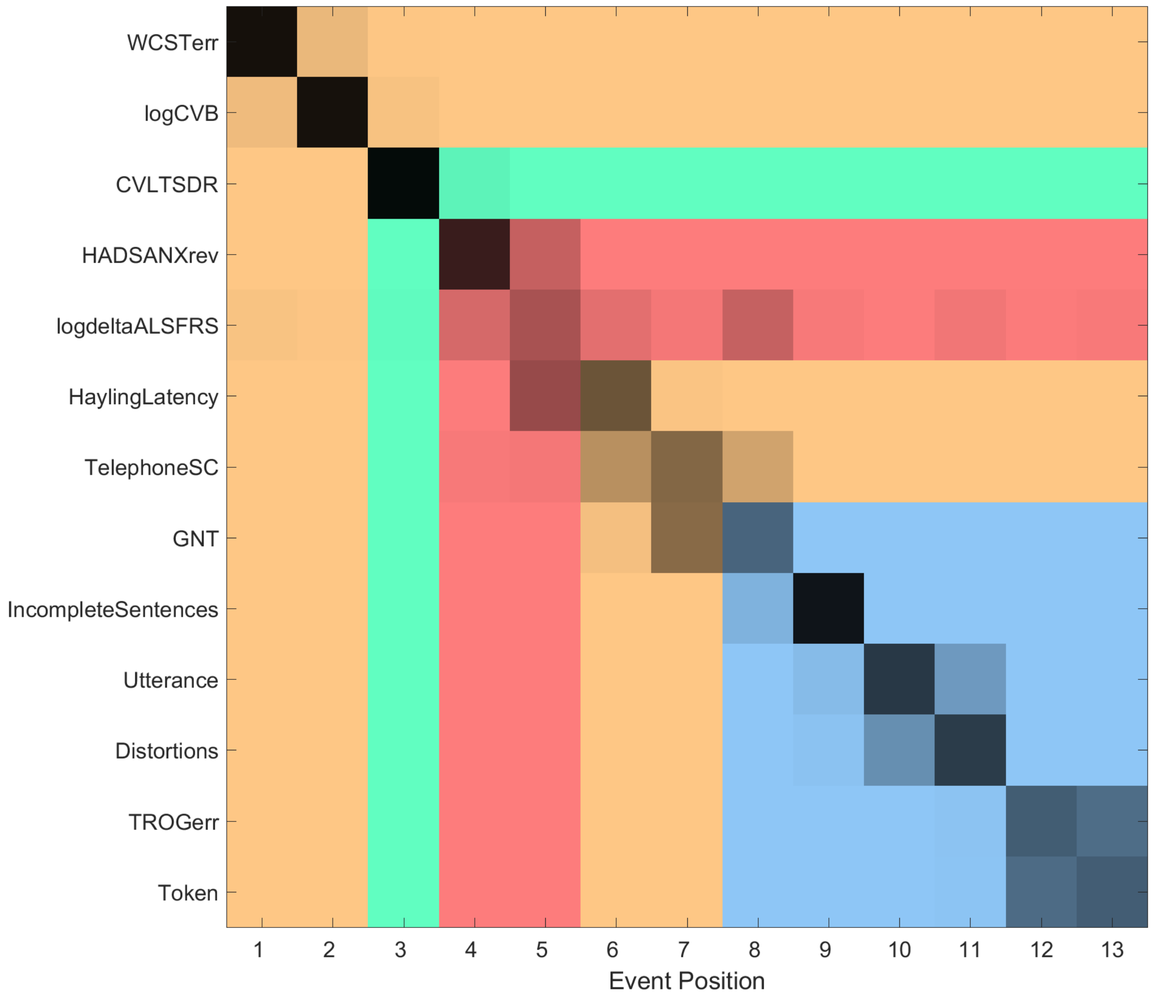


Figure 6.6: Positional variance diagram for Set K Clinical and Neuropsychological data. No constraints placed upon mixture fitting. The biomarkers have been colour-coded to indicate their classification into function and clinical, memory, executive function or speech and language.

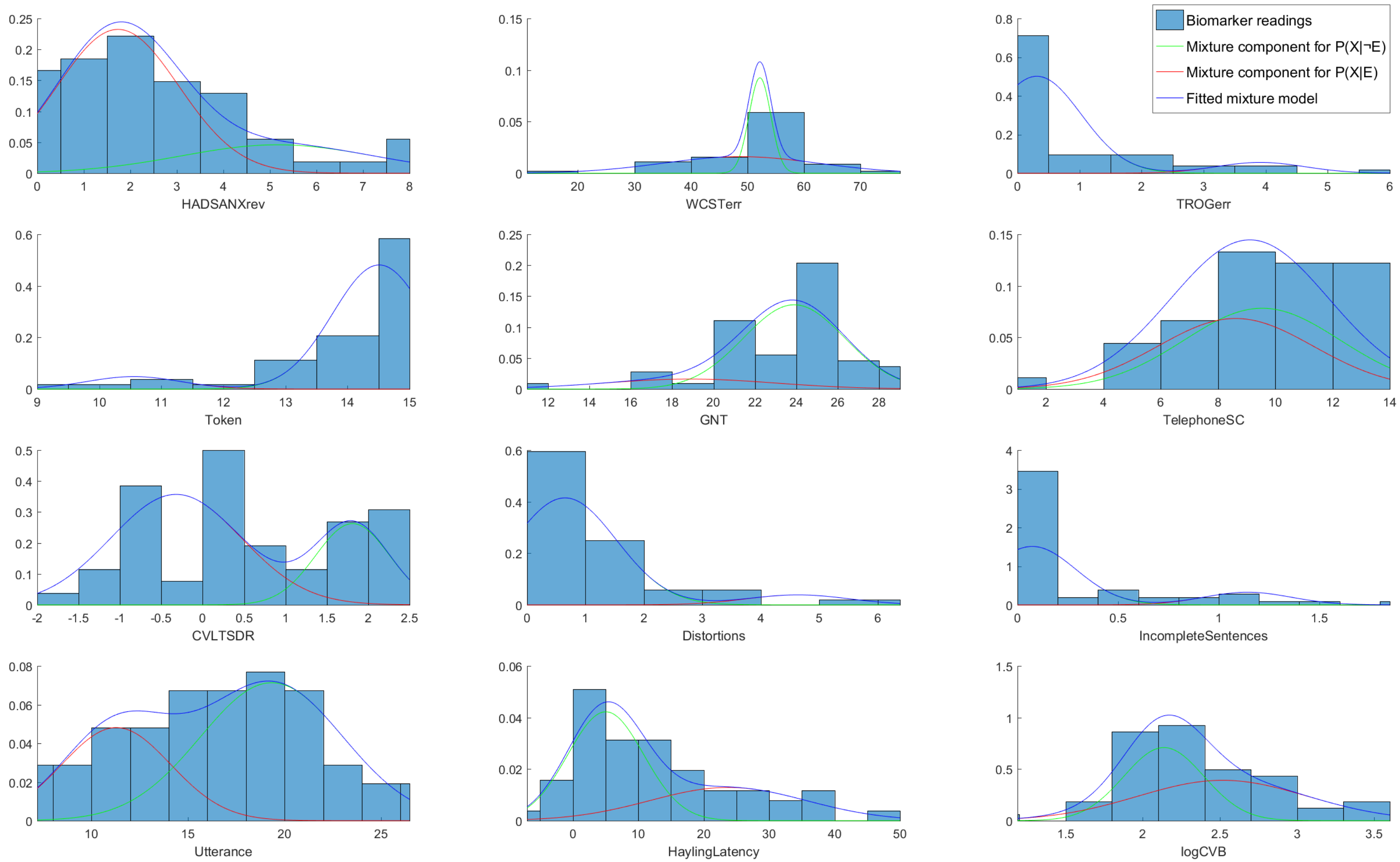


Figure 6.7: The fitted mixture models for Set K neuropsychological measures.

Chapter 7

Discussion

The aim of this study was to define disease progression in ALS, by applying the EBM to historical and contemporary data sets. However, given that the heterogeneity of ALS is reflected by the variation within the population subgroups, caution is warranted when considering extrapolating the cohort-specific EBM results to ALS in general.

In this chapter, I examine and discuss the model output for the three data modalities examined (FA, structural and neuropsychological). I also comment on some more general aspects of the modelling process.

7.1 FA results

7.1.1 Model output

Examining the τ correlations between the most likely order of progression (S_{ML}) for all groupings (as given in Figures 6.1 and 6.2) shows that only 4 of the 21 possible correlations are positive and significant at $p < 0.05$ (Table 7.1). This could be a reflection of the heterogeneity of ALS or a consequence of the small sample sizes. In either case, this subgroup variability means that we cannot necessarily assume that the EBM progression order is representative of ALS as a whole.

Subgroup variability may be a cause of the indistinct positional clusters seen in Figure 6.1. We note that if we perform *post hoc* analysis by removing outliers $> 2.5\ \sigma$ during mixture fitting and re-running the EBM (Figure 7.1), the progression order remains unchanged and the positional clusters become more distinct. The fact that outlier removal corresponds with reduced positional variance serves as a reminder of the importance of obtaining a sufficiently large sample size.

In contrast to the overall EBM progression order, the CST biomarkers are consis-

	K	N	FK	FN	KN	FKN
F	-0.0526	-0.2047	0.2398	0.0526	0.0292	-0.0058
K		-0.2982†	0.3099*	0.0526	-0.0409	0.3450*
N			-0.3099†	0.0409	0.2982*	-0.2047
FK				0.4152*	0.1813	0.1930
FN					0.1111	0.1462
KN						-0.1345

Table 7.1: τ correlation between S_{ML} of the population cohort groupings. EBM applied to FA data of Sets FKN. Constraints: mix > 0.25, outliers $> 2\sigma$ removed during mixture fitting for Sets F, K and N. * = positively correlated, $p < 0.05$. † = negatively correlated, $p < 0.05$.

tently staged with low positional variance across all population subgroups. We are therefore confident that the EBM demonstrates that the directionality of the FA changes within the CSTs (inferior preceding superior) is a core feature of ALS.

Due to the sample size and subgroup variability, we can only use the model output to provide us with a hypothesis of the temporal order of white matter tract degeneration in ALS (Figure 7.2). This characterisation of white matter degeneration is potentially compatible with the model proposed by Braak et al. (2013), although there are disagreements between the two systems.

Braak et al. characterise the progression of ALS as "induction and dissemination of pTDP-43 pathology chiefly from cortical neuronal projections, via axonal transport, through synaptic contacts to the spinal cord and other regions of the brain".

In contrast, the EBM model proposes that CST degeneration is an early feature of WM damage in ALS. Furthermore, the directional nature of FA changes suggests that ALS WM pathology is likely to be induced in the spinal cord or lower brain stem; this pathology ascends the CSTs and then spreads outwards (passing through the CC) in a "out-and-down-and-round" pattern. It is this secondary pattern that may be compatible with that proposed by Braak et al. The inclusion of spinal cord dMRI data into the EBM would provide a method by which to test this new hypothesis of WM pathology induction.

These differences in proposed progression may be due to the different data modalities employed: Braak et al. (2013) have derived their results from pTDP-43 pathology, and the EBM has modelled WM neurodegeneration as measured by changes in FA. It

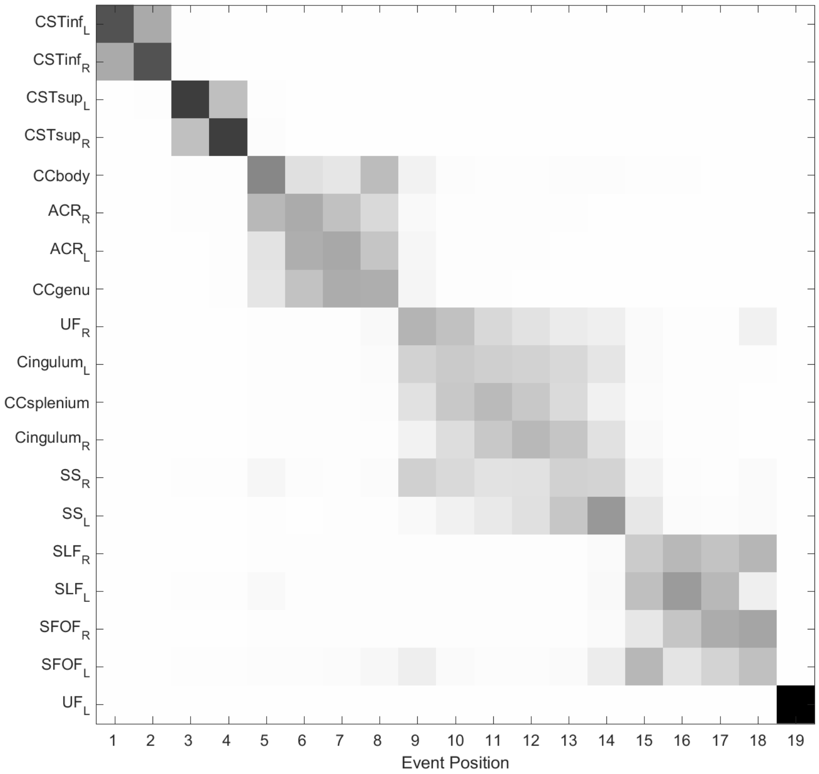


Figure 7.1: Positional variance diagram for FA biomarkers of Sets FKN. Mix > 0.25, outliers $> 2.5\,\sigma$ removed during mixture fitting.

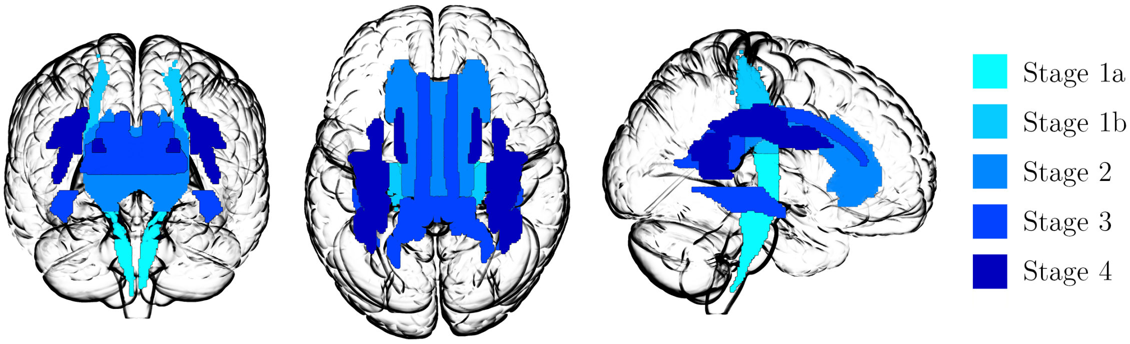


Figure 7.2: The proposed temporal order of white matter tract degeneration in ALS, as generated by the EBM. Stage 1a = Inferior CSTs; Stage 1b = Superior CSTs; Stage 2 = Corpus callosum Genu/Body, Anterior corona radiata; Stage 3 = Corpus callosum splenium, Cingulum (cingulate gyrus), Sagittal stratum; Stage 4 = Superior longitudinal fasciculus, Superior fronto-occipital fasciculus.

is possible that these two models of progression could be complementary, but further research with larger sample sizes and carefully selected regions of interest (ROIs) is needed. To this end, it would be desirable to replicate the 22 CNS regions used by Braak et al. in the form of a DTI atlas; however, we note that some regions may be too small to allow for accurate automatic segmentation, and as FA is not particularly sensitive to cortical abnormalities, it may not be possible to detect cortical changes preceding WM changes using only DTI measures.

7.1.2 The Uncinate Fasciculus

As noted previously, the mean FA value in ALS patients is lower than in controls (i.e. $\hat{\mu_a} < \hat{\mu_c}$) for all biomarkers except $\mathrm{UF_L}$. It seems likely that this exception is due to erroneous convergence of the EM mixture fitting algorithm, caused by noise within the data. The shape of the UF may also be responsible, as it is a highly curved and small fibre bundle, and both these factors can contribute to inaccurate tensor fitting.

A complete lack of positional variance within a set of MCMC samples can be an indicator of poor mixture fitting, in contrast to what Young et al. (2014) claim. This can be seen by contrasting the extremely low positional variance of the $\mathrm{UF_L}$ biomarker with its propensity to move between extremes of positions as biomarkers are added and removed or model parameters are altered. This appears to be linked to the problem of assessing biomarker directionality, as $\mathrm{UF_L}$ is the only FA biomarker for which $\hat{\mu_a} > \hat{\mu_c}$.

The UF is also different from the other modelled tracts in that its left and right sides do not progress in close proximity. For the purposes of experimental curiosity, let us suppose that the data are noisy, and that $\mathrm{UF_L}$ degeneration is actually more rapid in ALS patients than in healthy controls. We assume that the mixture components have "crossed over" during the fitting but are otherwise appropriate for the data. We then impose our desired directionality by swapping the mixture parameters so that $\hat{\mu_a} < \hat{\mu_c}$ for $\mathrm{UF_L}$, and recalculate $P(X|E)$ and $P(X|\neg E)$.

The effects of this exceptionally blunt manipulation of the model data are shown in Figure 7.3, and we find that the approximate shape of Figure 7.1 is retained. Interestingly, all biomarkers remain within their mini-clusters except for $\mathrm{UF_R}$, which moves to the very end to join $\mathrm{UF_L}$. We also note that the positional variance of $\mathrm{UF_L}$ is now above zero.

