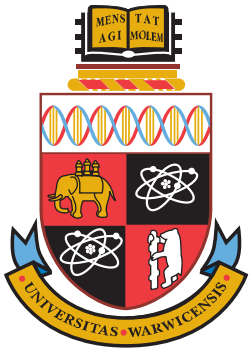

FUNCTIONAL DATA ANALYSIS APPROACHES FOR 3-DIMENSIONAL BRAIN IMAGES

Marco Palma

Thesis submitted for the degree of *Doctor of Philosophy*

University of Warwick
Department of Statistics

November 2021

Contents

List of Figures iii

List of Tables vi

Acknowledgements vii

Declaration x

Abstract xi

Notation xii

List of Abbreviations xiii

1 Introduction 1
- 1.1 Neuroimaging data analysis . . . 1
- 1.2 Thesis outline and contributions . . . 4

2 Mathematical aspects of functional data analysis 8
- 2.1 Mathematical aspects of functional data analysis . . . 8
  - 2.1.1 The space $L^2(\mathcal{T})$ . . . 9
  - 2.1.2 Random functions in $L^2(\mathcal{T})$ . . . 10
- 2.2 Smoothing . . . 13
  - 2.2.1 Smoothing by basis expansion . . . 13
  - 2.2.2 Kernel smoothing . . . 15
  - 2.2.3 Multidimensional smoothing . . . 16
- 2.3 Overview of functional regression . . . 17
- 2.4 Multidimensional functional data analysis . . . 20

3 Quantifying uncertainty in brain-predicted age using scalar-on-image quantile regression 24
- 3.1 Introduction . . . 24
- 3.2 Materials and Methods . . . 28
  - 3.2.1 Quantile regression . . . 28

3.2.2 Functional quantile regression 30
3.2.3 Data analysis workflow 33
3.3 Data 39
3.4 Results 41
3.4.1 Prediction accuracy 41
3.4.2 Correlation with cognitive decline measures 48
3.4.3 Sensitivity analysis 51
3.5 Discussion and further research directions 53

4 A whole-brain normative model for 3-dimensional morphometry images based on skewed functional data analysis 59
4.1 Introduction 59
4.2 Beyond case-control: normative modelling 65
4.3 Statistical framework 66
4.3.1 Voxelwise analysis: the skew-normal distribution 66
4.3.2 Computational aspects 69
4.4 Applications 71
4.4.1 Prediction for new observations 71
4.4.2 Subject-specific indices of “abnormality” 71
4.4.3 Z-maps as covariates in functional regression 73
4.5 Data and results 73
4.6 Conclusions 78

5 Function-on-function regression for large-scale brain imaging data 84
5.1 Introduction 84
5.2 Methods 87
5.2.1 PLS for functional regression 87
5.3 Data 90
5.3.1 Resting state fMRI (rfMRI) 91
5.3.2 Task fMRI 92
5.4 Results and discussion 92
5.4.1 Mask and basis expansion 92
5.4.2 Prediction with 4 covariates 93
5.5 Conclusions 100

6 Conclusions 102

Bibliography 106

List of Figures

3.1 Flowchart of the analysis from the brain images to the predicted intervals. 38

3.2 Histogram of age of the subjects in the sample, for each diagnosis. The number of bins has been fixed using the Freedman-Diaconis rule (Freedman and Diaconis, 1981). 40

3.3 Axial slices of the mean images for each diagnosis (from left to right: Control, MCI, AD). Slices are ordered from bottom to top. The colours are overlaid on the corresponding slice of the MDT. 42

3.4 Axial slices of the first 3 eigenfunctions for the control subset. Slices are ordered from bottom to top. The colours are overlaid on the corresponding slice of the MDT. The eigenfunctions account respectively for 15.43%, 13.95%, 6.87% of the total variability. The signs of the eigenfunctions are determined on the basis of clinical interpretation. 44

3.5 Plot of the brainPAD vs. predicted response. The coloured lines are local regression lines obtained with `loess` (locally estimated scatterplot smoothing) with span = 0.75 and 95% confidence bands. . . 45

3.6 Brain age 90% prediction intervals, relative to chronological age. There is one interval per subject, and subjects are sorted in descending order of predicted brain age (higher predicted ages at top). The black diamonds indicate the subjects for which chronological age does not fall into the prediction interval; the side indicates if the subject is in the $*$-negative (diamonds on the left) or $*$-positive group (diamonds on the right). 47

3.7 Left: distribution of the prediction interval width conditioned by diagnosis. Right: histogram of chronological age conditioned by $*$-positive indicator (equal to 1 if the chronological age is less than the prediction at $\tau = 0.05$, 0 otherwise). 48

3.8 Axial slices of the functional regression coefficient for $\tau = \{0.05, 0.5, 0.95\}$ (from left to right). Slices are ordered from bottom to top. The colours are overlaid on the corresponding slice of the MDT. For a unit increase (expansion) in the observed TBM image in a red voxel, there is an increase in predicted brain age, while in a blue voxel there is a decrease. 49
3.9 Left: association of *brainPAD* with ApoE4 value (Holm-corrected p-values) for different visits, with evidence of positive association. Right: (A) Correlation between baseline *brainPAD* and cognitive scores at different visits; (B) t-statistic for the comparisons of means of cognitive scores between $*$-positive group and the rest of the sample at different visits. The black lines are Student's t quantiles which correspond to different probabilities in the tails of the distribution. 50
3.10 Left: mean absolute error for control subjects as function of proportion of variance explained and knot spacing. Right: Coverage relative difference of prediction intervals induced by each choice of proportion of variance explained, knot spacing and nominal coverage. Points are jittered horizontally for visualisation purposes. . . . 52

4.1 Top: empirical cumulative distribution functions of TBM Jacobian values by diagnosis group (right) for a voxel in the lateral ventricles. Bottom: empirical cumulative distribution functions of TBM Jacobian values by diagnosis group (right) for a voxel outside the lateral ventricles. 61
4.2 2D histograms of voxelwise means and standard deviations by diagnosis group. The number of bins is fixed to 600. A smooth regression line is added in red. 62
4.3 2D histogram of the voxelwise mean and skewness across all subjects. The number of bins is fixed to 600. A smooth regression line is added in red. 63
4.4 Axial slices of the mean (left), standard deviation (centre) and skewness (right) parameter functions from skew-normal fitting. Slices are ordered from bottom to top. For the standard deviation, the colour white corresponds to the average standard deviation. 76
4.5 Top: Boxplots of u_3^{abs} by diagnosis group. Bottom: Plot of u_3^{abs} by ADAS13 and diagnosis group. 77

4.6 Axial slices of the first 3 eigenfunctions for the normative z-maps. They account respectively for 13.34%, 5.89%, 5.07% of the total variance. Slices are ordered from bottom to top. 79
4.7 Left: coefficient of functional logistic regression using the z-maps as covariates and the diagnosis group (CN vs. others) as outcome. Right: coefficient of functional linear regression using the z-maps as covariates and the logarithm of ADAS13 cognitive score as outcome. 80

5.1 Histogram of Pearson correlation between predicted and observed brain maps in the test set (100 subjects). The mean is denoted by the red diamond; the red line shows the variability ($\pm$ 1 standard deviation) 96
5.2 Top: axial slices of the observed (left) and predicted (right) task maps for one randomly selected subject in the test set. The Pearson correlation between them is equal to 0.415. Blue indicates negative values, while red is used for positive values. The midpoint is located at 0 (white) but the legend range of the actual image is larger. Bottom: axial slices of the observed (left) and predicted (right) quantile maps thresholded at the 95th percentile. Light brown indicates areas with values above the threshold. 97
5.3 Left: correlation heatmap between every pair of observed task activation maps in the test set. Right: correlation heatmap between every pair of predicted task activation maps from PLS in the test set. 98
5.4 Left: correlation heatmap between the observed and predicted task activation maps from the test set. The main diagonal shows correlation between predicted and observed images from the same subjects, while the elements outside the diagonal refer to cross-correlations between pairs of subjects. Right: boxplot of the on- and off-diagonal elements. The diamonds mark the group means. . 99

List of Tables

3.1 Summary statistics for each diagnosis group. N is the number of subjects in each group. The second part of the table shows selected quantiles of age. 39
3.2 Summary of the prediction results by diagnosis. Cor: correlation between predicted brain age and chronological age. CI_{Cor}: confidence interval for the correlation between predicted brain age and chronological age, obtained via Fisher-z transformation (Myers et al., 2013, Section 19.2). $\hat{\pi}$: sample coverage (proportion of cases for which the 90% prediction interval contain the chronological age). $*$-pos: proportion of cases for which the chronological age is less than the lower limit of the 90% prediction interval. 43
3.3 Cognitive decline measures used in the analysis. The arrows indicate the change in the measures associated to an increase in dementia severity. 50

5.1 Results from sensitivity analysis for the reconstruction of one random task activation map from basis functions. The metrics are root mean squared error (RMSE), mean absolute error (MAE) and correlation between the raw image and the one reconstructed as linear combination of basis functions. 93

Acknowledgements

I would like to thank my PhD supervisors Dr Shahin Tavakoli, Professor Thomas E. Nichols and Dr Julia Brettschneider. Your patience with my multiple and repeated writing mistakes, my incredibly long and detailed emails, my sometimes irrelevant questions would deserve a much bigger reward than what could be shown in a thesis. More importantly, your understanding and unfailing support throughout the academic and personal difficulties in the last 3 years has really given me the strength and the incentive to keep going during this journey.

I would also like to express my gratitude to Professor Ana-Maria Staicu, who hosted me at the Department of Statistics at North Carolina State University for 2 weeks and nevertheless kept working with me on one of the chapters for many months during the pandemic. Many thanks to Professor Xavier Didelot and Professor Jeff Goldsmith who gave their availability to examine this thesis, and to Professor David Firth, who gave with Professor Didelot very constructive feedback about this work.

The funding for my PhD has been provided by EPSRC. I would like to thank the Departments of Statistics at Warwick and Oxford for the great experience within the OxWaSP CDT and the amount of nice people that I have met in the program.

Among them, an honourable mention goes to Suzie Brown (also responsible for my improvement in the control of the length of my emails and for most of my—otherwise close to zero—physical activity at the squash court), Francesca Crucinio, Francesca Panero, Jack Carter, Giulio Morina, William Thomas and the others who never failed to convince me to spend so much time playing bridge, James Hodgson for the very interesting chats about a broad variety of different topics, Xin Zhi for the happiness that has brought in the fourth floor office in the last months of this thesis, Lionel Riou-Durand, Giorgos Vasdekis and Alice Corbella for their (very different, but all very effective) ways of supporting me in the last mile of this journey. Many others not nominated here have also contributed to many of the best memories of the last 4 years: games nights, Departmental

walks with too much rain, end-of-term celebrations with too much singing. All very much appreciated.

I would like to thank the people in Tom's group at the BDI (Sam Davenport, Petya Kindalova and all the others). Being not based in Oxford, I did not show up too often, but I have so many good memories of our OHBM in Rome (2019) thanks to you. Many thanks also to the people involved in OxWaSP with whom I worked and/or had fun.

Outside the life in the Statistics Departments, many people have made my life in Oxford and Coventry very much enjoyable. In addition to those already mentioned, a special thanks to Enrico Daviddi and Francesca Basini, whose friendship from Bologna continued fantastically throughout these years (in many cases involving so much good food that I would have not been able to make myself). I have been incredibly lucky to find nice flatmates, that I thank so much because they made the houses where I have lived home: Arianna Autieri (who survived for 3 years without complaining about me), Amul Gyawali, Rafael Vaquera Salazar, Harry Pitt Scott and Caterina Satiri. I would also like to say thank you to the organisers and the people playing poker at the Old Clarence: when I did not have to take very difficult decisions after your bets, I have had a lot of fun. Thank you for patiently dealing with my +25 bets at the poker table: the evenings spent with you have will be missed once I leave Coventry.

I want to thank also the people that I have known for so much time before my PhD experience. They deserve a lot of my gratitude and also a clear display of that, being this the first time I write an acknowledgement section in one of my written works.

It is not easy to say in words how much I thank Laura Bigoni, not only for the PhD years but for everything else. I hope that at least a bit of my gratitude for being such an important human being in my life shows throughout the time that we spend together (never enough). Now it's your turn! Thanks to you and Dimitrii Tanese for the great experience within the Alumni Association of Collegio Superiore (that is close to an end as I submit this thesis). Thanks to Caterina Severi, Luna Guglietta and Lucia Resca for reminding me of the nice years in Bologna via the many nice Skype calls of these years. Thanks to Alberto Antinoro, Luca Grementieri e Silvia Ferri for the good time spent together. Thanks to Massimiliano Paniccia for our in-person and Whatsapp chats, to Bellezze & Tavernelli (and all previous names—Silvia Colatosti, Francesca Fontana, Sara Guizzo, Giada Pasqualitto, Alessandro Pigliacelli, Chiara Iannarilli): you keep giving me so much happiness since an enormous number of years (and patiently bear with my "un-

conventional" Codenames style) and again to Silvia and Marta Marcoccia for our dinners in Veroli together. All those I have not thanked in the previous occasions are in my heart and thoughts: thank you for playing a part in my life.

A huge thanks to my relatives who stand by my side, and to Davide, my father and my mother. *Senza la vostra costante presenza tutto questo non avrebbe avuto né possibilità di esistere né senso nella mia vita.*

Marco Palma
22 November 2021

Declaration

I declare that this thesis is my own work, that I have carried out under the supervision of Dr Julia Brettschneider, Professor Thomas E. Nichols and Dr Shahin Tavakoli, for the degree of Doctor of Philosophy in Statistics. Chapter 4 of this thesis has also been supervised by Professor Ana-Maria Staicu from March 2020 to November 2020 as part of a research exchange (mostly run online because of the COVID-19 pandemic). I have not used sources or means without declaration in the text. I confirm that this thesis has not been submitted for a degree at any other university.

The contents of Chapter 3 and part of Chapter 2 are contained in the article "**Quantifying uncertainty in brain-predicted age using scalar-on-image quantile regression**" which I have written during my PhD in collaboration with Dr Julia Brettschneider, Professor Thomas E. Nichols and Dr Shahin Tavakoli. This article has been accepted for publication on Neuroimage (Palma et al., 2020).

Abstract

The analysis of brain images poses many challenges from a statistical perspective. First, these images are usually high-dimensional (sometimes millions of data points for each image), therefore a statistical analysis based on scalable techniques is often required. Second, these data exhibit clear spatial dependence due to the differences in structures and functions of the brain regions.

Functional data analysis is a modern branch of statistics aimed at analysing data that are in the form of functions. Many tools from multivariate analysis and nonparametric smoothing are used in functional data analysis to reduce noise and perform dimension reduction.

This thesis shows three applications of functional data analysis for large-scale 3-dimensional brain images, mainly focusing on prediction of scalar and imaging outcomes. A workflow for building prediction intervals for scalar outcomes from 3D covariates is devised and applied for the prediction of individual chronological age from brain anatomical images. Then, a framework for the analysis of functional data with spatially-dependent mean-variance relationship and skewness is described, with an application to structural imaging. At last, a functional imaging problem is studied: the prediction of a task-evoked response image from resting-state data is achieved through an image-on-image regression model.

The results discussed in this thesis are mostly comparable with more complicated machine-learning approaches available in the literature, while being more easily interpretable and often more computationally appealing. Functional data analysis might represent a valid option for the statistical analysis of brain images even in high-dimensional setting.

Notation

Within each chapter, the notation is explained when introduced for the first time. An attempt has been made to keep the notation consistent also across chapters, except for sum indices and math-mode accents. Vectors and matrices are denoted in boldface, as opposed to scalars, functions and operators. In Chapter 4, $v \in \mathcal{V}$ is used instead of the more common notation in $t \in \mathcal{T}$ to represent more immediately the function arguments as voxels of the brain images.

List of Abbreviations

3D	3-dimensional
AIC	Akaike Information Criterion
AD	Alzheimer's disease
ADAS	Alzheimer's disease Assessment Scale
ADNI	Alzheimer's disease Neuroimaging Initiative
ApoE-4	Apolipoprotein E4
brainPAD	brain predicted age difference
CN	cognitively normal
CP	centred parameterisation
DP	direct parameterisation
DR	dual regression
FEAT	fMRI Expert Analysis Tool
(F)PC	(functional) principal component
(F)PCA	(functional) principal component analysis
(F)PLS	(functional) partial least squares
fMRI	functional magnetic resonance imaging
FWHM	full width at half maximum
(G)CV	(generalised) Cross-Validation
GEV(D)	generalised extreme value (distribution)
GLM	generalised linear model
ICA	independent component analysis
IRLS	iteratively reweighted least squares
KS	knot spacing
LASSO	least absolute shrinkage and selection operator
loess	locally estimated scatterplot smoothing
MAE	mean absolute error
MCI	Mild Cognitive Impairment
MDT	minimal deformation template

MLE	maximum likelihood estimation
MNI	Montreal Neurological Institute
MMSE	Mini-Mental State Examination
OLS	ordinary least squares
PET	Position Emission Tomography
PVE	proportion of variance explained
RBF	radial basis function
REML	restricted maximum likelihood
ROI	region of interest
RMSE	root mean square error
sMRI	structural magnetic resonance imaging
SPM	Statistical Parametric Mapping
SVD	singular value decomposition
TBM	tensor-based morphometry

Chapter 1

Introduction

1.1 Neuroimaging data analysis

In the last decades, neuroimaging has provided great contributions towards a deeper understanding of the human brain, thanks to the large number of non-invasive techniques devised to display the anatomy or the functionality of the brain. In basic neuroscience, it has changed the way to identify the links between structural features and functional architecture of the brain. In the clinical practice, it represents an invaluable tool to support the diagnosis and monitor the progression of diseases which figure among the global health burdens in terms of deaths and years spent with disability, with substantial impact on social care, healthcare expenditures and on the quality of life of patients, caregivers and family members.

From the perspective of a statistician, the opportunities offered by neuroimaging are countless. Brain data come indeed in many fashions: high-dimensional multivariate measurements, longitudinal and time-to-event observations or multiple images recorded over time. The room for new research avenues in terms of

data analysis driven by neuroscientific questions is growing and multidisciplinary approaches are becoming common, if not necessary.

Restricting the attention only to imaging data, there are at least two important statistical issues that need to be addressed. Images come as large arrays of data which are recorded in pixels (or voxels, the analogous unit for 3-dimensional scans). The spatial characteristics encoded in the brain signal represent a major challenge for statistical modelling. Brains are indeed different between subjects, and even multiple scans on the same individual show distortions due to head motion artifacts, for example. These aspects are usually taken care in a series of preprocessing steps aimed at registering (or warping) the brain images to a common space and masking, i.e. subsetting the part of the image which refers to brain tissues and discarding those areas outside the brain. More important, the signal within the brain is spatially structured, meaning that the values recorded at neighbouring locations are often correlated. The partition in voxels is not indicative of physiologically compartments but it is a discretisation coming from the acquisition method; the underlying signal is actually considered to be smooth, showing gradual changes over different brain regions.

The other relevant issue in current neuroimaging studies is the data size. The technological advancement and the growth of worldwide research collaborations have reshaped the whole data collection process, moving in just a few decades from small studies of tens of subjects to multicenter initiatives with thousands of participants. This is the case for example of ADNI (Mueller et al., 2005), a repository of demographic, clinical and imaging data aimed at finding biomarkers of early Alzheimer's disease, or UK Biobank (Sudlow et al., 2015, Miller et al., 2016), a large biomedical database including lifestyle, genetic and health information which currently counts approximately half a million UK subjects (with the plan of obtaining brain scans for almost a fifth of them, as per Littlejohns et al., 2020). The improvement has also impacted on the image resolution, that in turn affects the number of data points recorded for each subject. The combination of higher sample size and higher resolution make even loading the data in memory in a standard laptop a difficult, if not unfeasible, step. This calls for new statistical approaches based on parallelisation, dimension reduction, batch modelling and federated learning. The worth of large-scale datasets in neuroimaging and other biomedical sciences is undoubted, as they allow to identify more refined statistical relationships otherwise undetectable in smaller samples, although at a higher risk of being subject to confounding effects (Smith and Nichols, 2018).

Among the different approaches to neuroimaging data analysis we can enumerate *mass-univariate* methods and machine-learning techniques. In the mass-univariate approach, a generalised linear model (GLM) is applied independently for each voxel. At each location, the voxel value is predicted using the same design matrix. The test statistic for the significance of the regression (t- or F-statistic) is then plotted spatially in the so-called *statistical parametric map*, then multiple testing corrections (using random field theory or nonparametric resampling methods) are introduced to account for the spatial structure. The main advantage of the mass-univariate analysis is scalability. Being developed several decades ago, this approach has been for a long time the first and only viable option in terms of computational efficiency as the regression models could be run independently (in parallel) for each voxel. The mass-univariate approach plays a primary role especially in estimation and inference problems, when comparison between groups are considered or in task activation studies.

On the other side, machine learning has provided a more flexible approach towards the analysis of brain imaging data, with a specific focus on individual prediction. In particular, artificial neural networks have been largely employed, thanks to their ability of learning about outcomes without choosing a specific model in advance. Deep neural-network architectures are indeed structured to return an output prediction from multiple layers of non-linear operations on an input. Although neural networks are scalable to high-dimensional settings, often the model is too complicated to get a complete understanding of all the parameters (Bzdok et al., 2019), at the cost of the interpretability of the model. To tackle this issue, recent work has focused on *explainable artificial intelligence* (see Krichmar et al., 2021 for a review). In addition, training deep neural networks (or ensemble models, where the predictions from multiple neural networks are combined) requires a large amount of data and is also computationally demanding. Some therefore have also studied whether the improvement in the predictive ability is worth the additional complications, or whether the prediction from simpler linear models could be preferable. When the problem of interest is intrinsically linear, more complex models will not outperform the linear alternatives (Davatzikos, 2019) and in neuroimaging applications it appears that even for moderate-to-high sample sizes linear approaches are not outperformed by deep learning models (Schulz et al., 2020).

Methods from multivariate statistics are becoming more popular in the neuroimaging field, especially in recent years. In particular, principal component analysis (PCA), canonical correlation analysis (CCA) and independent compo-

nent analysis (ICA) allow to detect data-driven relationships between different types of neuroscientific data. These methods have provided useful insights on the correlation between imaging, demographic and behavioural data (Smith et al., 2015) and have become the standard tools for achieving a dimension reduction which is not only data-driven, but also biologically meaningful (Smith and Nichols, 2018). Nonetheless, these multivariate techniques do not directly incorporate the spatial structure between voxels in the model and look at the brain scans just as vectors of voxels.

A modern approach that at the same time incorporates the spatial structure in a single model while keeping interpretability is represented by *functional data analysis* (FDA). In this context, the "atom" of interest (Wang et al., 2016) is a function, not the discrete collection of values which the function takes over a discretisation of its domain. Functional data are intrinsically defined on a continuous and their smoothness, although not strictly required in theory, is distinctive of the nature of the underlying phenomenon. An alternative definition describes functional data as an extension of multivariate data "with an ordering on the dimensions" (Wang et al., 2016), meaning that the observations lie in a space where invariancy to permutation does not hold. In a more theoretical perspective, functional data are no longer a finite collection of points but realisations of an infinite-dimensional stochastic process.

In the case of brain imaging, this definition has a clear application. The brain signal is recorded as a collection of voxels solely for the purpose of data collection, but actually the analyst would be more interested in the underlying, smooth signal function. In other words, the spatial structure of the brain introduces a topology in the vector of voxel values. The benefit of the spatial structure affects also the scalability of the statistical analysis, as smoothing techniques induces also a more parsimonious representation of the functional data. In other words, the smooth function can be represented via a number of coefficients often much smaller than the number of voxels.

1.2 Thesis outline and contributions

The main aim of this thesis is to provide FDA-based methods to analyse 3-dimensional imaging data, with an emphasis on computationally efficient approaches, for moderate-to-high sample sizes. An application of these methods to some questions in neuroscience is provided, using different imaging modalities.

The presentation of the methods in this thesis aims at being statistically principled and at the same time approachable for a reader whose main expertise is in biomedical science.

This thesis, dealing exclusively with 3D images, collects a suite of tools that are useful for the analysis of functional data with multidimensional domain and large sample size. In this setting, dimension reduction techniques are fundamental to tackle the high-dimensionality of the problem. Given the large number of images and their size, the FDA approaches presented in the central chapters of this thesis share a common structure which can be seen as made of at least two sections. The first one is the smoothing step, where the image array is reduced to a vector of coefficients from a basis expansion. This first step is run in parallel for each image, independently on the type of smoothing technique used (in this thesis the focus is on simple approaches like radial basis functions and B-splines tensor product). Once the matrix of coefficients for the whole set of images is created, along with the matrix of basis functions, then the second step of the analysis (model training) is performed. The application of several FDA methodologies is shown in the next chapters (functional principal component analysis, functional partial least squares, regression with scalar and functional outcome).

A key point of this thesis is the focus on individual predictions rather than estimation. The neuroimaging questions addressed in this work (all pertaining to currently open problems) require indeed low prediction errors, in order to be potentially applied to the everyday clinical practice. In this sense, often the comparison is run with machine-learning solutions available in the literature, where the flexibility of the tools is better designed for predictions. This thesis aims not necessarily at beating the performances of those approaches, but rather at the explanation of statistical workflows which could provide results which are not too dissimilar in a simpler and less computationally demanding way.

In Chapter 3, a functional quantile regression model is used to predict the chronological age of an individual from anatomical MRI scans, in order to provide a sensitive summary (generally called *brain age*) of brain changes which could be linked to different neurodegenerative diseases. The workflow designed for this project takes as input a set of tensor-based morphometry (TBM) images (which inform about regional volume changes in the brain) for a subset of cognitively normal (CN) subjects in the Alzheimer's Disease Neuroimaging Initiative (ADNI), and returns brain age predictions also for subjects with Mild Cognitive Impairment (MCI) and Alzheimer's Disease (AD). The main contribution of this work is the focus on individual prediction intervals (obtained by using quantile regres-

sion for different quantile levels), by which to account for the prediction uncertainty in a simple way. The approaches available in the literature of brain age prediction (mostly based on machine learning) are indeed returning only point predictions. In addition, this is the first work (to the best of our knowledge) which employs TBM imaging covariates for this purpose.

The same TBM dataset show interesting features from a statistical point of view. TBM images are smooth but they exhibit (especially in diseased groups) higher values in some brain regions called lateral ventricles. More specifically, a voxelwise analysis shows a mean-variance relationship in these areas and evidence of spatially dependent skewness which can be missed in the standard FDA settings, which focus only on the first two functional moments. Following the approach proposed by Staicu et al. (2012), Chapter 4 presents a statistical model for 3-dimensional functional data where mean, variance, and shape functions vary smoothly across brain locations. The voxelwise distribution is modelled as a skew-normal (Azzalini, 2013) and the spatial dependence is represented via a Gaussian copula. Each individual image is then represented in terms of a Gaussian process that captures the dependence between voxelwise distributions. The functional parameters are estimated on a reference population of cognitively normal subjects and the Gaussian maps can be obtained for subjects with unknown brain health condition. These subject-specific normative maps are used to derive indices of deviation from a healthy condition which could help to assess the individual risk of pathological degeneration or to cluster different disease groups. The use of the skew normal for TBM images represents a novelty in the TBM literature, offering the flexibility of a single family of distribution for the whole brain image.

Chapter 5 shows an application of linear regression with a 3D imaging outcome with multiple functional predictors based on functional partial least squares. The model handles moderate-to-high sample size in a computationally efficient way, without the need of high performance computing resources, by extending the approach proposed in Preda and Schiltz (2011) and Beyaztas and Shang (2020) to the case of 3D imaging data. The regression model is used to predict the subject-specific brain activation maps from data collected when the subject was at rest. This represents one of the very first instances in the literature of multivariate functional data analysis where the functions have a multidimensional domain, using a partial least squares approach. In terms of the neuroimaging application, the model does not differentiate between cortical and subcortical information and does not require to split the brain image into regions, as commonly done in the literature of task activation map prediction.

The thesis is structured in 6 chapters. After this introduction (Chapter 1), we provide in Chapter 2 a basic overview on the mathematical and statistical aspect of functional data analysis which are necessary to understand the following chapters. Chapter 3 presents the application of regression for functional data to the prediction of chronological age from brain anatomical imaging. In Chapter 4, the model which incorporates skewness is used to study some features of a 3D imaging dataset and to provide indices of deviation from a healthy condition. Chapter 5 describes the statistical model to predict the brain activation in a task from resting-state data where both the covariates and the outcome are images. Chapter 6 is dedicated to a list of the main questions that remain open, with the discussion of potential solutions.

Chapter 2

Mathematical aspects of functional data analysis

2.1 Mathematical aspects of functional data analysis

As discussed in Hsing and Eubank (2015, Chapter 7) the mathematical foundations of FDA rely on two different perspectives on functional data. On one side, functional data are seen as realisations of random variables taking values in a Hilbert space equipped with the Borel σ-algebra; on the other, the observations are realisations of continuous stochastic processes with smooth mean and covariance functions. The two perspectives provide both the theoretical background of concepts like mean and covariance in the abstract setting as well as the foundations of the tools used to analyse the variability of the sample. Hsing and Eubank (2015) provide an extensive illustration of the mathematical concepts underlying FDA; Kokoszka and Reimherr (2017) and Horváth and Kokoszka (2012) offer a more concise overview and will serve as the main resources which this section is built upon. Some paragraphs of this section are also borrowed from Palma et al. (2020).