This is a surprising outcome; by crudely altering the mixture fitting for $\mathrm{UF_L}$, we have primarily effected a change in the positional variance of $\mathrm{UF_R}$. Of course, as we

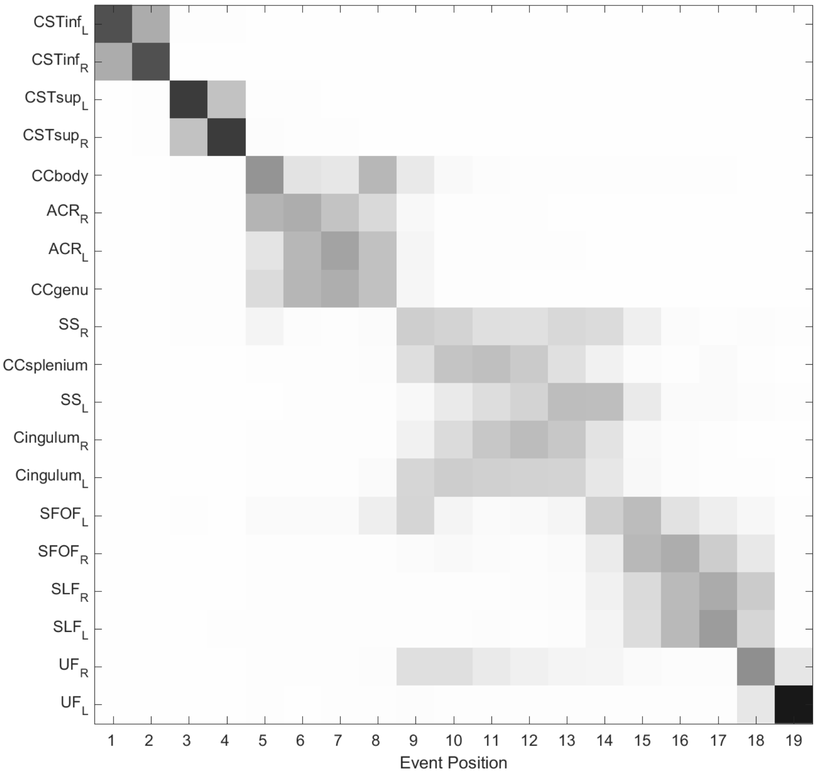


Figure 7.3: Positional variance diagram for FA biomarkers of Sets FKN. Mix > 0.25, outliers $> 2.5\,\sigma$ removed during mixture fitting. μ_a, σ_a swapped with μ_c, σ_c for $\mathrm{UF_L}$.

have enforced our own biased preconceptions on to the data, we cannot consider this model output as a significant result. We are therefore no wiser as to whether the UF's position should be in the middle or at the end of progression, although it seems likely that the left and right sides of the tract should progress together.

Overall, the unexpected nature of the effect serves as a reminder of why we employ statistical procedures to seek impartiality, and why model constraints and thresholds must be declared before modelling.

7.1.3 General observations on FA output

In this study, we have used FA as an index of neurodegeneration. We must therefore be mindful that although the EBM directly informs us about the temporal ordering of changes in FA, the model can only indirectly provide information about the progression of neurodegeneration in ALS. Thus, all results should be considered in the context of the likely biological causes of these changes in FA.

Bilateral symmetry is present in a majority of the results: the left and right halves of WM tracts seem to occur in close proximity to each other. It is possible that this indicates effective or accurate modelling; however, there is no consensus (and indeed little discussion) in the literature as to whether asymmetry of FA changes in WM tracts is considered to be a defining feature of ALS. In order to test this aspect of the modelling, it would be desirable to run the EBM on a disease with consistently asymmetric neurodegeneration of WM. Unfortunately, diseases of this sort are not plentiful; potential candidates include corticobasal degeneration (CBD) and frontotemporal lobar degeneration (FTLD), particularly the semantic dementia variant. A more feasible alternative may be to obtain a large dMRI data set of stroke patients and apply the EBM to subjects with WM damage primarily on one side.

Lastly, the effects of an outlier HC (as explored in subsection 5.1.3) show that even with a control cohort size of 70, the EM algorithm is sensitive to a single outlier. Alternative methods of fitting mixture models should be explored for future work, and we demonstrate one such possible method in subsection 7.4.1.

7.2 Structural results

Confidence in the model output for the structural data is lower than for the FA data, for multiple reasons:

- The presence of high positional variance across the EBM output.
- Gaussian mixture modelling may not be entirely appropriate for structural data, although normalisation of the data appears to partially mitigate this problem. Further investigation is required, once more data are available.
- The FreeSurfer parcellation and segmentation methods will have had a large effect on the results, particularly as manual surface pial editing was performed on only the structural data of Set N. Comparisons with other methods of parcellation and segmentation, such as those offered by ANTs, would be worthwhile.
- There is noticeably less bilateral symmetry of ROIs than in the FA results.

The model output for structural data possess slightly lower bilateral symmetry than when using FA data. However, FA is an index of microstructural abnormalities, while volume and thickness are “gross” indices of macroscopic tissue loss. Thus, it is possible that the structural markers are simply less appropriate for ALS, resulting in reduced bilateral symmetry across the most likely event order; further investigation is warranted.

Finally, it is worth noting that the comparison between the ALS and HC most likely orders of progression (MLOOPs), as employed in subsection 6.1.2, may be a useful tool. This is an unconventional use for the EBM output, but it appears to be a simple and effective method of visually extracting more information from the available results. Of course, this does have the drawback of reintroducing human bias into the system; it is possible that there are tools in the field of ranked data to assist with this.

7.3 Neuropsychological results

We reiterate that the neuropsychological results are included as proof of concept, and thus they are employed only to generate a hypothesis. Firstly, these results have been generated from a small data set, so the mixture distribution fitting is extremely vulnerable to outliers. Secondly, the assumption that the data are normally distributed seems much less likely to hold. While the EM algorithm has fitted mixtures that appear to approximate the underlying data (Figure 6.7), some biomarkers are clearly better suited to this type of modelling than others (e.g. compare the fitting for Utterance with that of Token).

Other mixture types, such as Poisson or Beta, may be appropriate for some neuropsychological biomarkers. A proof of concept inclusion of a Poisson mixture is given in subsection 5.1.4, although the simultaneous use of multiple mixture types requires further investigation.

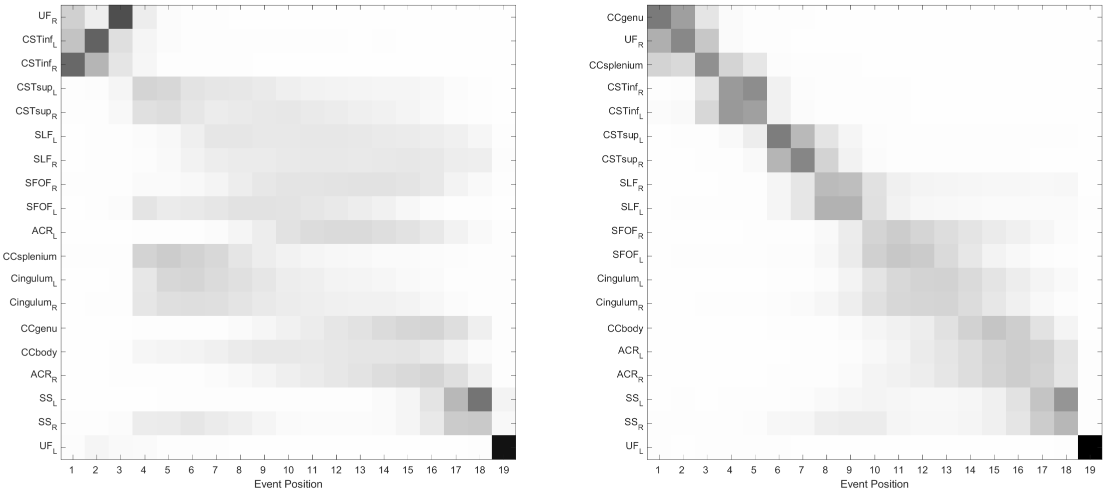
7.4 Aspects of the Modelling Process

7.4.1 Bootstrapping for mixture fitting

For each biomarker, we bootstrap the data by randomly sampling the control data with replacement; the size of the resample must be equal to the size of the original control sample. The ALS patient data are bootstrapped in the same manner. The mixture models are then fitted by EM, and the parameters of the mixtures are recorded.

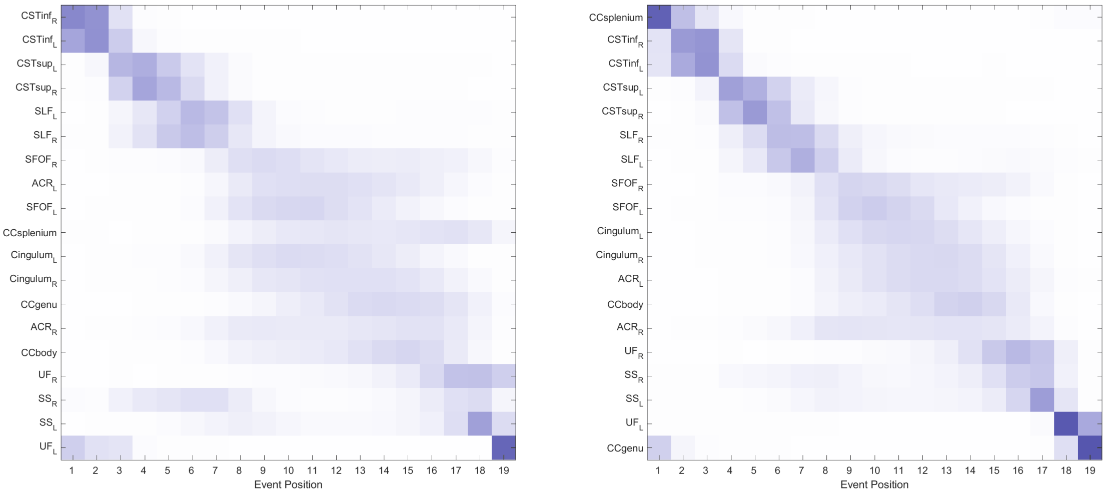
This process is repeated a large number of times. The median of the mixture parameters is then taken, giving a bootstrapped estimate of the parameters of the mixture models.

If repeated enough times, this bootstrapping process should reduce the impact of a single outlying biomarker reading (Figure 7.4).



(a) K_{24} – Outlier HC excluded

(b) K_{25} – Outlier HC included



(c) K_{24} – Outlier HC excluded*

(d) K_{25} – Outlier HC included*

Figure 7.4: PVDs comparing the MCMC samples from Set K_{24} with those of Set K_{25}. Constraints: mixing proportions > 0.25. * = EM fitted by 100 bootstrap iterations.

However, there is no simple method for deciding how many iterations are sufficient, due to the large number of distinct sets of resamples. To illustrate the magnitude of this problem, we will digress briefly: the **multiset coefficient** $\left(\!\binom{n}{k}\!\right)$ is defined as the number of distinct sets, when choosing k from n with replacement. The multiset coefficient is related to the binomial coefficient (proof omitted):

$$\left(\!\!\binom{n}{k}\!\!\right) = \binom{n+k-1}{k} \tag{7.1}$$

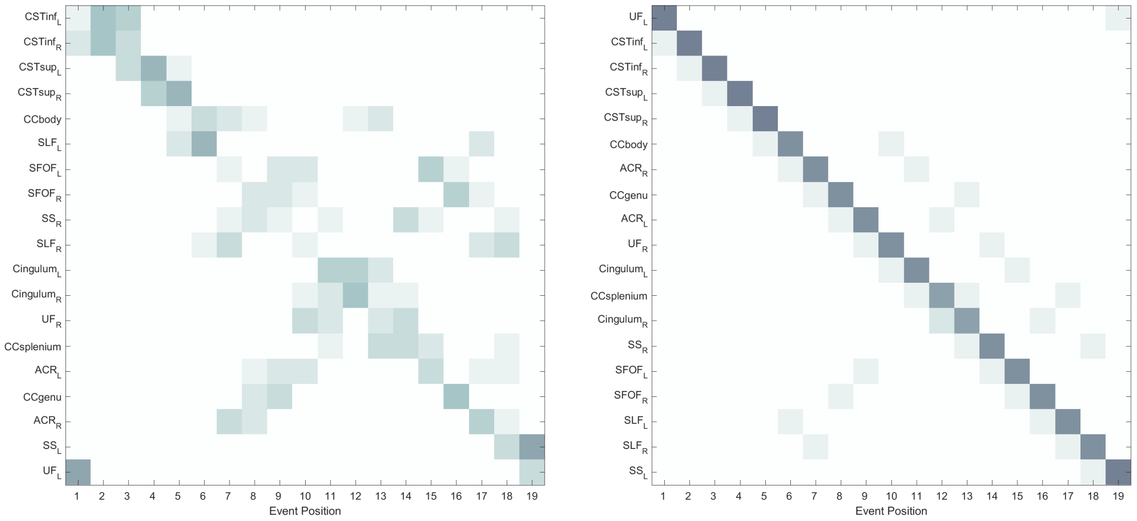
For 87 ALS patients and 70 HCs (the outlier HC must be included, as the point of the bootstrapping is to proof against outliers) there are $\left(\!\binom{87}{87}\!\right) \times \left(\!\binom{70}{70}\!\right)$ possible combinations; equivalently, there are $\binom{173}{87} \times \binom{139}{70} \approx 3.392 \times 10^{91}$ distinct possible bootstrap samples. Exhaustive bootstrapping is clearly not possible for this data set, nor is it possible to cover a significant fraction of the possibilities.

Given the vast numbers of possible bootstrap combinations, it seems necessary to assess the consistency of the bootstrapped EBM. We find that repeated application of the bootstrapped EBM to Sets FKN shows that model output is not entirely consistent. To investigate this further, we plot a "meta-positional variance diagram" of the multiple runs of the bootstrapped EBM (for both 100 and 1000 bootstrap iterations). This visualises how the model output varies across these runs (Figure 7.5), and we can see that i) the consistency of the model output increases with the number of EBM runs, and ii) the number of bootstrap iterations has a similar but much greater effect on model consistency.

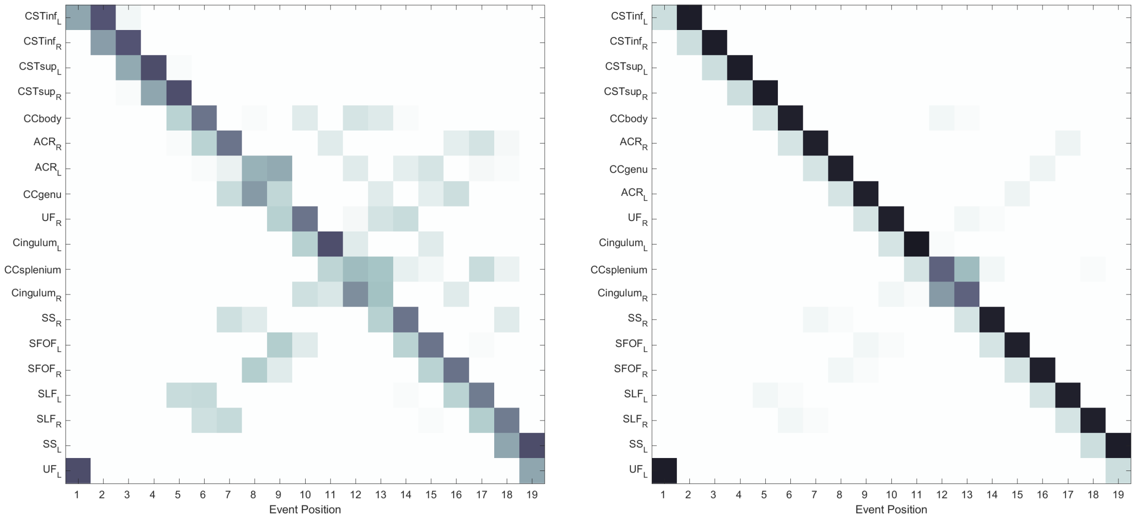
In other words, the more bootstrap iterations, the more likely it is that the model output will be "typical". However, performing large numbers of EBM runs at 1000 bootstrap iterations is extremely time consuming. If we assume these meta-PVD proportions to be accurate (a very large assumption), how many runs of how many iterations do we need to be sure of finding the "typical" order S_{mode}?

Let us declare the model output S_b across b bootstrapped iterations to be "correct" if $\tau(S_b, S_{\text{mode}})$ is significant at the $p < 0.1$ level. The distribution of the number of correct orders X can then be represented by a Binomial distribution: $X \sim \mathrm{B}(n, p)$, where n is the number of EBM runs, and p is the probability that an individual S_b is correct.

Say that we want a probability > 0.999 that at least half of all EBM runs are correct (i.e. that for at least $\lceil \frac{n}{2} \rceil$ EBM runs, $\tau(S_b, S_{\text{mode}})$ is significant). This is equivalent to $P(X \leq \lfloor \frac{X}{2} \rfloor) < 0.001$, which can be found using the Binomial cumulative distribution function, which is implemented in Matlab as *binocdf*.



(a) 10 runs, 100 bootstrapped EM iterations (b) 10 runs, 1000 bootstrapped EM iterations



(c) 100 runs, 100 bootstrapped EM iterations (d) 100 runs, 1000 bootstrapped EM iterations

Figure 7.5: Meta-PVDs, summarising the exploration of bootstrapped EM mixture fitting as a tool to reduce the impact of a single outlier HC. The EBM was applied to Sets FKN, including the outlier HC. The diagrams show how the model output, S_{ML}, varied across EBM runs for different numbers of bootstrapped EM iterations (they are not representations of MCMC samples).