2.1.1 The space $L^2(\mathcal{T})$

In FDA, the attention is restricted to the Hilbert space $L^2(\mathcal{T})$ of all the functions $f : \mathcal{T} \mapsto \mathbb{R}$ that are square-integrable,

$$\int_{\mathcal{T}} [f(t)]^2 \, dt < \infty. \tag{2.1}$$

Typically in FDA it is often assumed $\mathcal{T} \subseteq \mathbb{R}^d$ (Ramsay and Silverman, 2005, Ferraty and Vieu, 2006, Kokoszka and Reimherr, 2017). The space $L^2(\mathcal{T})$ is endowed with the inner product

$$\langle f, g \rangle = \int_{\mathcal{T}} f(t)g(t)dt, \tag{2.2}$$

and the L^2 norm

$$\|f\| = \left(\int_{\mathcal{T}} [f(t)]^2 \, dt \right)^{\frac{1}{2}}, \tag{2.3}$$

where $f, g \in L^2(\mathcal{T})$. A set of functions $\{e_1, e_2, ...\} \subseteq L^2(\mathcal{T})$ is called a basis in $L^2(\mathcal{T})$ if every square-integrable function $f \in L^2(\mathcal{T})$ admits a unique representation of the form

$$f(t) = \sum_{j=1}^{\infty} a_j e_j(t), \qquad \forall t \in \mathcal{T} \tag{2.4}$$

where $a_j \in \mathbb{R}$. The basis is orthonormal if $\langle e_j, e_{\check{j}} \rangle = 0$ for $j \neq \check{j}$ and $\|e_j\| = 1$.

The notion of operators (linear transformations on vector spaces) on a Hilbert space (and in particular $L^2(\mathcal{T})$) is largely used in FDA. Within the space $\mathcal{L}(L^2(\mathcal{T}))$ of linear operators with finite operator norm

$$\sup_{\|f\|=1} \|\Psi(f)\| < \infty, \tag{2.5}$$

let $\Psi \in \mathcal{L}(L^2(\mathcal{T}))$ be a compact (or completely continuous) operator if it admits a singular value decomposition of the form

$$\Psi(f) = \sum_{j=1}^{\infty} \lambda_j \langle f, w_j \rangle \check{w}_j \qquad f \in L^2(\mathcal{T}), \tag{2.6}$$

where $\{w_j\} \subseteq L^2(\mathcal{T})$ and $\{\check{w}_j\} \subseteq L^2(\mathcal{T})$ are orthonormal basis functions and $\{\lambda_j\} \subseteq \mathbb{R}_+$ is a real non-negative decreasing sequence converging to zero. Additionally, if $\sum_{j=1}^{\infty} \lambda_j^2 < \infty$, Ψ is called a *Hilbert–Schmidt* operator. If the Hilbert–

Schmidt operator is also symmetric, i.e.

$$\langle \Psi(f), g \rangle = \langle f, \Psi(g) \rangle \qquad f, g \in L^2(\mathcal{T}) \tag{2.7}$$

and positive-semidefinite

$$\langle \Psi(f), f \rangle \geq 0 \qquad f \in L^2(\mathcal{T}), \tag{2.8}$$

then Equation (2.6) reduces to

$$\Psi(f) = \sum_{j=1}^{\infty} \lambda_j \langle f, v_j \rangle v_j \qquad f \in L^2(\mathcal{T}), \tag{2.9}$$

and $\{w_j\}$ is the set of eigenfunctions of Ψ for which $\Psi(w_j) = \lambda_j w_j$.

Let us now consider an integral operator, defined as

$$\Theta(f)(t) = \int_{\mathcal{T}} \theta(t,s) f(s) ds \qquad s, t \in \mathcal{T} \tag{2.10}$$

for $f \in L^2(\mathcal{T})$ and for a bivariate square-integrable function θ on $\mathcal{T} \times \mathcal{T}$. The bivariate function θ is referred to as the kernel of the integral operator. The operator Θ is Hilbert–Schmidt if and only if

$$\int_{\mathcal{T}} \int_{\mathcal{T}} [\theta(t,s)]^2 \, dt ds < \infty, \tag{2.11}$$

The theory illustrated above allows us to state Mercer's theorem: a continuous symmetric positive semidefinite kernel admits the representation

$$\theta(t,s) = \sum_{j=1}^{\infty} \lambda_j w_j(t) w_j(s), \qquad t, s \in \mathcal{T} \tag{2.12}$$

where λ_j is the j-th eigenvalue corresponding to the j-th eigenfunction w_j of the integral operator Θ and the sum converges absolutely and uniformly in s and t.

2.1.2 Random functions in $L^2(\mathcal{T})$

In the FDA framework, an observation is viewed as a realisation of a random function X defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Given the sample space Ω, $X(\omega)$ is a deterministic function belonging to the space $L^2(\mathcal{T})$ of square-integrable

functions, that is

$$\|X(\omega)\|^2 = \int_{\mathcal{T}} [X(\omega)(t)]^2 \, dt < \infty, \qquad \forall \omega \in \Omega. \tag{2.13}$$

In the following paragraphs, the notation $X(\omega)(t)$ will be shortened as $X(t)$.

If $\mathbb{E}\|X\| < \infty$, the first order moment of X is the mean function $\mu(t) = \mathbb{E}[X(t)]$, $t \in \mathcal{T}$. If $\mathbb{E}\|X\|^2 < \infty$ the second order variations of X are encoded in the covariance function

$$\gamma(s,t) = \mathbb{E}\left[(X(s) - \mu(s))\,(X(t) - \mu(t))\right], \qquad s,t \in \mathcal{T} \tag{2.14}$$

of which the variance function is a special case ($s = t$).

The mean function can be estimated by the sample mean evaluated pointwise on the domain:

$$\hat{\mu}(t) = \frac{1}{N}\sum_{i=1}^{N} X_i(t), \qquad t \in \mathcal{T} \tag{2.15}$$

that corresponds to a pointwise average over all the functional observation. Likewise, the sample covariance function is the estimator of the covariance function:

$$\hat{\gamma}(s,t) = \frac{1}{N}\sum_{i=1}^{N} (X_i(s) - \hat{\mu}_X(s))\,(X_i(t) - \hat{\mu}_X(t)), \qquad s,t \in \mathcal{T}. \tag{2.16}$$

These estimators share the same inferential properties with their non-functional counterparts: $\hat{\mu}$ is an unbiased estimator for the mean function whereas for the covariance function

$$\mathbb{E}\left(\frac{N}{N-1}\hat{\gamma}(s,t)\right) = \gamma(s,t), \qquad s,t \in \mathcal{T} \tag{2.17}$$

where the scaling factor (due to the estimation of the mean function) tends to 1 asymptotically.

A central object when dealing with functional data is the covariance operator $\Gamma : L^2(\mathcal{T}) \mapsto L^2(\mathcal{T})$, which is the integral operator associated to the kernel function $\gamma(s,t)$. It is defined as

$$\Gamma(f) = \mathbb{E}\left[\langle X - \mu, f\rangle (X - \mu)\right], \quad \forall f \in L^2(\mathcal{T}). \tag{2.18}$$

The covariance operator transforms a function $f \in L^2(\mathcal{T})$ in another function $\Gamma(f) \in L^2(\mathcal{T})$ whose values are

$$\Gamma(f)(t) = \int_{\mathcal{T}} \gamma(t,s) f(s) ds, \qquad \forall t \in \mathcal{T}. \tag{2.19}$$

The covariance operator is a self-adjoint positive semi-definite Hilbert–Schmidt operator. Mercer's theorem therefore applies for the covariance function and the random function X admits the Karhunen–Loève expansion for square-integrable stochastic processes

$$X(t) = \mu(t) + \sum_{m=1}^{\infty} \nu_m \psi_m(t), \tag{2.20}$$

expressing X as an infinite linear combination of the deterministic eigenfunctions $\{\psi_m\} \in L^2(\mathcal{T})$ of Γ with random and uncorrelated weights ν_m. The eigenfunctions are the solutions of the eigendecomposition problem

$$\int_{\mathcal{T}} \gamma(t,s)\psi_m(s) ds = \lambda_m \psi_m(t), \qquad \forall t \in \mathcal{T}. \tag{2.21}$$

The eigenfunctions $\{\psi_m\}$ are orthogonal and rescaled to have unit norm, and their corresponding eigenvalues $\{\lambda_m\} \subseteq \mathbb{R}_+$ are in non-increasing order. In addition, the approximation error

$$\mathbb{E}\left(\left\| X - \sum_{m=1}^{M} \langle X, u_m \rangle u_m \right\|^2 \right), \qquad u_1, \ldots, u_M \in L^2(\mathcal{T}) \tag{2.22}$$

is minimised, for each $M \geq 1$, when $u_m = \psi_m, m = 1, \ldots, M$.

The results of the eigendecomposition of the covariance operator can be interpreted under the framework of functional principal component analysis (FPCA), which aims at studying the principal modes of variation of the random function X. FPCA is in practice an empirical version of the Karhunen–Loève decomposition, where the covariance operator is replaced by its sample version. The eigenvalue λ_m is the part of the variance of X explained by the m-th eigenfunction ψ_m, also called functional principal component. The random variables

$$\nu_m = \langle X - \mu, \psi_m \rangle \tag{2.23}$$

are called *scores*. The scores are uncorrelated and centered with variance λ_m.

2.2 Smoothing

In statistics, non-parametric smoothing techniques are aimed at modelling a function without imposing a specific functional form. Smoothing is extensively used for many applications, not only for data visualisation purposes, but also as building blocks of generalised additive models (see Wood, 2017, Hastie and Tibshirani, 2017 for monographs on the topic), where the relationship between some predictors and an outcome can be represented more flexibly as a smooth function. In FDA, smoothing is usually applied on the raw discretised data to obtain a noise-free estimate of the underlying function used in the following analysis. Smoothing is crucial especially in a big data setting, as the smooth underlying function can be represented more compactly with respect to the raw data, and brings additional insight on the data (for example, derivatives of the functions can also be considered, as in Ramsay and Silverman, 2005).

2.2.1 Smoothing by basis expansion

A common smoothing approach is to represent a function by a basis expansion. A system of basis functions is first chosen such that a function can be approximated by a linear combinations of them. Then, ordinary or weighted least squares allow to estimate the coefficients of this linear combination. This set of coefficients therefore provides a simpler representation of any general function. As shown in (2.4), the set of basis functions is defined to be infinite, but in practice the attention is restricted to a finite set. The choice of the type and number of basis functions controls how good the reconstruction of the original image is.

Splines are a common choice for basis functions, as they retain all the benefits of polynomial fitting while also ensuring good computational properties. Splines are piecewise polynomials which are joined together at some points $\kappa_1 \leq \kappa_2 \leq \ldots \leq \kappa_I$ in the domain of the function called *knots*. In mathematical terms, given the degree r (the power of each piecewise polynomial), a polynomial splines is $r-1$ times continuously differentiable. The flexibility of the polynomial splines is governed by the degree r (or analogously the order $r+1$, defined as the number of constants needed to define the polynomial), the number l of knots and their location. The number of free parameters is $r+l-1$. Natural cubic splines (well-suited for interpolation) are polynomial splines of degree 3 with null second derivative at the extremes of the domain.

Since the space of polynomial splines is a vector space for a given set of I knots and degree r (therefore linear combinations of splines are again splines), several types of basis functions have been used in the literature. B-splines (De Boor, 1978) are often considered among the most convenient ones. Given a set of I knots the B-spline basis of degree 0 is given by the functions $(B_1^0(x), \ldots, B_{I-1}^0(x))$ with

$$B_j^0(x) = \begin{cases} 1 & \kappa_j \leq x < \kappa_{j+1}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.24}$$

Given a set of I knots, the B-spline basis of degree $r > 0$ is given by the functions $(B_1^r(x), ..., B_{I-1}^r(x))$, defined recursively as

$$B_j^r(x) = \frac{x - \kappa_{j-r}}{\kappa_j - \kappa_{j-r}} B_{j-1}^{r-1}(x) + \frac{\kappa_{j+1} - x}{\kappa_{j+1} - \kappa_{j+1-r}} B_j^{r-1}(x). \tag{2.25}$$

Although defined in an iterative fashion (currently available in many software implementations), B-spline basis functions have interesting properties, such as the compact support: a B-spline basis function of degree r is positive over $r + 1$ adjacent intervals and 0 elsewhere. In addition, for any point within the domain $[\kappa_1, \kappa_I]$ the basis functions sum to 1 (this is also a result of the introduction of additional knots outside the domain of interest, which are needed just to construct the splines recursively). These features make B-splines (although not being orthogonal) appealing even in large data settings, because the matrix containing the inner products of these basis functions will be highly sparse.

While the choice of the degree is restricted to few options (usually there is no need to go more than $r = 3$), there is not a "one-size-fits-all" solution for the knot selection. Evenly spaced knots are often used, although placing more knots in a region of the domain rather than another might help to capture finer changes in the behaviour of the function. In addition, the number of effective degrees of freedom (which depends on the number of knots) can be reduced by applying roughness penalties on the function, in order to find a balance between the bias and the variance of the fit.

It is worth also mentioning other systems of basis functions which have found application in functional data analysis. Fourier series gives a basis expansion that is often used for periodic functions which are stable throughout the whole domain. Properties and derivatives of the Fourier series are well known and the fast Fourier transform makes the calculation of coefficients efficient. Wavelets basis functions are built as translations and dilations of an initial function called

mother wavelet and are used to obtain a parsimonious multiresolution expansion (Ramsay and Silverman, 2005).

2.2.2 Kernel smoothing

An alternative approach to smoothing and interpolation is based on kernel functions. These functions take as argument the distance between two points and return weights which are inversely proportional to this distance. Kernel functions are non-negative (and the weights are values in $[0, 1]$), have unit integral and are symmetric. Examples of popular kernels (with u being the distance between two points) are:

- the tri-cube kernel:

$$h(u) = (1 - |u|^3)^3 \qquad |u| \leq 1,$$

- the Gaussian kernel:

$$h(u, \check{\delta}_h^2) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u}{2\check{\delta}_h^2}\right)$$

  where the smoothing parameter $\check{\delta}_h^2 > 0$ is equal to the variance,

- the (inverse) multiquadric kernel

$$h(u) = \frac{1}{\sqrt{1 + (\varepsilon u)^2}}$$

  where ε is a constant,

- the Epanechnikov kernel:

$$h(u) = \begin{cases} \frac{3}{4}(1 - u^2) & |u| \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Before choosing the kernel function, it is necessary to specify the "centres", i.e. the points in the domain to compute the distance from, and the type of distance (for which the usual choice is the Euclidean one).

Depending on the task, the centres might be selected in different ways. In local smoothing, especially if the number of data points is small, every data point

might play as centre and the weighted average of the window is taken (as in the Nadaraya–Watson estimator, where the weights are proportional to the weights derived from the kernel function). In interpolation the centres are instead often taken on a regular grid, while the usual choice for distance is the Euclidean one.

The kernel function will assign non-zero weights only to the data points within a certain radius form the centres (for this reason, this is also known as *radial basis function*, RBF). For some kernels, the radius is also controlled by one or more additional parameters: for example, the scale of the Gaussian kernel is controlled by $\check{\delta}_h^2 > 0$ or the full width at half maximum (FWHM). The quality of the approximation is controlled by the number and locations of centers and the choice of the additional smoothing parameter(s). In most of the cases, there are no existing criteria to select the best value for this parameters.

2.2.3 Multidimensional smoothing

These considerations about smoothing can be also extended to the multidimensional case. For radial basis functions the extension is conceptually straightforward, as only the distance of a point from a centre is needed to compute the weights.

For B-splines, the immediate extension of the 1-dimensional setting is to define a tensor product of univariate B-splines defined in each dimension. At each point in the domain, the multivariate B-spline basis would correspond to the product of univariate B-splines for each coordinate. For the 3-dimensional domain (which will be used throughout this thesis), denote by $B_1^{(j)}(t^{(j)}), \ldots, B_{Q_j}^{(j)}(t^{(j)})$ the univariate basis functions for the j-th dimension ($j = 1, 2, 3$). The number of basis functions for each dimension is $Q_j = I_j + r - 1$, where I_j is the number of knots and r is the degree of the spline. We now define the set of basis functions

$$B_{q_1 q_2 q_3}(t^{(1)}, t^{(2)}, t^{(3)}) = B_{q_1}^{(1)}(t^{(1)}) B_{q_2}^{(2)}(t^{(2)}) B_{q_3}^{(3)}(t^{(3)}) \tag{2.26}$$

for $q_j = 1, \ldots, Q_j$, for $j = 1, 2, 3$.

In order to derive the projection of each image onto this set of basis functions, we define the following matrix of basis functions using the Kronecker product

$$\boldsymbol{\Phi} = \boldsymbol{S}^{(3)} \otimes \boldsymbol{S}^{(2)} \otimes \boldsymbol{S}^{(1)}. \tag{2.27}$$

where $\boldsymbol{S}^{(j)}$ is the $P_j \times Q_j$-dimensional matrix whose q_j-th column contains the evaluation of the function $B_{q_j}^{(j)}(t^{(j)})$ at each point in $t^{(j)}$ (for $j = 1, 2, 3$) and P_j is

the number of sampled points in the j-th dimension. The matrix $\mathbf{\Phi}$ has dimensions $P_1P_2P_3 \times Q_1Q_2Q_3$ (the number of rows is equal to the number of sampled points in the 3D function and the number of columns is equal to the number of basis functions). Once the basis set is determined, $\mathbf{\Phi}$ can be used as set of covariates where the original (vectorised) 3D function is the response variable. Estimation can be performed via ordinary least squares:

$$\hat{X}_i(t) = \sum_{k=1}^{K} \tilde{c}_{ik}\phi_k(t), \tag{2.28}$$

where $K = Q_1Q_2Q_3$, $\tilde{\boldsymbol{c}}_i$ is the K-dimensional vector containing the coefficients of the projection for the i-th 3D function and $\phi_k(t)$ is the k-th basis function corresponding to the the k-th column of $\mathbf{\Phi}$. In compact form, all the N 3D function are represented by the product of the $N \times K$ coefficient matrix $\tilde{\boldsymbol{C}}$ and the matrix of basis functions evaluations $\mathbf{\Phi}$.

Tensor products of univariate B-splines provide a highly flexible framework for multidimensional smoothing. Anisotropic smoothing can be easily accommodated by using a different smoothing parameter for each dimension. In addition, multidimensional penalties can be included, although the computational cost might become a relevant issue. The easier approach is to build roughness penalties for each dimension (in terms of difference matrices), then build other Kronecker products (see Wood, 2017, p.161). In the multidimensional penalised smoothing field, thin-plate regression splines are also very popular, as they are the result of applying natural cubic splines with a multidimensional roughness penalty based on second order partial derivatives (see Lindquist et al., 2010 for an application to neuroimaging). Smoothing on irregular domains is also of interest, but we shall not investigate this here.

2.3 Overview of functional regression

In addition to functional principal component analysis, regression models for functional data are among the most researched topics in FDA. The classical taxonomy of functional regression (Ramsay and Silverman, 2005, Morris, 2015) is based on the role of the functional data into the model, either as outcome/dependent variable/response or predictor/independent variable/covariate (we will use those terms interchangeably). We could therefore list:

- scalar-on-function regression, where at least one covariate is functional and the outcome is scalar;
- function-on-scalar regression, with functional response and scalar independent variables;
- function-on-function regression, where both the predictors and the response are functional objects.

Scalar-on function regression represents a natural extension of multiple regression as it is performed in multivariate statistics. The functional linear model indeed follows the same structure, replacing the inner product term with its functional counterpart:

$$\begin{aligned} y_i &= \alpha + \langle X_i, \beta \rangle + \varepsilon_i \\ &= \alpha + \int_{\mathcal{T}} X_i(t)\beta(t)dt + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2) \end{aligned} \tag{2.29}$$

where y_i denotes the scalar response, $X_i(t)$ the functional predictor and $\beta(t)$ the functional slope (belonging to the same space as $X_i(t)$).

The classical approach is to express both the functional covariate and the coefficient as linear combinations of basis functions:

$$\begin{aligned} X_i(t) &= \sum_{k=1}^{K_X} x_{ik}\phi_k(t) \\ \beta(t) &= \sum_{\check{k}=1}^{K_\beta} b_{\check{k}}\chi_{\check{k}}(t). \end{aligned} \tag{2.30}$$

Let us consider now the matrix forms for the basis coefficients $\boldsymbol{x}_i = [x_{i1}, \ldots, x_{iK_X}]^\mathsf{T}$ and $\boldsymbol{b} = \left[b_1, \ldots, b_{K_\beta}\right]^\mathsf{T}$, as well as for the basis functions $\boldsymbol{\phi}(t) = [\phi_1(t), \ldots, \phi_{K_X}(t)]^\mathsf{T}$ and $\boldsymbol{\chi}(t) = \left[\chi_1(t), \ldots, \chi_{K_\beta}(t)\right]^\mathsf{T}$. The inner product becomes a product of matrices (Ramsay and Silverman, 2005, Morris, 2015):

$$\begin{aligned} \int_{\mathcal{T}} X_i(t)\beta(t)dt &= \int_T \left[\sum_{k=1}^{K_X} x_{ik}\phi_k(t)\right]\left[\sum_{\check{k}=1}^{K_\beta} b_{\check{k}}\chi_{\check{k}}(t)\right] \\ &= \boldsymbol{x}_i^\mathsf{T}\boldsymbol{W}_{\phi,\chi}\boldsymbol{b}. \end{aligned} \tag{2.31}$$

where $\boldsymbol{W}_{\phi,\chi}$ is the matrix of inner products

$$\boldsymbol{W}_{\phi,\chi} = \int_T \phi(t)\chi(t)^\mathsf{T} dt. \tag{2.32}$$

The only parameter is $\boldsymbol{b}$ and the estimates can be obtained using standard multivariate regression techniques. For suitable choices of basis functions (e.g. orthogonal) the matrix $\boldsymbol{W}_{\phi,\chi}$ reduces to the identity matrix.

On one side, fixed basis functions can be selected a priori. Splines, Fourier, wavelets for X and β can be used to turn the regression model in Equation (2.29) into an ordinary linear model. On the other side, basis functions can be built using available data. This is the case of functional principal component analysis (FPCA) or functional partial least squares (FPLS) basis functions, which rely on the variability within the functional predictor or the covariance with the outcome. Hybrid approaches are also available. The same machinery virtually applies to a large share of regression models originally designed for multivariate data. There are simple extensions to generalised linear models (Müller and Stadtmüller, 2005, Crainiceanu et al., 2009), quantile regression (Cardot et al., 2005, Kato, 2012, Yao et al., 2017), longitudinal (Yao et al., 2005) and survival analysis (Gellar et al., 2015, Kong et al., 2018).

Function-on-scalar regression of the form

$$Y_i(t) = \beta_0(t) + \boldsymbol{Z}_i\beta(t) + \varepsilon_i(t), \tag{2.33}$$

with error term generally assumed to be normally distributed with mean zero and within-function covariance

$$\mathrm{Cov}(Y_i(t), Y_i(\check{t})) = \sigma^2(t,\check{t}), \qquad t,\check{t} \in \mathcal{T} \tag{2.34}$$

is of interest in several application, for example growth curves (Morris, 2015), where covariates might be either categorical or continuous. The pointwise approach illustrated by Ramsay and Silverman (2005) consists in considering a grid on the domain on the function and run a series of multiple regression. The resulting regression coefficients are then interpolated. Other approaches are instead based on the smoothing of the functional outcome first, although it has been pointed out that the resolution of the outcome should be kept high (or higher than the one of the functional regression coefficient) in order not to lose informative features which could improve the quality of the fit. Basis expansion approaches

(either with fixed or data-driven basis) could be used to make the computation more efficient. In the setting of regression with functional predictors, simultaneous inference on the functional outcome (especially in the setting of mean differences) is also a topic of interest (see Morris, 2015 for further references).

In function-on-function regression models, the relationship between a covariate and the functional outcome is denoted by a functional coefficient $B(t,s)$:

$$Y_i(s) = \beta_0(s) + \int_{\mathcal{T}} X_i(t)B(t,s)dt + \varepsilon_i(s), \qquad s \in \mathcal{S}. \tag{2.35}$$

Suppose both $X_i(t)$ and $Y_i(s)$ can be represented as linear combinations of basis functions $\{\phi_X(t)\}$ and $\{\phi_Y(s)\}$, respectively. The coefficient $B(t,s)$ simplifies to $\phi_X(t)^\mathsf{T}\boldsymbol{B}\phi_Y(s)$, where $\boldsymbol{B}$ is the matrix of coefficients in the basis space. In this setting, FPCA for both $X_i(t)$ and $Y_i(s)$ independently has been used to reduce the dimensionality of the functional data (Yao et al., 2005). FPLS approaches are also of large interests as the components directly take into account the joint variability of the outcome and predictors (Preda and Schiltz, 2011, Beyaztas and Shang, 2020). Special cases of the function-on-function regression where the predictors and the outcome are defined on the same domain $\mathcal{T}$ are the concurrent or pointwise model (Ramsay and Silverman, 2005), when the prediction of $Y_i(t)$ depends only on the value of the predictor at the same point t, and the historical model, when the prediction of $Y_i(t)$ depends on the value of the predictor for $s \in [0,t]$ (Morris, 2015).

2.4 Multidimensional functional data analysis

A great deal of work in functional regression is focused on 1-dimensional functional data. i.e. smooth curves $X(t)$ observed on some grid of points (as $t \in [0,1]$). Nevertheless, a growing interest is on more complex settings, where images with multidimensional domain are considered. In this direction, the natural approach would be to repurpose methods originally designed for 1D functional data by adopting for example a different system of basis functions which would incorporate and solve the issue of having more than one dimension. This is conceptually grounded and in many cases also practically feasible, but new technical challenges arise when the combination of the number of measurements (e.g. pixels/voxels) per subject and the sample size makes even loading all the data in memory at one time an unviable option (Reiss et al., 2017).

In terms of functional principal component analysis, Zipunnikov et al. (2011) propose a high-dimensional multilevel FPCA aimed for densely-observed images recorded at multiple visits for each subject. The proposed solution relies on block-partitioning the matrix containing all the subject-specific measurements and then perform SVD sequentially on each block. A best linear unbiased prediction then returns the estimates for scores at cross-sectional and longitudinal level. The method is recommended for balanced designs with a moderate number of subjects and visits.

In the setting of multivariate FDA, when each subject has 2 or more functional elements, FPCA with multidimensional imaging data is presented in Happ and Greven (2018). The relationship between univariate and multivariate Karhunen–Loève decomposition is derived. After computing a univariate FPCA for each element, the matrix that contains the univariate scores undergo an additional eigenanalysis. The eigenvectors obtained in this way are then used as weights to compute multivariate scores and eigenfunctions. This approach works for functional data of different dimensionality (e.g. a curve and an image) and comes with a `R` package called `MFPCA` (Happ, 2018) which offers also a great deal of basis functions for multidimensional data (especially or 2D images, ranging from penalised splines to the functional higher-order PCA in Allen, 2013 and tensor cosine basis).

The field of scalar-on-function regression has seen many more contributions pertaining imaging covariates. Reiss and Ogden (2010) introduced a mixed approach for dimension reduction in functional principal component regression (FPCR) with a roughness penalty. The functional regression coefficient is first projected onto the span of a B-spline basis and then reduced by considering the first M^* principal components. The images are smoothed with a radial cubic B-spline basis with a thin plate penalty. The fitting method for generalised functional linear models is the iteratively reweighted least squares (IRLS) method as in common GLM. The smoothing parameter is selected via GCV, AIC, corrected AIC or REML. Reiss and Ogden (2010) propose also some simultaneous testing by inverting the simultaneous confidence bands derived by nonparametric bootstrap. This method is claimed to have a conceptual advantage over Statistical Parametric Mapping (SPM) in those cases in which the observed quantity in the image is supposed to be causing the disease (Reiss and Ogden, 2010); moreover, it offers a straightforward way to predict the disease given the image and to assess the contribution of each predictor.

Wang et al. (2014) consider the case of functional regression where a 3D brain image is the predictor. A tensor product of 1D Haar wavelets basis expansion pro-

vides a flexible and parsimonious way to obtain sparsity while taking account of the spatial correlation at different levels of smoothness. In a similar setting, Wang et al. (2017) study a relaxation of functional regression models where the coefficient image is assumed to be piecewise smooth image with unknown jumps and edges, by recurring to the notion of total variation. The approach by Park et al. (2016) focuses on finding a principled way to partition the domain of the regression coefficient. A sequential segmentation procedure based on an approximation of the spatial correlation is provided, then the selection algorithm is applied until the improvement in the cross-validation prediction error becomes negligible.

Bayesian approaches have been also proposed in the scalar-on-images literature. Goldsmith et al. (2014) select some specific prior distributions to get a sparse and smooth estimate. Given a latent binary indicator which detects those locations that are predictive of a scalar outcome, an Ising prior distribution is applied to estimate contiguous predictive regions and an intrinsic Gaussian Markov random field prior distribution controls the smoothness of the non-null coefficients. A fast single-site Gibbs sampler is used to fit the model. The parameters of the Ising prior and the variances are tuned via cross-validation instead of being derived using hyperpriors in order to improve the computational speed. The simulations show that this model works particularly well in detecting nonpredictive regions (true negative values). Spatial variable selection is also addressed by Kang et al. (2018) by using a soft-thresholding function for a latent Gaussian process. This choice is claimed to be more efficient with respect to the Ising model, especially for large datasets, and enforces a gradual transition between predictive and non-predictive regions. Metropolis–Hastings within Gibbs is used to sample from the posterior distribution.

Smoothness and sparsity assumptions of several scalar-on-image regression models are evaluated in Happ et al. (2018). The models are categorised in *fixed basis functions* expansion (penalised B-splines or wavelets), *data-driven basis functions* expansion (principal component regression), *combined methods* (such as FPCR by Reiss and Ogden, 2010), PCR and PLS in wavelet space), *random field methods* (like in Goldsmith et al., 2014). Smoothness, sparsity and projection (the assumption that a set of basis functions is a space where the coefficient image lies) are defined *underlying assumptions*; in the modelling framework, smoothness is governed by penalties or prior distribution (in a Bayesian setting) while sparsity translates into a variable selection method or restrictions on the number of principal components included. The findings show that while different assumptions

produce quite different results in terms of estimation accuracy, the predictive performance seems not to vary dramatically across models.

In the image-on-scalar literature, the recent contribution of Yu et al. (2021) proposes an expansion of the outcome in terms of flexible multivariate splines over triangulations, in order to deal efficiently with the irregular domain of the images. It is worth mentioning also some innovative applications of functional data analysis on more complex domains: for example, Lila and Aston (2020) consider the setting where functions are evaluated on surfaces and the variability between subjects arises also in terms of differences in the domain of the functions and devise a specific FPCA for this application.