From the data used to create Figure 7.5, we establish that 100 EBM runs at 100 bootstrap iterations produces 70% correct runs and 100 EBM runs at 1000 bootstrap iterations produces 96% correct runs. Matlab's *binocdf* function can then be used to find that we require 54 EBM runs at 100 bootstrap iterations or 5 EBM runs at 1000 bootstrap iterations to ensure that S_{mode} will have < 0.001 probability of being atypical. Note that although these numbers are only estimates, they are bound to be overly harsh, as the incorrect orders are likely to have varying levels of disagreement with S_{mode}.

In this section, we have demonstrated that the mixture fitting process can be successfully bootstrapped, and briefly investigated the minimum number of EBM runs and bootstrap iterations required to obtain a given level of consistency of EBM output.

7.4.2 Bimodality of Progression

We have shown that bootstrapping the EM mixture fitting is a possible solution to the problem of excessive outlier impact, but that the results do not appear completely consistent. The bimodality visible in Figure 6.1 starts to reappear in the bootstrap trials (Figures 7.5 - 7.6), and we propose that this could suggest the outline of two different patterns of ALS progression.

This bimodality – a second potential MLOOP – appears with diminishing frequency as the number of bootstrap iterations increases. Could this be a consequence of a subpopulation or phenotype? Investigation is needed to see if there are any commonalities in the ALS subjects that are causing the different MLOOPs. For example, take the list of subjects sampled for 100 EBM runs of 100 bootstrap iterations, divide them into "early ACR" and "late ACR" groups, and then perform analyses (such as latent cluster analysis, principal component analysis or latent Dirichlet allocation) to see if there any traits distinguishing the two groups. Any common traits could then be compared with established clinical phenotypes.

7.4.3 Assessing the EBM's potential as a diagnostic tool

If the EBM were able to completely define the progression of ALS, then clinical uses would include: i) diagnosing phenotypes of ALS, ii) being able to distinguish ALS from other neurodegenerative diseases, thus no longer having to diagnose the disease by exclusion, and iii) earlier diagnosis of ALS. These challenges are beyond the capabilities of the results produced in this study, so we instead consider a far simpler task: assuming

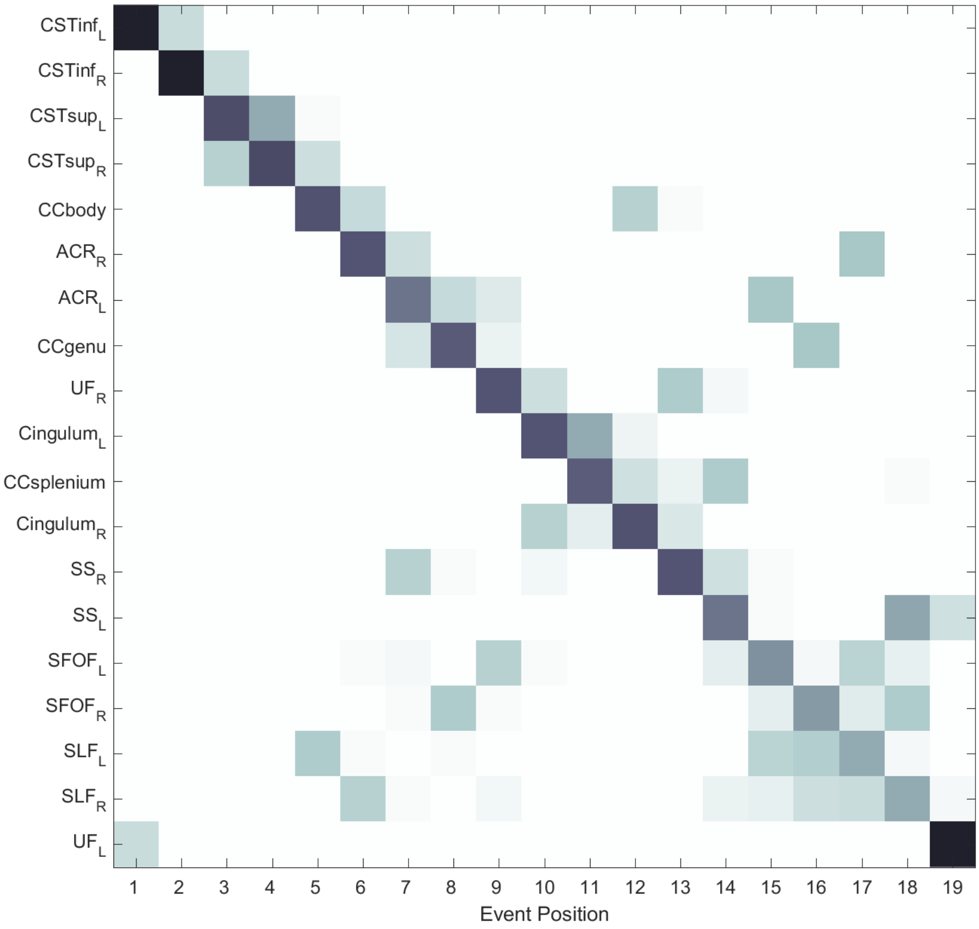


Figure 7.6: Meta-PVD for Sets FKN, excluding the outlier HC. The diagram shows how the model output, S_{ML}, varied across 100 EBM runs of 100 bootstrapped EM iterations, and is not a representation of MCMC samples. Constraints: mixing proportions > 0.25

the CST progression to be a primary feature of ALS, can we use only the EBM to distinguish between patients and controls?

The classifying stages are defined as the four biomarkers of the CST as they occur in the most likely progression order (inferior left, inferior right, superior left, superior right). We assign each patient j to the stage k which is given the highest probability by the EBM, by solving the likelihood of that patient's data given S_{ML} (Equation 2.1) for argmax_k. Note that this is only possible after the MCMC process, as we have then defined S_{ML}, $P(x_{ij}|E_i)$ and $P(x_{ij}|\neg E_i)$ for our cohort:

$$\begin{aligned}\max k_j &= \mathrm{argmax}_k\, P(X_j|S_{ML}, k) \\ &= \mathrm{argmax}_k \prod_{i=1}^{k} P(x_{ij}|E_i) \prod_{i=k+1}^{N} P(x_{ij}|\neg E_i)\end{aligned} \tag{7.2}$$

The distribution of stages assigned to patients and controls is shown in Figure 7.7. Defining a subject's classification as "healthy" if they are assigned stage 0, and in a "disease state" if they are assigned stage 1 or above, gives sensitivity of 0.782 and specificity of 0.594. Plotting the ROC curve (Figure 7.8) for all possible thresholds gives the area under the curve as 0.713.

The FA values of the CSTs have previously been investigated as a candidate for an ALS diagnostic test. Foerster et al. (2012) performed a meta-analysis of 30 DTI studies in order to estimate the diagnostic test accuracy measures of CST FA. With a combined cohort of 561 ALS subjects and 530 HCs, Foerster et al. found pooled sensitivity of 0.65 (95% CI 0.61-0.69), pooled specificity of 0.67 (95% CI 0.63-0.72) and the pooled area under the ROC curve was 0.76 (95% CI 0.71-0.81). The accuracy measures produced by the EBM are comparable to those of the meta-analysis, although they certainly cannot be said to be substantially better.

As currently modelled, measuring the FA of the CSTs is not sensitive or specific enough for diagnosis of ALS. However, if biomarkers of varying types and distinct temporal positions are added to the progression order, it should be possible to improve the discriminatory capabilities of the EBM.

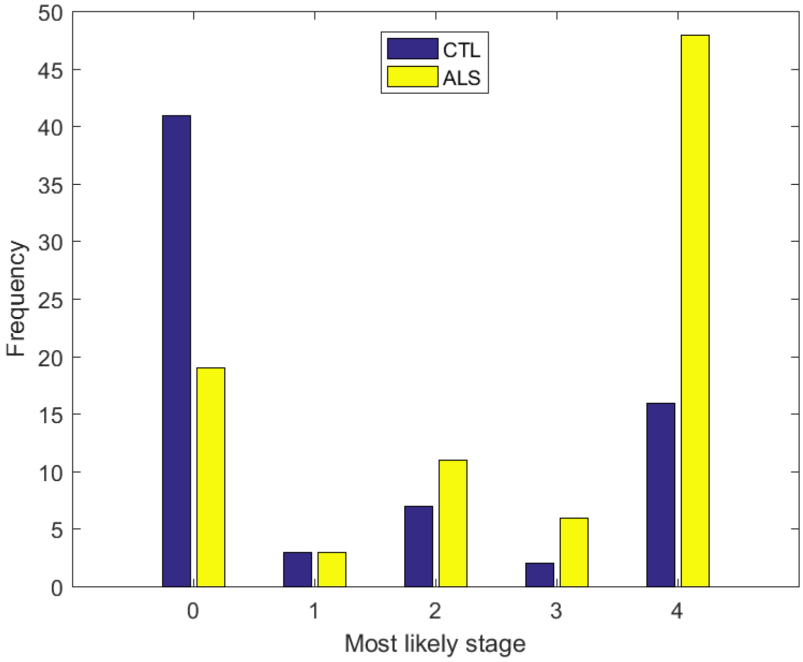


Figure 7.7: The distribution of the most likely stages assigned to all 156 subjects, as generated from the four CST biomarkers. Stage 0 is equivalent to being in a healthy state for all four CST regions. Here, the proposed diagnostic event order is CST $\text{inferior}_{\text{L}}$, $\text{inferior}_{\text{R}}$, $\text{superior}_{\text{L}}$ and $\text{superior}_{\text{R}}$.

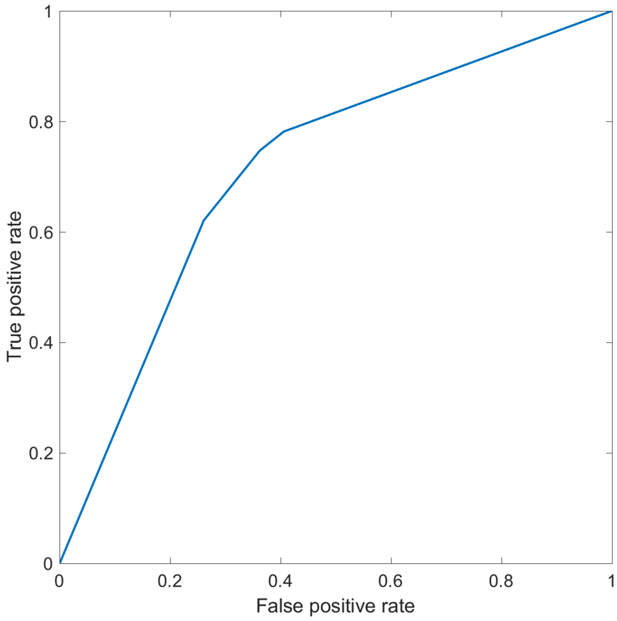


Figure 7.8: ROC curve for the EBM as a diagnostic tool for ALS, using only the four CST biomarkers. The discrimination threshold was varied to cover all possible stages shown in Figure 7.7. The AUC is 0.713.

7.4.4 CST pathology

Evidence for CST changes acting as primary driver of model likelihood

In simplistic terms, finding the most likely event order for a data set is straightforward: calculate the probability of every possible order, and select the order with the greatest log-likelihood (LLH). For n biomarkers, there are $n!$ possible permutations, and thus the exhaustive approach becomes excessively computationally expensive as n increases. To overcome this problem, the EBM employs MCMC methods to sample from the posterior distribution. This gives the EBM one of its main strengths: the ability to handle large numbers of biomarkers (even more so after the variable precision adaptation described in section 2.5).

However, if we select only a few biomarkers, brute force methods can be employed in place of MCMC techniques: for 8 biomarkers, there are only 40 320 possible permutations. We are then able to exhaustively compute the LLH for every permutation, thus covering the entire sample space, Ω.

To carry out an exhaustive exploration of Ω for an eight-biomarker model, we combined the left and right sides of eight white matter tracts into new ROIs. As before, the mean FA values were taken for each ROI, and then Gaussian mixture models were fitted to each biomarker (Figure 7.9). We elected to examine the inferior and superior thirds of the CSTs, as they should have greater distinction than if divided into halves. These new ROIs are respectively designated CST inf 3 and CST sup 3. Due to the larger volumes of the ROIs, we chose to impose less strict constraints upon mixing proportions (only greater than 0.15). Interestingly, all mixing proportions assigned by the EM algorithm were greater than our standard threshold of 0.25. This suggests that averaging across entire tracts, giving larger regions and greater volumes, guides the mixture fitting away from unrealistically unbalanced models; we recall that increasing the sample size has a similar effect.

As can be seen in Figure 7.10, the posterior distribution of the LLHs is bimodal, with a clear and distinct separation into two groups. If we define $A = \{$permutations in lower LLH group$\}$ and $B = \{$permutations in higher LLH group$\}$, then A and B form a partition of Ω. Plotting the positional variance diagrams (PVDs) (Figure 7.11) for A and B reveals that the only differences between these two groups are the order in which the inferior and superior portions of the CST are affected: we note that B, the group with the higher LLH, possess the inferior before superior CST directionality.

We also find that the ordering of the other six biomarkers are evenly distributed

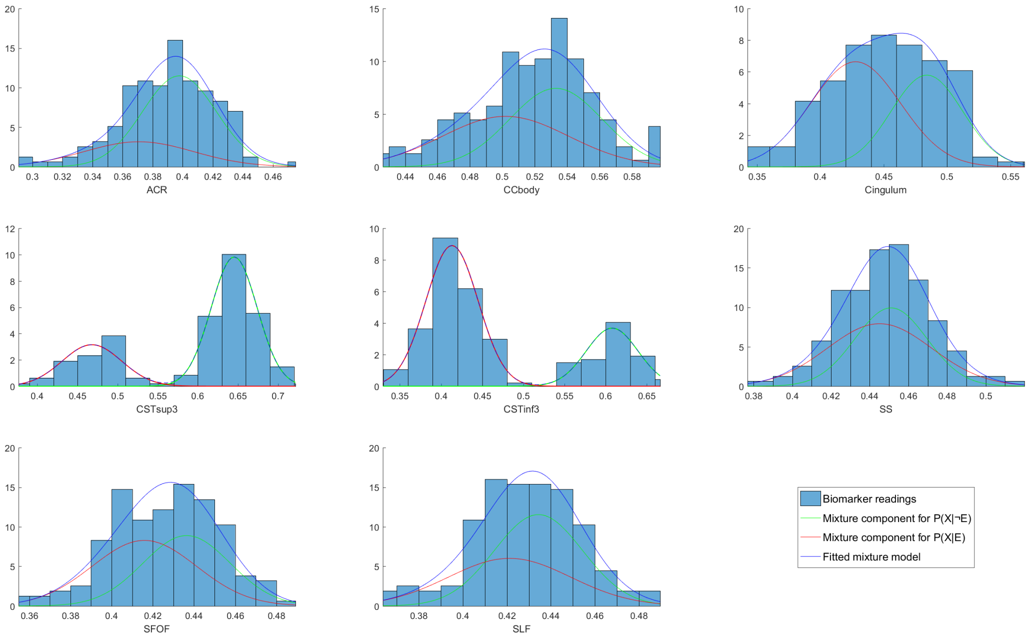


Figure 7.9: The fitted mixture models for FA data of Sets FKN. Each biomarker reading is the mean FA value across the whole tract (left and right sides combined). Constraints: mixing proportion > 0.15, no outliers removed.

across A and B (as indicated by the uniform grey shading), suggesting that these six biomarkers have no effect on whether a permutation falls into the low or high LLH group. For these eight biomarkers, it appears as though the CST changes are acting as the primary driver of model likelihood.

However, this posterior bimodality is not present if modelling the inferior and superior halves of the CST; we can conclude that the huge and distinct component separation of the CST outer thirds is responsible for the strong temporal distinction found. This is supportive of the hypothesis of CST directional degeneration: if ALS neurodegeneration progresses along the CSTs from inferior to superior, then (by the simple mathematics of arithmetic means) we would expect the differences between the outer thirds to be greater than when divided into halves.

Closer examination of the mixture models fitted to the CST inferior and superior thirds reveals (as shown in Table 7.2) that both mean separation and variance separation differ in a manner consistent with the results of section 4.2. In other words, these two sources of TI appear to be acting as predicted. The distribution of the mixing proportions is noteworthy, as the proportions are almost exactly inverted between inferior and superior. This could suggest that there is some useful information in the mixing proportion, which is currently being discarded.