Chapter 3

Quantifying uncertainty in brain-predicted age using scalar-on-image quantile regression

3.1 Introduction

The process of brain ageing is known to be associated to a general decline in cognitive functions and higher risk of neurodegenerative diseases (Yankner et al., 2008, Denver and McClean, 2018). In some cases, both ageing and dementia affect the same areas in the brain (Lockhart and DeCarli, 2014). For these reasons, a deeper understanding of brain ageing in healthy conditions could potentially improve the diagnosis of neurodegeneration at early stages.

Neuroimaging provides a non-invasive and safe way to study brain structure and functioning. A large part of the research in neuroimaging data analysis has

been focused on explanatory analyses aimed at describing the relationship between the brain and some variables of interest (such as neurodegenerative diseases, sex, physical activity). With the advent of imaging databases with larger size, a prediction-oriented focus has been also considered, in order to detect individual differences among subjects that could be used in clinical practice (for example Yoo et al., 2018, Zhou et al., 2019).

The study of brain ageing has recently gained attention in the neuroscientific community thanks to the availability of this large amount of data and of computational tools for their analysis. A growing body of research employs neuroimaging to develop a biomarker of individual brain health, called "brain age" (Franke and Gaser, 2019, Cole et al., 2017). In the absence of a clear definition and assessment of biological brain age, a brain-derived prediction of chronological age is considered. In order to be integrated in clinical practice, a brain age biomarker should be easily accessible from brain data (or better, images), harmless for the subjects, computationally not demanding and correlated with other brain health indicators (Franke and Gaser, 2019). In addition, since there is a high variability between subjects in terms of their brain ageing, a useful biomarker should predict cognitive decline better than the chronological age itself.

In this work we propose a statistically grounded workflow that produces brain age individual predictions from 3-dimensional brain images. Furthermore, we go beyond simple point predictions by also providing prediction intervals of the brain age to quantify the uncertainty. Our model is trained on a control group with no ongoing brain diseases in order to avoid spurious effects due to other conditions. The same model can be used to predict age in neurodegenerative diseases, in order to provide a "baseline" or "normative" brain age, whose difference from the individual chronological age (brain-predicted age difference or *brain-PAD* as in Cole et al., 2017) might inform about the extent of the effect induced by the pathology.

In addition, reporting a prediction interval alongside a point estimate offers another potential binary biomarker (whether the chronological age falls within it). Since the width of the prediction interval is different for each subject, the same brainPAD could be interpreted in different ways in light of its location with respect to the individual prediction limits. The joint use of point and interval brain age predictions could therefore be employed to easily assess departures from a typical ageing profile.

The approach developed in this paper is based on modern statistical tools. In order to use 3D brain images without the need to summarise information by re-

gions of interest, a functional data analysis (FDA) framework is adopted (Ramsay and Silverman, 2005, Horváth and Kokoszka, 2012). Functional data get this name because the observation for each statistical unit is a function[1] (a curve, surface, or image). These data are usually considered as infinite dimensional and intrinsically continuous, even if the data collection process reduces them to a discrete series of observed points (Ramsay and Silverman, 2005, Section 3.2). In other words, the whole function is considered as the object of interest, and not only the specific value observed at a discrete location for each image. A common model in FDA is scalar-on-function regression (see Morris, 2015, Reiss et al., 2017 for reviews), which provides an effective way to predict a scalar quantity of interest from a functional observation, by fitting a regression model using the whole function as a covariate. In our context we call it *scalar-on-image regression*. The non-identifiability problem (Happ et al., 2018) arising from having sample size lower than the number of voxels for each image can be attenuated by imposing some assumptions on the data generating process (for example smoothness).

We obtain prediction intervals by integrating the FDA framework with quantile regression (Koenker and Bassett, 1978, Koenker and Hallock, 2001), a model that is largely used in fields such as economics (Fitzenberger et al., 2013) and ecology (Cade and Noon, 2003) to derive a more complete picture of the relationship between a covariate and the response variable. Quantile regression does not model the expected value (or a function of it) of the outcome of interest given the predictors, but some selected quantiles of the conditional distribution (for example the median). This model can be adapted for functional covariates: in a functional quantile regression model we explore the linear relationship between a certain quantile of the outcome and the 3D image. By fitting several quantile regression models we can build the prediction intervals given the covariates. Prediction intervals from quantile regression (or similar models) have received some attention in recent decades (Zhou and Portnoy, 1996, Meinshausen, 2006, Mayr et al., 2012), but not within the framework of functional data. In addition, the scalar-on-image quantile regression generates a regression coefficient with the same dimensionality as the brain image, providing an interpretable map that shows how the changes in each brain structure are related to the predicted age.

Our FDA-based approach departs considerably from other methods that are commonly used in the neuroimaging literature. The state-of-the-art method in neuroimaging data analysis is the so-called *mass-univariate* approach and it is

[1] the word "functional" in this case is used in a mathematical sense and is not related to functional MRI.

implemented in the *Statistical Parametric Mapping* software (Ashburner et al., 2014). A model is fitted to predict the signal at each voxel independently using the clinical or demographic information as covariate, then a significance map is produced (see for further details Friston et al., 1994, Penny et al., 2011). Although computationally efficient, this approach does not explicitly model the spatial correlation of adjacent pixels and is not tailored for prediction purposes (Reiss and Ogden, 2010). The functional data approach allows instead the incorporation of the spatial structure by using smoothing techniques and in this way the fit of a global model for a scalar outcome given the entire brain image.

Another popular approach is based on machine learning algorithms. Franke and Gaser (2019) review a collection of studies published in the last decade based on a technique called relevance vector regression. They review a number of studies that examine associations with brain age, including effects of meditation and playing an instrument. Cole et al. (2019) collects a larger number of studies dealing with brain age prediction conducted from 2007 to 2018 with different imaging modalities and pathologies. Many of them adopt support vector regression (as the ones listed in Franke et al., 2012, Franke and Gaser, 2019 or Sone et al., 2019) or more recently Gaussian processes and convolutional neural networks (Cole et al., 2017, Cole, 2017, Varatharajah et al., 2018, Wang et al., 2019). A comparison between the predictive performances of these methods is difficult due to the use of different datasets and different age ranges, but according to Cole et al. (2019) the choice of the algorithm does not seem to play a fundamental role. However, these approaches provide only a point prediction with little knowledge of the internal procedure that returned it, and in particular deep learning methods are often criticised as "black boxes". Our approach attempts to provide a better picture of the set of information on which brain age is based, introducing a straightforward quantification of uncertainty and at the same time producing a visual display of the regions that are most relevant for the prediction. In addition, the features of each step of the workflow proposed here can be evaluated, therefore improving the interpretability of the results. This last aspect is crucial in medical sciences and is particularly welcome for predictive modelling in neuroscience (Scheinost et al., 2019).

Another important distinction with the available literature on brain age prediction relates to the imaging techniques used. Although several models use functional imaging or multiple modalities, a large share of studies focused on structural magnetic resonance imaging (MRI), in particular T1-weighted images, usually segmentated into grey and white matter. Unprocessed MR images have also

been employed with success (Cole et al., 2017). In this work we still remain in the family of structural imaging but we use tensor-based morphometry (TBM) images, that are obtained after a transformation of standard MRI images. TBM images give information about relative volumes of brain structures with respect to a common template; for this reason the images are all spatially registered. TBM quantifies volumetric differences in brain tissue for each voxel and is therefore specifically aimed at assessing the level of local cortical atrophy which might help to study brain degeneration for different diseases (Hua et al., 2008). To the best of our knowledge, this is the first study addressing brain age prediction from TBM images. The dataset used in this manuscript comes from the Alzheimer's Disease Neuroimaging Initiative (ADNI, Mueller et al., 2005).

The work is structured as follows. Section 3.2 gives an overview of functional data analysis and quantile regression. Section 3.2.3 introduces the plan of the analysis and discusses details of the implementation. The main characteristics of the ADNI dataset are described in Section 3.3, while the results of the analysis are reported in Section 3.4 in terms of the predictions, their robustness with respect to the choices of the parameters in the model and their correlation with standard cognitive measures. Finally, Section 3.5 discusses the main findings, summarises the work and briefly introduces further research directions.

3.2 Materials and Methods

3.2.1 Quantile regression

Regression models are used to study the relationship between some fixed and known predictors $\boldsymbol{Z} = (\boldsymbol{z}_1, ..., \boldsymbol{z}_{\check{P}})^\mathsf{T} \in \mathbb{R}^{\check{P}}$ and an outcome variable Y. For example, linear models are used to evaluate the change in the expected value of the continuous outcome conditioned on the values of the predictors, under specific assumptions on the error term. Nevertheless, there are occasions in which either these assumptions do not hold (for example, when there is heteroskedasticity in the residuals) or simply the main interest is to model specific quantiles of the conditional distribution of the response variable in order to produce a deeper analysis of the randomness of $Y|Z$ that goes beyond the conditional mean[2]. Quantile regression (Koenker and Bassett, 1978) can effectively deal with these cases by

[2]From Mosteller and Tukey (1977): "Just as the mean gives an incomplete picture of a single distribution, so the regression curve gives a correspondingly incomplete picture for a set of distributions."

specifying the model:

$$\mathsf{Q}_\tau(Y|\boldsymbol{Z}) = \alpha_\tau + \sum_{\check{p}=1}^{\check{P}} z_{\check{p}}\delta_{\check{p},\tau}, \qquad \tau \in (0,1), \tag{3.1}$$

where $\mathsf{Q}_\tau(Y|\boldsymbol{Z})$ is the τ-th conditional quantile of $Y|Z$ defined as

$$\mathsf{Q}_\tau(Y|\boldsymbol{Z}=\boldsymbol{z}) = \inf\{y : F_{Y|\boldsymbol{Z}}(y|\boldsymbol{z}) \geq \tau\} \tag{3.2}$$

and

$$F_{Y|Z}(y|\boldsymbol{z}) = \Pr(Y \leq y|\boldsymbol{z}) \tag{3.3}$$

is the conditional cumulative distribution function of $Y|Z$. For example, $\mathsf{Q}_{0.5}(Y|Z)$ is the median of the conditional distribution of $Y|Z$. The interpretation of $\delta_{\check{p},\tau}$ is similar to the one in linear models: it corresponds to the marginal effect on the conditional quantile due to a one-unit increment in the $\check{p}$-th covariate.

Given N observations, the estimation procedure for the model in Equation (3.1) is based on the following minimisation problem:

$$(\hat{\alpha}_\tau, \hat{\delta}_{1,\tau}, \ldots, \hat{\delta}_{\check{P},\tau}) = \underset{a,\delta_1,\ldots,\delta_{\check{P}}}{\arg\min} \left[\sum_{i=1}^{N} \rho_\tau \left(y_i - \alpha - \sum_{\check{p}=1}^{\check{P}} z_{i\check{p}}\delta_{\check{p}} \right) \right], \tag{3.4}$$

where $\rho_\tau(u) = \left[\tau - \mathbb{1}_{\{u\leq 0\}}\right] u$ is the check (or quantile loss) function (Koenker and Bassett, 1978). There is a relationship between the linear formulation $Y = Z\delta + \varepsilon$ and the quantile formulation in Equation (3.1). Under a linear data generating process $Y = \alpha + Z\delta + \varepsilon$ with known α and δ, we can write the conditional quantile restriction

$$\mathsf{Q}_\tau(Y|\boldsymbol{Z}) = \alpha + \boldsymbol{Z}\delta + F_\varepsilon^{-1}(\tau), \qquad \tau \in (0,1) \tag{3.5}$$

with ε being the mean zero random term of the model with cumulative distribution function (CDF) F_ε. In this simple setting, the marginal effect of the covariate is constant across quantiles. Note that the result in Equation (3.5) holds for any distribution of the error term. Quantile regression can nonetheless accommodate more complicated data generating processes, like for example the location-scale model where ε is replaced by $\sigma(Z)\varepsilon$, with $\sigma(Z) > 0$ and $\varepsilon \perp\!\!\!\perp Z$. In this case the variance of the random term depends on Z and it can be shown that the estimated slope in the quantile regression model will be governed by the quantiles of ε.

All the quantile regression models return as output a prediction at a specific quantile level. For example, the model with $\tau = 0.5$ gives the conditional median prediction for each experimental unit given particular values of the covariates. Predictive accuracy of the conditional median can be measured through the mean absolute error (MAE) and the root mean square error (RMSE) between the point predictions and the observed responses. By fitting a model for several values of τ, we can also build prediction intervals for new observations $(y^*, \boldsymbol{z}^*)$ (Davino et al., 2013, Mayr et al., 2012). For example, if we fit a model on the same data for two quantile levels $\tau_1 = \tilde{w}/2$ and $\tau_2 = 1 - \tilde{w}/2$ (with $\tilde{w} \in (0,1)$), the interval

$$\text{PI}_{1-\tilde{w}}(\boldsymbol{z}^*) = \left(\hat{\text{Q}}_{\tau_1}(Y|Z = \boldsymbol{z}^*), \hat{\text{Q}}_{\tau_2}(Y|\boldsymbol{Z} = \boldsymbol{z}^*)\right) \tag{3.6}$$

should contain the observed response value for new data $(1-\tilde{w})100\%$ of the time (provided Equation (3.1) is true). For example, a 90% prediction interval can be obtained by fitting a model for $\tau_1 = 0.05$ and $\tau_2 = 0.95$. This prediction model can effectively handle heteroskedasticity or skewness, since in quantile regression there are no assumptions on the response distribution: using simulated data Davino et al. (2013) provide examples in which prediction intervals obtained via quantile regression achieve the nominal levels where ordinary least squares prediction intervals fail. This is also confirmed theoretically in Zhou and Portnoy (1996): the coverage probability tends to $1 - \tilde{w}$ with an error of $O(N^{-1/2})$, as the sample size of the training set $N \to \infty$.

3.2.2 Functional quantile regression

A large body of literature has been developed in order to translate regression models into the functional data framework, where the observations are no longer considered to lie in $\mathbb{R}^d$, but they are realisations of a random function $X \in L^2(\mathcal{T})$, the space of square-integrable functions.

For example, functional GLMs are now well established in the theory, both in the frequentist and Bayesian approaches (see for example Müller and Stadtmüller, 2005 and Crainiceanu et al., 2009). Quantile regression (Koenker and Bassett, 1978) has also been extended in the functional data paradigm: first with Cardot et al. (2005), then with Kato (2012) and Yao et al. (2017), the model has been readapted for the case of functional covariates with scalar response. The model illustrated in Kato (2012) (which provides the main reference for this section) shares the main characteristics with the scalar-on-function regression of Müller

and Stadtmüller (2005), except for the assumption that the conditional quantile is a linear function of the (centered) covariates. In particular, the conditional quantile of the response is expressed as a linear function of the scalar product between the functional data and a coefficient function $\beta_\tau \in L^2(T)$:

$$\mathsf{Q}_\tau(Y|X) = \alpha_\tau + \int_{\mathcal{T}} X(t)\beta_\tau(t)dt, \qquad \tau \in (0,1). \tag{3.7}$$

The functional nature of the coefficient makes its interpretation less straightforward than in standard regression. In the regions where $\beta_\tau(t) = 0$ any increment in the covariate produces no marginal change on the quantile of the conditional distribution $Y|X$. On the other hand, if $\beta_\tau(t)$ is constant over a region $\mathcal{T}^* \subset \mathcal{T}$ and null elsewhere, then only the region $\mathcal{T}^*$ plays a role in the prediction of the conditional quantile. Despite the differences between quantile and linear scalar-on-function regression, the same difficulties of the interpretation of the functional coefficients discussed in James et al. (2009) apply. The model can easily accommodate scalar covariates $z_1, ..., z_{\check{P}}$ (see for example Yao et al., 2017):

$$\mathsf{Q}_\tau(Y|X, \boldsymbol{Z}) = \alpha_\tau + \int_{\mathcal{T}} X(t)\beta_\tau(t)dt + \sum_{\check{p}=1}^{\check{P}} z_{\check{p}}\delta_{\check{p},\tau}, \qquad \tau \in (0,1). \tag{3.8}$$

In order to estimate the parameters in Equation (3.7), both the predictors and the coefficient functions are represented in the truncated Karhunen–Loève expansion in Equation (2.20):

$$\begin{aligned} X_i(t) &\approx \sum_{m=1}^{M} \nu_{im}\psi_m(t) \\ \beta_\tau(t) &\approx \sum_{\check{m}=1}^{M} b_{\check{m},\tau}\psi_{\check{m}}(t). \end{aligned} \tag{3.9}$$

Thanks to the orthonormality of the eigenfunctions ψ_m,

$$\begin{aligned} \int_{\mathcal{T}} X_i(t)\beta_\tau(t)dt &\approx \sum_{m=1}^{M}\sum_{\check{m}=1}^{M} \nu_{im}b_{\check{m},\tau}\int_{\mathcal{T}} \psi_m(t)\psi_{\check{m}}(t)dt \\ &= \sum_{m=1}^{M} \nu_{im}b_{\check{m},\tau}. \end{aligned} \tag{3.10}$$

Thus the functional model in (3.7) becomes a standard quantile regression problem of the form

$$\mathsf{Q}_\tau(Y|X) = \alpha_\tau + \sum_{m=1}^{M} \nu_{im} b_{m,\tau}, \tag{3.11}$$

where α_τ and $b_{1,\tau}, \ldots, b_{m,\tau}$ are estimated as in Equation (3.4). The estimated functional coefficient is then reconstructed by computing

$$\hat{\beta}_\tau(t) = \sum_{m=1}^{M} \hat{b}_{m,\tau} \psi_m(t); \tag{3.12}$$

for a given τ the estimated value for the quantile function is obtained by plugging in the estimated coefficient into (3.7):

$$\hat{\mathsf{Q}}_\tau(Y|X) = \hat{\alpha}_\tau + \int_{\mathcal{T}} X(t)\hat{\beta}_\tau(t)dt. \tag{3.13}$$

In this functional principal components regression (FPCR) setting, the number of principal components M to be used as regressors controls the smoothness and the approximation error with respect to the real images. The choice of M could be automated by using information criteria or percentage of variance explained; nevertheless, there is no guarantee that the first M components (which explain the most of the variability of X) are also able to capture effectively the relationship between the functional predictor and the scalar response (Febrero-Bande et al., 2017, Delaigle and Hall, 2012). For this reason, a simple option could be to select M such that a very large share of explained variability is represented and then use LASSO regularisation within the quantile regression model (Belloni and Chernozhukov, 2011, Wang, 2013). The regularisation might produce a different subset of selected variables across different quantile levels τ. Since for each τ a different model has to be fitted, the plug-in estimator $\hat{\mathsf{Q}}_\tau(Y|X)$ is not guaranteed to be monotonically increasing in τ as the conditional quantile function $\mathsf{Q}_\tau(Y|X)$ is by construction.

It must be considered that the bias introduced by the penalised estimation could harm the interpretability of the coefficients for each covariate. A way to solve this issue is the post-ℓ_1 quantile regression, where LASSO is used only for model selection and then a vanilla quantile regression model is fitted using only the covariates selected. This approach guarantees better convergence rates and could reduce the bias (Belloni and Chernozhukov, 2011).

3.2.3 Data analysis workflow

3.2.3.1 Imaging

The brain images are acquired using structural MRI. This workflow does not depend on any specific preprocessing stages, except for intersubject registration to an atlas image, such that voxels from different images are aligned.

More transformations can be operated on the structural MR images. For example, the analysis can be based on tensor-based morphometry (TBM) images. TBM is an image technique that aims at showing local differences in brain volume from structural imaging. In a cross-sectional setting (one image for each subject), each image is aligned to a common MRI template called *minimal deformation template* (MDT). The deformation induced by this alignment can be represented by a function that maps a 3-dimensional point in the template to the corresponding one in the individual image. The Jacobian matrix of the deformation can be used to inform about volume differences in terms of shearing, stretching and rotation. The determinant of the Jacobian matrix for each voxel is then a summary of local relative volumes compared to the MDT: a value greater than 1 indicates expansion, while a value less than 1 means contraction. Further details about TBM are available in Ashburner and Friston (2004).

In order to reduce the dimensionality of the problem, the voxels outside the brain can be excluded from the analysis imposing a mask on the images. We used FSL (through its R interface `fslr`, Muschelli et al., 2015) to obtain a mask on the template image with smooth boundaries.

3.2.3.2 Basis expansion

A common assumption in FDA is that the observed data are a noisy, discretised version of the true underlying signal function that is of interest in the analysis. In other words, the values observed at a specific voxel may be contaminated with some measurement error that could have an impact on the spatial correlation structure within the images. Removing this measurement error leads therefore then to smoother images, improving the performances of FPCA.

For this reason, nonparametric basis expansion techniques such as B-splines or wavelets are usually employed. The latter are chosen mainly when the underlying function is thought to be characterised by rapid changes in behavior (Ramsay and Silverman, 2005); B-splines are instead preferred for their properties (compact support, unit sum) when less abrupt changes in the function are expected.

In this case, TBM images are already smooth by construction, so we can use B-spline basis functions with the main aim to obtain a parsimonious representation (under the fairly safe assumption that the main sources of error have been already removed).

In order to get a 3-dimensional basis function, a tensor product of univariate B-spline basis functions is built as described in Section 2.2.3. In compact form, all the N images are represented by the product of the $N \times K$ coefficient matrix $\tilde{\boldsymbol{C}}$ and the matrix of basis functions $\boldsymbol{\Phi}$. We center the projected data (equivalent to centering the raw data since the projection is linear). This apparently negligible aspect is actually very relevant in the big data context as it allows to parallelise the basis expansion stages without the need to import and store simultaneously all the images. We call the centered coefficient matrix $\boldsymbol{C}$.

In this work we used a 3D tensor product of quadratic B-spline univariate basis functions with equidistant knots. The number of knots (or analogously their spacing) can be fixed in advance, but a poor choice might heavily affect the number of basis functions that are needed to represent the functions and consecutively the computational time and the quality of projection. For this reason a preliminary study on a subset of the data is recommended. Outcomes of interest for this preliminary study could be the number of non-zero basis functions within the masked image, the average time needed for the projection of an image and the R^2 value obtained from the regression of each image using as design matrix the matrix of basis functions. The latter value can be interpreted as a proportion of variance explained. At this stage, it is highly recommended to retain as much variability as possible: a 0.95 threshold for R^2 should work for many applications and should ensure a manageable set of basis functions. Alternative criteria could be established in terms of full width at half maximum (FWHM).

3.2.3.3 Functional PCA

The coefficients of the projection are the quantities needed to solve the eigen-decomposition problem in Equation (2.21). In this section, we rely heavily on Ramsay and Silverman (2005, Section 8.4.2), with minor modifications to make this high dimensional problem computationally feasible. The procedure is described also in Chen et al. (2018).

The (corrected) sample variance-covariance function can be written as

$$\hat{\gamma}(s,t) = \frac{1}{N-1}\boldsymbol{\phi}(s)^{\mathsf{T}}\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C}\boldsymbol{\phi}(t) \tag{3.14}$$

using the same decomposition in (2.28). Suppose then that the eigenfunctions in Equation (2.21) can be expressed as linear combinations of the same basis functions $\boldsymbol{\Phi}$:

$$\psi(s) = \sum_{k=1}^{K} \xi_k \phi_k(s) = \boldsymbol{\phi}(s)^\mathsf{T} \boldsymbol{\xi}. \tag{3.15}$$

Then the eigenanalysis of the covariance operator described in Equation (2.21) takes the following form:

$$\int_T \left[\frac{1}{N-1} \boldsymbol{\phi}(s)^\mathsf{T} \boldsymbol{C}^\mathsf{T} \boldsymbol{C} \boldsymbol{\phi}(t) \right] \left[\boldsymbol{\phi}(\boldsymbol{t})^\mathsf{T} \boldsymbol{\xi} \right] dt = \lambda \boldsymbol{\phi}(s)^\mathsf{T} \boldsymbol{\xi}. \tag{3.16}$$

Denoting by $\boldsymbol{W}_\phi$ the $K \times K$ symmetric basis product matrix with elements

$$w_{kl} = \langle \boldsymbol{\Phi}_k, \boldsymbol{\Phi}_l \rangle, \tag{3.17}$$

Equation (3.16) can be rewritten as

$$\frac{1}{N-1} \boldsymbol{\phi}(\boldsymbol{s})^\mathsf{T} \boldsymbol{C}^\mathsf{T} \boldsymbol{C} \boldsymbol{W}_\phi \boldsymbol{\xi} = \lambda \boldsymbol{\phi}(s)^\mathsf{T} \boldsymbol{\xi}. \tag{3.18}$$

The entries in $\boldsymbol{W}_\phi$ are usually computed with some numerical quadrature rules (Ramsay and Silverman, 2005) but these procedures are computationally demanding in our 3D context. The cross product, although less accurate at the boundaries with respect to the trapezoidal rule, offers a good result in shorter time. Simplifying both sides of Equation (3.18) by $\boldsymbol{\phi}(s)^\mathsf{T}$ (the relationship must hold for all s) we obtain

$$\frac{1}{N-1} \boldsymbol{C}^\mathsf{T} \boldsymbol{C} \boldsymbol{W}_\phi \boldsymbol{\xi} = \lambda \boldsymbol{\xi}. \tag{3.19}$$

In order to get orthonormal eigenfunctions, some constraints must be imposed:

$$\xi_j^\mathsf{T} \boldsymbol{W}_\phi \xi_j = 1 \qquad \text{and} \qquad \xi_j^\mathsf{T} \boldsymbol{W}_\phi \xi_k = 0. \tag{3.20}$$

These are fulfilled by setting $\boldsymbol{u} = \boldsymbol{L}^\mathsf{T} \boldsymbol{\xi}$, where $\boldsymbol{L}$ is obtained through the Cholesky decomposition $\boldsymbol{W}_\phi = \boldsymbol{L}\boldsymbol{L}^\mathsf{T}$ (Ramsay and Silverman, 2005, p. 181); solving the equivalent problem

$$\frac{1}{N-1} \boldsymbol{L}^\mathsf{T} \boldsymbol{C}^\mathsf{T} \boldsymbol{C} \boldsymbol{L} \boldsymbol{u} = \boldsymbol{\lambda} \boldsymbol{u}, \tag{3.21}$$

the original eigenfunctions are obtained using $\boldsymbol{\xi} = \left(\boldsymbol{L}^\mathsf{T}\right)^{-1} \boldsymbol{u}$.

We note that for $\boldsymbol{A} = (N-1)^{-1/2} \boldsymbol{C}\boldsymbol{L}$ the eigendecomposition problem consists in finding the eigenvalues and eigenvectors of $\boldsymbol{A}^\mathsf{T}\boldsymbol{A}$. These can be obtained in

a computational efficient way by using the SVD of the matrix $\boldsymbol{A}$. In particular, the non-zero eigenvalues $\{\lambda_m\}$ are equal to the squared non-zero singular values, whereas the eigenvalues $\{\boldsymbol{u}_m\}$ of $\boldsymbol{A}^\mathsf{T}\boldsymbol{A}$ are equal to the right singular vectors of $\boldsymbol{A}$. The m-th score for the i-th image is then

$$\begin{aligned}\nu_{im} &= \langle X_i - \mu, \psi_m \rangle \\ &= \int_{\mathcal{T}} \left[\sum_l c_{il}\phi_l(t)\right]\left[\sum_{\check{l}} \xi_{m\check{l}}\phi_{\check{l}}(t)\right] dt \\ &= \boldsymbol{c}_i^\mathsf{T}\boldsymbol{W}_\phi\boldsymbol{\xi}_m. \end{aligned} \tag{3.22}$$

3.2.3.4 Functional Quantile Regression

The scores obtained after FPCA are plugged into a standard quantile regression problem. We create the design matrix for the quantile regression model using the first M scores for each image such that the first M eigenfunctions represent at least 80% of the variability within the sample (see Section 3.4.3 for a sensitivity analysis). LASSO regularisation can be applied within the quantile regression framework. The minimisation problem in Equation (3.4) can be readapted therefore to our situation by writing

$$(\hat{\alpha}_\tau, \hat{b}_{1,\tau}, ..., \hat{b}_{M,\tau}) = \underset{\alpha, b_1, ..., b_M}{\arg\min} \left\{ \sum_{i=1}^{n} \rho_\tau \left(y_i - \alpha - \sum_{m=1}^{M} \nu_{im} b_m \right) + \iota_{LASSO} \sum_{m=1}^{M} |b_m| \right\} \tag{3.23}$$

where ι_{LASSO} is the LASSO tuning parameter. For a specific value of ι_{LASSO}, a solution path is found, where the LASSO penalty will induce the shrinkage of the estimates towards zero, but also sparsity, as some estimates are exactly zero (Tibshirani, 1996).

Several R packages offer built-in functions that perform automatic selection of the tuning parameter. For this purpose, we use the package `rqPen` (Sherwood and Maidman, 2017), that produces penalized quantile regression models for a range of tuning parameters and then selects the one with minimum cross-validation error.

3.2.3.5 FPCA and functional quantile regression in a prediction setting

The scores are obtained by taking an inner product of each image with the eigenfunctions estimated on the training set. For this reason, they can be obtained for images from other datasets with the same formula, even if the properties of zero mean and variance equal to the eigenvalues apply only for the training dataset. The scores are in turn produced within the FPCA step, where the estimation of the eigenfunctions depends on the training data as well.

This workflow is aimed at deriving brain age prediction intervals for healthy individuals. This means that FPCA and functional quantile regression should be based on a dataset of control subjects. In order to get predictions for this dataset, 10-fold cross validation can be used, reducing in this way the risk of overfitting. Age predictions for subjects with neurodegenerative diseases can be obtained from the same normative model. In this case the full dataset of control subjects can be used for FPCA and functional quantile regression and the brain age is to be interpreted as the equivalent brain age of a healthy individual having the same brain image.

The R code implementing the workflow is available at `https://github.com/marcopalma3/neurofundata`.

3.2.3.6 Alternative models

The degree of smoothing in the basis expansion step can be controlled in different ways, by changing either the location or the numbers of knots. When the number of knots is equal to the number of voxels, we recover the original data, where the coefficient of the basis functions are just the observed values at each voxel. The analysis of the "unsmoothed" images can still be based on standard multivariate analysis techniques such as PCA and quantile regression, but it requires an increased computational effort. The data matrix containing the images as rows is indeed large (in our case the memory needed to store it is more than 6.4GB) and high performance computing tools are required to fit models on these data. In addition, quantile regression under memory constraints is receiving attention only recently (Chen et al., 2019), therefore the calculation of the prediction interval is not straightforward. A small amount of smoothing is recommended to reduce both the storage issues and the computational time required to train the model.

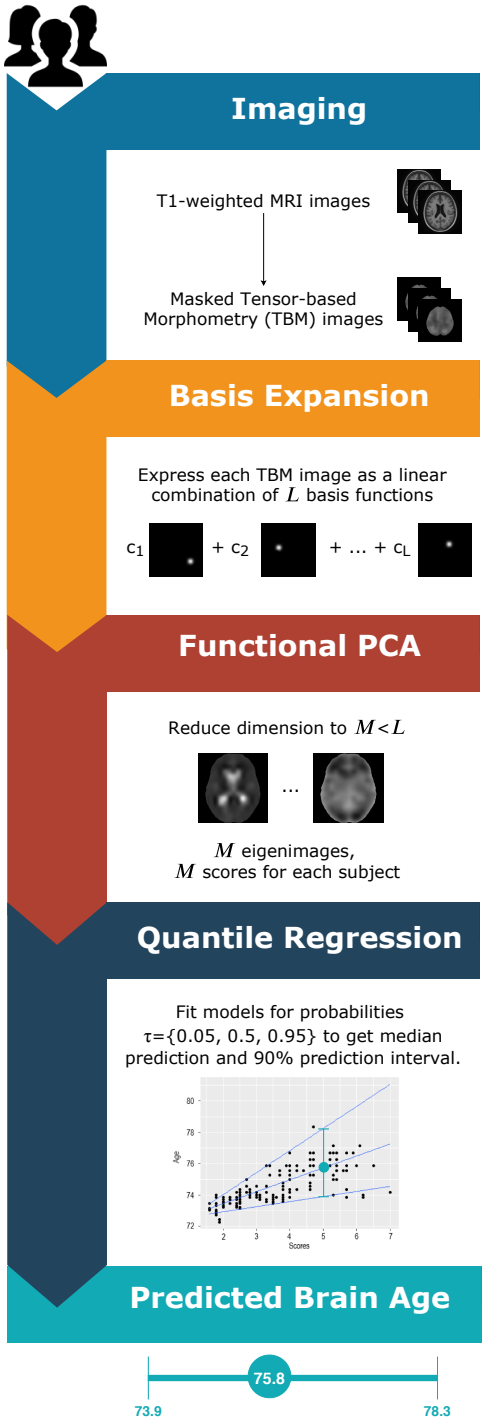

Figure 3.1: Flowchart of the analysis from the brain images to the predicted intervals.

3.3 Data

The workflow proposed in Section 3.2.3 is applied on a dataset coming from the Alzheimer's Disease Neuroimaging Initiative (ADNI, Mueller et al., 2005), that supports the investigation about biological markers to be employed to detect signs of Alzheimer's Disease (AD) in the brain at early stages. The sample used in this paper is made of 796 subjects, identified through an ID code, for which several demographic and clinical variables are measured. In this analysis, we will consider only the chronological age at the entry of the study (ranging from 59.90 to 89.60 years; mean age 75.60 ± 6.29) and their diagnosis: 180 subjects were diagnosed with AD, 387 with MCI (Mild Cognitive Impairment, considered as an intermediate stage between healthy condition and AD) and 229 people were belonging to a control group of cognitively normal (CN) subjects. The histogram of age by diagnosis group is displayed in Figure 3.2.

Diagnosis	N	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Control	229	59.90	72.30	75.60	75.87	78.50	89.60
MCI	387	60.10	70.85	75.60	75.30	80.40	89.30
AD	180	59.90	70.98	76.15	75.90	81.58	89.10

Table 3.1: Summary statistics for each diagnosis group. N is the number of subjects in each group. The second part of the table shows selected quantiles of age.

The functional part of the dataset consists of tensor-based morphometry images taken at the baseline of the study for each subject. In this dataset, the threshold 1 is rescaled to 1000 for computer number format reasons. Information about the preprocessing stages for the ADNI TBM dataset is available in Hua et al. (2013).

The analysis is based on the original 3D TBM scans ($220 \times 220 \times 220$, with voxel size equal to 1 mm^3). The conventional neurological orientation ("right is right") is used: the (x, y) axes of the images are set such that x increases from left to right and y increases from posterior to anterior.

The mean functions for each diagnosis are shown in Figure 3.3. MCI and AD patients share similar average brain volumes patterns (namely, expansion of the lateral ventricles and shrinkage almost everywhere else) even if the intensity of the expansion is higher for people with dementia. The expansion of the lateral ventricles is also visible in the healthy control mean function, but it is less pronounced. Conversely, the healthy control mean function shows other slightly expanded brain areas, such that the cerebellum and several regions in the posterior and frontal lobes. Further analyses based on the voxelwise variance functions per

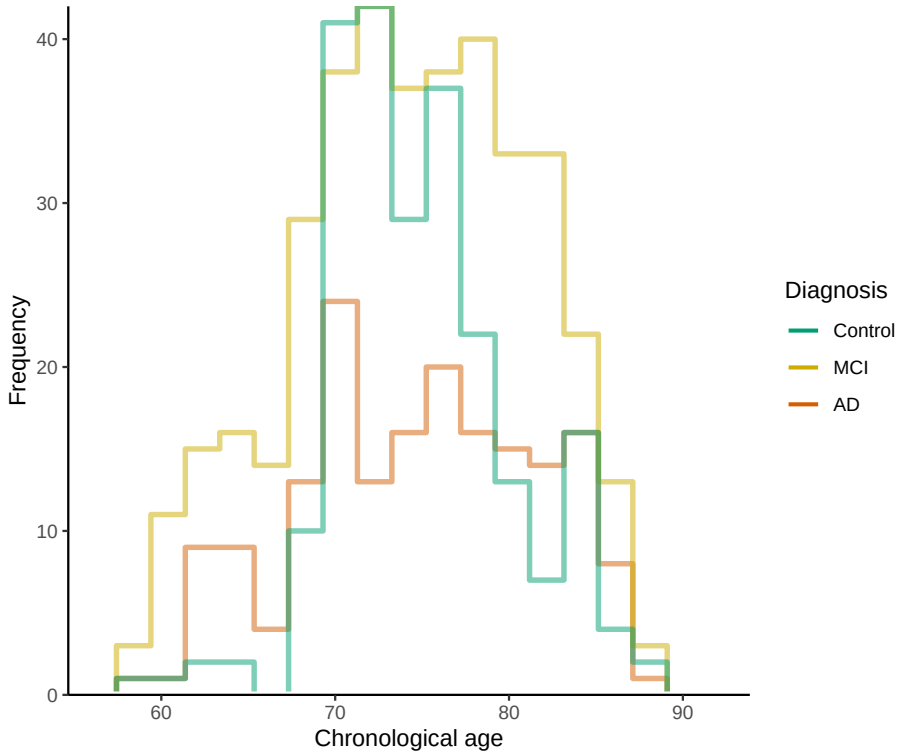

Figure 3.2: Histogram of age of the subjects in the sample, for each diagnosis. The number of bins has been fixed using the Freedman-Diaconis rule (Freedman and Diaconis, 1981).

each group show that the lateral ventricles are the areas with the highest variability in terms of volume expansion.

3.4 Results

3.4.1 Prediction accuracy

The preprocessed images are masked to remove unnecessary voxels for the analysis. A 3D smooth mask is obtained by smoothing the raw mask with a Gaussian kernel with standard deviation equal to 2 voxels (FWHM 4.7 mm) and thresholding it at 0.5, to regularise the boundary, producing just over 2 million nonzero voxels.

For the dataset at hand the B-splines projection with equidistant knots every 12 mm (equivalent to FWHM ≈ 15.33 mm) for each dimension allows to represent each image with R^2 approximately equal to 0.96. The number of B-spline functions in the tensor product that fall within the mask is 2694. In the current implementation, the process of importing one image into R and obtaining its B-spline coefficients takes approximately 30 seconds.

The eigendecomposition problem in Equation (2.21) solved for the dataset of healthy control subjects returns M $= 54$ eigenfunctions (which represent at least 80% of the variability within the sample) of which the first 3 are plotted in Figure 3.4. In analogy with standard PCA, a basic interpretation can be provided. The first eigenfunction clearly distinguishes the lateral ventricles from the rest of the brain. Subjects with high scores for this eigenfunctions will show stronger expansion within the lateral ventricles with respect to the mean function. Due to the similarities with the observed patterns in the mean function for the subjects with disease, it is likely that the scores for this eigenfunction computed for all the 796 subjects in the dataset are correlated with the diagnosis and with the chronological age, for the known interplay of the effects of these two factors. The second mode of variation refers instead to a more general expansion across the whole brain: in other words, it discriminates between individuals with bigger brains and those with smaller ones. For this reason, this component might account for some sex-related effects, as males have on average larger overall absolute brain than females (Ruigrok et al., 2014). The third eigenfunction weights negatively some of the internal parts of the brain. This component might therefore roughly distinguish white matter from the cortex, even if this interpretation is not very clear and can be influenced by the smoothing induced by the projection onto the ba-

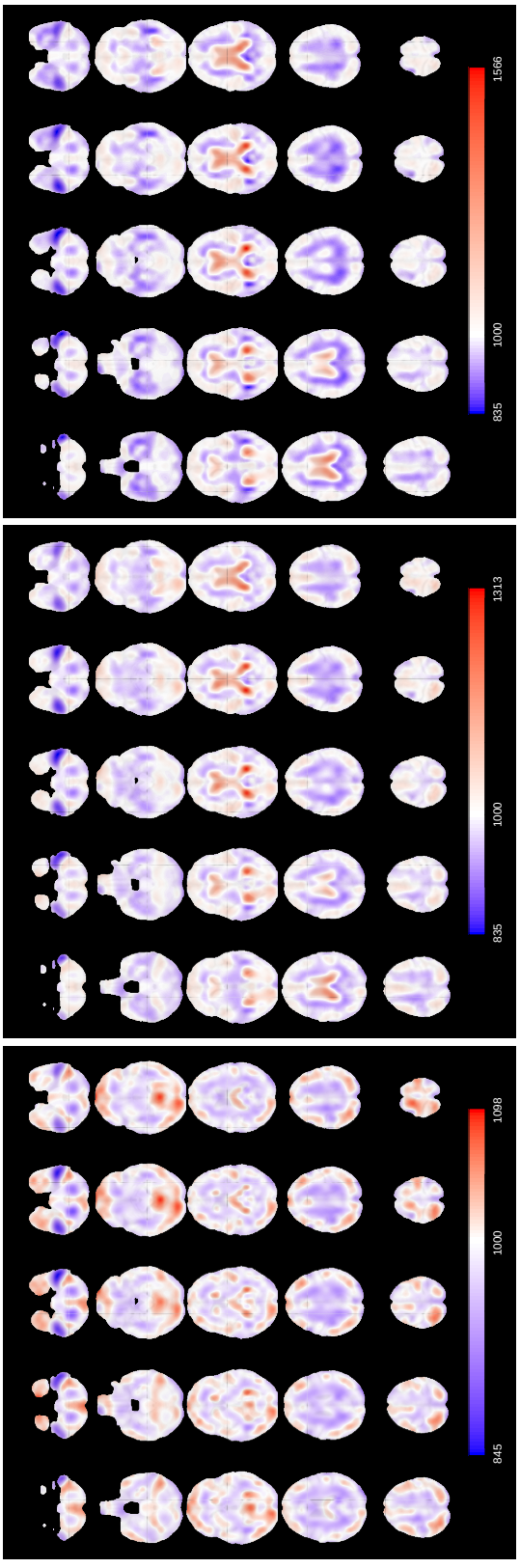

Figure 3.3: Axial slices of the mean images for each diagnosis (from left to right: Control, MCI, AD). Slices are ordered from bottom to top. The colours are overlaid on the corresponding slice of the MDT.

sis functions. The first 3 components account for 36.25% of the variance of the images of the healthy control group.

We compute the scores for MCI and AD individuals as the product of the centered images and the eigenfunctions in Figure 3.4. For the control subjects, we use 10-fold cross validation (with check function as loss function) to run FPCA, produce scores and fit the models such the predictions are obtained on held-out data. Quantile regression models for $\tau \in \{0.05, 0.5, 0.95\}$ are considered. Table 3.2 shows that the MAE and RMSE based on the difference between median brain-predicted age and chronological age are lower for control subjects than the other groups. This result is expected under the choice of a normative model that predicts brain age in absence of any diseases and indicates that the two subpopulations (controls vs. cases) show different ageing characteristics (if they were belonging to the same population, the MAE and RMSE would have been similar).

Diagnosis	N	MAE	RMSE	Cor	95% CI$_{Cor}$	$\hat{\pi}$	$*$-pos
Control	229	3.49	4.43	0.48	[0.37, 0.57]	0.86	0.05
MCI	387	4.99	6.12	0.46	[0.38, 0.54]	0.68	0.24
AD	180	5.16	6.27	0.38	[0.25, 0.50]	0.64	0.28

Table 3.2: Summary of the prediction results by diagnosis. Cor: correlation between predicted brain age and chronological age. CI$_{Cor}$: confidence interval for the correlation between predicted brain age and chronological age, obtained via Fisher-z transformation (Myers et al., 2013, Section 19.2). $\hat{\pi}$: sample coverage (proportion of cases for which the 90% prediction interval contain the chronological age). $*$-pos: proportion of cases for which the chronological age is less than the lower limit of the 90% prediction interval.

The MAE observed for the control group is 3.49, in line with other results obtained in the literature for other MRI datasets and different age ranges (Cole et al., 2019). In addition, as shown in Figure 3.5, the smoothed regression line for control subjects indicates that the average *brainPAD* (difference between predicted and chronological age) is close to zero for the whole age range, while it departs from it for the other groups in the predicted age range between 73 and 75. The statistical and clinical relevance of the age threshold after which the regression lines of the 3 groups overlap should be further evaluated. Prediction metrics do not improve after debiasing using post-ℓ_1 quantile regression.

We focus now our attention on the features of the 90% prediction intervals and the sample coverage. We observe that the actual sample coverage for control subjects is slightly lower than the nominal level. The groups with cognitive impairment show lower coverage with respect to the control group: the chronological

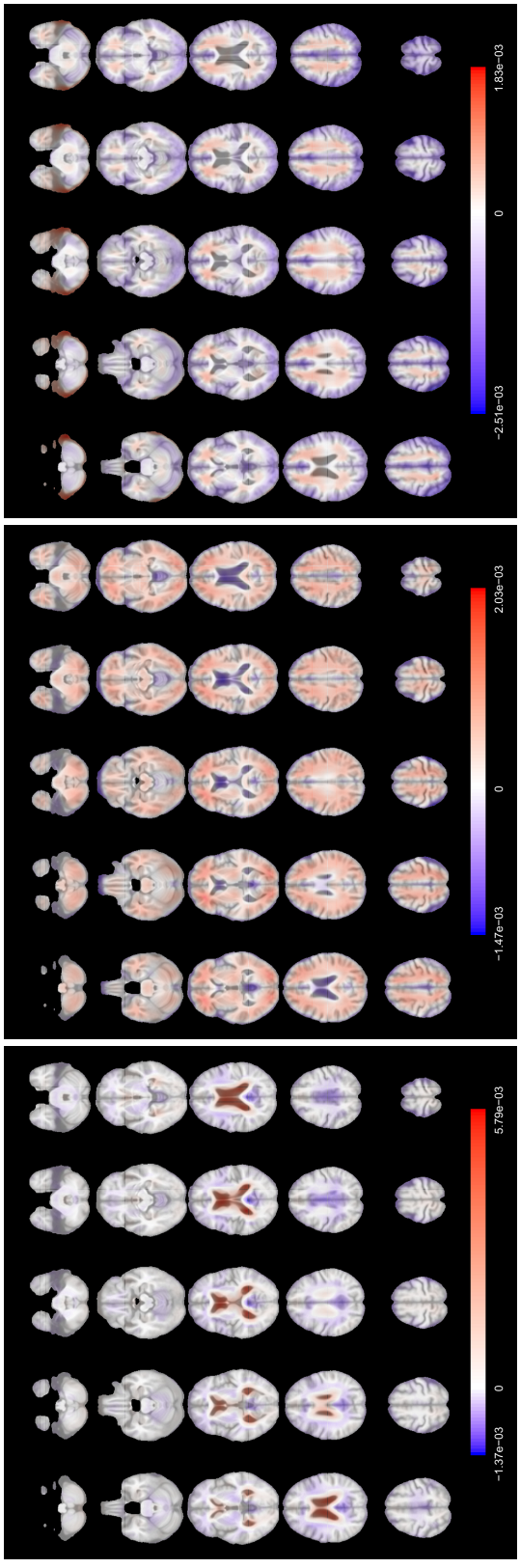

Figure 3.4: Axial slices of the first 3 eigenfunctions for the control subset. Slices are ordered from bottom to top. The colours are overlaid on the corresponding slice of the MDT. The eigenfunctions account respectively for 15.43%, 13.95%, 6.87% of the total variability. The signs of the eigenfunctions are determined on the basis of clinical interpretation.

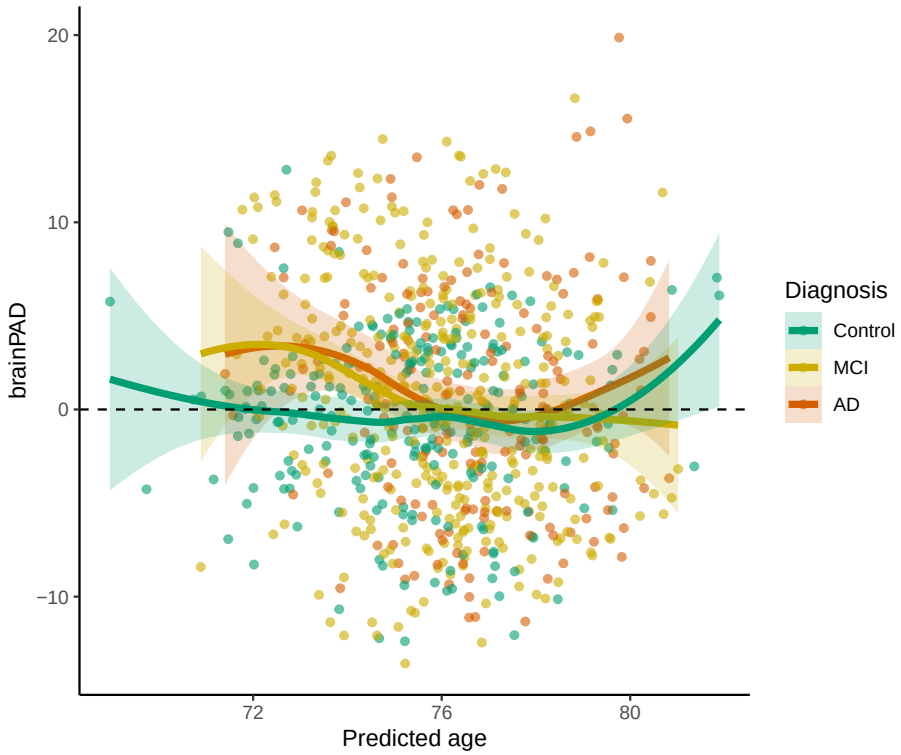

Figure 3.5: Plot of the brainPAD vs. predicted response. The coloured lines are local regression lines obtained with `loess` (locally estimated scatterplot smoothing) with span = 0.75 and 95% confidence bands.

ages of around 1 in 3 subjects with diseases do not fall in the prediction intervals obtained under the normative model. When we further analyse the direction of the discrepancy, we can define a "$*$-positive brainPAD" group (for which the chronological age is lower than the lower limit of the prediction interval, or equivalently with positive brainPAD and chronological age outside the prediction interval) and a "$*$-negative brainPAD" one (composed of those subjects with negative brainPAD and chronological age outside the prediction interval). While the share of $*$-negative subjects is approximately constant across the diagnosis, the percentage of $*$-positive subjects for MCI and AD groups is approximately 5 times the one for the control subjects. This result aligns with the literature, where it has been shown that MCI and AD patients show higher apparent brain age (Cole et al., 2019, Franke et al., 2012): for this reason the $*$-positive group is more interesting for their potential correlation with other disease indicators. All the prediction intervals are plotted in Figure 3.6, stratified by diagnosis and sorted by predicted age. The prediction intervals for the control subjects are scattered closer to the line of identity between predicted and chronological age and there are no relevant trends in the residuals that are left unexplained by the regression models. The variability of the width of the 90% prediction intervals is displayed in Figure 3.7: the average width is similar for the 3 diagnosis groups, but there is higher variability in the width distribution of the MCI and AD subjects. Moreover, $*$-positive brainPAD is mainly observed in the lower part of the age domain covered in the dataset. This could be just a consequence of our regression approach, or it might be due to the low number of subjects in the training set with chronological age less than 70, which might produce issues in the estimation of extreme quantiles of the conditional distribution of the outcome.

The brain maps displayed in Figure 3.8 are the functional coefficients obtained from the scalar-on-image quantile regression trained on the whole control dataset. They can be used to identify the regions that are responsible for the age prediction for the different quantiles. The functional coefficient for $\tau = 0.05$ shows that the expansion of the lateral ventricles is the principal factor that leads to higher predicted age (Preul et al., 2006, Apostolova et al., 2012) in the lower tail of the chronological age distribution. Other areas seem to have more limited impact on the prediction. In the coefficient obtained from the median regression, the lateral ventricles still play a role in the prediction (especially the posterior part) but expansion in several other areas is correlated to higher predicted age. Among them we point out the central sulcus (perpendicular to the median longitudinal fissure that divides the two hemispheres) that separates the primary motor cortex

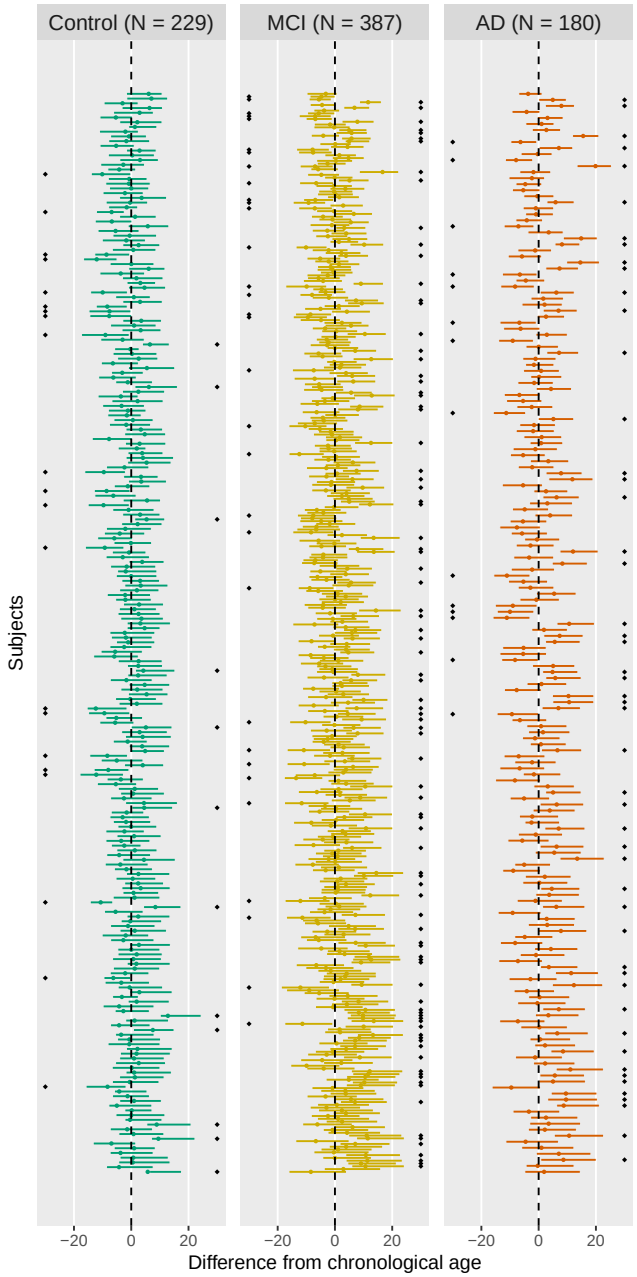

Figure 3.6: Brain age 90% prediction intervals, relative to chronological age. There is one interval per subject, and subjects are sorted in descending order of predicted brain age (higher predicted ages at top). The black diamonds indicate the subjects for which chronological age does not fall into the prediction interval; the side indicates if the subject is in the $*$-negative (diamonds on the left) or $*$-positive group (diamonds on the right).

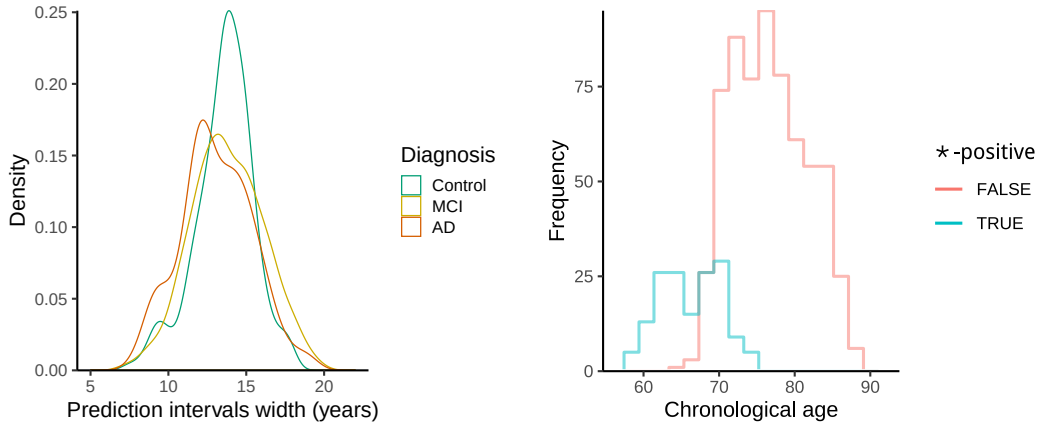

Figure 3.7: Left: distribution of the prediction interval width conditioned by diagnosis. Right: histogram of chronological age conditioned by $*$-positive indicator (equal to 1 if the chronological age is less than the prediction at $\tau = 0.05$, 0 otherwise).

and the primary somatosensory cortex. In addition, the frontal lobe shows negative values for the functional coefficient, meaning that expansion in this part of the brain is linked to a lower predicted age. This agrees with the literature: age-related atrophy is more pronounced in the frontal lobe (Cabeza and Dennis, 2013, Fjell et al., 2014, MacPherson and Cox, 2016) and less in the occipital lobe (Dennis and Cabeza, 2011). For $\tau = 0.95$, the brain map indicates that the upper part of the cortex and the cerebellum are related to higher predicted age, while a larger left temporal lobe (in blue in the lower axial slices, it plays a role in memory and language control) is associated to younger brain age. Especially for these last two maps, asymmetry between hemispheres appears in the relationship with brain age.

3.4.2 Correlation with cognitive decline measures

A small number of cognitive decline measures available in ADNI has been used to evaluate the clinical utility of the predictions obtained. The list of measures reported in Table 3.3 includes genetic assessments (ApoE4) and various evaluations of writing and speaking skills, visual attention and task switching. The outcomes of interest in this section are both the brain-predicted age difference (*brainPAD*, difference between predicted and chronological age, as defined in Cole et al., 2017) and the binary $*$-positive indicator (equal to 1 if the chronological age is less than the prediction at $\tau = 0.05$, 0 otherwise).

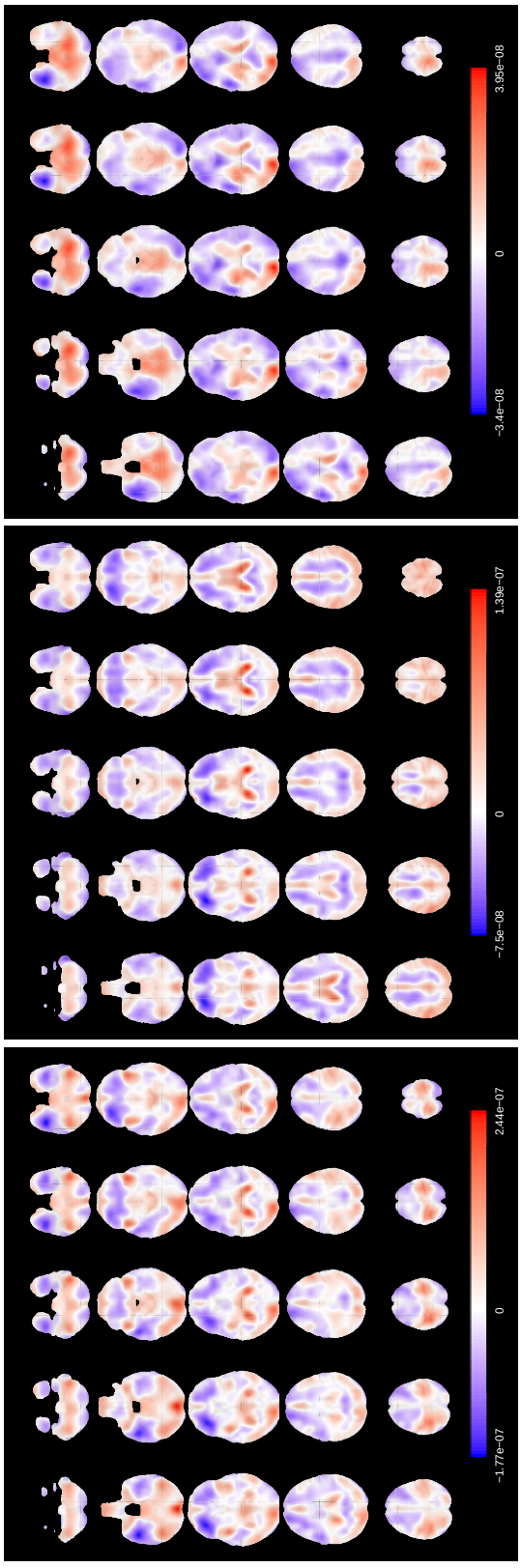

Figure 3.8: Axial slices of the functional regression coefficient for $\tau = \{0.05, 0.5, 0.95\}$ (from left to right). Slices are ordered from bottom to top. The colours are overlaid on the corresponding slice of the MDT. For a unit increase (expansion) in the observed TBM image in a red voxel, there is an increase in predicted brain age, while in a blue voxel there is a decrease.