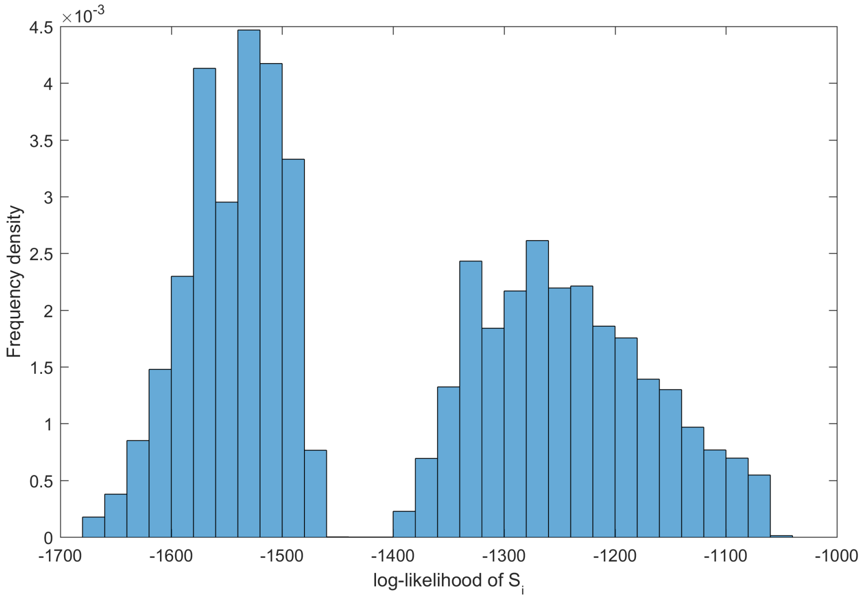
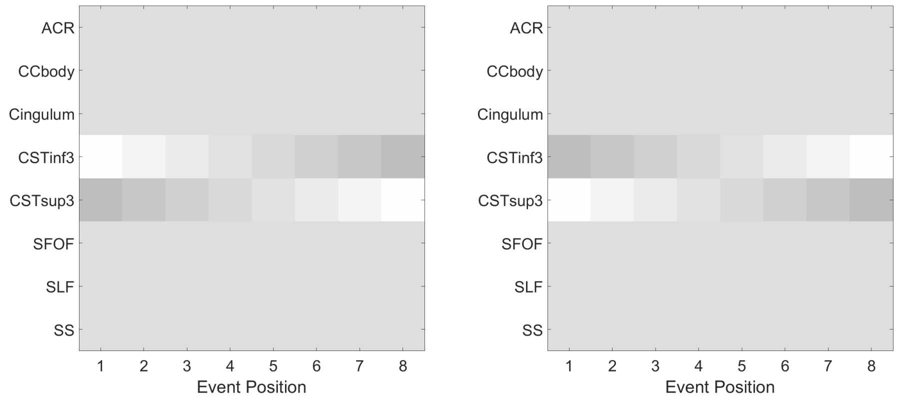


Figure 7.10: Log-likelihood calculated for all 40 320 possible event orders of an eight-biomarker model, revealing that the posterior distribution is bimodal.



(a) Set A, lower LLH group (b) Set B, higher LLH group

Figure 7.11: Positional variance diagrams for groups A and B, which form a partition of the sample space Ω of an eight-biomarker model (as seen in Figure 7.10). $A = \{\text{permutations in lower LLH group}\}$, $B = \{\text{permutations in higher LLH group}\}$.

	$\lvert\mu_a - \mu_c\rvert$	$\sigma_a - \sigma_c$	Mix HC:ALS
CST inf 3	0.1945	0.0083	0.2899 : 0.7101
CST sup 3	0.1774	0.00036	0.7071 : 0.2929

Table 7.2: Temporal information (TI) parameters and mixing proportions for the outer thirds of the CSTs. The TI parameters vary in a manner consistent with the EBM progression order: the earlier event has greater mean and variance separation.

7.4.5 The mixing proportion

In this section, I examine further the role of the mixing proportion in the EBM, and its relationship to the encoding of temporal information.

Simple implementation

For each biomarker, let us assume that the mixing proportion represents the relative "weights" of the healthy and diseased populations. This assumption appears reasonable, as we have fitted two-component mixtures, assigning one component to $P(x|E)$ and the other to $P(x|\neg E)$.

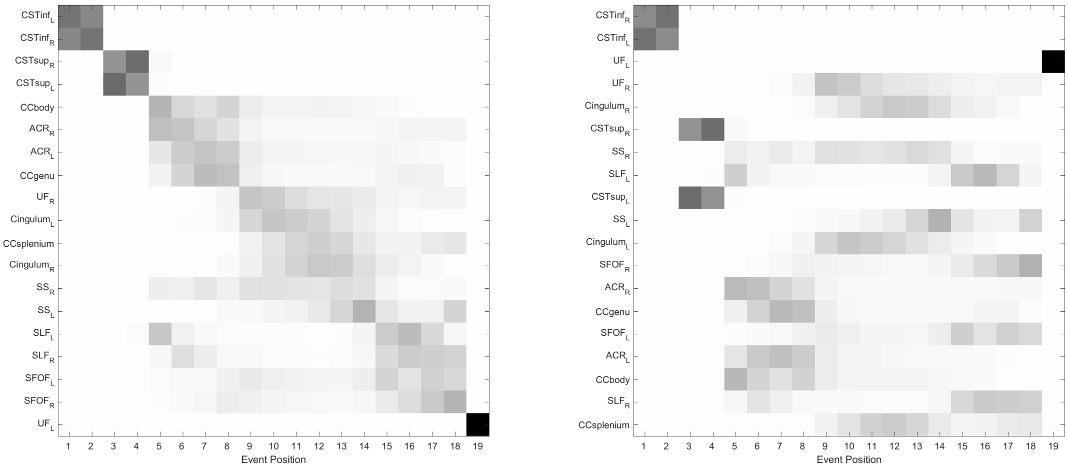
Here, subjects for whom an event has not occurred are effectively assigned to the healthy population, irrespective of whether they are an HC or a patient (and the converse is also true). Thus, the mixing proportion should indicate what fraction of the patient population have a biomarker reading in the diseased state. We can then rank the biomarkers by mixing proportion: a high ratio of diseased to healthy subjects suggests an early event, while a low ratio suggests a late event. This ranking gives us a theoretical event order, which we designate T_{mix}.

This implementation can be tested by applying the ranking process to the FA data of Sets FKN; if the mixing proportion does encode TI in the manner proposed above, then S_{ML} and T_{mix} should be similar. Visual assessment of the accuracy of T_{mix} is performed by replotting the MCMC samples of Figure 6.1, using T_{mix} as the most likely event order (Figure 7.12).

From the positional variance diagrams, it is clear that S_{ML} and T_{mix} differ substantially; the Kendall tau distance is 0.4211, meaning that over 42% of event pairs are discordant. Unsurprisingly, we find that the Kendall rank correlation coefficient for the two orders is not significant: $\tau(S_{ML}, T_{\text{mix}}) = 0.1579, p = 0.1840$. However, the directional degeneration of the CSTs is preserved

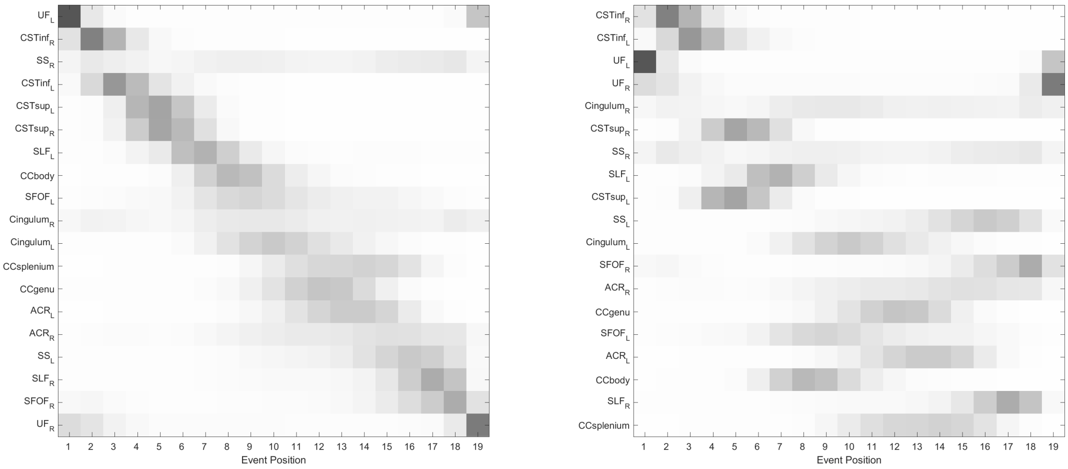
As we have previously shown that CDFs and PDFs affect the model output in different ways (subsection 4.3.1), we re-run the EBM, calculating the probabilities using PDFs (Figure 7.13). Here, we see slightly greater agreement between S_{ML} and T_{mix} (Kendall tau distance is 0.4035), and their correlation is also positive and not significant: $\tau(S_{ML}, T_{\text{mix}}) = 0.1930, p = 0.1334$.

Although we have not found significant positive correlation between S_{ML} and T_{mix} when using PDFs, this implementation is simplistic, and thus a different approach may prove more effective.



(a) S_{ML} (b) T_{mix}

Figure 7.12: Positional variance diagrams comparing S_{ML} and T_{mix}, the theoretical event order obtained by sorting the fitted biomarker mixing proportions ("simple implementation"). CDFs were used to calculate probabilities. S_{ML} and T_{mix} are not significantly correlated: $p = 0.1840$.



(a) S_{ML} (b) T_{mix}

Figure 7.13: Positional variance diagrams comparing S_{ML} and T_{mix}, simple implementation. PDFs were used to calculate probabilities. S_{ML} and T_{mix} are not significantly correlated: $p = 0.1334$.

Exploitation of bimodality in ALS patients

Here, we shall use the control data in a more aggressive manner, making use of the clearly defined control population. This approach relies upon the assumption of bimodality within the patient data, and so we shall refer to it as the bimodal implementation:

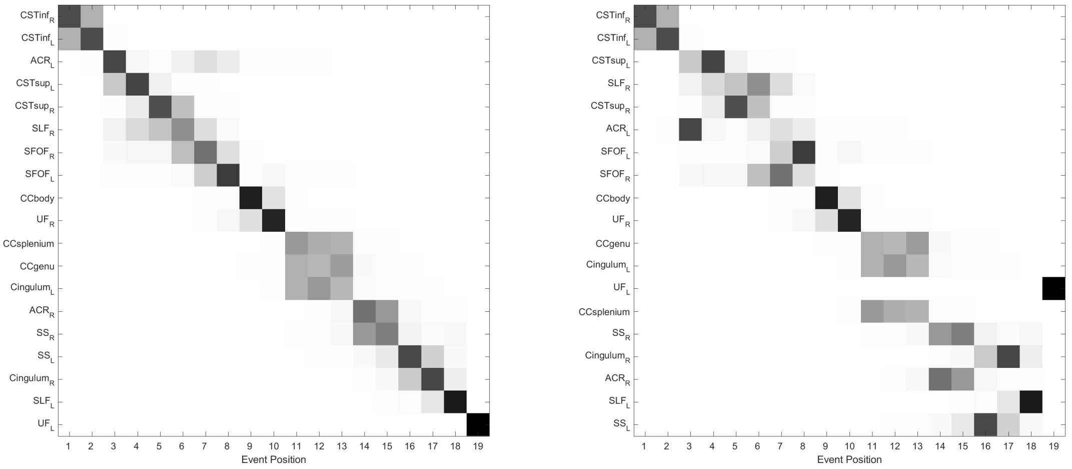
- Fit a single Gaussian distribution to control data, using Matlab's maximum likelihood function, *mle*
- Fix the resulting control distribution parameters, μ_c and σ_c
- Fit constrained two-component Gaussian mixture to **ALS patient data only**, using Matlab's *fmincon*
  - model has five parameters: $\mu_c, \sigma_c, \mu_a, \sigma_a$ and mix (the mixing proportion)
  - μ_c, σ_c are fixed
  - μ_a, σ_a and mix are free parameters

In practice, it proved necessary to set $\mu_c, \sigma_c \in$ (95% confidence interval of control distribution parameters), otherwise fitting consistently failed. Weak constraints were applied to the remaining three parameters in order to increase the speed of fitting: $\mu_a \in$ (min/max of all data), $\sigma_a \in (0, 2\sigma_a)$ and $mix \in (0.01, 0.99)$.

This implementation directly links the mixing proportion to the ratio of the healthy and diseased populations that exist within the patient data for each biomarker. Thus, the T_{mix} obtained here should be more closely related to S_{ML} than in the previous simplistic implementation.

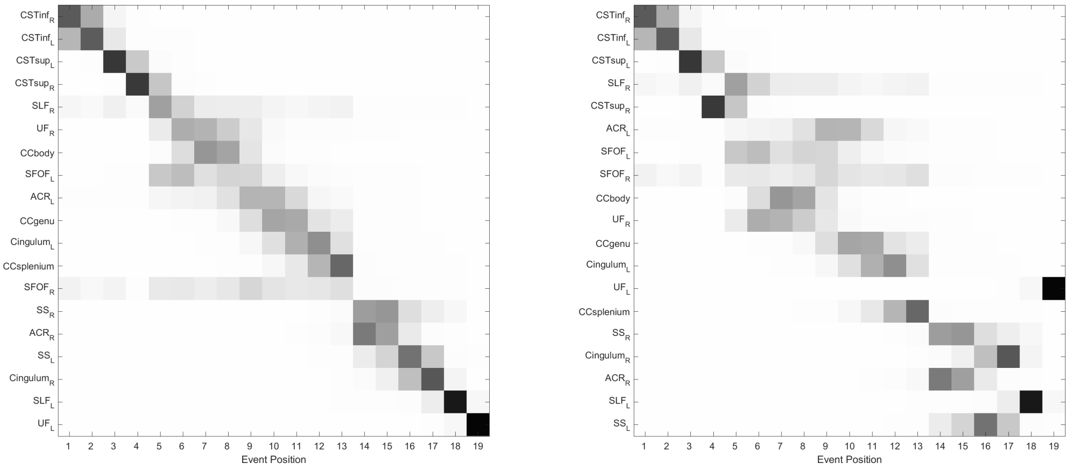
This assertion is tested by re-running the EBM and generating new MCMC samples. We begin by using CDFs to calculate probabilities; Figure 7.14 shows that this results in a strong similarity between S_{ML} and T_{mix} (Kendall tau distance = 0.3216), as well as a significant positive correlation between the two orders: $\tau(S_{ML}, T_{\text{mix}}) = 0.3567$, $p = 0.0172$.

As before, we repeat this process using PDFs to calculate probabilities (Figure 7.15). This time, we find a lesser agreement between S_{ML} and T_{mix} (Kendal tau distance = 0.4152), and their correlation is positive but non-significant: $\tau(S_{ML}, T_{\text{mix}}) = 0.1696$, $p = 0.1660$. However, we note that the positional variance diagram for T_{mix} appears far more coherent than that of the simple implementation.



(a) S_{ML} (b) T_{mix}

Figure 7.14: Positional variance diagrams comparing S_{ML} and T_{mix}, where the theoretical order has been obtained by the "bimodal implementation", using CDFs to calculate probabilities. S_{ML} and T_{mix} are significantly correlated: $p = 0.0172$.



(a) S_{ML} (b) T_{mix}

Figure 7.15: Positional variance diagrams comparing S_{ML} and T_{mix}, bimodal implementation. PDFs were used to calculate probabilities. S_{ML} and T_{mix} are not significantly correlated: $p = 0.1660$.

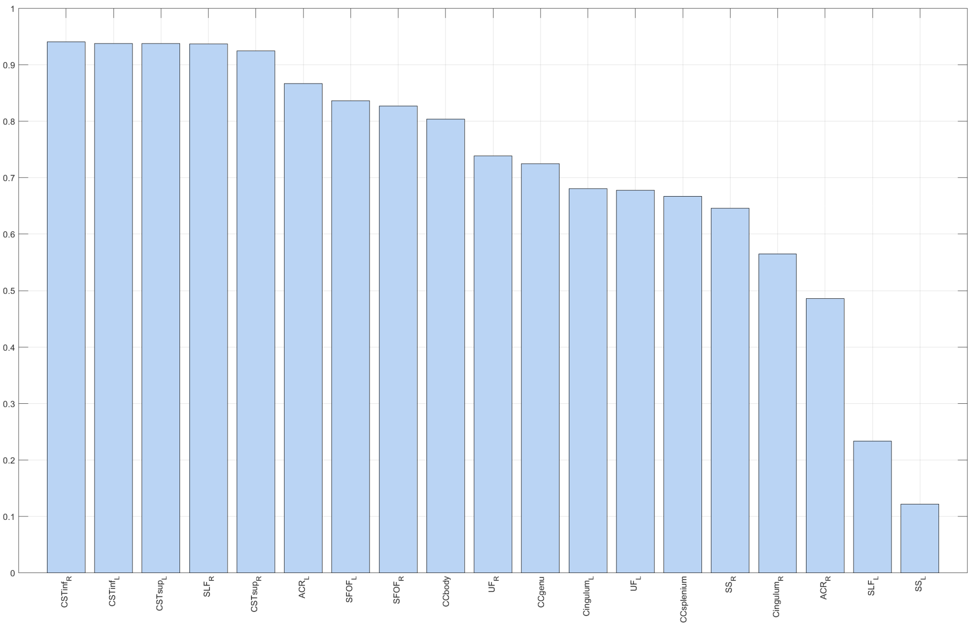


Figure 7.16: Estimates of the duration of each biomarker stage, given as a proportion of the entire length of the disease process. The duration of each stage is derived from the mixing proportion for the corresponding biomarker, assuming this to be equivalent to the fraction of patients in a "disease" state. The mixing proportions plotted here are those of the bimodal implementation, applied to the FA data of Sets FKN.

Length of stages

The mixing proportion can be used to give an estimate of the relative length of each biomarker stage, assuming that *mix* is proportional to the number of patients in a "disease" state for that stage. Plotting the mixing proportions from the Bimodal implementation as a bar graph (Figure 7.16) reveals that there is very little difference in the length of the CST biomarker stages. This suggests that they are involved throughout almost the entire course of the disease, and occur within a relatively short space of time. In contrast, under this model, the body of the corpus callosum does not become involved until approximately 20% through the disease course, and the genu and splenium are delayed further still.

Conclusion

This brief investigation of the relationship between the mixing proportion and the model output shows that the mixing proportion does indeed encode TI, but that the strength of the relationship appears to vary inconsistently when CDFs or PDFs are used to calculate probabilities. I conclude that the role of the mixing proportion in the EBM should be reassessed in depth in future work.