Variable		Values	
ApoE4	Apolipoprotein E - Number of $\varepsilon 4$ alleles	$\{0, 1, 2\}$	↗
ADAS11	AD Assessment Scale - 11-item variant	$\{0, 1, \ldots, 70\}$	↗
ADAS13	AD Assessment Scale - 13-item version	$\{0, 1, \ldots, 85\}$	↗
ADASQ4	AD Assessment Scale - Delayed Word Recall	$\{0, 1, \ldots, 10\}$	↗
MMSE	Mini-Mental State Examination	$\{0, 1, \ldots, 30\}$	↘
DIGITSCOR	Digit Symbol Substitution Test	$\{0, 1, \ldots, 83\}$	↘
TRABSCOR	Trails B Making Test	$\{0, 1, \ldots, 996\}$	↗

Table 3.3: Cognitive decline measures used in the analysis. The arrows indicate the change in the measures associated to an increase in dementia severity.

Figure 3.9 summarises the main findings in this validation analysis. A higher ApoE4 value—linked to higher risk of dementia—is also related to higher predicted age difference on average (the p-values refer to one-sided tests). In addition, for the group with the highest ApoE4, more than 75% of the individuals show higher predicted age than chronological.

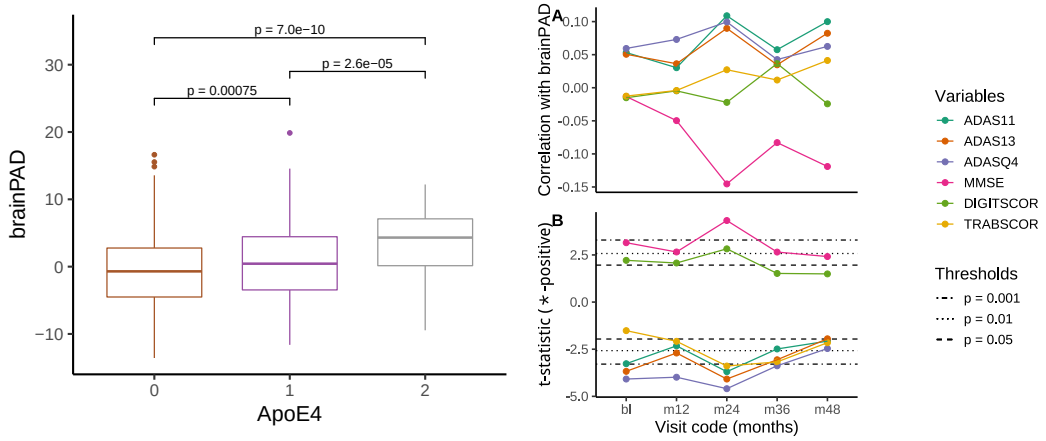

Figure 3.9: Left: association of *brainPAD* with ApoE4 value (Holm-corrected p-values) for different visits, with evidence of positive association. Right: (A) Correlation between baseline *brainPAD* and cognitive scores at different visits; (B) t-statistic for the comparisons of means of cognitive scores between $*$-positive group and the rest of the sample at different visits. The black lines are Student's t quantiles which correspond to different probabilities in the tails of the distribution.

The correlation between baseline *brainPAD* and cognitive scores at different visits shows some association (uncorrected) for several measures, with ADAS measures and MMSE showing the strongest associations after 2 years. Nevertheless,

no cognitive measure recorded at baseline is associated with the difference between predicted and chronological age. On the other hand, there is some evidence that the average of the cognitive measures is different between the $*$-positive group and the rest of the subjects across different time points. Also in this case the direction of the relationship is consistent with the numerical definition of the measures.

3.4.3 Sensitivity analysis

The prediction results are obtained under specific choices of several parameters. In order to assess how these choices might affect the results, we perform a sensitivity analysis using different values of the following parameters:

- PVE: proportion of variance explained (criterion to decide the number of fPC to be included in the quantile regression models), PVE $\in \{0.65, 0.8, 0.95\}$;
- KS: knot spacing, KS $\in \{6, 9, 12, 15\}$;
- nominal coverage: desired width of the prediction intervals. Values considered:
  - $\tau \in \{0.1, 0.5, 0.9\}$ for a 80% nominal coverage,
  - $\tau \in \{0.05, 0.5, 0.95\}$ for a 90% nominal coverage.

For each combination of values, we get the projections for each image and then fit the LASSO quantile regression. For the cases with KS $= 6$, the standard procedure did not work because of a failure in the Cholesky decomposition of the weight matrix $\boldsymbol{W}_\phi$ in Section 3.2.3, due to numerical tolerance issues. In these cases, the pivoted Cholesky decomposition can be applied: due to the fact that the matrix $\boldsymbol{W}_\phi$ is symmetric semipositive definite by construction, there is a permutation matrix $\boldsymbol{P}$ for which $\boldsymbol{P}^\top \boldsymbol{W}_\phi \boldsymbol{P}$ can be factorised with an upper triangular matrix (see Higham, 2009 for an introduction).

We report as main outcomes the mean absolute error and the actual relative coverage ($1 - \iota_{cov}$, where ι_{cov} is the ratio between observed and nominal coverage) obtained for the control subjects in Figure 3.10.

The MAE refers to the predictions obtained with $\tau = 0.5$, so it is not affected by the choice of nominal coverage. In general, the MAE remains rather stable across combinations of PVE and knot spacing, suggesting that our results are robust to the choices of these parameters. The lower MAE is always achieved for PVE $= 0.8$:

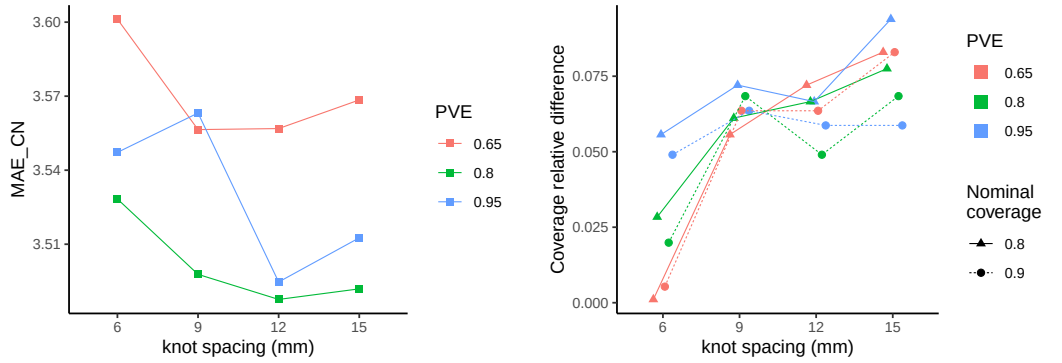

Figure 3.10: Left: mean absolute error for control subjects as function of proportion of variance explained and knot spacing. Right: Coverage relative difference of prediction intervals induced by each choice of proportion of variance explained, knot spacing and nominal coverage. Points are jittered horizontally for visualisation purposes.

this might suggest that a low PVE neglects important sources of variation while a higher one introduces too many useless variables in the models. In terms of knot spacing, 12 mm gives in almost all the cases the best results across PVE values.

Looking at the coverage for each setting of knot spacing, PVE and nominal coverage, we first observe that there are no cases in which the observed coverage is higher than the nominal level. This phenomenon of undercoverage gets more pronounced for higher knot spacing values. Except for KS $= 6$, when the coverage relative difference increases as the number of components in the quantile regression increases, for the other KS values no clear pattern is visible. The relative difference seems not to be influenced by the prespecified nominal coverage.

The table in the Supplementary Material in Palma et al. (2020) includes also a sanity check based on non-monotonic prediction intervals—those for which the predicted age at the upper τ level is smaller than the one at the lower level. The number of occurrences of this phenomenon is negligible in almost all the cases.

As an additional analysis, we have explored the prediction performances in terms of MAE for the control group in two models which do not use the basis expansion step, using the R packages `bigmemory` (Kane et al., 2013) and `bigstatsr` (Privé et al., 2018). The first model (M1) is a sparse linear regression with LASSO regularisation applied on the unsmoothed data (represented by 1 column per voxel in the data matrix). The second model (M2) is closer to our approach: a PCA is performed on the covariance of the matrix of unsmoothed images, then the scores corresponding to the first principal components selected (using a pro-

portion of variance explained of at least 0.8) are plugged into a penalised quantile regression model. M2 can be interpreted as a special case of our functional approach when the distance between adjacent knots is equal to 1 mm.

The difference in computational time between our approach (M0) and the models M1 and M2 is not substantial. On one hand, the smoothing step in M0 is performed independently for each image in a parallelised setting therefore it requires only a few minutes in total. On the other hand, in M1 and M2 we need to load the matrix (6.4 GB in our case) in memory and run PCA with a large matrix (which also requires inversion), which could take several minutes. In this case there is a dependency on the number of basis functions used. The quantile regression step takes also less than a minute per model (the cross validation procedure to find the LASSO parameter is the most computationally demanding aspect). For what concerns the prediction performances, M0 achieves lower MAE for the control group with respect to M1 (MAE $= 3.63$) and M2 (MAE $= 3.65$).

3.5 Discussion and further research directions

The functional data paradigm represents a useful approach to the analysis of complex data such as brain scans and offers a way to fit a global model for 3D images. In this work we have discussed the basic aspects of functional data and presented an application of quantile scalar-on-image regression (as extensions of classical quantile regression) in the field of brain age studies. Following the existing literature, we have devised an efficient workflow that takes as input a tensor-based morphometry image and returns a prediction interval. The advantages of employing the whole images as covariates are that some common preprocessing steps might be avoided (e.g. brain tissue segmentation) and there is no need to summarise information at the ROI (regions of interest) level. In addition, quantile regression gives a more detailed picture of the relationship between the covariate and the response and returns an interval with the desired coverage when the distribution of the dependent variable departs from normality. In contrast with other existing models coming from a machine learning perspective, our method outputs not only a point estimate but also a prediction interval. In addition, the model allows to investigate the functional coefficient estimated, in order to visualise the brain regions that influence most the predicted age.

Our modelling strategy introduces new features with respect to the standard prediction-oriented approaches in the literature. While other approaches focus

only on maximising prediction accuracy, we emphasise the detection of individual atypical ageing: the prediction intervals give a simple and preliminary assessment of the relevance of the observed brainPAD. In other words, the same brainPAD could be indicative of potential neurodegenerative diseases for one subject, while being less linked to such disease for another subject.

The results from the analysis of ADNI data are encouraging: the point (median) prediction performances in terms of MAE and RMSE for the control subjects are comparable with the literature on the topic—even with deep learning approaches applied on bigger ADNI datasets (Varatharajah et al., 2018)—while being also more principled and interpretable. The correlation between chronological and predicted age results to be lower than the one found with other methods. The model trained on the control group highlights differences with respect to the MCI and AD groups: individuals with cognitive impairment are predicted to be older on average than their observed age, as observed in the literature (Franke et al., 2012, Cole et al., 2017).

The model proposed is an example of penalised functional regression. In this respect, some degree of regularisation can be applied at different stage of functional data analysis, starting from smoothing (Ramsay and Silverman, 2005). At the same time, the choice of the number of functional principal components to be used in regression (by using the proportion of variance explained) is itself a penalisation. On top of this we added a further penalisation, driven this time by the relationship between outcome and predictors, to account for the potential high number of covariates given the sample size (following the indication provided in Heinze et al., 2018). Our model represents a novelty in the literature as it easily accommodates this aspect into a quantile regression model with 3D functional covariates.

In addition to the bias induced by the regularisation, another potential issue related to the functional coefficient is its sensitivity to the modelling strategy used. As extensively studied in Happ et al. (2018), the smoothness induced by splines could lead to different estimates with respect to other approaches (e.g. wavelet basis expansion or random field methods). Further work can be done to confirm the contribution of each brain region to the final prediction. Nevertheless, the predictive ability - which is the first focus of our model - does not seem to be harmed by this modelling choice.

Our approach is competitive in terms of speed compared to existing methods (Franke et al., 2012, Cole, 2017). In particular, for a new image the model returns the predicted interval in approximately a minute and the training phase of the

model is expected to be shorter and less computationally intensive than training a neural network, especially because the basis expansion step runs in parallel for each image.

The modelling approach illustrated in this paper can be extended in multiple ways, from both theoretical and practical perspectives. For what concerns the key points of the workflow, in this paper we have chosen to project the images (and the functional coefficients) using B-spline basis functions and sketched a possible strategy to select knot spacing. We have shown that some degree of smoothing produces slightly better predictions with respect to no smoothing at all with negligible computational cost. The benefit of this approach could more easily appreciated when the number of images is much larger, in which case loading the whole unsmoothed data into memory can be unfeasible.

The quantile regression approach is a technically easy-to-implement strategy to build prediction intervals without assuming normality. Since we consider only the best fit for each of the regression models, it could be of interest to study how the uncertainty about the coefficients and the models could play a role in the calculation of individual prediction intervals. The observed coverage in the control group could also depend on the bias/variance trade-off introduced by the cross-validation procedure (and in particular on the type of penalty and the number of folds chosen). Further simulation study can be done to assess the extent of this relationship.

In addition, further extensions of quantile regression could be considered. Additive terms might be introduced in order to explore nonlinear effects of the imaging covariate. Moreover, quantile boosting (Mayr et al., 2012) could provide better prediction intervals by reducing the bias due to the estimation at extreme quantiles. This approach has a higher computational cost but keeps the advantage of interpretability, which is no longer available with other approaches such as quantile regression forests described in Meinshausen, 2006. A potential issue for the current formulation of our approach is the phenomenon of *quantile crossing*, that occurs when the predicted quantiles are not monotonically increasing in τ as the conditional quantile function is by construction. Although in 90% prediction intervals the problem arises rarely (in our application it has been reported for only 1 case out of 796), still this could introduce some bias. Monotonicity can be forced after the estimation by using rearrangement or isotonic regression (see e.g. Kato, 2012, Chernozhukov et al., 2010). An alternative modelling strategy for quantile regression that ensures monotonicity of the function is provided in Chen and Müller (2012): the quantile function is obtained indirectly by first es-

timating the entire CDF of the response variable and then inverting it to recover the quantile function at the level of interest. The key idea is to use a generalised functional linear model to model the conditional distribution of $Y|X$ as conditional expected values of indicator functions. This "indirect" model is claimed to provide better estimation of the quantile function with respect to the classical quantile regression at extreme quantile levels for non-gaussian response variables (Chen and Müller, 2012), although the flexibility induced by considering different predictors at different quantile levels is lost. In addition, generalised additive models for location, scale and shape (GAMLSS, Rigby and Stasinopoulos, 2005) can also provide a detail picture of the conditional distribution of the outcome of interest. In GAMLSS the parameters of the distribution (not only the location, as in GLM) can be written as (smooth) functions of the covariates. GAMLSS can handle functional covariates (Brockhaus et al., 2018) and ensures monotonocity of the quantile predictions, but the family of the conditional distribution of the outcome must be specified in advance.

From the application point of view, it is currently very difficult to provide a sensible comparison between different models. This is due to the large range of possible approaches (from multivariate statistics to deep learning) applied to a plethora of datasets with different sizes, age ranges and imaging modalities (T1-weighted MRI to PET or fMRI). Cole et al. (2019) uses a MAE weighted by the age range in the training set as a measure of comparison. That approach might be too simplistic, as a 1-year absolute error for a 6-year child should probably be weighted more than the same error for a 70-year old individual. A more adaptive measure should be devised, or alternatively there should be an incentive towards the use of a specific dataset as a benchmark. Big databases such as UK Biobank (Sudlow et al., 2015) seem the right testing ground for all the methods available in the literature. Our model could be applied on different imaging modalities, for example voxel-based morphometry, in order to specify potential differences in the effects due to white and gray matter.

Coming to more specific modelling-related issues, as observed from the plots concerning the prediction intervals, a non negligible correlation is noticed between chronological age and the brain age differences (predicted minus chronological, called *brainPAD* in Cole et al., 2017, *brainAGE*—brain age gap estimate—in Franke and Gaser, 2019 or δ in Smith et al., 2019). This undesirable effect arises from the simple fact that by construction the residuals (which become the objects of interest when we want to explore the relationship with other variables such as disease conversion) in a regression model are uncorrelated with respect

to the predicted values, but not with the observed ones. Similar issues are also reported in the deep learning approaches to brain age prediction (Cole et al., 2017, Varatharajah et al., 2018). The work by Smith et al. (2019) identifies potential reasons for this phenomenon and proposes some solutions. Among others, a viewpoint that is conceptually grounded and at the same time can be embedded in our model could be rephrasing the whole problem in terms of an errors-in-variables framework. In particular, this accounts for the imaging covariate (consistently with the functional data perspective) or its scores representation being measured with some errors. At the same time, the response itself (chronological age) can be considered as a noisy proxy for biological brain age (for which it is difficult or even impossible to define a reference measure).

Another aspect left for future research is to extend the analysis of the clinical utility of the prediction intervals obtained with our workflow by using a larger battery of cognitive measures. The first basic measures selected in this work show interesting and sensible results, especially for the correlation with the $*$-positive binary variable. A desired feature of this indicator in a prognostic context should be its correlation with conversion to dementia, in order to provide a sensible way to early detect neurodegenerative diseases. Furthermore, a similarly defined "$*$-negative indicator" could be also explored in the same way in order to show potential aspects of a healthy ageing process.

In addition, introducing other covariates in the model (such as sex, years of education or physical activity measures) is rather straightforward and it could improve the detection of discrepancies from normative ageing. On the other hand, these covariates might potentially introduce confounding effects: the variability due to non-imaging information could be already captured by one or more functional principal components. Our approach can be also easily incorporated in a longitudinal model where brain age trajectories could provide evidence of stable or accelerated brain ageing. This setting might be especially beneficial if applied to tensor-based morphometry images in ADNI, where data at latter visits can be easily interpreted as changes in regional volumes with respect to the previous scans.

ADNI Acknowledgements

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, National Institute of Biomed-

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

Chapter 4

A whole-brain normative model for 3-dimensional morphometry images based on skewed functional data analysis

4.1 Introduction

The study of shapes and volumes of brain regions represents a valid approach to highlight differences between subjects (Ashburner and Friston, 2004). Many phenomena, either non-pathological (like ageing) or pathological (e.g. Alzheimer's disease), are characterised by increasing atrophy at differential rates throughout the brain lobes that can be observed (cross-sectionally between subjects or longitudinally) through deformation of structural magnetic resonance images (sMRI).

Within the family of brain morphometry methods, tensor-based morphometry (TBM) is used to identify regional differences with respect to a common tem-

plate (Hua et al., 2013). Each voxel in TBM images is associated to the relative volumetric differences with respect to the template which can be interpreted as a factor of expansion or shrinkage of the brain area. In particular, values above the threshold of 1 in a brain area indicate that the subject shows an expanded volume with respect to the common template: for example, a TBM value of 1.1 means that the volume in the voxel of the subject image is 10% higher than the volume in the same voxel in the common template. This multiplicative factor of expansion/contraction corresponds to the determinant of the Jacobian matrix that for each voxel encodes the deformation that maps the points in the template to the original MRI scan of the subject (Ashburner and Friston, 2004, Chung, 2013).

The Alzheimer's Disease Neuroimaging Initiative (ADNI) provides a dataset containing TBM brain images of adults to study the differences between a control group and two groups of people with different levels of neurodegeneration (Mild Cognitive Impairment, MCI or Alzheimer's Disease, AD). An exploratory analysis of 817 TBM images reveals spatial heterogeneity in the voxelwise distributions. Figure 4.1 shows the empirical cumulative distributions of TBM values by diagnosis groups for two voxels selected at random in the brain, one of which belongs to the lateral ventricles and the other is outside this region. The patterns observed are very different: for the voxel in the ventricles, the probability of observing higher extreme values (the comparison threshold with the template in this dataset is 1000) increases towards the group with AD. The distributions for all the groups do not appear to be symmetric. On the contrary, the voxel outside the ventricles shows no clear differences between the cumulative distributions.

When looking at the summary statistics for all the voxels across all subjects, the patterns between diagnosis groups are even more evident. Figure 4.2 shows the relationship between voxelwise means and standard deviations for each group. While most voxels show a mean around 1000, for some others higher mean and variances are observed, especially for the groups with diseases. But even for the cognitively normal subjects, the variability of the standard deviations is not constant as the means increases. When we compute Pearson's coefficient of skewness on the whole dataset, Figure 4.3 suggests that, for the regions with mean higher than 1000, the mean-standard deviation relationship is also linked to higher asymmetry. Brain areas with higher mean (among which the lateral ventricles) tend to exhibit more skewed distributions.

These considerations show that the characteristics of the voxelwise distributions are highly heterogeneous across the brain and could play a role in the statistical models for brain images. The statistical properties of the Jacobian values

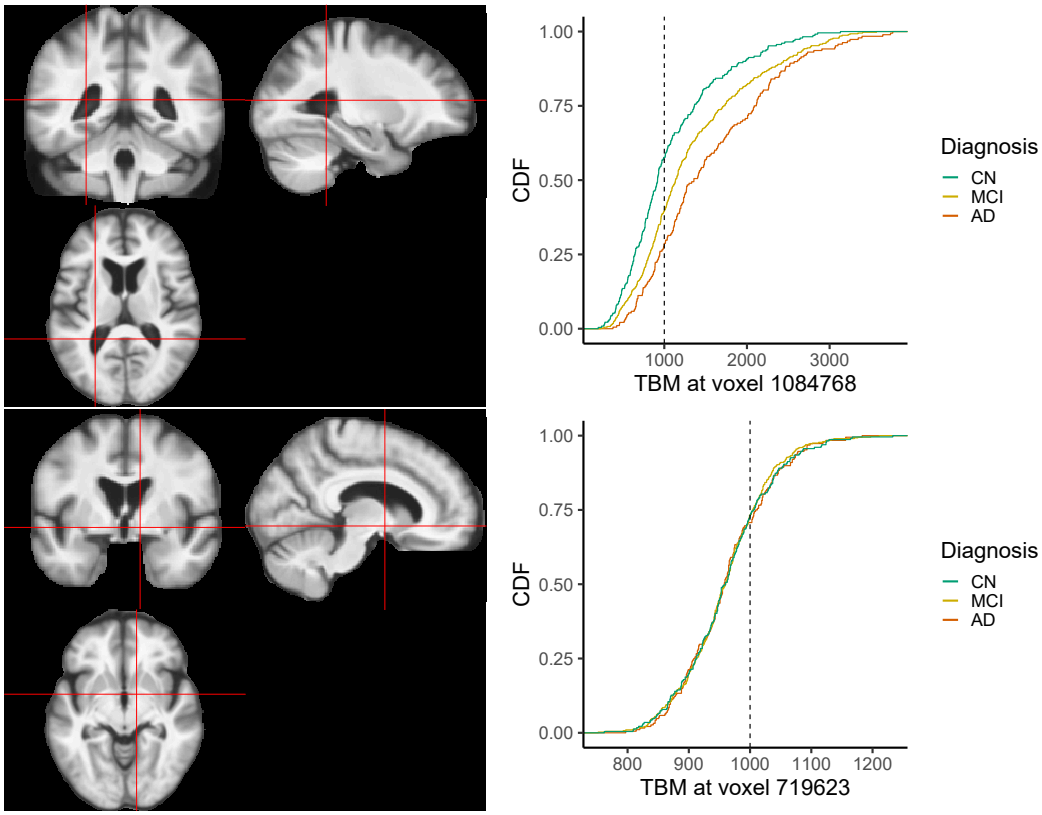

Figure 4.1: Top: empirical cumulative distribution functions of TBM Jacobian values by diagnosis group (right) for a voxel in the lateral ventricles. Bottom: empirical cumulative distribution functions of TBM Jacobian values by diagnosis group (right) for a voxel outside the lateral ventricles.

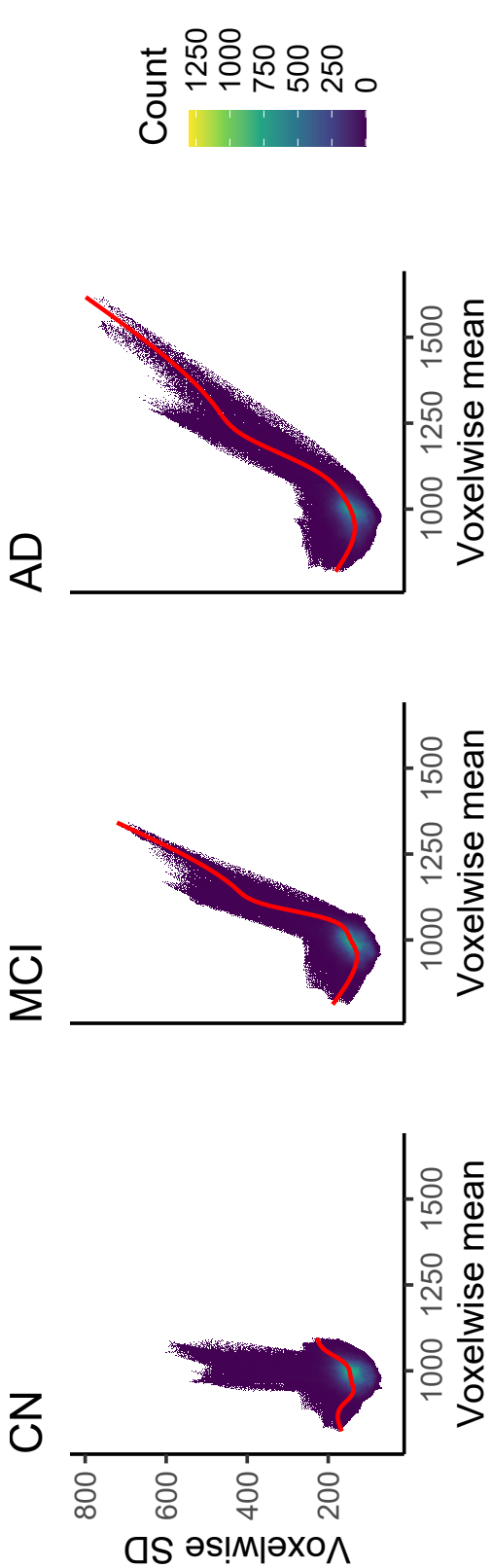

Figure 4.2: 2D histograms of voxelwise means and standard deviations by diagnosis group. The number of bins is fixed to 600. A smooth regression line is added in red.

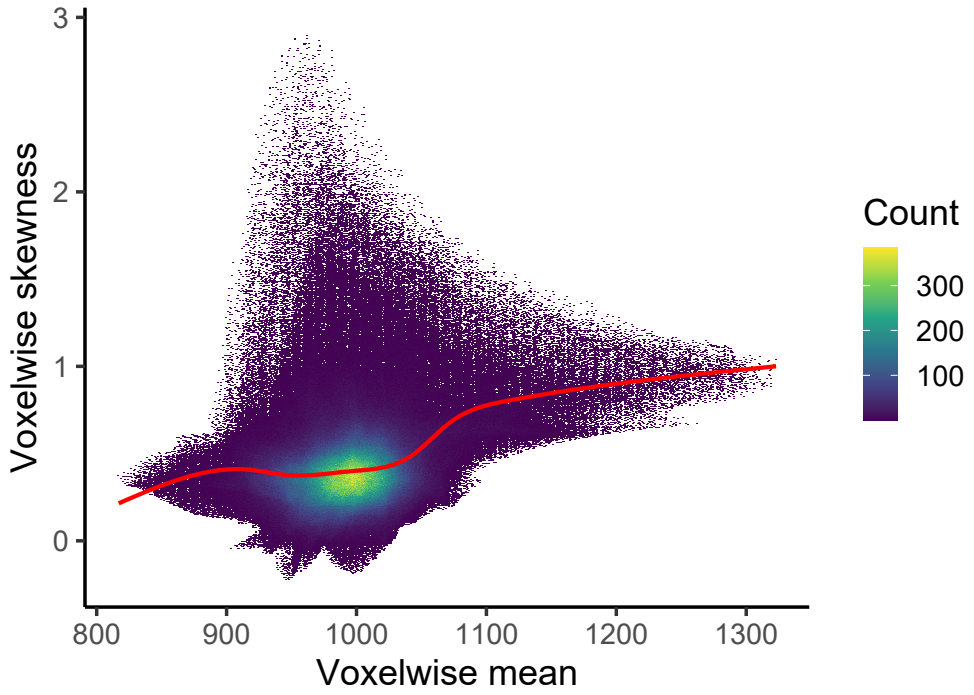

Figure 4.3: 2D histogram of the voxelwise mean and skewness across all subjects. The number of bins is fixed to 600. A smooth regression line is added in red.

in TBM have been studied in the literature. In Chung (2013) two alternatives are presented: normal distributions were traditionally assumed (Chung et al., 2003), while more recently in the literature (Leow et al., 2007, among others) have discussed mathematical arguments in favour of the log-normal distribution.

In this work, we provide a modelling strategy for 3D brain data showing skewness in the voxelwise distributions. Following the approach illustrated in Staicu et al. (2012) and extendend in Li et al. (2015), we integrate the analysis of voxelwise distributions into a functional data analysis (FDA, Ramsay and Silverman, 2005) framework, by using copulas to model the dependence structure between voxels. Each voxelwise distribution is modelled using a skew-normal distribution (Azzalini, 2013) whose flexibility allows to bypass the choice between normal and log-normal distribution by means of a single skewness parameter. In practice, the main objective of our analysis is to transform the original TBM images into "z-maps" (Section 4.3) which are based on standard normal processes.

To use the z-maps for predictions of cognitive dysfunctions at the individual level, we embed these approach into a normative model based on the reference population of healthy subjects. In this way, we expect subjects with relevant neurodegeneration to appear as extremes with respect to the normative population. Several indices are also proposed to summarise the information of z-maps into a single value (Section 4.4), which would ideally capture the extent of the departure from the normative population and potentially represent a score of the severity of the neurodegeneration.

The normative z-maps proposed here carry a more informative evaluation of the original images. First, they directly encode not only the relationship between the single image and the template, but also the one between the single image and the mean image in the reference population. This represents a great advantage in terms of the readability of the z-maps, where values closer to 0 indicate brain volumes closer to the expected pattern observed in the reference population. Second, observing for example a high Jacobian value does not indicate by itself whether the corresponding brain area has an "outlying" expansion that raises suspects of a disease, while through the z-maps this could be more easily detected.

These z-maps can be used for multiple modelling scopes. Being not linked to any specific scalar variable, these could be used for an exploratory analysis of the normative population or to determine measures of deviation from the centre of the population. The normative z-maps could also be corrected for other scalar variables such as age and sex. In contrast with other normative models, with our

approach we openly enforce smoothness of the z-maps, that represents both a sensible choice in the brain imaging setting and at the same time brings a useful computational gain.

The work is structured as follows. After a brief description of the benefits of a normative model in the framework of personalised medicine in Section 4.2, we describe the features of the statistical model in Section 4.3, breaking it into the choice of a copula, a voxelwise distribution and the prediction for new subjects. A range of application of the z-maps for different tasks is presented in Section 4.4, accompanied by some results (Section 4.5) computed for a subset of ADNI data. Finally, we discuss potential further developments of the model in Section 4.6.

4.2 Beyond case-control: normative modelling

Normative models are a novel statistical approach aimed at parsing the heterogeneity within a neuroimaging cohort. They can be used to produce individual predictions based on the choice of a reference population (Marquand et al., 2016a, 2019).

The key point of normative modelling is that subjects with a certain disease do not necessarily fit into one single group. For example, different subjects might show some aspects of the disease which could require the definition of subgroups of the disease or even a broader continuous spectrum of the pathology. In some diseases, it could also be argued that the group with disease cannot be clearly separated from the healthy population (Marquand et al., 2016b). In other words, a disease could be studied as a deviation from an expected healthy pattern.

These different views about a disease as a condition with high heterogeneity are not captured in the usual case-control approach, which is useful for comparing the averages in the two clinical groups, but does not focus on the individual variation (which is often seen as "residual" under that framework). An important assumption in case-control studies is also that the diagnostic label for each subject is given—in other words, a clear-cut separation between the two groups is considered. Normative models using healthy subjects as the reference population rely instead just on the labels for the training set. The symptoms of a certain disease can be used not to define thresholds between cases and controls, but as an extreme value with respect to the reference population. Subjects with no previous diagnosis obtain in this framework a risk score that could be used to perform inferences about the individual labels.

In the neuroimaging setting, the normative modelling paradigm could bring new insights on brain-related outcomes. First, in this setting the variability in the healthy cohort is taken into account too, potentially highlighting differences within that group too (while often the differences are overlooked in the group of patients with diseases only). Second, this might be especially useful in AD studies, where the neurodegeneration could be seen as a continuous process rather than a step function with 2 or 3 levels. In addition, the normative approach could also be used to highlight clusters of clinical interest: in particular for MCI, researchers have been looking at different subgroups of the dysfunctions (Hanfelt et al., 2011, Clark et al., 2013).

The normative modelling pipeline proposed in Marquand et al. (2016a) works for any scalar measure of brain functions or structure and any covariate of interest. The predictive confidence for each point is estimated, therefore accounting for both the error within the normative population and the model uncertainty. Then for each subject a normative probability map is computed, by means of a z-score which quantifies the deviation from the normative model for each voxel. The normative probability map is then summarised via an index based on extreme value theory, which will highlight underestimation or overestimation, and thresholded to identify abnormal brain region with respect to the healthy condition.

4.3 Statistical framework

4.3.1 Voxelwise analysis: the skew-normal distribution

Consider a reference sample $Y_1, \ldots, Y_N$. Let Y_i be a realisation of a random function for the i-th subject ($i = 1, ..., N$), with $Y_i = \{Y_i(v), v \in \mathcal{V}\}$. We assume that Y_i is a square integrable random function on the closed cube $\mathcal{V}$.

For $v \in \mathcal{V}$, suppose that

$$U_i(v) = F_{\text{SN}}(Y_i(v); \mu(v), \sigma^2(v), \gamma_1(v)) \tag{4.1}$$

where F_{SN} is the skew-normal cumulative distribution function with mean parameter $\mu(v)$, variance parameter $\sigma^2(v)$ and skewness parameter $\gamma_1(v))$ for the reference population. The probability integral transform in Equation (4.1) returns $U_i(v) \in (0,1)\ \forall v \in \mathcal{V}$: it is a latent uniform process based on the F_{SN} for the i-th subject.

In the skew-normal distribution literature (Azzalini, 2013, Arellano-Valle and Azzalini, 2008) the set of parameters of F_{SN} is called centred parameterisation (CP) to distinguish it from the original direct parameterisation (DP); the two can be mapped from one to the other. The centred parameterisation is easier to interpret: the parameters are functions of the first three moments in population. CP is also the standard choice in estimation, because it removes the problem of singularity of the Fisher information matrix when the DP shape parameter is equal to 0, which in turn harms the asymptotic normality of the MLE estimates (Azzalini, 2013). The likelihood function for CP gets closer to a quadratic function and produces estimators which are less correlated than the DP estimators (Monti et al., 2003). An iterative procedure is needed to produce maximum likelihood estimates of the three parameters. The sample moments could be used as starting points for the procedure.

In Azzalini, 2013 it is noted that the skewness parameter γ_1 is constrained within the set $(-c_1, c_1)$, with

$$c_1 = \frac{\sqrt{2}(4-\pi)}{(\pi-2)^{3/2}} \approx 0.9953, \tag{4.2}$$

while other distributions such as skew-t might be more appropriate for higher observed sample skewness. The direction of the skewness is determined by the sign of $\gamma_1(v)$: if positive, the distribution is skewed to the right. For skewness (and equivalently shape parameter in DP) equal to 0, SN reduces to a normal distribution with the same mean and variance.

In practice, the domain $\mathcal{V}$ is discretised into V voxels $v_1, \ldots, v_V$, therefore we refer to the observed data for the i-th subject as $Y_i(v_j), v_j = 1, \ldots, V$.

4.3.1.1 Gaussian copula

Staicu et al. (2012) propose to use a copula approach to model dependencies in the pointwise distribution of functional observations. The uniform marginal distributions can be used within a copula framework. A copula is a joint cumulative distribution of a random vector with uniform marginals. In our setting, the random vector $(Y_i(v_1), ..., Y_i(v_V))$ is transformed into $(U_i(v_1), ..., U_i(v_V))$ using the skew-normal distribution. Sklar's theorem states that any multivariate distribution can be expressed as $\mathsf{C}(U_i(v_1), \ldots, U_i(v_V))$, where C is the copula and U is a uniform random variable obtained using the probability integral transform (Sklar, 1959). This theorem allows to break the model into two separate components: the

marginal univariate distributions and the copula which explains the dependencies between them.

We can use a Gaussian copula to model the dependencies between voxel random variables. A Gaussian copula for a random vector is the copula of some multivariate Gaussian distribution, without necessarily implying that the random vector itself is Gaussian.

In formal terms, let us consider the new Gaussian process for the i-th subject and a voxel v

$$Z_i(v) = \eta^{-1}(U_i(v)) \tag{4.3}$$

where η^{-1} is the inverse CDF of a standard normal. The Gaussian process is discretised into the V voxels and goes into the Gaussian copula

$$\mathsf{C}_{K_i}^{\text{Gaussian}}(u) = \eta_K(Z_i(v_1), ..., Z_i(v_V)) \tag{4.4}$$

where η_K is a joint multivariate normal distribution with zero mean. Given that the variable $Z_i(v_j)$ are distributed as a standard normal, the matrix $\boldsymbol{K}$ with elements $K_i(v_j, v_l) = \text{Cov}(Z_i(v_j), Z_i(v_l)) \;\; \forall j, l = 1, \ldots, V$ is both the covariance and Pearson correlation matrix of the multivariate distribution. $\boldsymbol{K}$ is therefore the only parameter which defines the copula. This can be estimated with the method of moments by using the estimated latent Gaussian process, that is $\hat{K}_i(v_j, v_l) = \text{Cov}(\hat{Z}_i(v_j), \hat{Z}_i(v_l))$.

4.3.1.2 Functional principal component analysis

The covariance operator of the functional sample obtained with this procedure (that is now made of zero-mean Gaussian processes) is the only parameter of interest. Functional principal component analysis (FPCA, Ramsay and Silverman, 2005) can be used to represent it: each functional observation can be approximated using a (truncated) linear combination of eigenfunctions of the covariance operator weighted by some scalar quantities called scores. The scores are uncorrelated with zero mean and variance equal to the eigenvalues of the covariance operator. The number of functional principal components used to reconstruct each Gaussian process $Z_i(v)$ is usually determined using the proportion of variance explained criterion. More theoretical and implementation details in a 3D imaging application are available in Palma et al. (2020).

4.3.2 Computational aspects

The maximisation of the likelihood of the skew-normal distribution can be performed in a parallel setting as the spatial dependence is captured at a later stage using the copula. Nevertheless, the number of voxels might still be large enough to slow down the calculation of the parameters at the voxelwise level.

Instead of running the computation for every voxel, a grid of preselected voxels can be used. Let $\{\kappa_1, \ldots, \kappa_{V^*}\}$ be a subset of voxels ($V^* \ll V$). For these voxels, the likelihood for the skew-normal distribution is maximised and the mean, standard deviation and skewness are computed. For the i-th subject in the reference sample, the latent process U_i and subsequently the standard Gaussian z-values can be computed as in (4.1) and (4.12).

The z-values for voxels outside the set with cardinality V^* can be estimated using smoothing basis functions. A tensor product of univariate B-splines could be considered. For each of the three dimensions, the degree of B-splines and the number and position of knots must be determined. The simplest choice is to keep the degree fixed for all the dimensions and set in advance a regular grid with the same distance between knots. A Kronecker product of the basis functions will return a matrix where each 3D basis function is reported as a column vector and it is evaluated for every voxel in the brain mask. Further details about implementation for 3D brain data are reported in Chapter 2.

Radial basis functions (RBF) could also be employed to overcome the separation of the 3 dimensions. Given a selected voxel ("centre"), the input of a radial basis function is not the location in space, but just the Euclidean distance of another voxel from the centre. Radial basis functions are generally used for approximation or interpolation of functions (Carr et al., 2001).

Radial basis function interpolation requires the choice of the basis functions and the definition of the centres. The basis function h depends on the (Euclidean) distance d between a centre and another voxel and it is symmetric around the centre. The value of the basis function decreases as the distance from the centre increases. For example, the (inverse) multiquadric $h(d) = \dfrac{1}{\sqrt{1+(\varepsilon d)^2}}$ or a Gaussian kernel $h(d) = \exp\left((-\varepsilon d)^2\right)$ could be used. The bandwidth of the basis function is controlled by one or more tuning parameters (ε for multiquadric, standard deviation for Gaussian).

The standard choice for centres is the same grid of preselected voxels $\{\kappa\}_{k=1}^{V^*}$ where we have carried the likelihood estimation out. For the sake of simplicity, we recommend to use a regular grid, where the distance (in the 3 directions) is pre-

specified. This approach would guarantee that the distance between any voxels and the closest centre is within a certain range which depends on the grid spacing. Depending on the specific application, a grid with irregular spacing as in Chebyshev discretisation could also be chosen to capture finer changes in some areas. We recall that the observed value of the function at the k-th centre is $Y(\kappa_k)$.

Following Carr et al. (2001), we define the interpolant s as a function with constraints $s^*(\kappa_k) = Y(\kappa_k)$. To define s^*, we build the matrix

$$\boldsymbol{G} = \begin{pmatrix} \boldsymbol{H}^* & \mathbf{1} \\ \mathbf{1}^T & 0 \end{pmatrix} \tag{4.5}$$

where the $V^* \times V^*$ symmetric matrix $\boldsymbol{H}^*$ contains the evaluation of the radial basis functions for any distance d between any pair of centres

$$H^*_{k\check{k}} = h(d(\kappa_k, \kappa_{\check{k}})) \qquad k, \check{k} = 1, ..., V^*. \tag{4.6}$$

and $\mathbf{1}$ is the V^*-dimensional vector whose elements are equal to 1. The problem is now phrased in terms of a linear system: we are interested in finding the V^*-dimensional vector $\boldsymbol{b}$ and the scalar b_0 such that

$$G \begin{bmatrix} \boldsymbol{b} \\ b_0 \end{bmatrix} = \begin{bmatrix} Y(\boldsymbol{\kappa}) \\ 0 \end{bmatrix} \tag{4.7}$$

or analogously

$$Y(\kappa_k) = b_0 + \sum_{\check{k}=1}^{V^*} b_{\check{k}} H^*_{k\check{k}}. \tag{4.8}$$

The solution is now used to predict a value for all the voxels $\{v_j\}_{j=1}^{V}$:

$$\begin{pmatrix} \boldsymbol{H} & \mathbf{1} \end{pmatrix} \begin{bmatrix} \boldsymbol{b} \\ b_0 \end{bmatrix} \tag{4.9}$$

where V is the number of voxels and

$$H_{jk} = h(d(v_j - \kappa_k)) \qquad j = 1, \ldots, V;\ k = 1, \ldots, V^*. \tag{4.10}$$

The gain in computational efficiency that stems from applying basis functions on the grid instead of using all the voxels in the brain comes at a price. First, the performance of smoothing basis functions relies on some shape parameters (such as the standard deviation for Gaussian RBF) for which it is not easy to determine

optimality criteria. In Fasshauer and Zhang (2007) it is observed that the best parameter is chosen by trial-and-error or ad-hoc solutions in many cases; in addition, there is a trade-off principle for which choosing a small value for the shape parameter increases accuracy but also the condition number of the interpolation matrix. Furthermore, for high values of the shape parameter the so-called "bed-of-nails" interpolant is obtained: the function sharply peaks at the centres but decreases to 0 elsewhere. In the 3D grid case, we suggest to use a value for the shape parameter that is below the grid spacing.

Another aspect of interpolation using radial basis function and polynomials is Runge's phenomenon, i.e. the approximation errors further from the centres are higher at the boundary of the domain (Fasshauer and Zhang, 2007, Boyd, 2010). In the 3D brain imaging setting, although the brain mask has irregular boundaries in the three dimensions, this issue is not likely to be relevant, especially when the grid spacing (and consequently the maximum distance between a voxel and the closest centre) is moderate.

4.4 Applications

4.4.1 Prediction for new observations

Let Y^* be a realisation of a random function for a new subject who could either belong or not to the reference population. We could use the parameters of the reference population to compute

$$U^*(v) = F_{\text{SN}}(Y^*(v); \mu(v), \sigma^2(v), \gamma_1(v)) \tag{4.11}$$

and then using the normal CDF

$$Z^*(v) = \eta^{-1}(U^*(v)). \tag{4.12}$$

4.4.2 Subject-specific indices of "abnormality"

The normative z-maps (which are, in a broad sense, derived from the covariate-free normative probability maps $U(v)$, not based on any relationship between the function and clinical predictors) give information about how the subject image compares to the reference population. Voxel values closer to zero indicate that the volume observed in the subject is close to the mean value observed in the refer-

ence population, whereas more extreme voxel values are potentially informative of non-healthy expansion/shrinkage of brain regions.

The z-maps could be therefore used for each subject to locate those brain areas that depart from the mean of the normative population. This approach gives an alternative way to read through the original voxel values, because it is now possible to assess whether high TBM values are an indication of individual abnormality or they are actually not far from the average signal observed in the reference population.

Various scalar indices could be built in order to partially summarise the information carried by the z-maps into a single value.

Dealing with Gaussian voxelwise distributions, the easiest approach is to count the number (or proportion) of voxels for which the normative z-value is greater in absolute value than a certain threshold. For example, n_3 could be the number of voxels for which $|Z(v)| \geq 3$. A high value for this index suggests that in a relevant part of the brain the observed TBM values are very far from the mean in the reference population and therefore might show evidence of deviation from the healthy pattern. This index could also be turned into a proportion by dividing for V, the total number of voxels.

An alternative approach (described in Marquand et al., 2016a) is based on extreme value statistics. Each individual z-map is summarised by the (robust) mean of an extreme block (e.g. from the 99th percentile of the distribution of z-values for each subject). A parametric distribution from the generalised extreme value (GEV) family is then fitted on the normative sample using these averages, then for every subject the cumulative probability under the GEV distribution is used to quantitatively assess the extent of the deviation. Under this approach, the specific application will drive the choice of the percentile defining the extreme block and the choice of the tail (whether to deal with the highest, lowest or highest in absolute value). In our setting, where the enlargement of the ventricles is balanced by the shrinkage of the cortex, it seems likely that extremes in both sides are carrying information, therefore the absolute block maxima approach seems more appropriate.

Further indication can be obtained from the histogram of the individual z-maps. By construction, the z-values are drawn from a standard normal distribution. In the grid-based approach, this holds for the voxels of the grid and through smoothing for the rest of the brain, approximately. A histogram for a subject belonging to the normative population would therefore be looking like a standard normal distribution too. Any departure from this distribution could be linked

to a departure from the average in the normative population. If we exclude the mean (if the difference between the reference population and other groups is localised in a small subset of voxels, these would probably not drive the mean far from zero), the variance and skewness are likely to increase if extreme voxels are observed. In addition, evidence of multimodality or quantities like the test statistics for normality tests like Jarque–Bera, Anderson–Darling, Cramér–von Mises or Lilliefors could help in detecting far-from-normal behaviours.

Finally, drawing from the functional data literature, the L^2 norm of the z-map

$$\|Z_i\|_2 = \sqrt{\int Z_i(v)^2 dv} \tag{4.13}$$

can be employed an alternative summary quantity.

4.4.3 Z-maps as covariates in functional regression

Normative z-maps could also be used within a regression framework to predict quantities of interest. For scalar outcomes, scalar-on-function regression (Morris, 2015) employs the whole function (for example through its basis representation or FPCA scores) as the independent variable. In this way not only is smoothness of functional slope coefficient maintained, but also a sparse solution can be achieved. It is worth mentioning that the results of the regression model are based on the reference population under study, as all z-maps are constructed on that. Nevertheless, in the normative setting we can use the model trained on the reference sample to predict the scalar outcome for any other subject. In this case, the prediction is to be interpreted as the equivalent outcome of a healthy individual having the same predictors (and as in (Palma et al., 2020) and the brain age literature cited in Chapter 3, then check if the difference with respect to the observed outcome could give insights on the disease status).

4.5 Data and results

We use the normative model to analyse a dataset coming from the Alzheimer's Disease Neuroimaging Initiative (ADNI), which consists of 817 adults (with age ranging between 54.4 and 90.9 years). A diagnosis is provided for each of them: 229 subjects were considered as cognitively normal (CN), whereas 400 subjects were showing mild cognitive impairment (MCI) and 189 were diagnosed with

Alzheimer's Disease. The sample used in Palma et al. (2020) represents a subset of the data analysed in this work.

The imaging data used in this work are tensor-based morphometry images. In a cross-sectional setting, a common MRI template called *minimal deformation template* (MDT) is obtained by averaging several anatomical MRI scans (Hua et al., 2013), then each MRI scan is aligned to it. The deformation induced by this alignment is mathematically described by a function that maps a 3-dimensional point in the template to the corresponding one in the individual image. To evaluate volume differences with respect to the minimal deformation template in terms of shearing, stretching and rotation, the Jacobian matrix of the deformation is considered. Its determinant evaluated at each voxel is a summary of local relative volumes compared to the MDT. Further details about TBM are available in Ashburner and Friston (2004).

A 3D preprocessed tensor-based morphometry (TBM) image taken at the baseline of the study is available for each of the individuals in the sample. The dimensions of the images are $220 \times 220 \times 220$, with voxel size equal to 1 mm^3. The threshold that determines equality with respect to the template is 1000: values higher than this threshold indicate that expanded volume with respect to the minimal deformation template is observed in that specific voxel.

The mask used to subset only the part of the image that displays the brain is built with the same characteristics as described in Palma et al. (2020): we use a Gaussian kernel with standard deviation equal to 2 voxels (FWHM 4.7 mm) and threshold it at 0.5. Each masked image is made of approximately 2 million nonzero voxels.

We consider first the normative sample (training set) to be used to define the skew-normal parameter estimates: it is made of 183 CN subjects (approximately 80% of the CN group), selected after stratification by age group and sex. We then define the grid of voxels in which to carry out the skew-normal fitting procedure: in this setting we use a regular grid with 8mm spacing in the three dimensions. This returns 3949 voxels within the mask, approximately equal to 0.2% of all the voxels within the mask. For these voxels, the skew-normal likelihood optimisation (with centred parameterisation) is carried using the R package sn (Azzalini, 2020).

Radial basis functions (RBF) with Gaussian kernel and standard deviation equal to 5.33 mm (66.67% of the grid spacing) are used to interpolate the SN parameter functions across the rest of the brains. For other values below 8mm tried on the same dataset, the bed-of-nails behaviour is not reported, while for higher grid spacing the interpolation quality is poor. A tensor product with univariate

B-splines spaced every 8mm has been also used on the same training set. Some analyses (not shown here) show that the number of B-splines functions is higher than for RBF (10920 against 3949 basis functions) and seem to suggest also that they do not improve the quality of the fit, especially at the boundaries of the mask where approximation errors are higher. In this procedure, fitting the skew normal parameters on the grid takes approximately 3 minutes on a standard laptop. Turning the original brain scans into z-maps based on the parameter function takes approximately 1 minute per image: this step can be run in parallel.

The parameter functions are plotted in Figure 4.4. The mean and the standard deviation are higher in the lateral ventricles than the rest of the brain. The skewness is greater than 0 across almost the whole mask.

The z-maps computed using the skew-normal parameter values at the grid and then smoothed across the rest of the brain are obtained for both the training and test sets. The indices of deviation are then computed. For example, Figure 4.5 shows the boxplots of the u_3^{abs}, an index of deviation obtained by taking the mean of the top 1% z-values in absolute values. Some evidence of a trend between this index and the severity of disease status is observed, although the results are not confirmed in terms of statistical significance. In addition, this index does not depend on a covariate like the ADAS13, a neuropsychological test often used in AD studies to assess cognitive dysfunctions (the higher the score, the higher is the disease severity, see Kueper et al., 2018): the regression lines for the diseased groups remains significantly above the one for cognitively normal subjects for lower values of ADAS13 in the study.

Focusing now on the normative population, we perform functional principal component analysis on the z-maps of the training sample. The first 3 eigenfunctions (Figure 4.6) highlight some areas which are usually linked to neurodegeneration within the control group as well. In particular, in the first eigenfunction the cingulate cortex stands out with respect to the rest of the brain. This area is involved in many executive and cognitive functions (Mann et al., 2011, Leech and Sharp, 2014) and its atrophy has been documented as one feature of memory deterioration (Lin et al., 2017) in healthy ageing (Fjell et al., 2009). Pronounced gray matter loss in this region (and especially in the posterior cingulate area) is also associated with higher risk of dementia (Choo et al., 2010, Peng et al., 2016). The second eigenfunction clearly distinguishes the lateral ventricles (whose enlargement is linked to shrinkage in the surrounding areas and it is of wide interest when studying AD) and the third eigenfunction seems to highlight external cor-

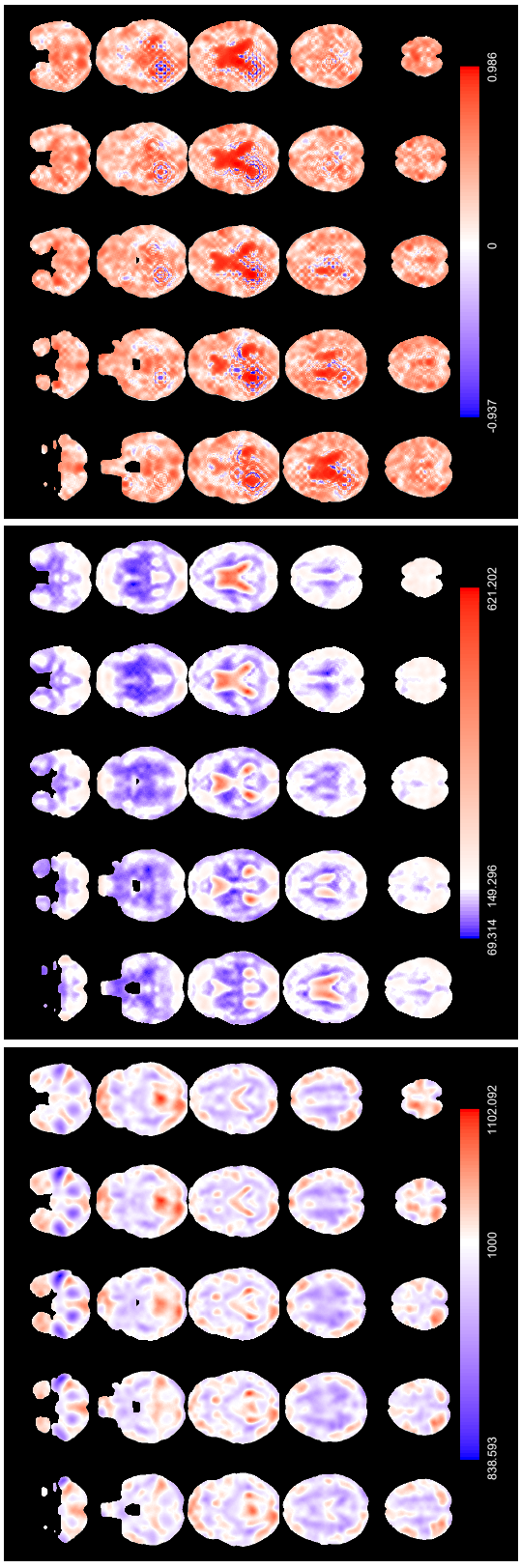

Figure 4.4: Axial slices of the mean (left), standard deviation (centre) and skewness (right) parameter functions from skew-normal fitting. Slices are ordered from bottom to top. For the standard deviation, the colour white corresponds to the average standard deviation.

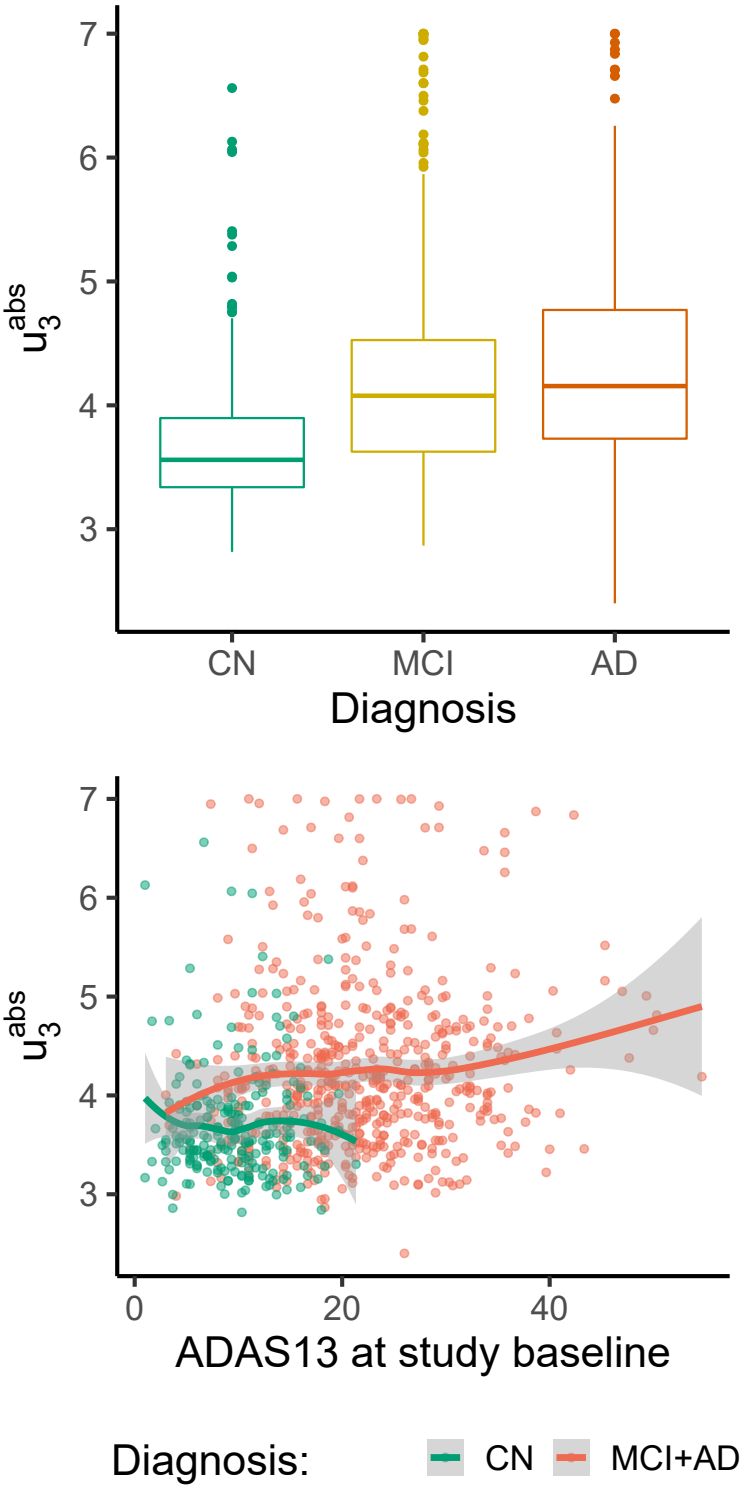

Figure 4.5: Top: Boxplots of u_3^{abs} by diagnosis group. Bottom: Plot of u_3^{abs} by ADAS13 and diagnosis group.

tex as opposed to internal brain areas (it could also be broadly linked to the tissue difference between white and gray matter).