7.4.6 Data harmonisation

We have deliberately employed minimal data harmonisation in this study, although basic statistical tests were performed to ensure that there were no significant group differences between the three data sets. This was partly to assess whether the EBM could function when applied to combined data sets without extensive harmonisation, but also to account for the possibility that over-stringent harmonisation could remove interesting facets of the data by classifying them as noise. However, one of the largest multi-centre DTI ALS studies to date (Müller et al., 2016) found that their results increased in significance and had larger regions of significance after data harmonisation; we suggest that if the EBM can find results within minimally-harmonised data, then its functionality should be improved after extensive harmonisation.

Chapter 8

Future Work

8.1 Validation and extension of findings

8.1.1 Validation

Having shown that the model can function when applied to ALS data, the next step is to extend and confirm the findings using other, larger, data sets.

Fractional Anisotropy data

Validation of the hypothesis of white matter degeneration is the most urgent task. Clarification regarding the tracts that display some bimodality of staging (such as ACR, SS and UF) would also be desirable; to that end, it would probably be beneficial to update the selection of white matter ROIs with reference to the current literature. Finally, the inclusion of spinal cord dMRI would be an excellent opportunity to establish parity or otherwise with the directional degeneration of the CSTs. This last point is essential, as firmly establishing the directional degeneration would provide new insight into the biological mechanisms of ALS.

Structural data

The sample size for the structural data needs to be increased, in order to be able to generate robust hypotheses of cortical thinning and volumetric atrophy in ALS patients. A change of atlas may be beneficial, as the use of the Desikan-Killiany atlas is somewhat limited outside of studies that use the FreeSurfer pipelines. If an atlas can be found that uses brain divisions closer to those typically found in the literature for healthy aging, then it could be possible to validate the EBM progression order for

HCs against this literature. Ideally, such an atlas would be based on obvious physical landmarks, thus allowing for automatic segmentation.

Neuropsychological data

There is much work to be done regarding the neuropsychological data. The first step should be a revision of the current hypothesis, after acquiring considerably more data. Again, these data have been included only as proof of concept.

8.1.2 Directions for Extension

Clinical staging systems

The main priority must be to integrate my findings with the clinical ALS staging systems (King's, MiToS and Braak), as this would in principle allow a far richer level of detail in analysing the phenotypic composition of the various stages (in terms of motor, cognitive and neuroimaging measures). The combination of the subjective staging systems with the data-driven model should inform the former and assess the validity of the latter, as well as stimulating new hypotheses and developments that could be used to further refine the EBM and test its wider usefulness.

Integration with the King's and MiToS systems would require subjects' clinical stage, as well as death date. After assessing the clinical and EBM staging systems for concordance, the EBM stages could – in principle – be assigned to the clinical stages.

Integration with the Braak system appears more difficult, as the two systems are measuring very different metrics of disease progression (pTDP-43 pathology and neurodegeneration as characterised by changes in FA). However, an obvious starting point would be to replicate the 22 ROIs of Braak et al. as a DTI atlas. Difficulties with this approach include the small size of some regions, which can introduce errors into the automatic registration process, as well as the lowered sensitivity of FA to changes in the grey matter (particularly cortical) of ALS patients.

Different MRI modalities

There are other important modalities of MRI data that could be included in the EBM process. We have not yet made use of resting-state functional MRI (rsfMRI) data, which has been used to show increased functional connectivity in ALS patients (Douaud et al., 2011; Heimrath et al., 2014). These increases in functional connectivity correlate with cognitive deficits, and there is evidence that the connectivity changes in ALS patients resemble patterns of pTDP-43 pathology (Schulthess et al., 2016). The inclusion of rsfMRI data would therefore not only allow modelling of another aspect of the physical processes of ALS, but also provide a method of linking to the Braak staging system.

Longitudinal data would also be a valuable addition to any data set. The most obvious use is for validation of model consistency (i.e. ensuring that a patient's longitudinal scans are staged in the correct temporal order), but it may also be possible to adapt the EBM for repeated measures. This could be as simple as using the longitudinal data to inform us about the prior distribution, or as complex as constraining a subject's position during the modelling process.

Finally, it is well-established that although DTI is sensitive to tissue microstructure, it is an overly simplistic model of diffusion: DTI suffers from partial volume effects due to its relatively large voxel sizes, and its metrics (FA, MD) are lacking in specificity for individual tissue microstructure features (Pierpaoli et al., 1996). For example, a local reduction of mean FA in a single voxel can be caused by lower neurite density or by the crossing of fibre bundles. To combat the inherent non-specificity of DTI metrics, various higher-order models of diffusion (e.g. diffusional kurtosis imaging (Jensen et al., 2005), CHARMED (Assaf & Basser, 2005), NODDI (Zhang et al., 2012) and its successor Bingham-NODDI (Tariq et al., 2016)) continue to be proposed, each with different methods of attempting to directly measure tissue microstructure features. However, these models typically require either multiple b-shells or larger number of diffusion directions, which are more time-consuming and complex than standard DTI sequences; also, protocols vary between imaging centres, increasing the difficulties of data harmonisation. In spite of these drawbacks, the inclusion of advanced dMRI data in the EBM could provide greater insight into the low-level biological processes of ALS.

ALS phenotypes and subgroups

There are two immediately apparent methods by which the EBM could be used to investigate differences between ALS phenotypes or subgroups:

- **Hypothesis-driven approach**: Run the EBM on different cohorts selected by phenotype, and examine differences of progression between phenotypes. This requires large quantities of data.

- **Exploratory approach**: The reverse of the previous approach. Run the EBM on all available data, and look for hidden subgroups of subjects within the progression output; then assess clinical differences between the subgroups. This would allow data-driven classification of subjects, potentially revealing subgroups different from those assigned by the current (biased) clinical systems.

Possible subgroups and phenotypes include ALS/ALS-FTD, genetic classification (including familial and sporadic ALS), site of onset, rate of progression (assessed by ALSFRS-R scores or otherwise), and geographical origin (although this last comparison would require truly enormous quantities of data).

8.2 Model development

8.2.1 Mixture fitting

Given that the mixture fitting has such a large impact on the modelling results (see subsection 5.1.3), it is extremely important that the mixture fitting process is developed further. We have shown that the EM appears to be sensitive to outliers, and so other methods of mixture fitting should be explored. Bootstrapping the EM algorithm appears to be a promising option for reducing the effects of outliers; it would also be worthwhile to revisit Matlab's *fmincon*, due to its inherent versatility.

Support for mixture types other than Gaussians is greatly desirable, due to the high strength of the assumption of data normality. Matlab does not currently have native support for non-Gaussian EM mixture fitting, so this would require a from-scratch implementation. Further investigation into the impact of multiple mixture types on the EBM output is also required.

The issue of imposing constraints on the mixture fitting will doubtless reoccur. We must consider how many constraints we impose, and how strict those constraints should be. For example, it appears obvious that mixing proportion should be constrained to avoid wildly unrealistic fitting, but it is perhaps less clear how the directionality of a biomarker's progression should be assessed. How much weight should we place on the individual distributions fitted to ALS patients and HCs? Should we assume that the directionality given by these single distributions is accurate, or instead rely on the output of the mixture fitting? A balance must be found between avoiding unrealistic outcomes and imposing too much information.

Finally, for the reasons discussed in subsection 2.5.5, the use of CDFs or PDFs should be reassessed once larger data sets are available.

8.2.2 Disparate data modalities and collinearity

During this study, the accuracy of mixture distribution fitting has repeatedly been shown to be highly impactful on the output of the EBM. Thus, there is a need to study the effects of employing data that are parameterised by different underlying distributions. For example, if all data being considered are from two-component Gaussian mixtures, then it is plausible that inaccuracies in mixture fitting will be comparable between biomarkers; this may not hold true for data from disparate types of mixture distributions.

Given the impact of mixture fitting accuracy, it is not clear if we can meaningfully apply the EBM to data from different modalities. Further study is required, either through large numbers of simulations, or validation in a patient population.

The impact of data collinearity must also be considered. Collinearity between controls and patients has been partially addressed in this study by the use of the regularisation term in Matlab's *fitgmdist* when fitting the mixture distributions. Linear dependency can cause ill-conditioned covariance, leading to failure of mixture fitting; the use of the regularisation term can help ensure that the algorithm converges to a solution.

However, accounting for collinearity between biomarkers (both within and across modalities) is an area that requires further investigation. It is possible that the use of multivariate distributions could provide a solution; they could be employed by fitting distributions to both the left and right sides of individual WM tracts, or by fitting mixtures to multiple biomarkers at once. These higher-dimensional mixture models are more difficult to fit, requiring longer processing time, but could help account for biomarker covariance.

8.2.3 The Bayesian EBM

The current implementation of a Bayesian approach to event-based modelling described in section 2.6 has been included as proof of concept, and multiple aspects of this method require further development.

The Metropolis algorithm may not be the optimal choice when working with directional data, and so changes to the sampling methods may be necessary (see subsection 8.2.5). Even if the sampling methods are appropriate, the Bayesian EBM currently assumes that the posterior distribution is unimodal and can therefore be described by a single vMF distribution. We have previously shown the existence of

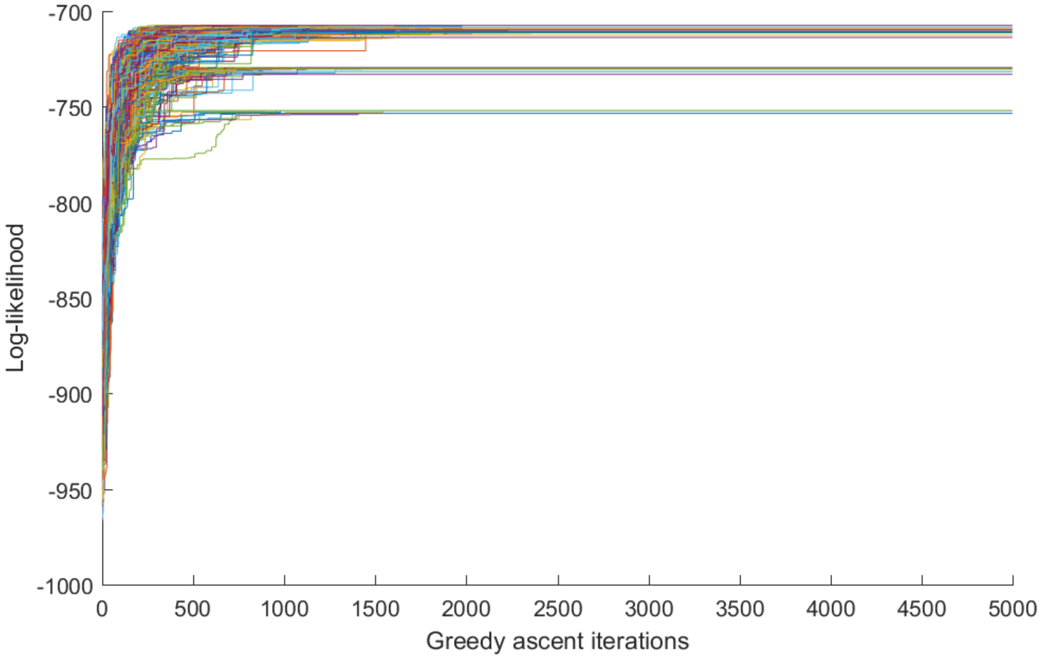


Figure 8.1: Mapping Ω with the greedy ascent algorithm: 1000 greedy runs were performed on the FA data of Sets FK. Plotting log-likelihood against greedy ascent iterations reveals that the posterior distribution appears to have three local maxima.

a bimodal posterior distribution within ALS FA data (Figure 7.10), and so a single vMF distribution may well be too simplistic. Alternatives include more complicated distributions and mixtures of distributions such as vMF (Banerjee et al., 2005), Watson (Bijral et al., 2007) or Bingham (Mardia & Jupp, 2000).

The greedy ascent algorithm (section 2.4) could potentially be used to map out Ω, revealing the number of modal peaks (Figure 8.1). This gives a possible mechanism for assessing the required number of components for a vMF mixture model, which could be adapted for mixtures of other distributions.

To take full advantage of the Bayesian EBM, we require a prior distribution. As there are currently no other comparable studies from which to generate a prior distribution, Bayesian best practice requires dividing the data being studied into a training set and a validation set. The posterior distribution (generated with flat priors) from the training set can then be used as the prior distribution for the validation set. This approach requires a larger sample size than is presently available, but should be employed whenever possible in future EBM studies.

8.2.4 Fourier Transforms

The current proof of concept implementation of the Bayesian EBM is made possible by a core idea: that permutations, a form of ranked data, can be transformed into directional data. However, this is not the only method by which a predictive model can be learnt from a data set of permutations: the generating probability distribution of a given sample of permutations can also be compactly represented by the "low-frequency" terms of a Fourier transform over the permutations (Huang et al., 2009; Irurozki et al., 2011).

A comparison between the directional data approach and the use of Fourier transforms is required to determine which one is best suited to the EBM. In general, further investigation is required into the subject of probability distributions over permutations.

8.2.5 MCMC methods

Autocorrelation is a general concern when working with MCMC methods, as a set of nearby MCMC samples must, by definition, be correlated with each other. Given this, it is possible that the Metropolis samples generated by the EBM are not appropriate for the maximum likelihood parameter estimation of a vMF distribution, as this requires independent and identically distributed samples (Sra, 2012). Investigation into more sophisticated alternatives to Metropolis sampling is therefore worthwhile; these could include Adaptive Rejection Metropolis Sampling (Gilks et al., 1995) and Hamiltonian Monte Carlo (Duane et al., 1987).

If we are more interested in the most likely event order than the actual posterior distribution, then "simulated annealing for global optimisation" is a sampling technique that could be appropriate (Andrieu et al., 2003):

> Let us assume that instead of wanting to approximate $p(x)$, we want to find its global maximum. For example, if $p(x)$ is the likelihood or posterior distribution, we often want to compute the ML and maximum a posteriori (MAP) estimates. As mentioned earlier, we could run a Markov chain of invariant distribution $p(x)$ and estimate the global mode... This method is inefficient because the random samples only rarely come from the vicinity of the mode. Unless the distribution has large probability mass around the mode, computing resources will be wasted exploring areas of no interest. A more principled strategy is to adopt *simulated annealing*.

8.3 General improvements

A priority must be the implementation of EBM support for covariates such as age, gender and handedness. These are currently not taken account of by the model, but are straightforward measures known to affect imaging parameters. These covariates can already be used to allocate subjects to cohorts, but this is only practical in large data sets, and is a work-around rather than a solution to the problem.

Another direction for model development is to improve the efficiency of the Variable Precision implementation, in order to reduce the time needed for processing. Although the processing time is not currently overly onerous, even when working in quadruple precision, this may change as the model is applied to larger data sets.

The NiSALS data harmonisation protocol (Müller et al., 2016) is due to be implemented in the EBM pipeline in the near future. Data harmonisation is generally necessary in ALS research, due to the small sizes of the data sets that are available. However, the effects of extensive data harmonisation require investigation: overly strict harmonisation can result in interesting facets of data being treated as noise, thereby producing results that are unnecessarily simplistic.

If possible, the creation of an EBM data harmonisation protocol would be beneficial. The purpose of this protocol would be to allow external research centres to harmonise their data and extract biomarker readings to be shared. This would greatly simplify collaboration and reduce time costs, as it would allow data preprocessing to be divided between multiple centres. However, this could potentially require the creation of "gold-standard reference datasets" for various biomarkers, particularly in the case of dMRI data; this would be a significant task to undertake.

Finally, I would like to publish a standalone version of the EBM. Ideally, this would be open source, to allow for collaborative development of the model. This might require converting the current code to a different language, due to dependencies on specific Matlab toolboxes which are not included as standard. Alternatively, Matlab will publish a standalone executable, albeit with obfuscated code.

Bibliography

Al-Chalabi, A., Hardiman, O., Kiernan, M. C., Chiò, A., Rix-Brooks, B. & van den Berg, L. H. Amyotrophic lateral sclerosis: moving towards a new classification system. *The Lancet Neurology* **15,** 1182–1194 (2016).

Alexander, A. L., Lee, J. E., Lazar, M. & Field, A. S. Diffusion tensor imaging of the brain. *Neurotherapeutics* **4,** 316–329 (2007).

Alsultan, A. A., Waller, R., Heath, P. R. & Kirby, J. The genetics of amyotrophic lateral sclerosis: current insights. *Degenerative Neurological and Neuromuscular Disease* **6,** 49–64 (2016).

Andrieu, C., De Freitas, N., Doucet, A. & Jordan, M. I. An introduction to MCMC for machine learning. *Machine learning* **50,** 5–43 (2003).

Arthur, K. C., Calvo, A., Price, T. R., Geiger, J. T., Chiò, A. & Traynor, B. J. Projected increase in amyotrophic lateral sclerosis from 2015 to 2040. *Nat Commun* **7,** 12408 (2016).

Assaf, Y. & Basser, P. J. Composite hindered and restricted model of diffusion (CHARMED) MR imaging of the human brain. *NeuroImage* **27,** 48–58 (2005).

Axler, S. *Linear Algebra Done Right* 2nd ed. (Springer-Verlag New York, Inc., 1997).

Balendra, R., Jones, A., Jivraj, N., Knights, C., Ellis, C. M., Burman, R., Turner, M. R., Leigh, P. N., Shaw, C. E. & Al-Chalabi, A. Estimating clinical stage of amyotrophic lateral sclerosis from the ALS Functional Rating Scale. *Amyotroph Lateral Scler Frontotemporal Degener* (2014).