The z-maps are then used as the only functional covariates in a regression setting. We add to the training normative sample a subset of MCI and AD subject and predict both the diagnosis and ADAS13. The basis coefficients for each z-map are used as covariates; in both models, an elastic-net penalty (Zou and Hastie, 2005) with mixing parameter 0.5 is applied (whereas λ is chosen via cross-validation). The functional coefficients obtained are displayed in Figure 4.7. Both of them identify a positive slope coefficient in the lower lateral ventricles: increased expansion in these areas are linked with a higher probability of having the disease and of showing higher values of ADAS13. As expected, large regions of the brain appear to be not relevant in the models, therefore do not provide support for classification of subjects into the CN group or the others.

In the binary classification problem high sensitivity (91.5% of the subjects with diseases are correctly identified as such) is achieved but the specificity is poor (only 37% of CN subjects is classified as having no diseases). Additional analyses made on z-maps corrected for age and sex (not shown here) provide similar results.

4.6 Conclusions

The analysis of brain morphometry images is of large interest due to its ability to show and quantify signs of atrophy within different brain regions. We have shown using a dataset of tensor-based morphometry images that the voxelwise distributions of TBM values exhibit interesting patterns in terms of mean, standard deviation and skewness. In this work we have proposed a method to take into account these characteristics by using a skew-normal distribution at the voxelwise level and a Gaussian copula to model the spatial dependence between brain locations. We have linked this approach to a normative model to study brain volumes in absence of neurodegeneration. The normative approach provides then a set of reference parameters on which to build individual brain maps, which can then be summarised into single indices. By using this approach, we aim at observing subjects with cognitive impairment as "extremes" with respect to the reference population, providing individual risk scores rather than focussing on the group differences between cognitively normal control subjects and patients with neurodegenerative diseases.

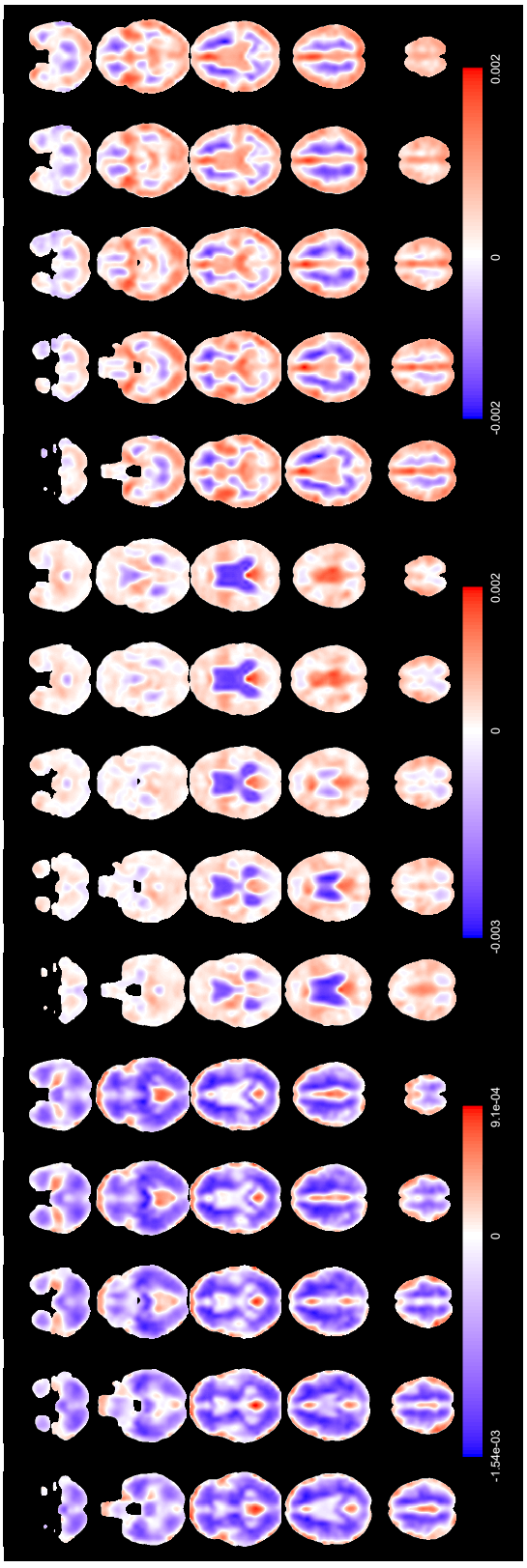

Figure 4.6: Axial slices of the first 3 eigenfunctions for the normative z-maps. They account respectively for 13.34%, 5.89%, 5.07% of the total variance. Slices are ordered from bottom to top.

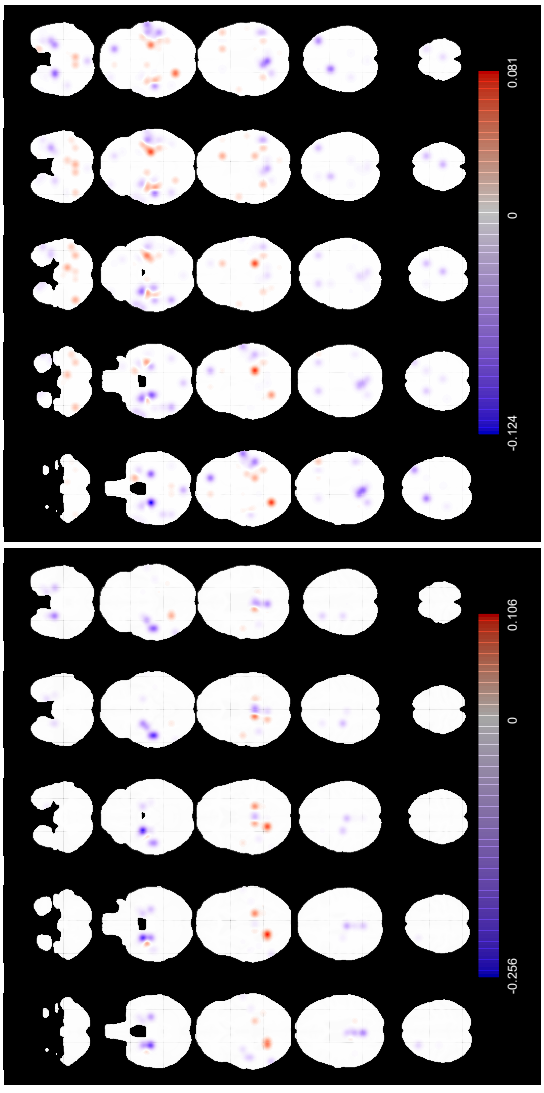

Figure 4.7: Left: coefficient of functional logistic regression using the z-maps as covariates and the diagnosis group (CN vs. others) as outcome. Right: coefficient of functional linear regression using the z-maps as covariates and the logarithm of ADAS13 cognitive score as outcome.

To manifest the strengths of our model, we could further explore other research questions. Within a binary classification exercises (presence vs. absence of disease) we can look at the diagnosis given later in time, such as 12 or 24 months after the image was taken, to see if the z-maps or the indices of abnormality at baseline are able to capture early changes in advance with respect to what done by e.g. neuropsychological scores. This would have a great impact especially in a clinical setting, where prediction of future disease stages is currently an open question.

Additional work can be also done for example on the indices of deviations. The ones proposed so far all refer to the z-values at the voxelwise levels—in other words, they can be defined just looking at the histograms of the z-maps. They are extremely simple to build, but unfortunately they represent just a rough summary of the z-maps, as they do not convey the important information about spatial distributions: knowing for example if the higher voxel values are observed in the lateral ventricles rather than in other areas might be beneficial, especially when assessing the risk of neurodegeneration due to Alzheimer's disease. Spatially-aware indices of deviation might be built by reusing the voxelwise standard deviations as weights to be multiplied to the z-maps. In our case, this would give more relevance to outlying values in the ventricles, which are linked to higher risk of Alzheimer's disease. Alternatively, context-dependent prior information should be incorporated within the definition of those indices.

The choice of the voxelwise distributions is a crucial part of the analysis where improvement can be achieved. We have proposed the skew-normal distribution as its high flexibility is paid for only with an additional parameter with respect to normal or log-normal distributions (which are the distributions that have been usually proposed in the neuroimaging literature for this type of data), but the copula modelling strategy works for any continuous distribution. For example, a gamma distribution might be considered to specify the relationship between mean and variance. The properties of this distribution family are likely to make it a good candidate for the imaging variable at hand and would hint towards an explanation in terms of multiplicative errors (just as the log-normal distribution, see Firth, 1988). If high skewness is a concern, skew-t distribution could provide a good solution, although the analyst needs to take into account the fact that it requires the estimation of four parameters. Another different approach could be based on avoiding to define a distribution family and use a nonparametric approach—as long as it allows to track the characteristics of the voxelwise distributions in a parsimonious way.

The problem of the voxelwise distribution can be tackled in a different way by, for example, defining a parcellation scheme that could be used to separate different areas of the brain to which to fit different probability distribution. In this setting, higher flexibility is gained at the price of breaking the spatial relationships between brain regions. The parcellation scheme can be either prespecified or data-driven: for example, the voxelwise distribution for voxels with values less than or equal to a certain threshold can be different to the one for the values higher than the same one. In other words, two types of variation (relationship between summary statistics) can be considered. From a computational point of view, this would mainly require a more flexible handling on the smoothing step over the grid.

This normative model could also be extended in a longitudinal framework, to track the evolution of both the normative population and the subject-specific map. In brain imaging studies, the longitudinal evaluation of the normative population allows to account the changes due only to ageing, which represents a major driver for non-pathological degeneration. To this aim, the parameter functions could be estimated at each time point and smoothed over time. The subject-specific indices of deviation would therefore be able to capture whether an individual is showing additional changes that could be interpreted as signs of diseases. Such a model would also require careful handling of the subjects within the normative samples who are lost at follow-up or convert to disease. A more extended version of the dataset used in this analysis is available on ADNI, containing multiple longitudinal measurements (both for the imaging part and the neuropsychological scores) that could be used in a similar framework.

Finally, other methodological extensions for functional data with spatial heterogeneity could arise from our approach. The indices of deviation from the mean of the reference population call for a centre-outward ordering of the subjects. Some tools already available in the literature such as functional data depth (López-Pintado and Romo, 2009, Mosler and Polyakova, 2012) could be extended to the case of 3D imaging data and applied in this context. This would also open a new avenue for the use of this tools which has so far been explored mainly from a theoretical perspective, although in this sense the computational efficiency becomes a much more crucial aspect of the whole analysis (obtaining depths for such a high-dimensional setting would require relevant computing resources).

In addition, the normative model explored in this work could hint towards a potential framework for 3D data simulation in the reference population. Indeed, given the summary statistics functions computed on the real data, the z-maps

could be simulated via random number generation from a standard normal. More work could be done to find strategies that would ensure the simulated images to be both smooth and plausible from a neuroscientific perspective.

ADNI Acknowledgements

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies;Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (`www.fnih.org`). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Chapter 5

Function-on-function regression for large-scale brain imaging data

5.1 Introduction

Biomedical research is increasingly focused on individualised prediction of complex outcomes which often cannot be summarised by a single number. In neuroscience studies, often the attention is focused on objects like curves and images which do not only show additional internal structure (such as spatial correlation) but also require computationally efficient statistical analysis due to their high-dimensionality.

These aspects play an even more fundamental role once the prediction of one image is based on other brain images. This is the case highlighted in Tavor et al. (2016), Parker Jones et al. (2017), Zheng et al. (2021) and other papers, where researchers used brain functional connectivity recorded when the subject was at rest to predict the brain activation of the same subject while performing an experimental task. Rather than studying the average activation in a group of subjects, their goal is to predict individual differences in the response to tasks, which could

hint towards potential different cognitive processes involved in solving the task itself. Those results obtained in a large selection of tasks from different cognitive domains could enhance the research on more difficult-to-study tasks for which activation localisation is not yet clear. In addition, this approach could open new ways to study task-evoked responses in those patients which cannot collaborate during the task itself (Tavor et al., 2016). A similar analysis carried on a population of pre-surgical patients (where higher individual variability is observed with respect to healthy controls) has also shown that the neural tissues associated with a specific language task can be identified using resting-state data (Parker Jones et al., 2017). The model proposed in Tavor et al. (2016) predicts the individual task activation based on the parcellation of the brain cortex in non-overlapping regions, using as distinct covariates information at the cortical as well as the subcortical level. The choice of a parcellation over another (and the implicit choice of modelling the signal with sharp boundaries instead of smooth transitions across brain regions) can play a relevant role, which is in part reduced by aggregating predictions obtained over multiple overlapping parcellation schemes (Dohmatob et al., 2021). A generalised linear model (GLM) is then fitted to each parcel of the cortex. The choice of the GLM, driven by computation efficiency, has been shown to be only slightly outperformed by more complicated methods such as neural networks and random forest bootstrap aggregation (Cohen et al., 2020).

We propose here an alternative approach based on functional data analysis (FDA) to predict task activation maps from resting-state data. In functional data analysis, the unit of interest is the whole "function" (in our case, a 3-dimensional brain image) and not the single value at each voxel. In other words, we aim at representing the underlying smooth brain signal and not its discretisation in voxels which could be potentially an additional source of noise. In particular, the setting of function-on-function regression is considered in this instance (Morris, 2015), being both the covariates and the outcome in our prediction model observed on the 3D domain.

Several approaches proposed for function-on-function regression rely on some form of dimension reduction, as even stacking the functions as 1-dimensional vectors becomes unfeasible when both the number of functions and the number of data points for each function are high (and this is standard in current neuroimaging applications, where different research projects are pushing towards the collection of imaging data at very high resolution for an increasing number of subjects). To this aim, basis expansions offer a useful solution. Splines or Fourier basis functions are often used to express the functional response and predictors

into smaller sets of coefficients. The functional regression coefficient is then obtained as a matrix of coefficients weighted by the same basis functions (Ramsay and Silverman, 2005). The model could also be enriched by applying roughness penalties on the basis expansion of the functional objects as well as regularisation on the functional regression coefficient (Ivanescu et al., 2015).

The basis expansion can also be used as a preliminary step to express functional data as a linear combination of data-driven, potentially interpretable functions. This happens for example with functional principal component analysis (FPCA, Ramsay and Silverman, 2005) which aims at identifying the main modes of variation in the covariance operator in a functional dataset. Yao et al. (2005) proposed to model both the predictor and the outcome using FPCA, then to regress the response FPC scores on the predictor ones. FPCA requires to select an approximation truncation level (namely the number of FPC scores to be retained) but in the regression setting there is no guarantee that the components which explain most of the variability in the functional covariates will be playing a role in the prediction of the response (Febrero-Bande et al., 2017).

An alternative technique which jointly takes into account the outcome and the covariates is functional partial least squares (FPLS). FPLS-based regression coefficients are obtained by means of an iterative procedure which produces a sequence of orthogonal functions maximising the covariance between the response and the predictors. Studied by Preda and Saporta (2005), Reiss and Ogden (2007), Aguilera et al. (2010), and Delaigle and Hall (2012) in the setting of functional regression with scalar response, PLS for function-on-function regression has been shown by Preda and Schiltz (2011) to be linked to the multivariate PLS on the basis expansion coefficients. Beyaztas and Shang (2020) extended this framework to the case of multiple functional predictors. Functional PLS, given its focus on predictive power of the covariates with respect to the response, usually leads to a lower-dimensional expansion with respect to FPCA (Delaigle and Hall, 2012).

In this work we apply FPLS function-on-function regression to predict the task activation using multiple 3D imaging covariates which contain resting-state information. The model is applied on a subset of 521 subjects from the UK Biobank repository. To the best of our knowledge, this is the first application of FPLS in a 3-dimensional setting, as well as in a neuroimaging context. In the task activation prediction application, this approach differs from the existing literature as it works on the whole-image level and does not require to choose a parcellation scheme. This modelling strategy also allows us to retain the spatial structure of

the signal by using smooth 3D basis functions without distinguishing cortical and subcortical information.

The work is structured as follows. In Section 5.2, an overview of PLS for function-on-function regression is provided, along with some aspects pertaining to the choice of the 3D basis functions. The predictor and response images drawn from the UK Biobank dataset are described in Section 5.3, while Section 5.4 illustrates the main findings in terms of the predictions of the task-evoked response maps. Lastly, Section 5.5 presents the conclusions and highlights further research directions.

5.2 Methods

5.2.1 PLS for functional regression

Functional PLS has gained attention in the field of functional data analysis thanks to its ability to model in a reduced dimensionality the covariance between the predictors and the outcome of interest. The functional random variables take values in $L^2(\mathcal{T})$, the space of square-integrable functions with domain $\mathcal{T}$ and finite second-order moments (Ramsay and Silverman, 2005).

The standard function-on-function regression model is of the form

$$Y_i^*(t) = \sum_{m=1}^{M} \int_{\mathcal{S}} X_{im}^*(s)\beta_m(s,t)ds + \epsilon_i^*(t), \tag{5.1}$$

where $*$ indicates centered variables.

The main structure of PLS for functional data illustrated here has been studied and applied in several papers, such as Preda and Schiltz (2011) and Beyaztas and Shang (2020). The main idea is to decompose both $X(t)$ and $Y(t)$ in terms of a series of uncorrelated random variables with mean zero, such that they attain maximum predictive ability. To this aim, the squared covariance between the outcome and the predictors is maximised, given the unit-norm constraints:

$$\max_{\zeta_X,\zeta_Y \in L^2, \|\zeta_X\|=\|\zeta_Y\|=1} \mathrm{Cov}^2\left(\int_{\mathcal{S}} X^*(s)\zeta_X(s)ds, \int_{\mathcal{T}} Y^*(t)\zeta_Y(t)dt\right). \tag{5.2}$$

The solution to this problem is given by an iterative procedure, of which we give a sketch here (see Preda and Schiltz, 2011 for the details and Febrero-Bande et al., 2017 for a more general version). At the ι-th step, the weight function $\zeta_X^{(\iota)}(s)$ is

computed by taking as input the covariance between $X^{*(\iota-1)}(s)$ and $Y^{*(\iota-1)}(t)$ (for $\iota = 1$, these correspond to the original functions $X^*(s)$ and $Y^*(t)$). Let $\rho^{(\iota)}$ be the inner product between $\zeta_X^{(\iota)}(s)$ and $X^{*(\iota-1)}(s)$ (this object is equal to the eigenvector corresponding to the largest eigenvalue of a product of certain operators derived from $X^{*(\iota-1)}(s)$ and $Y^{*(\iota-1)}(t)$). Then the linear regression models of $X^{*(\iota-1)}(s)$ and $Y^{*(\iota-1)}(t)$ on $\rho^{(\iota)}$ are computed. The weight function $\zeta_Y^{(\iota)}(t)$ is the OLS coefficient of the regression model for $Y^*(t)$. The error functions of these regression models become the starting point of the next iteration.

Both the outcome and predictor functions are recorded in practice on a discrete grid of points of cardinality J_X and J_Y within the spaces $\mathcal{S}$ and $\mathcal{T}$. Let therefore $\boldsymbol{Y}^*(t) = [\boldsymbol{Y}_i^*(t)]$, $i = 1, \ldots, N$ be the $N \times J_Y$ matrix that contains the discretised version of the N functional responses. Let also $\boldsymbol{X}^*(s) = [\boldsymbol{X}_{im}^*(s)]$, $i = 1, \ldots, N;\ m = 1, \ldots, M$ be the $(NM) \times J_X$ matrix of the discretised version of the M functional predictors (in this work we assume that all the predictor functions share the same domain).

A common step in functional regression is to approximate the functional observations as a linear combination of basis functions. Let $\{\phi_j(t)\}$ and $\{\chi_{lm}(s)\}$ be basis functions for $Y^*(t)$ and $X_m^*(t)$ respectively. We can approximate each individual observation using the following basis expansion:

$$\begin{aligned} \hat{y}_i^*(t) &= \sum_{j=1}^{K_Y} c_{ij}\phi_j(t) \\ \hat{x}_{im}^*(s) &= \sum_{l=1}^{K_{X_m}} d_{ilm}\chi_{lm}(s), \qquad m = 1, \ldots, M \end{aligned} \tag{5.3}$$

which in matrix form can be expressed as

$$\begin{aligned} \boldsymbol{Y}^*(t) &= \boldsymbol{C}\boldsymbol{\phi}(t) \\ \boldsymbol{X}_m^*(s) &= \boldsymbol{D}_m\boldsymbol{\chi}_m(s), \qquad m = 1, \ldots, M. \end{aligned} \tag{5.4}$$

The matrices $\boldsymbol{C}$ and $\boldsymbol{D}$ contain the coefficients of the basis expansion which could be estimated via ordinary least squares; the number of rows is N, the number of subjects, while the number of columns is K_Y and K_{X_m}, respectively. Let us also define the matrices $\boldsymbol{W}_\phi$ and $\boldsymbol{W}_{\chi_m}$ as the inner products of the basis functions for

$Y(t)$ and $X_m(t)$, namely

$$\begin{aligned} \boldsymbol{W}_\phi &= \int_{\mathcal{T}} \phi(t)\phi(t)^\mathsf{T} dt \\ \boldsymbol{W}_{\chi_m} &= \int_{\mathcal{S}} \chi_m(s)\chi_m(s)^\mathsf{T} ds, \qquad m = 1, \ldots, M. \end{aligned} \tag{5.5}$$

To exploit the main result in Preda and Schiltz (2011), we need to introduce the matrices

$$\boldsymbol{\Pi} = \boldsymbol{C}\boldsymbol{W}_\phi^{1/2} \qquad \text{and} \qquad \boldsymbol{\Lambda}_m = \boldsymbol{D}_m \boldsymbol{W}_{\chi_m}^{1/2}, \tag{5.6}$$

where the coefficient matrices $\boldsymbol{C}$ and $\boldsymbol{D}$ are postmultiplied by the square root matrices of the matrices defined in Equation (5.5). Given

$$\boldsymbol{\Lambda} = \left[\boldsymbol{\Lambda}_1^\mathsf{T} \ldots \boldsymbol{\Lambda}_M^\mathsf{T}\right]^\mathsf{T}, \tag{5.7}$$

Beyaztas and Shang (2020) show how to extend the result from Preda and Schiltz (2011) to the case of regression with multiple functional predictors. The main result is that the (functional) PLS regression of Y on X is equivalent to the PLS regression of $\boldsymbol{\Pi}$ on $\boldsymbol{\Lambda}$. This means that the functional PLS regression coefficient $\beta_{\text{PLS}}(s,t)$ at each PLS step is a function of the PLS regression coefficient $\boldsymbol{\Xi}$ in the regression model

$$\boldsymbol{\Pi} = \boldsymbol{\Lambda}\boldsymbol{\Xi} + \boldsymbol{\varepsilon}. \tag{5.8}$$

In other words, a multivariate PLS algorithm can be used in a functional context.

The prediction of new observations from the test set can be achieved using $\hat{\boldsymbol{\Xi}}$. Given a new observation with covariates $\tilde{X}(s)$, we express those as linear combination of the same basis functions $\boldsymbol{\chi}_m(s)$. We then compute the blocks of $\tilde{\boldsymbol{\Lambda}}$ of the form $\tilde{\boldsymbol{D}}_m \boldsymbol{W}_{\chi_m}^{1/2}$, $m = 1, \ldots, M$ and predict

$$\tilde{\boldsymbol{\Pi}} = \tilde{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Xi}}. \tag{5.9}$$

To obtain the uncentered version of $\tilde{\boldsymbol{\Pi}}$ (needed to obtain $\tilde{Y}(t)$ in the same scale of the original images) we add the mean of $\boldsymbol{\Pi} = \boldsymbol{C}\boldsymbol{W}_\phi^{1/2}$ in the training set.

The matrix $\tilde{\boldsymbol{\Pi}}$ — given the matrix $\boldsymbol{W}_\phi$ defined above — can be used to reconstruct $\tilde{Y}(t)$, using the formula:

$$\begin{aligned} \tilde{Y}(t) &= \tilde{\boldsymbol{\Pi}} \boldsymbol{W}_\phi^{-1/2} \boldsymbol{\phi}(t) \\ &= \tilde{\boldsymbol{C}} \boldsymbol{\phi}(t). \end{aligned} \tag{5.10}$$

Therefore, using the same decomposition in Equation (5.6) we are able to obtain the matrix $\tilde{C}$ of basis function coefficients, then the predicted outcomes $Y(t)$ is obtained as the product of the same matrix with the basis functions matrix.

5.2.1.1 Choice of basis functions

The validity of the theoretical framework of functional PLS described above is established regardless of the dimensionality of the functional data, which is instead taken care by choosing basis functions which allow for a more parsimonious representation of the data.

In the context of 3D brain images, we use the tensor product of univariate B-splines as described in Palma et al. (2020). A set of univariate B-splines is considered for each of the 3 dimensions. Depending on the needs of the analysis, the location and the number of knots (the points at which the local polynomials are joined together) as well as the degree of the splines can be selected separately for each dimension.

In this application, B-splines with a prespecified knot spacing are built in each dimensions, then a Kronecker product is taken to collect them into a matrix with dimensions V (number of voxels) and B (number of basis functions). A further check is made to keep only the voxels which fall within the mask and remove those basis functions which take nonzero values only outside the mask. In this formulation, the parameters to be selected are the degree of splines and the distance between knots. These two parameters together control the smoothness of the reconstructed image (that is, the prediction outcome of a regression model where the basis functions are used as covariates and the original—vectorised—image as the dependent variable).

5.3 Data

At the start of this research, a group of 521 subjects has been randomly sampled from the UK Biobank repository (Miller et al., 2016, Alfaro-Almagro et al., 2018) to conduct this analysis. For these subjects, both resting state and task activation images have been extracted. The acquisition information and the preprocessing steps are described in the UK Biobank documentation (Smith et al., 2018). In brief, for both resting-state and task functional MRI the images were acquired with identical scanners with image resolution equal to $2.4 \times 2.4 \times 2.4$ mm. The resting-state scans had a duration of 6 minutes, with 490 time points recorded for

each subject, and the data were preprocessed and registered to the standard MNI space (the Montreal Neurological Institute template which is commonly used to display brain imaging results).

5.3.1 Resting state fMRI (rfMRI)

In this study we use for each subject the dual regression images corresponding to the group Independent Component Analysis (ICA) decomposition (with $M_{temp} = 25$ components).

The ICA spatial maps constitute a data-driven parsimonious parcellation of the brain grey matter (Beckmann et al., 2009) that is common for an entire population. For UK Biobank, the ICA components were computed on a subset of more than 4000 subjects from the repository. These maps, used to identify the most relevant resting-state networks, were then used with a dual regression procedure (Nickerson et al., 2017), also known as back-projection (Calhoun et al., 2002), to obtain noisy estimates of subject-specific maps corresponding to the independent component maps. A brief overview of the procedure (which produces the covariates used in this work) is presented here, based on the extensive description in Nickerson et al., 2017 and on the compact version in Mejia et al., 2020.

Given the estimated template ICA data matrix $\hat{\boldsymbol{S}}_{temp}$ of dimensions $V \times M_{temp}$ whose columns are the vectorised ICA components, and the individual fMRI $V \times W_{time}$ data matrix $\boldsymbol{Z}_i$ (where V is the number of voxels and W_{time} is the number of time points), the first stage of dual regression returns an estimate of $\boldsymbol{B}_{TC}$, the subject-specific time series for each component network by means of a linear regression with multivariate outcome:

$$\boldsymbol{Z}_i = \hat{\boldsymbol{S}}_{temp}\boldsymbol{B}_{TC} + \boldsymbol{E}_i^{(1)}. \tag{5.11}$$

In the second stage of dual regression, $\hat{\boldsymbol{B}}_{TC}$ is used as design matrix for another linear regression model where the dependent variable is the dependent variable is the same fMRI data matrix used previously (transposed):

$$\boldsymbol{Z}_i^{\mathsf{T}} = \hat{\boldsymbol{B}}_{TC}^{\mathsf{T}}\boldsymbol{X}_i + \boldsymbol{E}_i^{(2)}. \tag{5.12}$$

The estimates of the second-stage dual regression images $\boldsymbol{X}_i$ (rearranged into M_{temp} 3D arrays) used in this study as covariates of the functional regression model are available in the `rfMRI_25.dr` folder within the UK Biobank repository. For each subject, the set of dual regression images is provided with a mask to to mark in–

brain voxels for the resting fMRI acquisition. The same mask across all dual regression maps is used, restricting the number of voxels to 228,483.

5.3.2 Task fMRI

Task fMRI time series for the 521 subjects is available on UK Biobank. Each individual time series is made of 332 points (for a duration of 4 minutes). In the preprocessing phase, a Gaussian kernel of FWHM 5mm is applied to get smoother images (Smith et al., 2018).

The contrast used in this analysis is "Faces-Shapes" (also defined as contrast 5 or `cope5`). It is classified as an emotion task (Hariri et al., 2002, Barch et al., 2013), where the subject is asked to match shapes or faces with the same expressions at the top and at the bottom of a screen. The faces have either angry or fearful expressions.

In this analysis we use the fixed-effect z-statistic maps (obtained using FEAT, fMRI Expert Analysis Tool, Woolrich et al., 2001) for each subjects. These maps contain z-values based on the task-induced activation of brain areas. We use the un-thresholded maps, standardised in MNI space. Each individual map comes with a mask already designed to mark those voxels which lie within the brain (therefore the ones useful for the analysis).

5.4 Results and discussion

5.4.1 Mask and basis expansion

To have the same set of basis functions across all images, we need to make sure that the same mask applies within each imaging modality. For the task fMRI images, we have built a common mask by selecting those voxels which appear in at least 90% of the masks (so in 472 or more subjects). The number of voxels in the common mask is 266,879. The voxels outside the common mask are excluded from further analysis.

Once a mask for each imaging modality is defined, we proceed with the basis expansion step. We follow the same procedure described in Palma et al. (2020) to produce a 3-dimensional tensor product of B-splines basis set within the mask. Given the matrix of evaluations of univariate B-splines for each dimension, the Kronecker product is considered and the basis functions whose non-zero values insist on the area outside the brain are discarded.