Balendra, R., Jones, A., Jivraj, N., Steen, I. N., Young, C. A., Shaw, P. J., Turner, M. R., Leigh, P. N. & Al-Chalabi, A. Use of clinical staging in amyotrophic lateral sclerosis for phase 3 clinical trials. *J Neurol Neurosurg Psychiatry* **86,** 45–49 (2014).

Banerjee, A., Dhillon, I., Ghosh, J. & Sra, S. Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research* **6,** 1345–1382 (2005).

Basser, P. J. & Pierpaoli, C. Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor MRI. *J Magn Reson B* **111,** 209–219 (1996).

Basser, P. J., Mattiello, J. & LeBihan, D. MR diffusion tensor spectroscopy and imaging. *Biophys J* **66,** 259–67 (1994).

Baumer, D., Talbot, K. & Turner, M. R. Advances in motor neurone disease. *J R Soc Med* **107,** 14–21 (2014).

Beaulieu, C. & Allen, P. S. Water diffusion in the giant axon of the squid: Implications for diffusion-weighted MRI of the nervous system. *Magn Reson Med* **32,** 579–583 (1994).

Bijral, A., Breitenbach, M. & Grudic, G. Mixture of Watson distributions: A generative model for hyperspherical embeddings. *Journal of Machine Learning Research* **2,** 35–42 (2007).

Bozzali, M., Parker, G. J. M., Serra, L., Embleton, K., Gili, T., Perri, R., Caltagirone, C. & Cercignani, M. Anatomical connectivity mapping: A new tool to assess brain disconnection in Alzheimer's disease. *NeuroImage* **54,** 2045–2051 (2011).

Braak, H., Brettschneider, J., Ludolph, A. C., Lee, V. M., Trojanowski, J. Q. & Tredici, K. D. Amyotrophic lateral sclerosis—a model of corticofugal axonal spread. *Nat Rev Neurol* **9,** 708–714 (2013).

Broad, R., Gabel, M. C., Cercignani, M. & Leigh, P. N. Neurite orientation dispersion and density imaging (NODDI) demonstrates microstructural changes associated with Amyotrophic Lateral Sclerosis. *Amyotroph Lateral Scler Frontotemporal Degener* **16,** 97–114 (2015).

Handbook of Markov Chain Monte Carlo (eds Brooks, S., Gelman, A., Jones, G. L. & Meng, X.-L.) (Taylor & Francis, 2011).

Cedarbaum, J. M., Stambler, N., Malta, E., Fuller, C., Hilt, D., Thurmond, B. & Nakanishi, A. The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function. *J Neurol Sci* **169,** 13 –21 (1999).

Cercignani, M., Embleton, K., Parker, G. J. & Bozzali, M. Group-averaged anatomical connectivity mapping for improved human white matter pathway visualisation. *NMR Biomed* **25,** 1224–33 (2012).

Charcot, J. & Joffroy, A. Deux cas d'atrophie musculaire progressive. *Arch Physiol Neurol Pathol* **2,** 744 (1869).

Chiò, A. & Traynor, B. J. Motor neuron disease in 2014: Biomarkers for ALS-in search of the Promised Land. *Nat Rev Neurol* **11,** 72–4 (2015).

Chiò, A., Hammond, E. R., Mora, G., Bonito, V. & Filippini, G. Development and evaluation of a clinical staging system for amyotrophic lateral sclerosis. *J Neurol Neurosurg Psychiatry* **86,** 38–44 (2013).

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S. & Killiany, R. J. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* **31,** 968–980 (2006).

Douaud, G., Filippini, N., Knight, S., Talbot, K. & Turner, M. R. Integration of structural and functional magnetic resonance imaging in amyotrophic lateral sclerosis. *Brain* **134,** 3470–3479 (2011).

Duane, S., Kennedy, A., Pendleton, B. J. & Roweth, D. Hybrid Monte Carlo. *Phys Lett B* **195,** 216–222 (1987).

Fang, T., Khleifat, A. A., Stahl, D. R., Torre, C. L. L., Murphy, C., Young, C., Shaw, P. J., Leigh, P. N. & Al-Chalabi, A. Comparison of the King's and MiToS staging systems for ALS. *Amyotroph Lateral Scler Frontotemporal Degener* **5,** 1–6 (2017).

Ferraro, D., Consonni, D., Fini, N., Fasano, A., Giovane, C. D. & Mandrioli, J. Amyotrophic lateral sclerosis: a comparison of two staging systems in a population-based study. *Eur J Neurol* **23,** 1426–1432 (2016).

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B. & Dale, A. M. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron* **33,** 341–55 (2002).

Foerster, B. R., Dwamena, B. A., Petrou, M., Carlos, R. C., Callaghan, B. C. & Pomper, M. G. Diagnostic Accuracy Using Diffusion Tensor Imaging in the Diagnosis of ALS: A Meta-analysis. *Acad Radiol* **19,** 1075–1086 (2012).

Fonteijn, H. M., Clarkson, M. J., Modat, M., Barnes, J., Lehmann, M., Ourselin, S., Fox, N. C. & Alexander, D. C. An event-based disease progression model and its application to familial Alzheimer's disease. *Inf. Process Med. Imaging* **22,** 748–59 (2011).

Fonteijn, H. M., Modat, M., Clarkson, M. J., Barnes, J., Lehmann, M., Hobbs, N. Z., Scahill, R. I., Tabrizi, S. J., Ourselin, S., Fox, N. C. & Alexander, D. C. An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. *NeuroImage* **60,** 1880–1889 (2012).

Fowler, T., Scadding, J. W. & Marsden, C. D. *Clinical neurology* 3rd ed. (Arnold, London, 2003).

Franchignoni, F., Mandrioli, J., Giordano, A. & and, S. F. A further Rasch study confirms that ALSFRS-R does not conform to fundamental measurement requirements. *Amyotroph Lateral Scler Frontotemporal Degener* **16,** 331–337 (2015).

Franchignoni, F., Mora, G., Giordano, A., Volanti, P. & Chiò, A. Evidence of multidimensionality in the ALSFRS-R Scale: a critical appraisal on its measurement properties using Rasch analysis. *J Neurol Neurosurg Psychiatry* **84,** 1340–1345 (2013).

Frigge, M., Hoaglin, D. C. & Iglewicz, B. Some Implementations of the Boxplot. *The American Statistician* **43,** 50–54 (Feb. 1989).

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B. & Vehtari, A. *Bayesian Data Analysis* (Taylor & Francis Ltd, 2013).

Geyer, C. J. Practical Markov Chain Monte Carlo. *Statistical Science* **7,** 473–483 (1992).

Gilks, W. R., Best, N. G. & Tan, K. K. C. Adaptive Rejection Metropolis Sampling within Gibbs Sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **44,** 455–472 (1995).

Gordon, P. Amyotrophic Lateral Sclerosis: An update for 2013 Clinical Features, Pathophysiology, Management and Therapeutic Trials. *Aging and Disease* **4,** 295–310 (2013).

Gordon, P. H., Miller, R. G. & Moore, D. H. ALSFRS-R. *Amyotroph Lateral Scler & Other Motor Neuron Disord* **5,** 90–93 (2004).

Grolez, G., Moreau, C., Danel-Brunaud, V., Delmaire, C., Lopes, R., Pradat, P. F., Mendili, M. M. E., Defebvre, L. & Devos, D. The value of magnetic resonance imaging as a biomarker for amyotrophic lateral sclerosis: a systematic review. *BMC Neurology* **16** (2016).

Heimrath, J., Gorges, M., Kassubek, J., Müller, H.-P., Birbaumer, N., Ludolph, A. C. & Lulé, D. Additional resources and the default mode network: Evidence of increased connectivity and decreased white matter integrity in amyotrophic lateral sclerosis. *Amyotroph Lateral Scler Frontotemporal Degener* **15,** 537–545 (2014).

Hoff, P. D. *A First Course in Bayesian Statistical Methods* (Springer, 2010).

Holoborodko, P. *Advanpix Multiprecision Computing Toolbox* version 4.3.2.12144. Jan. 4, 2017.

Huang, J., Guestrin, C. & Guibas, L. *Efficient inference for distributions on permutations* in *Advances in neural information processing systems* (2007), 697–704.

Huang, J., Guestrin, C. & Guibas, L. Fourier theoretic probabilistic inference over permutations. *Journal of machine learning research* **10,** 997–1070 (2009).

Irurozki, E., Calvo, B. & Lozano, J. A. in *Advances in Artificial Intelligence: 24th Canadian Conference on Artificial Intelligence, Canadian AI 2011, St. John's, Canada, May 25-27, 2011. Proceedings* (eds Butz, C. & Lingras, P.) 186–191 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2011).

Jansons, K. M. & Alexander, D. C. Persistent Angular Structure: new insights from diffusion MRI data. *Inf Process Med Imaging* **18,** 672–83 (2003).

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W. & Smith, S. M. FSL. *NeuroImage* **62,** 782–790 (2012).

Jensen, J. H., Helpern, J. A., Ramani, A., Lu, H. & Kaczynski, K. Diffusional kurtosis imaging: The quantification of non-gaussian water diffusion by means of magnetic resonance imaging. *Magn Reson Med* **53,** 1432–1440 (2005).

Jones, D. K., Williams, S. C., Gasston, D., Horsfield, M. A., Simmons, A. & Howard, R. Isotropic resolution diffusion tensor imaging with whole brain acquisition in a clinically acceptable time. *Hum Brain Mapp* **15,** 216–30 (2002).

Jones, D. K. & Cercignani, M. Twenty-five pitfalls in the analysis of diffusion MRI data. *NMR Biomed* **23,** 803–820 (2010).

Kassubek, J., Muller, H.-P., Tredici, K. D., Brettschneider, J., Pinkhardt, E. H., Lule, D., Bohm, S., Braak, H. & Ludolph, A. C. Diffusion tensor imaging analysis of sequential spreading of disease in amyotrophic lateral sclerosis confirms patterns of TDP-43 pathology. *Brain* **137,** 1733–1740 (2014).

Kendall, M. G. A New Measure of Rank Correlation. *Biometrika* **30,** 81–93 (1938).

Kendall, M. G. *Rank Correlation Methods* (Griffin & Co, London, 1970).

Labra, J., Menon, P., Byth, K., Morrison, S. & Vucic, S. Rate of disease progression: a prognostic biomarker in ALS. *J Neurol Neurosurg Psychiatry* (2015).

Laferriere, F. & Polymenidou, M. Advances and challenges in understanding the multifaceted pathogenesis of amyotrophic lateral sclerosis. *Swiss Med Wkly* **145** (2015).

Lattante, S., Ciura, S., Rouleau, G. A. & Kabashi, E. Defining the genetic connection linking amyotrophic lateral sclerosis (ALS) with frontotemporal dementia (FTD). *Trends Genet* **31,** 263–273 (2015).

Mandrioli, J., Biguzzi, S., Guidi, C., Sette, E., Terlizzi, E., Ravasio, A., Casmiro, M., Salvi, F., Liguori, R., Rizzi, R., Pietrini, V., Borghi, A., Rinaldi, R., Fini, N., Chierici, E., Santangelo, M., Granieri, E., Mussuto, V., De Pasqua, S., Georgoulopoulou, E., Fasano, A., Ferro, S. & D'Alessandro, R. Heterogeneity in ALSFRS-R decline and survival: a population-based study in Italy. *Neurol Sci* **36,** 2243–2252 (2015).

Mardia, K. V. & Jupp, P. E. *Directional statistics* (Wiley, 2000).

McLachlan, G. J. & Peel, D. *Finite mixture models* (Wiley, 2000).

Menke, R. A. L., Agosta, F., Grosskreutz, J., Filippi, M. & Turner, M. R. Neuroimaging Endpoints in Amyotrophic Lateral Sclerosis. *Neurotherapeutics* **14,** 11–23 (2016).

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics* **21,** 1087–1092 (1953).

Mori, S., Wakana, S., van Zijl, P. C. M. & Nagae-Poetscher, L. M. *MRI Atlas of Human White Matter* (Elsevier Science, 2005).

Morrison, K. E., Dhariwal, S., Hornabrook, R., Savage, L., Burn, D. J., Khoo, T. K., Kelly, J., Murphy, C. L., Al-Chalabi, A., Dougherty, A., Leigh, P. N., Wijesekera, L., Thornhill, M., Ellis, C. M., O'Hanlon, K., Panicker, J., Pate, L., Ray, P., Wyatt, L., Young, C. A., Copeland, L., Ealing, J., Hamdalla, H., Leroi, I., Murphy, C., O'Keeffe, F., Oughton, E., Partington, L., Paterson, P., Rog, D., Sathish, A., Sexton, D., Smith, J., Vanek, H., Dodds, S., Williams, T. L., Steen, I. N., Clarke, J., Eziefula, C., Howard, R., Orrell, R., Sidle, K., Sylvester, R., Barrett, W., Merritt, C., Talbot, K., Turner, M. R., Whatley, C., Williams, C., Williams, J., Cosby, C., Hanemann, C. O., Iman, I., Philips, C., Timings, L., Crawford, S. E., Hewamadduma, C., Hibberd, R., Hollinger, H., McDermott, C., Mils, G., Rafiq, M., Shaw, P. J., Taylor, A., Waines, E., Walsh, T., Addison-Jones, R., Birt, J., Hare, M. & Majid, T. Lithium in patients with amyotrophic lateral sclerosis (LiCALS): a phase 3 multicentre, randomised, double-blind, placebo-controlled trial. *The Lancet Neurology* **12,** 339–345 (2013).

Müller, H.-P., Turner, M. R., Grosskreutz, J., Abrahams, S., Bede, P., Govind, V., Prudlo, J., Ludolph, A. C., Filippi, M. & Kassubek, J. A large-scale multicentre cerebral diffusion tensor imaging study in amyotrophic lateral sclerosis. *J Neurol Neurosurg Psychiatry* **87,** 570–579 (2016).

Pierpaoli, C, Jezzard, P, Basser, P. J., Barnett, A & Di Chiro, G. Diffusion tensor MR imaging of the human brain. *Radiology* **201,** 637–648 (3 Dec. 1996).

Plis, S. M., Lane, T. & Calhoun, V. D. *Permutations as angular data: efficient inference in factorial spaces* in *2010 IEEE International Conference on Data Mining* (2010), 403–410.

Roche, J. C., Rojas-Garcia, R., Scott, K. M., Scotton, W., Ellis, C. E., Burman, R., Wijesekera, L., Turner, M. R., Leigh, P. N., Shaw, C. E. & Al-Chalabi, A. A proposed staging system for amyotrophic lateral sclerosis. *Brain* **135,** 847–852 (2012).

Rokem, A., Takemura, H., Bock, A. S., Scherf, K. S., Behrmann, M., Wandell, B. A., Fine, I., Bridge, H. & Pestilli, F. The visual white matter: The application of diffusion MRI and fiber tractography to vision science. *J Vis* **17,** 4 (2017).

Rooney, J., Burke, T., Vajda, A., Heverin, M. & Hardiman, O. What does the ALSFRS-R really measure? A longitudinal and survival analysis of functional dimension subscores in amyotrophic lateral sclerosis. *J Neurol Neurosurg Psychiatry* (2016).

Schulthess, I., Gorges, M., Müller, H.-P., Lulé, D., Tredici, K. D., Ludolph, A. C. & Kassubek, J. Functional connectivity changes resemble patterns of pTDP-43 pathology in amyotrophic lateral sclerosis. *Sci Rep* **6,** 38391 (2016).

Sra, S. A short note on parameter approximation for von Mises-Fisher distributions: and a fast implementation of $I_s(x)$. *Comput Stat* **27,** 177–190 (2012).

Tariq, M., Schneider, T., Alexander, D. C., Wheeler-Kingshott, C. A. G. & Zhang, H. Bingham-NODDI: Mapping anisotropic orientation dispersion of neurites using diffusion MRI. *NeuroImage* **133,** 207–223 (2016).

Thompson, G. L. Generalized Permutation Polytopes and Exploratory Graphical Methods for Ranked Data. *Ann Statist* **21,** 1401–1430 (1993).

Tsermentseli, S., Leigh, P. N., Taylor, L. J., Radunovic, A., Catani, M. & Goldstein, L. H. Syntactic processing as a marker for cognitive impairment in amyotrophic lateral sclerosis. *Amyotroph Lateral Scler Frontotemporal Degener* **17,** 69–76 (2015).

Turner, M. R. & Verstraete, E. What Does Imaging Reveal About the Pathology of Amyotrophic Lateral Sclerosis? *Curr Neurol Neurosci Rep* **15** (2015).

Turner, M. R., Agosta, F., Bede, P., Govind, V., Lulé, D. & Verstraete, E. Neuroimaging in amyotrophic lateral sclerosis. *Biomark Med* **6,** 319–337 (2012).

Wheeler-Kingshott, C. A., Barker, G. J., Steens, S. C. & van Buchem, M. A. *D: the Diffusion of Water* in *Quantitative MRI of the Brain: Measuring Changes Caused by Disease* (ed Tofts, P.) (John Wiley & Sons, 2003), 203–256.

Wijesekera, L. C. & Leigh, P. N. Amyotrophic lateral sclerosis. *Orphanet Journal of Rare Diseases* **4,** 3 (2009).

Young, A. L., Oxtoby, N. P., Daga, P., Cash, D. M., Fox, N. C., Ourselin, S., Schott, J. M. & Alexander, D. C. A data-driven model of biomarker changes in sporadic Alzheimer's disease. *Brain* **137,** 2564–77 (2014).

Zhang, H., Schneider, T., Wheeler-Kingshott, C. A. & Alexander, D. C. NODDI: Practical in vivo neurite orientation dispersion and density imaging of the human brain. *NeuroImage* **61,** 1000–1016 (2012).

Appendix A

Matlab code

A.1 Fitting mixtures by EM

```
function [ EM_con,phat ] = fit_EM_con( labels,CTL,ALS,varargin )
% Parse inputs
p = inputParser;
addOptional(p,'mix_thresh',0.15);
% Fitting will satisfy mix_thresh < ComponentProportion < 1 - mix_thresh
addOptional(p,'outlier_thresh',7);
% Number of SDs to use to determine outliers. Set to 2 for the usual 95.45%
% CI, or 3 for 99.73% CI. Default is 7 SDs; this should give no outliers.
addOptional(p,'plot_figure',1);
addOptional(p,'verbose',1);
p.parse( varargin{:} )

oCTL = CTL;
oALS = ALS;
threshold = p.Results.mix_thresh;
outlier_thresh = p.Results.outlier_thresh;
plot_figure = p.Results.plot_figure;
verbose = p.Results.verbose;

for j = 1:size(labels,2)
    biomarker = matlab.lang.makeValidName(char(labels(j)));

    nCTL = CTL(~isnan(CTL(:,j)),j); % Remove NaNs
    nALS = ALS(~isnan(ALS(:,j)),j);

    reg_value = min(var(nCTL),var(nALS)); % Set regularisation value
    reg_thresh = 0.01;
```

```
[mu_c, sig_c] = normfit(nCTL); % For assigning biomarker directionality

% Remove outliers
oCTL( (CTL(:,j) < (mu_c - outlier_thresh*sig_c)),j ) = NaN;
oCTL( (CTL(:,j) > (mu_c + outlier_thresh*sig_c)),j ) = NaN;

% Print the biomarkers for which CTL data are not Gaussian
if lillietest(oCTL(:,j)) == 1 && verbose == 1
    fprintf('%s CTL not Gaussian\n',char(labels(j)))
end

[mu_a, sig_a] = normfit(nALS);

oALS( (ALS(:,j) < (mu_a - outlier_thresh*sig_a)),j ) = NaN;
oALS( (ALS(:,j) > (mu_a + outlier_thresh*sig_a)),j ) = NaN;

if lillietest(oALS(:,j)) == 1 && verbose == 1
    fprintf('%s ALS not Gaussian\n',char(labels(j)))
end

% Create status vector for EM algorithm
status = [ones(sum(~isnan(oCTL(:,j))),1); ...
    2*ones(sum(~isnan(oALS(:,j))),1)]';
% Fit Gaussian mixtures
EM_con.(biomarker) = fitgmdist([oCTL(~isnan(oCTL(:,j)),j); ...
    oALS(~isnan(oALS(:,j)),j)],2,...
    'Regularize',reg_value*reg_thresh,'Start',status,...
    'Options',statset('MaxIter',5000));

% While mixing proportion is below threshold, increase regularisation
% value by 1% and refit
while EM_con.(biomarker).ComponentProportion(1) < threshold ...
        || EM_con.(biomarker).ComponentProportion(1) > (1 - threshold)
    % reg_value = min(var(CTL(:,j)),var(ALS(:,j)));
    reg_thresh = reg_thresh + 0.01;
    EM_con.(biomarker) = fitgmdist([oCTL(~isnan(oCTL(:,j)),j); ...
        oALS(~isnan(oALS(:,j)),j)],2,...
        'Regularize',reg_value*reg_thresh,'Start',status,...
        'Options',statset('MaxIter',5000));
end
```

```
end

% Extract mu_c, sig_c, mu_a, sig_a and mixing proportion
phat = NaN(size(labels,2),5); % Initialise phat
for j = 1:size(labels,2)
    biomarker = matlab.lang.makeValidName(char(labels(j)));
    phat(j,1) = EM_con.(biomarker).mu(1);
    phat(j,3) = EM_con.(biomarker).mu(2);

    phat(j,2) = sqrt(EM_con.(biomarker).Sigma(1));

    if EM_con.(biomarker).SharedCovariance == 0
        phat(j,4) = sqrt(EM_con.(biomarker).Sigma(2));
    else
        phat(j,4) = phat(j,2); % as shared covariance
    end

    phat(j,5) = EM_con.(biomarker).ComponentProportion(1);
end

% Plot histograms and fitted mixtures if requested
if plot_figure == 1
    plot_fitted_mixtures([CTL;ALS],EM_con,labels)
end

end
```

A.2 Plotting fitted mixtures

```
function [] = plot_fitted_mixtures(data,fitted_mixtures,labels)
% Plot histograms of data with the fitted mixture model parameters overlaid
figure;

for i = 1:size(data,2)
    % Initialise subplot
    subplot(ceil(sqrt(size(data,2))),round(sqrt(size(data,2))),i)
    hold on;

    % Select histogram bin method
    if (sum(rem(data(~isnan(data(:,i)),i),1)) == 0) ...
            && (range(data(:,i)) < 11) % If integer data and range is small
```

```
        histogram(data(:,i),'BinMethod','integer','Normalization','pdf')
    else
        histogram(data(:,i),'BinMethod','sqrt','Normalization','pdf')
    end

    % Make valid names from list of biomarkers
    biomarker = matlab.lang.makeValidName(char(labels(i)));

    % Create data points for x-axes
    ix = min(data(:,i)):1e-2*range(data(:,i)):max(data(:,i));

    % If Guassian mixtures exist as Matlab objects...
    if isstruct(fitted_mixtures)

        % ...then extract the mixture parameters directly
        mu1 = fitted_mixtures.(biomarker).mu(1);
        sd1 = sqrt(fitted_mixtures.(biomarker).Sigma(1,1));
        % as gmdistributions store variance rather than SD
        mix1 = fitted_mixtures.(biomarker).ComponentProportion(1);
        iy1 = normpdf(ix,mu1,sd1);

        mu2 = fitted_mixtures.(biomarker).mu(2);
        if fitted_mixtures.(biomarker).SharedCovariance == 1
            sd2 = sd1;
        else sd2 = sqrt(fitted_mixtures.(biomarker).Sigma(1,2));
        end
        mix2 = fitted_mixtures.(biomarker).ComponentProportion(2);
        iy2 = normpdf(ix,mu2,sd2);

    else % Otherwise, read the parameters from the given matrix
        mu1 = fitted_mixtures(i,1);
        sd1 = fitted_mixtures(i,2);
        % already converted to SD
        mix1 = fitted_mixtures(i,5);
        iy1 = normpdf(ix,mu1,sd1);

        mu2 = fitted_mixtures(i,3);
        sd2 = fitted_mixtures(i,4);
        mix2 = 1 - fitted_mixtures(i,5);
        iy2 = normpdf(ix,mu2,sd2);
    end

    % Plot CTL Gaussian component
    plot(ix,iy1*mix1,'g');
    % Plot ALS Gaussian component
    plot(ix,iy2*mix2,'r');

    try % if mixtures are Matlab objects, plot using exact pdf
        plot(ix,pdf(fitted_mixtures.(biomarker),ix'),'b');
    catch % otherwise, plot approximation
        approx_cov(:,:,1) = sd1^2;
        approx_cov(:,:,2) = sd2^2;
        approx_dist = gmdistribution([mu1; mu2],approx_cov,[mix1; mix2]);
        plot(ix,pdf(approx_dist,ix'),'k');
    end

    % Label each histogram with biomarker name
    xlabel(char(labels(i)))
    xlim([min(data(:,i)) max(data(:,i))])
end

end
```

A.3 Calculating probabilities

A.3.1 Cumulative Distribution Functions

```
function [ log_xE,log_xnotE ] = calc_cdf_probs( phat,data,CTL,ALS )
% calc_cdf_probs returns log of p(x|E) and p(x|~E) for all patients on all
% events, with probabilities calculated by CDF

% If mu_als < mu_ctl, then:
% p(x|E) = 1 - normcdf(x,mu_als,sig_als)
% p(x|~E) = normcdf(x,mu_ctl,sig_ctl)

% If mu_als > mu_ctl, then:
% p(x|E) = normcdf(x,mu_als,sig_als)
% p(x|~E) = 1 - normcdf(x,mu_ctl,sig_ctl)

% % Calculate p(x|E) and p(x|~E) for speed

% % p(x|E)
```

```
x_given_E = NaN(size(data,1),size(data,2)); % Initialise x_given_E
for j = 1:size(data,1) % number of patients
    for n = 1:size(data,2) % number of events
        if  nanmean(ALS(:,n)) < nanmean(CTL(:,n)) % mu_als < mu_ctl
            x_given_E(j,n) = 1 - normcdf(data(j,n),phat(n,3),phat(n,4));
        elseif nanmean(ALS(:,n)) > nanmean(CTL(:,n)) % mu_als > mu_ctl
            x_given_E(j,n) = normcdf(data(j,n),phat(n,3),phat(n,4));
        end
    end
end

% % p(x|~E)
x_given_notE = NaN(size(data,1),size(data,2)); % Initialise x_given_notE
for j = 1:size(data,1) % number of patients
    for n = 1:size(data,2) % number of events
        if nanmean(ALS(:,n)) < nanmean(CTL(:,n)) % mu_als < mu_ctl
            x_given_notE(j,n) = normcdf(data(j,n),phat(n,1),phat(n,2));
        elseif nanmean(ALS(:,n)) > nanmean(CTL(:,n)) % mu_als > mu_ctl
            x_given_notE(j,n) = 1 - normcdf(data(j,n),phat(n,1),phat(n,2));
        end
    end
end

% % log(p(x|E)) and log(p(x|~E))
log_xE = log(x_given_E);
log_xnotE = log(x_given_notE);

% % % % Replace -Inf log probs
log_xE(log_xE == -Inf) = log(1e-251);
log_xnotE(log_xnotE == -Inf) = log(1e-251);

% % Replace missing data with 0.5 prob
log_xE(isnan(log_xE)) = log(0.5);
log_xnotE(isnan(log_xnotE)) = log(0.5);

end
```

A.3.2 Probability Density Functions

```
function [ log_xE,log_xnotE ] = calc_pdf_probs( phat,data )
% calc_pdf_probs returns log of p(x|E) and p(x|~E) for all patients on all
% events, with probabilities calculated by PDF

% p(x|E) = normpdf(x,mu_als,sig_als)
% p(x|~E) = normpdf(x,mu_ctl,sig_ctl)

% % p(x|E)
x_given_E = NaN(size(data,1),size(data,2)); % Initialise x_given_E
for j = 1:size(data,1) % number of patients
    for n = 1:size(data,2) % number of events
        x_given_E(j,n) = normpdf(data(j,n),phat(n,3),phat(n,4));
    end
end

% % p(x|~E)
x_given_notE = NaN(size(data,1),size(data,2)); % Initialise x_given_notE
for j = 1:size(data,1) % number of patients
    for n = 1:size(data,2) % number of events
        x_given_notE(j,n) = normpdf(data(j,n),phat(n,1),phat(n,2));
    end
end

% % log(p(x|E)) and log(p(x|~E))
log_xE = log(x_given_E);
log_xnotE = log(x_given_notE);

% % % % Replace -Inf log probs
log_xE(log_xE == -Inf) = log(1e-251);
log_xnotE(log_xnotE == -Inf) = log(1e-251);

% % Replace missing data with 0.5 prob
log_xE(isnan(log_xE)) = log(0.5);
log_xnotE(isnan(log_xnotE)) = log(0.5);

end
```

A.4 Calculating ln(p(X|S))

A.4.1 Double-precision implementation

```
function [ LSE ] = LSE( current_S,log_x_given_E,log_x_given_notE )
% calculate_p_X_given_S returns ln(p(X|S)) for a given order S.
N = size(log_x_given_E,2);
J = size(log_x_given_E,1);

% % p(Xj|S,k) (Equation 2.1) % % % % % % % % % % % % % % % % % % % % % % %

log_Xj_given_S_k = NaN(J,N+1); % Initialise Xj_given_S_k

% (p(x|~E) for all events i.e. no events have occurred
log_Xj_given_S_k(:,1) = sum(log_x_given_notE,2);
% Want ((p(x|E) for S(1:k) events) * (p(x|~E) for S(k+1:N) events))
for k = 1:N-1
    log_Xj_given_S_k(:,k+1) = sum(log_x_given_E(:,current_S(1:k)),2) + ...
        sum(log_x_given_notE(:,current_S(k+1:N)),2);
end
% and (p(x|E) for all events) i.e. all events have occurred
log_Xj_given_S_k(:,N+1) = sum(log_x_given_E,2);

% % p(Xj|S) (Equation 2.2) % % % % % % % % % % % % % % % % % % % % % % % %
Xj_given_S = sum(exp(log_Xj_given_S_k),2);

% % p(X|S) (Equation 2.3)% % % % % % % % % % % % % % % % % % % % % % % % %
LSE = sum(log((1/N)*Xj_given_S));

end
```

A.4.2 Variable-precision implementation, Advanpix Toolbox

```
function [ mp_LSE ] = mp_LSE( current_S,log_x_given_E,log_x_given_notE )
% calculate_p_X_given_S returns ln(p(X|S)) for a given order S.
N = size(log_x_given_E,2);
J = size(log_x_given_E,1);

% % p(Xj|S,k) (Equation 2.1) % % % % % % % % % % % % % % % % % % % % % % %

log_Xj_given_S_k = NaN(J,N+1); % Initialise Xj_given_S_k
```

```
% (p(x|~E) for all events i.e. no events have occurred
log_Xj_given_S_k(:,1) = sum(log_x_given_notE,2);
% Want ((p(x|E) for S(1:k) events) * (p(x|~E) for S(k+1:N) events))
for k = 1:N-1
    log_Xj_given_S_k(:,k+1) = sum(log_x_given_E(:,current_S(1:k)),2) + ...
        sum(log_x_given_notE(:,current_S(k+1:N)),2);
end
% and (p(x|E) for all events) i.e. all events have occurred
log_Xj_given_S_k(:,N+1) = sum(log_x_given_E,2);

% % p(Xj|S) (Equation 2.2) % % % % % % % % % % % % % % % % % % % % % % % %
Xj_given_S = sum(exp(mp(log_Xj_given_S_k)),2);

% % p(X|S) (Equation 2.3)% % % % % % % % % % % % % % % % % % % % % % % % %
mp_LSE = sum(log((1/N)*Xj_given_S));

end
```

A.4.3 Variable-precision implementation, native Matlab

```
function [ vpa_LSE ] = vpa_LSE( current_S,log_x_given_E, ...
    log_x_given_notE,varargin )
% calculate_p_X_given_S returns ln(p(X|S)) for a given order S.

% Parse inputs
p = inputParser;
addOptional(p,'precision',NaN);

p.parse( varargin{:} )
precision = p.Results.precision;

% Set precision of vpa if requested
if isnan(precision) == 0
    digits(precision);
end

N = size(log_x_given_E,2);
J = size(log_x_given_E,1);

% % p(Xj|S,k) (Equation 2.1) % % % % % % % % % % % % % % % % % % % % % % %

log_Xj_given_S_k = NaN(J,N+1); % Initialise Xj_given_S_k
```

```
% (p(x|~E) for all events i.e. no events have occurred
log_Xj_given_S_k(:,1) = sum(log_x_given_notE,2);
% Want ((p(x|E) for S(1:k) events) * (p(x|~E) for S(k+1:N) events))
for k = 1:N-1
    log_Xj_given_S_k(:,k+1) = sum(log_x_given_E(:,current_S(1:k)),2) + ...
        sum(log_x_given_notE(:,current_S(k+1:N)),2);
end
% and (p(x|E) for all events) i.e. all events have occurred
log_Xj_given_S_k(:,N+1) = sum(log_x_given_E,2);

% % p(Xj|S) (Equation 2.2) % % % % % % % % % % % % % % % % % % % % % % % %
Xj_given_S = sum(exp(vpa(log_Xj_given_S_k)),2);

% % p(X|S) (Equation 2.3)% % % % % % % % % % % % % % % % % % % % % % % % %
vpa_LSE = sum(log((1/N)*Xj_given_S));

end
```

A.5 Comparing multiple greedy ascent runs

```
function [ greedy_S_ML ] = compare_greedy( log_xE,log_xnotE,varargin )
% Parse inputs
p = inputParser;
addOptional(p,'iterations',5000);
addOptional(p,'startpoints',100);
addOptional(p,'plot_figure',1);
p.parse( varargin{:} )

iterations = p.Results.iterations;
startpoints = p.Results.startpoints;
plot_figure = p.Results.plot_figure;

greedy_S_ML_temp = NaN(startpoints,size(log_xE,2)+1);
greedy_plot_temp = NaN(startpoints,iterations)';

greedy_legend = sprintf('%d greedy ascent runs',startpoints);
% Call greedy_ascent function to perform the requested number of runs
parfor i = 1:startpoints
    S_initial = randperm(size(log_xE,2));
    [ ~, S ] = greedy_ascent( S_initial,iterations,log_xE,log_xnotE );
    greedy_S_ML_temp(i,:) = S(end,:);
    greedy_plot_temp(:,i) = S(2:end,end);

end

% Plot log-likelihoods of greedy iterations for each run
if plot_figure == 1
    figure;
    hold on

    for i = 1:startpoints
        plot(1:iterations,greedy_plot_temp(:,i))
        title(char(greedy_legend));
    end
end

% Find the most likely greedy ascent output
[~, idx] = max(greedy_S_ML_temp(:,end));
greedy_S_ML = greedy_S_ML_temp(idx,1:end);

end

% % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % %
function [ greedy_S_ML, S ] = greedy_ascent( S_initial,iterations, ...
    log_x_given_E,log_x_given_notE )

N = size(log_x_given_E,2);

S = NaN(iterations+1,N+1);

S(1,1:N) = S_initial;
S(1,N+1) = LSE(S_initial,log_x_given_E,log_x_given_notE);

for k = 1:iterations

    swap = randi(N,2,1);
    %  To ensure that we are not swapping an event with itself:
    while swap(1) == swap(2)
        swap = randi(N,2,1);
    end
    % Generate S'
    S(k+1,1:N) = S(k,1:N); % Copy S_t
    S(k+1,swap(1)) = S(k,swap(2)); % Swap random event 1
    S(k+1,swap(2)) = S(k,swap(1)); % Swap random event 2

    % Calculate p(X|S')
    S(k+1,N+1) = LSE(S(k+1,1:N),log_x_given_E,log_x_given_notE);

    % If p(X|S') < p(X|S_t), then set S_t+1 = S_t
    if S(k+1,N+1) < S(k,N+1)
        S(k+1,1:N+1) = S(k,1:N+1);
    end
    % otherwise, leave S_t+1 = S'
end

% Find the order with the maximum likelihood:
[row, ~] = find(S == max(S(:,end)));
greedy_S_ML = S(max(row),1:N+1);

end
```

A.6 Performing MCMC iterations

```
function [ S_ML, S_mcmc ] = mcmc_phase( greedy_S_ML, ...
    log_xE,log_xnotE,varargin )
% Parse inputs
p = inputParser;
addOptional(p,'iterations',1e6);
addOptional(p,'burnin',1e5);
p.parse( varargin{:} )

burnin = p.Results.burnin;
iterations = p.Results.iterations + burnin;
N = size(log_xE,2);
S_mcmc = NaN(iterations+1,N+1);
S_mcmc(1,:) = greedy_S_ML;

for k = 1:iterations

    swap = randi(N,2,1);
    %  To ensure that we are not swapping an event with itself:
```

```
    while swap(1) == swap(2)
        swap = randi(N,2,1);
    end

    % Generate S'
    S_mcmc(k+1,1:N) = S_mcmc(k,1:N); % Copy S_t
    S_mcmc(k+1,swap(1)) = S_mcmc(k,swap(2)); % Swap random event 1
    S_mcmc(k+1,swap(2)) = S_mcmc(k,swap(1)); % Swap random event 2

    % Calculate p(X|S')
    S_mcmc(k+1,N+1) = LSE(S_mcmc(k+1,1:N),log_xE,log_xnotE);

    % Calculate a = p(X|S')/p(X|S_t)
    % i.e.      a  = e^( log(p(X|S') - log(p(X|S_t) ),
    % This is because calculate_p_X_given_S returns log(p(X|S)).
    a = exp(S_mcmc(k+1,N+1) - S_mcmc(k,N+1));

    % Leave S_t+1 = S' with probability min(a,1)
    if min(a,1) < rand(1)
        % ((1-a)*100)% of the time, set S_t+1 = S_t
        S_mcmc(k+1,1:N+1) = S_mcmc(k,1:N+1);
        % otherwise, leave S_t+1 = S'
    end
end

S_mcmc(1,:) = []; % Remove S_ML
if burnin > 0
    S_mcmc(1:burnin,:) = []; % Remove burn-in iteratons
end

% Find order with maximum likelihood [i.e. the maximum of the posterior
% distribution, S_hat = max p(S|X) ]:
[row, ~] = find(S_mcmc == max(S_mcmc(:,end)));
S_ML = S_mcmc(max(row),1:N+1);

end
```

A.7 Plotting a positional variance diagram

```
function [] = positional_variance_diagram(S_ML,S_mcmc,labels,varargin)
% Parse inputs
p = inputParser;
addOptional(p,'font_size',12);
addOptional(p,'new_figure',true);
p.parse( varargin{:} )

% Positional variance diagram:
if p.Results.new_figure == true
    figure;
end

N = size(labels,2);
confus_matrix = zeros(N,N);

% For each biomarker, find the proportion of MCMC iterations spent in each
% position
for i = 1:N
    confus_matrix(i,:) = sum(S_mcmc(:,1:N)==S_ML(i));
end
confus_matrix = confus_matrix/max(size(S_mcmc));
yaxislabels = cell(1,N);

for i = 1:N
    yaxislabels{i} = char(labels(S_ML(i))); % Order labels by S_ML
end

image((1-confus_matrix)*255); % Invert confusion matrix
colormap(gray(256)); % Apply grey colourmap
set(gca,'XTick',1:N);
set(gca,'XTickLabel',(1:N));
set(gca,'YTick',1:N);
set(gca,'YTickLabel',yaxislabels); % Label each
set(gca,'FontSize',p.Results.font_size);
set(gca,'DataAspectRatio',[1 1 1]); % Fix Aspect Ratio
xlabel('Event Position'); % Number each event

end
```

A.8 The Bayesian EBM

A.8.1 Embedding a permutation onto a hypersphere

```
function [ embedded_perms ] = embed_on_hypersphere(permutations)
% Permutations need to be entered as row vectors.
j = size(permutations,1);
n = size(permutations,2);
permutations = permutations'; % Transpose to column vectors

cM = 0.5*(n + 1) * ones(n,1); % Find the centre of mass of the hypersphere

[Q_tilde,~] = qr(cM); % QR factorisation of mass vector cM

[~,row] = max(abs(sum(Q_tilde))); % Find the row Q_ort
Q_basis = Q_tilde;
Q_basis(row,:) = []; % Remove the row Q_ort, which is collinear with cM

p_tilde = permutations - repmat(cM,1,j); % Recenter the origin at 0

phat_S = zeros(n-1,j);
for i = 1:j
    phat_S(1:n-1,i) =   Q_basis * p_tilde(:,i);
end

% Transpose back to row vectors
embedded_perms = (phat_S / sqrt((n^3 - n)/12))';
end
```

A.8.2 Maximum likelihood estimation of parameters of von Mises-Fisher distribution

```
function [mu, kappa] = mle_vonMisesFisher(samples)
% mle_vonMisesFisher returns the MLE of mu and kappa.
% Let X = {x_1,...,x_n} be a set of points drawn from p(x;mu,kappa).
% We wish to estimate mu and kappa via maximising the log-likelihood:
% L(X;mu,kappa) = log c_p(kappa) + sum_i (kappa * mu' * x_i)
% subject to the condition that mu' mu = 1 and kappa >= 0.

% See Sra, S. A short note on parameter approximation for von Mises-Fisher
% distributions: and a fast implementation of I_s (x). Comput Stat 2011
% mu = sum_i(x_i) / || sum_i(x_i) ||

% let x_i = "samples"

N = size(samples,1); % Number of sampled points
n = size(samples,2); % Number of elements in each permutation
p = n - 1; % Plis et al. and Sra et al. use different notation.

mu = sum(samples,1)/norm(sum(samples,1));

R_bar = norm(sum(samples,1)) / N;

% Simplest kappa_hat (Banerjee et al. 2005):
kappa_hat = (R_bar * (p - R_bar^2)) / (1 - R_bar^2);

% Truncated Newton Approximation for kappa_hat (Sra et al. 2011):
kappa_0 = kappa_hat; % initialise with kappa from Banerjee et al.

Ap_kappa_0 = (besseli(p/2,kappa_0) / besseli((p/2)-1,kappa_0));

kappa_1 = kappa_0 -  ( Ap_kappa_0 -  R_bar) / ...
    (1 - Ap_kappa_0^2 - (((p - 1)/kappa_0) * Ap_kappa_0) );

Ap_kappa_1 = (besseli(p/2,kappa_1) / besseli((p/2)-1,kappa_1));

kappa_2 = kappa_1 -  ( Ap_kappa_1 -  R_bar) / ...
    (1 - Ap_kappa_1^2 - (((p - 1)/kappa_1) * Ap_kappa_1) );

kappa = kappa_2;
end
```

A.8.3 Projecting from a hypersphere

```
function [ projected_perms ] = project_from_hypersphere(embedded_perms)
% Embedded permutations need to be entered as row vectors.
j = size(embedded_perms,1);
n = size(embedded_perms,2) + 1;
embedded_perms = embedded_perms'; % Transpose to column vectors

cM = 0.5*(n + 1) * ones(n,1); % Find the centre of mass of the hypersphere

[Q_tilde,~] = qr(cM); % QR factorisation of mass vector cM
```

```
[~,row] = max(abs(sum(Q_tilde))); % Find the row Q_ort
Q_tilde = [Q_tilde; Q_tilde(row,:)]; % Append Q_ort
Q_tilde(row,:) = []; % Remove the original row Q_ort

rho = ( sqrt((n^3 - n)/12) * Q_tilde' * ...
    [embedded_perms; zeros(1,j)] ) + repmat(cM,1,j);

if (isequal(ceil(rho),floor(rho)) == 1 )
    % If rho consists of integers, then done
    projected_perms = rho';
else
    % Otherwise, need to find the permutation vector closest to the
    % transformed point. This needs to be done one permutation vector
    % (column of rho) at a time
    projected_perms = NaN(n,j);
    for i = 1:j
        % Construct the cost matrix W_ij = (rho_i - j)^2
        W = ( repmat(rho(:,i),1,n) - repmat([1:n],n,1) ).^2;
        [~,I] = min(W,[],2); % Location of the min of each row
        if size(unique(I),1) == n
            % If the row mins are in different columns, then we have a
            % simple solution
            projected_perms(:,i) = I;
        else
            % The permutation p closest to the point rho is found by the
            % Hungarian algorithm. This implementation,
            % ("assignmentoptimal"), is by Markus Buehren:
            % http://www.mathworks.com/matlabcentral/...
            % fileexchange/loadFile.do?objectId=6543
            projected_perms(:,i) =  assignmentoptimal(W);
        end
    end
    projected_perms = projected_perms'; % Transpose back to row vectors
end

end
```

A.8.4 The Bayesian EBM – incorporating a prior distribution

```
function [ S_mcmc_ML, S_mcmc ] ...
    = bayesian_mcmc( S_mcmc_ML,S_mcmc_perms,log_xE,log_xnotE,varargin )
% Parse inputs
p = inputParser;
addOptional(p,'iterations',1e6);
addOptional(p,'mu',NaN);
addOptional(p,'kappa',NaN);
p.parse( varargin{:} )

% Initialise variables:
iterations = p.Results.iterations + 1e5;
N = size(log_xE,2);
S_mcmc = NaN(iterations+1,N+1);
S_mcmc(1,:) = S_mcmc_ML;
% mcmc_perms = S_mcmc_perms(:,1:N);

% % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % %
% Update the uniform prior distribution to a von Mises-Fisher distribution:

% Embed the permutations of S_mcmc onto a (N-2)-hypersphere:
embedded_perms = embed_on_hypersphere(S_mcmc_perms(:,1:N));

% Find the MLE for the parameters of the von Mises-Fisher distribution:
[mu, kappa] = mle_vonMisesFisher(embedded_perms);
clear embedded_perms
% % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % %

% If mu and kappa are supplied, overwrite the von Mises-Fisher fitting:
if ~isnan(sum(p.Results.mu)) && ~isnan(p.Results.kappa) == 1
    mu = p.Results.mu;
    kappa = p.Results.kappa;
end

% % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % %
% Calculate the log likelihood of p(S_1):
S_hypersphere = embed_on_hypersphere(S_mcmc(1,1:N));
llh_S = log(vMFpdf(S_hypersphere,mu,kappa));

% Update log(p(X|S_1)) to log(p(X|S_1)*p(S_1)):
S_mcmc(1,N+1) = S_mcmc(1,N+1) + llh_S;
```

```
% % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % %

    for k = 1:iterations

        swap = randi(N,2,1);
        %  To ensure that we are not swapping an event with itself:
        while swap(1) == swap(2)
            swap = randi(N,2,1);
        end

        % Generate S'
        S_mcmc(k+1,1:N) = S_mcmc(k,1:N); % Copy S_t
        S_mcmc(k+1,swap(1)) = S_mcmc(k,swap(2)); % Swap random event 1
        S_mcmc(k+1,swap(2)) = S_mcmc(k,swap(1)); % Swap random event 2

        % Calculate p(X|S')
        S_mcmc(k+1,N+1) = LSE(S_mcmc(k+1,1:N),log_xE,log_xnotE);

        % Calculate p(S')
        S_hypersphere = embed_on_hypersphere(S_mcmc(k+1,1:N));
        llh_S = log(vMFpdf(S_hypersphere,mu,kappa));

        % Update log(p(X|S')) to log(p(X|S')*p(S')):
        S_mcmc(k+1,N+1) = S_mcmc(k+1,N+1) + llh_S;

        % Calculate a = (p(X|S')*p(S'))) / (p(X|S_t)*p(S_t))
        % i.e.      a  = e^( log(p(X|S')*p(S') - log(p(X|S_t)*p(S_t) ),
        % This is because calculate_p_X_given_S returns log(p(X|S)).
        a = exp(S_mcmc(k+1,N+1) - S_mcmc(k,N+1));

        % Leave S_t+1 = S' with probability min(a,1)
        if min(a,1) < rand(1)
            % ((1-a)*100)% of the time, set S_t+1 = S_t
            S_mcmc(k+1,1:N+1) = S_mcmc(k,1:N+1);
            % otherwise, leave S_t+1 = S'
        end
    end

S_mcmc(1,:) = []; % Remove S_ML
S_mcmc(1:100000,:) = []; % Remove burn-in iteratons
```

```
% Find order with maximum likelihood [i.e. the maximum of the posterior
% distribution, S_hat = max p(S|x) ]:
[row, ~] = find(S_mcmc == max(S_mcmc(:,end)));
S_mcmc_ML = S_mcmc(max(row),1:N+1);

end
```

A.8.5 PDF for von Mises-Fisher distribution

```
function y = vMFpdf(x,mu,kappa)
%   Y = VMFPDF(X,MU,KAPPA) returns the pdf of the von Mises-Fisher
%   probability distribution with mean direction MU and concentration
%   parameter KAPPA, evaluated at the values in the p-dimensional vector X.
%   MU and X must be row vectors.

% y = Cp(kappa) * exp(kappa * mu' * x)
% where kappa >= 0, ||mu|| = 1 and the normalisation constant is given by
% Cp(kappa) = [kappa^((p/2)-1)] / [((2*pi)^(p/2))*I_((p/2)-1) (kappa)]
% I_v denotes the modified Bessel function of the first kind at order v.

p = size(x,2);

norm_const = (kappa^((p/2)-1)) / (((2*pi)^(p/2)) * besseli((p/2)-1,kappa));

y = norm_const * exp(kappa * mu * x');

end
```

A.9 Tools for permutations

A.9.1 Calculating Kendall's tau distance

```
function [ Ktau ] = Ktau(ref_perm,samples)
% Kendall tau distance between two permutations

% Remove log-likelihood column from ref_perm:
if ( ceil(ref_perm(end)) ~= floor(ref_perm(end)) ) ...
        || ref_perm(end) < 0
    ref_perm = ref_perm(1,1:end-1);
end
```

```
N = length(ref_perm); % Number of biomarkers
n = size(samples,1); % Number of samples

all_poss_pairs = nchoosek(1:N,2);
% Number of possible of pairs, N choose 2  = N!/((N-k)!*k!) = N*(N-1)/2
num_pairs = N*(N-1)/2;

% Find number of discordant pairs:
Ktau = zeros(n,1);
for i = 1:n
    for j = 1:num_pairs
        if ( ref_perm(1,all_poss_pairs(j,1)) < ...
                ref_perm(1,all_poss_pairs(j,2)) ) ...
                ~= ( samples(i,all_poss_pairs(j,1)) < ...
                samples(i,all_poss_pairs(j,2)) )
            Ktau(i,1) = Ktau(i,1) + 1;
        end
    end
 end

% Normalise Kendall tau distance by num_pairs:
Ktau(:,1) = Ktau(:,1) / num_pairs;

end
```

A.9.2 Calculating Kendall's tau correlation coefficient

```
function [ KtauCoef ] = KtauCoef(perm1,perm2)
% Kendall tau rank coefficient  between two permutations

% Remove log-likelihood column from perm1 and/or perm2:
if ( ceil(perm1(end)) ~= floor(perm1(end)) ) ...
        || perm1(end) < 0
    perm1 = perm1(1,1:end-1);
end
if ( ceil(perm2(end)) ~= floor(perm2(end)) ) ...
        || perm2(end) < 0
    perm2 = perm2(1,1:end-1);
end

N = length(perm1); % Number of biomarkers
```

```
all_poss_pairs = nchoosek(1:N,2);
% Number of possible of pairs, N choose 2  = N!/((N-k)!*k!) = N*(N-1)/2
num_pairs = N*(N-1)/2;

% Find number of concordant pairs:
concordant = 0;
    for j = 1:num_pairs
        if ( perm1(all_poss_pairs(j,1)) < ...
                perm1(all_poss_pairs(j,2)) ) ...
                == ( perm2(all_poss_pairs(j,1)) < ...
                perm2(all_poss_pairs(j,2)) )
            concordant = concordant + 1;
        end
    end

% Find number of discordant pairs:
discordant = 0;
    for j = 1:num_pairs
        if ( perm1(all_poss_pairs(j,1)) < ...
                perm1(all_poss_pairs(j,2)) ) ...
                ~= ( perm2(all_poss_pairs(j,1)) < ...
                perm2(all_poss_pairs(j,2)) )
            discordant = discordant + 1;
        end
    end

% Kendall tau rank coefficient = ...
% (concordant pairs - discordant pairs) / N*(N-1)/2
KtauCoef(:,1) = (concordant - discordant) / num_pairs;

end
```