For the task fMRI maps, we first define the univariate B-splines functions degree to 3 (cubic B-splines). We then perform a sensitivity analysis for a random task activation map to display the effect of different choices of knot spacing onto the image reconstruction (keeping the same degree). The metrics used are the Pearson correlation and the mean absolute error (MAE) between the observed and reconstructed task-related image. Results are shown in Table 5.1. As the

Knot spacing	Number of bases	RMSE	MAE	Corr
2	47240	6.22	3.26	0.99
4	8560	16.34	9.65	0.93
6	3463	22.00	13.09	0.86
8	1910	25.66	15.16	0.81

Table 5.1: Results from sensitivity analysis for the reconstruction of one random task activation map from basis functions. The metrics are root mean squared error (RMSE), mean absolute error (MAE) and correlation between the raw image and the one reconstructed as linear combination of basis functions.

knot spacing increases (and therefore the number of basis functions decreases) the predicted image get slightly further from the original one. In order to find a good compromise between the complexity of subsequent analysis (driven by the number of basis functions) and the quality of the reconstruction, we decide to place knots in each dimensions every 6 voxels, for a total of 3463 basis functions.

A similar approach could be employed for the DR maps. For these images, we use a tensor product of univariate cubic B-splines with knot spacing equal to 6 voxels. The number of voxels within the mask for each DR map is 228,483. The matrix of basis functions is made of 3095 columns.

It is worth mentioning that being two different imaging modalities (with different mask dimensions as well) there is no real advantage in picking the same basis functions parameters for task and resting state images. This could instead enhance the interpretability of the results in case the same mask between two types of imaging is used.

5.4.2 Prediction with 4 covariates

We approximate the square root matrices built as in Beyaztas and Shang (2020) by deriving the spectral decomposition of the matrix of inner products, then taking the square roots of the eigenvalues. A real valued symmetric matrices $\boldsymbol{G}$ can

be raised to the power p by applying

$$\boldsymbol{G}^p = \boldsymbol{H}_G \boldsymbol{S}_G^p \boldsymbol{H}_G^\top, \tag{5.13}$$

where $\boldsymbol{S}_G$ is the diagonal matrix with the eigenvalues of G and $\boldsymbol{H}_F$ contains the eigenvectors of G. By storing in memory the eigendecomposition of both $\boldsymbol{W}_\phi$ and $\boldsymbol{W}_{\chi_m}$ (and setting to 0 the negative eigenvalues which arise for numerical reasons) it is possible to compute both the square roots ($p = \frac{1}{2}$) and the inverse square roots ($p = -\frac{1}{2}$) of these matrices by reusing the same eigenvectors and eigenvalues twice.

We select $M = 4$ DR maps as covariates in the functional PLS model. This selection is based on previous knowledge on the phenomenon under study. In the list of 25 ICA components[1], we have selected a few referring to the default mode network (component 1) and the visual network (component 2, component 9, component 20), which are more relevant for the task under study. For each DR map corresponding to the ICA components mentioned above, we extract the $K_{X_m} = 3095$ B-splines coefficients. The matrix $\boldsymbol{D}_m$ containing these coefficients is then multiplied by the square root matrix $\boldsymbol{W}_{\chi_m}^{1/2}$ to obtain $\boldsymbol{\Lambda}_m, m = 1, \ldots, 4$.

The sample is split in training (421 subjects) and test set (100 subjects). We use the SIMPLS algorithm as coded in the function `pls.regression` of the `R` package `plsgenomics` (Boulesteix et al., 2018) and we consider the coefficient obtained from the first 10 PLS component.

We use Pearson correlation between predicted and observed task maps for each subject in the test set as the main prediction quality measure. The histogram in Figure 5.1 shows that the average correlation is equal to 0.37 (with 75% of the correlations above 0.30), with only a few of them being close to 0 or negative.

Although it does not clearly include the spatial structure of the data, the Pearson correlation between predicted and observed maps can be used as a single summary of prediction quality. In Figure 5.2, for a subject in the test set with high correlation between predicted and actual task map, we observe some agreement in terms of the sign of the z-values in different areas in the brain. In particular, in both images higher positive values are observed in the occipital lobe (where primary visual areas are located, see second rows in Figure 5.2), while negative values appear in the posterior parietal lobe.

[1]The list of the group-ICA resting state networks is available on `https://www.fmrib.ox.ac.uk/ukbiobank/`. In this work, the indices of the components refer to the whole sequence of 25 components, not to the subset of "good" components which does not include the 4 "artifactual" components.

We also carried out further analyses to evaluate the change in the prediction performance after increasing the number of functional PLS components selected to build the functional PLS regression coefficient. The mean and the standard deviations of correlations in the test set appear to be stable after increasing the number of components (results not shown here).

In addition, we extend our consideration to all the pairs of predicted-observed task activation images in the test set. Ideally the individual differences between task-activated brain maps should emerge when for the i-th subject the i-th predicted map agrees more with the corresponding observed image rather than the observed images for other subjects. We first consider the cross-correlation of all the observed images and the cross-correlation of all the predicted ones in Figure 5.3. Outside of the main diagonal, the cross-correlation of the functional PLS predicted images is much higher than in the observed ones (on average, this cross-correlation is approximately equal to 0.80, while for the observed images is equal to 0.19). In Figure 5.4, we also look at the 100×100 matrix of cross-correlations between actual images (columns) and predicted ones from functional PLS (rows). The main diagonal of this matrix corresponds to the values in the histogram in Figure 5.1. We clearly observe vertical stripes which are due to the combined effect of the higher cross-correlation between predictions and the higher variability in the actual maps rather than in the predicted ones. This agrees with the results described in the same task (and many others) in Tavor et al. (2016). Nonetheless, differently from other work in the literature, the distribution of the values outside the main diagonal is not different from the one of those in the main diagonal. The histogram of prediction ranks (the position of the correlation on the main diagonal within the list of sorted correlations for each subject), used in the Supplementary Material in Parker Jones et al. (2017) to assess goodness of fit, does not provide evidence of skewness towards the higher ranks (plot not reported here).

In terms fo computational time, there are large differences between the different steps of the analysis. Obtaining the B-splines coefficients takes less than a minute per image (and this step is parallelisable). The full spectral decomposition of both the matrices of inner product to obtain the weights matrices W take approximately 3 minutes for the task data and 2 minutes for the resting state data and the computational time depends on the number of basis chosen. The PLS fitting step took approximately 7 minutes per PLS component in a standard laptop.

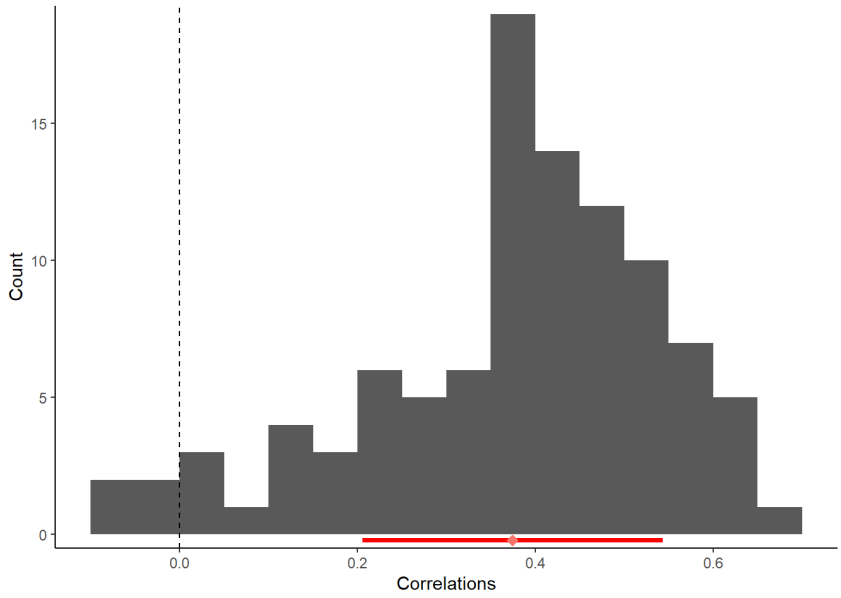

Figure 5.1: Histogram of Pearson correlation between predicted and observed brain maps in the test set (100 subjects). The mean is denoted by the red diamond; the red line shows the variability ($\pm$ 1 standard deviation)

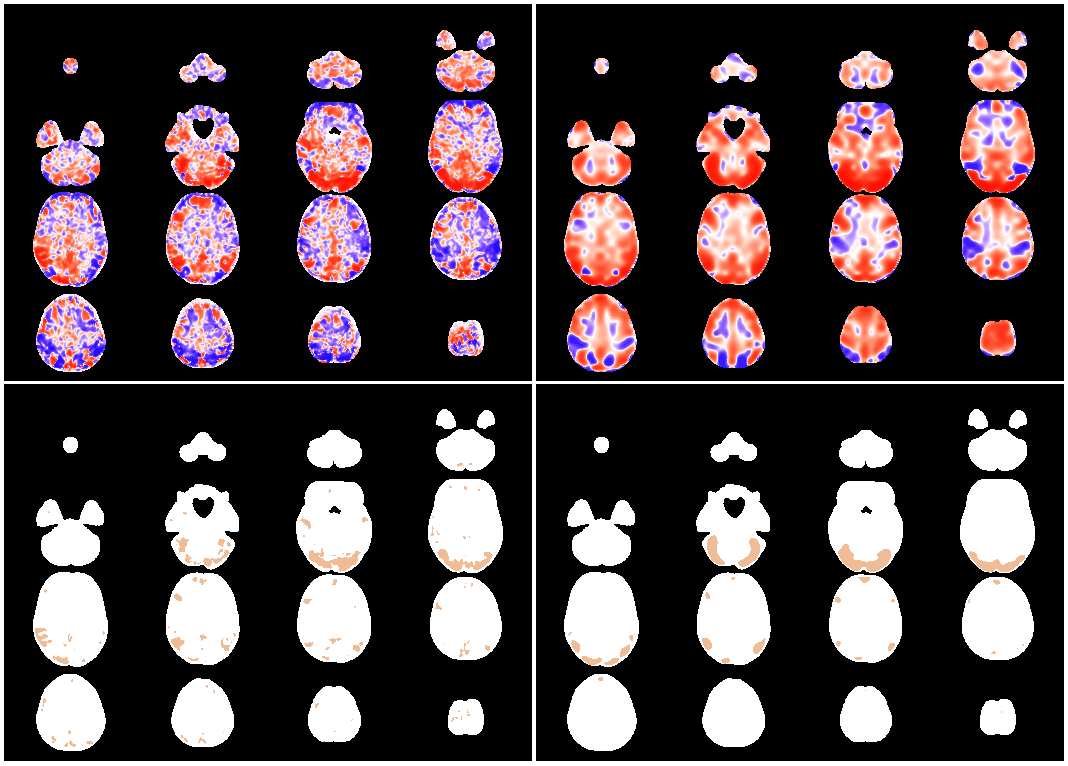

Figure 5.2: Top: axial slices of the observed (left) and predicted (right) task maps for one randomly selected subject in the test set. The Pearson correlation between them is equal to 0.415. Blue indicates negative values, while red is used for positive values. The midpoint is located at 0 (white) but the legend range of the actual image is larger.
Bottom: axial slices of the observed (left) and predicted (right) quantile maps thresholded at the 95th percentile. Light brown indicates areas with values above the threshold.

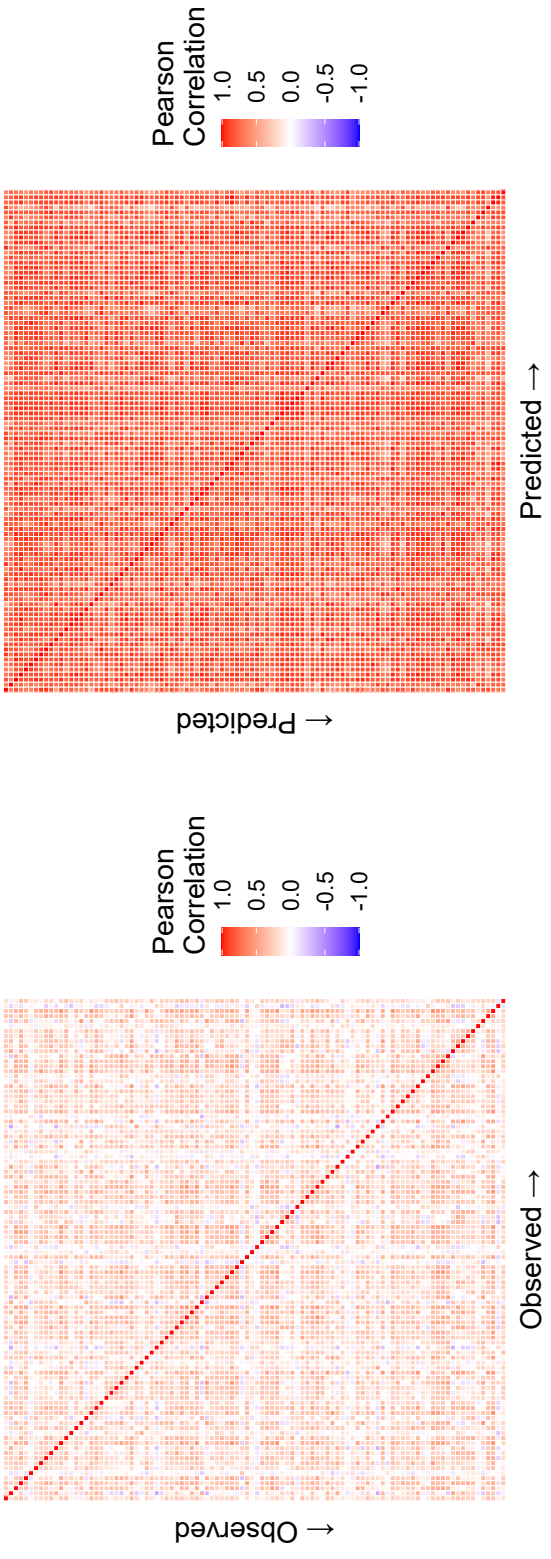

Figure 5.3: Left: correlation heatmap between every pair of observed task activation maps in the test set. Right: correlation heatmap between every pair of predicted task activation maps from PLS in the test set.

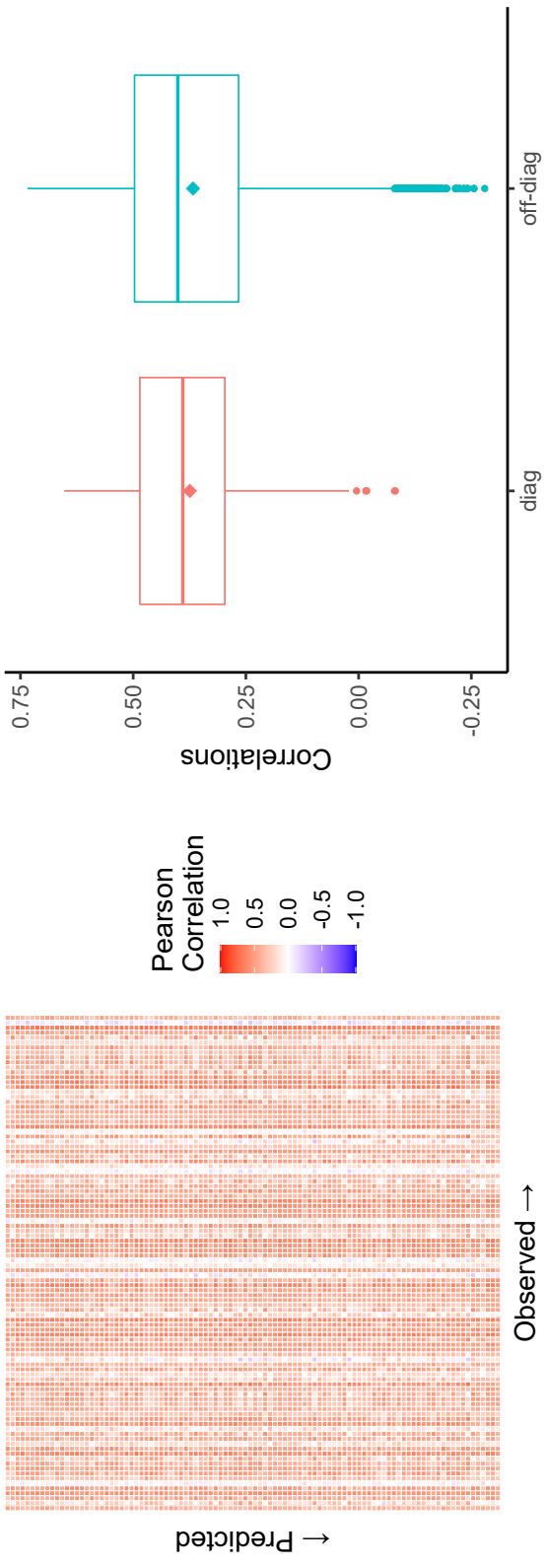

Figure 5.4: Left: correlation heatmap between the observed and predicted task activation maps from the test set. The main diagonal shows correlation between predicted and observed images from the same subjects, while the elements outside the diagonal refer to cross-correlations between pairs of subjects. Right: boxplot of the on- and off-diagonal elements. The diamonds mark the group means.

5.5 Conclusions

In this work we have employed partial least squares function-on-function regression to predict a 3-dimensional outcome from multiple imaging covariates. The regression model proposed can accommodate a large number of data points per image by using spatially-informed basis functions such as tensor products of univariate B-splines. The coefficients obtained from this expansion are then used in a multivariate PLS algorithm and the functional regression coefficient is built using PLS components. To predict the image outcome for a new subject, the functional regression coefficient is simply multiplied by the observed covariates.

In our application, we aimed at predicting task activation images using functional connectivity at rest, which was represented through dual regression maps linked to 4 preselected ICA components. For the "Faces-Shapes" contrast, we have observed moderate correlation between predicted and observed task activation maps for the large part of subjects in the test set. When cross-correlations between an individual predicted map and the maps observed for other subjects were considered, no large differences with respect to the individual observed maps were noticed. Predicted maps show in general a lower range and voxelwise variability (as also reported in Tavor et al., 2016), although the shape of brain regions with high and low z-values showed a good matching with the corresponding observed maps. To summarise, we have noticed some correlation between the observations and the outcomes of our model but not all the individual differences in patterns of activation were detected.

This work can be extended in multiple ways. From a methodological point of view, the evaluation of the choice of the parameters in the analysis (from the knot spacing for the basis expansion of the predictors and the response, to the number of PLS components that constitute the functional regression coefficient) needs further sensitivity analysis. At the moment, several criteria exist in the literature but there is no clear advantage of one over the others. The choice of the analyst remains crucial to ensure a good quality of the results.

Furthermore, alternative basis functions could also be considered, in order to improve the fitting on irregular domains as those observed in brain images (in this direction, see the multivariate splines on triangulations in Yu et al., 2021) or reduce the number of coefficients to be retained for the PLS step by means of roughness penalties. In this sense, the computational efficiency is a core concept that needs to be considered.

Functional PLS is drawing increasing attention from theoretical and applied perspectives thanks to the recent evolution of this subject in the statistical literature. Extending results to its multivariate counterpart is still beneficial to the functional data approaches as well. In particular, alternative versions of functional PLS (as in Delaigle and Hall, 2012, then extended in Zhou, 2021) have been proposed, which are also based on non-iterative procedures. This would be highly beneficial to speed up the computation.

In addition, there is room for work in automatic selection and sparsity constraints in PLS-based regression (following for examples the approaches proposed in the multivariate statistics literature in de Micheaux et al., 2019, Sutton et al., 2018). The core idea, translated in the 3D functional setting, is that regularisation might help to discard entire covariates (or regions of them) when they are non-informative for the prediction of the response image. This could represent an interesting PLS algorithm with computationally efficient implementations able to handle big data such as those from neuroimaging repositories like UK Biobank and ADNI.

From the application point of view, many questions remain open. The phenomenon of the much higher variability in the observed rather than in the predicted task-evoked response maps, reported in this work as well as in others with different parcellation-based regression modelling approaches (as for example in Tavor et al., 2016 and Parker Jones et al., 2017), seems not to have a clear immediate explanation. On a positive side, often the main goal of the analysis is to provide a thresholded task-activation map (which can be driven by a voxelwise ranking, rather than the absolute values), but nevertheless this aspect needs further investigation.

Our approach can be enriched with other kind of neuroimaging covariates, both scalar or images. Zheng et al. (2021) approach the same task activation prediction problem by using a different set of features with respect to dual regression and focusing on residualised images. The functional PLS approach could easily include these alternative covariates. Furthermore, structural MRI scans, which inform about the morphology of different brain regions, can be used as additional covariates (although in Tavor et al., 2016 they do not appear to play a relevant role in the prediction). Our modelling approach can be also followed by subsequent analysis (for instance the thresholding of activation maps) and can be replicated for other tasks and more generally other high-dimensional image-on-image regression settings.

Chapter 6

Conclusions

Functional data analysis is currently a fast-developing area of statistics which benefits from advancement in multivariate analysis and nonparametric smoothing. It represents a valid option for the analysis of high-dimensional data with complex correlation structure, as commonly observed in neuroimaging settings.

In this work we have described several approaches based on functional data analysis for 3-dimensional data. We have illustrated some workflows for the analysis of large-scale brain imaging data in real-world neuroscience problems. We have also implemented those approaches in R in a flexible and customisable set of functions. We have shown that computation efficiency can be achieved without the need of imposing strict normality assumptions when either the data suggest they may be inappropriate (skewed functional data analysis) or the specific application might benefit from a more flexible modelling choice (functional quantile regression). Even in a setting of multivariate functional data analysis (such as functional partial least squares), we develop fast and scalable modelling strategies in a case where loading all the data in memory can be prohibitive.

In terms of neuroimaging applications, this thesis has proposed various models for structural as well as functional magnetic resonance imaging and addressed questions which are currently of large interest in the brain science community.

This work contributes to the literature of the brain age problem by introducing a simple way of considering prediction variability and opens new potential avenues in the direction of detecting early signs of diseases. The workflow proposed takes as input a tensor-based morphometry image, but the application to other imaging modalities is straightforward. The output of the model (scalar-on-image quantile regression) is a prediction interval whose features are customisable to suit different needs in the clinical practice. We have shown that the predictions obtained using this workflow are plausible if compared to the relationship between chronological age and neurodegenerative diseases and the prediction quality achieved is in line with more computationally intensive modelling strategies. The main features of the regression coefficients estimated are also coherent with the effect of ageing over the brain structure which are commonly reported on the literature.

A further exploration of tensor-based morphometry images presented in this work has allowed us to highlight some interesting features of this modality, which have been modelled by means of a skewed probability distribution and a Gaussian copula. In the normative modelling setup, for each subject a new normative map has been generated, in order to detect spatial patterns of deviation with respect to the healthy population. Some simple indices of deviations are useful to detect some hints of neurodegeneration in the subjects in the test set, providing the basis for the construction of individual risk scores.

This thesis shows also an application with functional imaging, i.e. tackling the problem of assessing individual task response from resting-state data. A regression model based on functional partial least squares is able to accommodate several imaging covariates with high dimensionality, via a B-splines basis expansion. The result of this analysis show that moderate correlation is achieved between predicted and observed task activation maps for a large number of subjects in the test set. In addition, the areas of the brain which are predicted to be activated in response to the task largely agree with what observed.

The analysis of 3D imaging data using FDA is a wide open research field and this thesis has provided only an overview on a limited range of potential problems and solutions. In this thesis we have not explored the whole range of basis functions available in the literature for nonparametric smoothing. We have relied only on two simple approaches (B-splines tensor products and radial basis func-

tions) that could give a good enough approximation of the original image to be used in the next steps of the analysis. In principle, three main features play a role in choosing how to smooth brain images: the multidimensional domain, the high resolution and the irregularity of the mask. A good smoother should handle the 3D domain (for example as a product of univariate functions) in a computationally efficient way (as the number of voxels is in the order of millions) and ideally not lead to large reconstruction errors at the boundaries of the mask.

While many appealing solutions to this issue exist already in the literature, they might be demanding even for a moderate number of voxels and therefore not appealing as a first step in FDA. In addition, it would be also beneficial to have parsimonious basis functions representations to speed up the model fitting procedures that come after the smoothing step. In this thesis the focus on simple techniques often reduced the task to control the smoothness of the function (and therefore the bias-variance trade-off of the image reconstruction) by imposing a common discretisation step (or knot spacing) across all the domain of the images. It is worth mentioning that, depending on the imaging modalities and acquisition, the preprocessing phase of brain images often already contains a smoothing step, and the signal does not necessarily show the same level of smoothness across the whole brain. This leaves a lot of room for improvement in this field: appealing approaches in this directions might be penalised and adaptive smoothing (Marx and Eilers, 2005, Lindquist et al., 2010), or splines over triangulations (Yu et al., 2021) which better describe the irregular domains, at whose boundaries tensor product splines or kernel smoothing might perform poorly. These alternative smoothing techniques can then be embedded in the methods applied in this thesis (and mainly the whole conceptual framework of FDA is valid for any choice related to smoothing).

In addition to the sample size, the choice of the common discretisation step plays a role in the computational aspects of the models proposed in this work. The number of basis function coefficients has been kept in the order of a few thousands in order to perform matrix operations (for example matrix inversions and spectral decompositions) faster without being forced to use high performance computing resources. Especially for functional partial least squares and in general all multivariate FDA applications, this issue limits the choices of the statistician, at the cost of higher reconstruction error or lower number of imaging elements to be included. Functional variable selection could prove to be a useful resource in this setting.

Functional variable selection might be intended in different ways. When dealing with more than one functional covariate, we could be interested in checking whether an image has an impact on the estimation of a certain outcome, but we could as well want to detect which regions of the image play a role in the model. The applications of FDA addressed in this thesis were more prediction-oriented, but actually large part of the literature in this field deals with estimation and inferential problems. Functional domain selection (James et al., 2009, Park et al., 2016), local inference for functional data (Olsen et al., 2020, Pini and Vantini, 2017) and points of impact (Kneip et al., 2016) are just some examples that could help in this direction. Several neuroimaging applications are aimed at establishing thresholds to detect regions with significant effects as well (via random field theory for example) while controlling the family wise error rate or the false discovery rate. If similarities in scope exist, cross-fertilisation would be highly beneficial for both the disciplines.

Lastly, multidimensional functional data analysis is still an increasing area of research. While some modelling aspects are already under development, the need for innovative visualisation tools is still alive: even a concept as simple as correlation between 3D object requires at least some kind of interactive visualisation. This issue would be decisive in view of a future wide-ranging use of FDA tools in a clinical setting. In this direction, a higher engagement of statisticians with these data coming from biomedical applications would be welcomed towards facing one of the greatest battles of the next decades and building a better health for all.

Bibliography

A. M. Aguilera, M. Escabias, C. Preda, and G. Saporta. Using basis expansions for estimating functional PLS regression: applications with chemometric data. *Chemometrics and Intelligent Laboratory Systems*, 104(2):289–305, 2010.

F. Alfaro-Almagro, M. Jenkinson, N. K. Bangerter, J. L. Andersson, L. Griffanti, G. Douaud, S. N. Sotiropoulos, S. Jbabdi, M. Hernandez-Fernandez, E. Vallee, et al. Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *Neuroimage*, 166:400–424, 2018.

G. I. Allen. Multi-way functional principal components analysis. In *2013 5th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 220–223. IEEE, 2013.

L. G. Apostolova, A. E. Green, S. Babakchanian, K. S. Hwang, Y.-Y. Chou, A. W. Toga, and P. M. Thompson. Hippocampal atrophy and ventricular enlargement in normal aging, mild cognitive impairment and alzheimer's disease. *Alzheimer disease and associated disorders*, 26(1):17, 2012.

R. B. Arellano-Valle and A. Azzalini. The centred parametrization for the multivariate skew-normal distribution. *Journal of multivariate analysis*, 99(7):1362–1382, 2008.

J. Ashburner and K. Friston. Morphometry. In R. S. J. Frackowiak, K. J. Friston, C. D. Frith, R. J. Dolan, C. J. Price, S. Zeki, J. T. Ashburner, and W. D. Penny, editors, *Human Brain Function*, chapter 36, pages 707–722. Elsevier, 2 edition, 2004.

J. Ashburner, G. Barnes, C. Chen, J. Daunizeau, G. Flandin, K. Friston, S. Kiebel, J. Kilner, V. Litvak, R. Moran, et al. SPM12 manual. *London: Wellcome Trust*, 2014.

A. Azzalini. *The skew-normal and related families*, volume 3. Cambridge University Press, 2013.

A. Azzalini. *The R package `sn`: The Skew-Normal and Related Distributions such as the Skew-t (version 1.5-5)*. Università di Padova, Italia, 2020. URL `http://azzalini.stat.unipd.it/SN`.

D. M. Barch, G. C. Burgess, M. P. Harms, S. E. Petersen, B. L. Schlaggar, M. Corbetta, M. F. Glasser, S. Curtiss, S. Dixit, C. Feldt, et al. Function in the human connectome: task-fMRI and individual differences in behavior. *Neuroimage*, 80:169–189, 2013.

C. F. Beckmann, C. E. Mackay, N. Filippini, and S. M. Smith. Group comparison of resting-state FMRI data using multi-subject ICA and dual regression. *Neuroimage*, 47(Suppl 1):S148, 2009.

A. Belloni and V. Chernozhukov. ℓ1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130, 2011.

U. Beyaztas and H. L. Shang. On function-on-function regression: partial least squares approach. *Environmental and Ecological Statistics*, 27(1):95–114, 2020.

A.-L. Boulesteix, G. Durif, S. Lambert-Lacroix, J. Peyre, and K. Strimmer. *plsgenomics: PLS Analyses for Genomics*, 2018. R package version 1.5-2.

J. P. Boyd. Six strategies for defeating the Runge phenomenon in Gaussian radial basis functions on a finite interval. *Computers & Mathematics with Applications*, 60(12):3108–3122, 2010.

S. Brockhaus, A. Fuest, A. Mayr, and S. Greven. Signal regression models for location, scale and shape with an application to stock returns. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(3):665–686, 2018.

D. Bzdok, T. E. Nichols, and S. M. Smith. Towards algorithmic analytics for large-scale datasets. *Nature Machine Intelligence*, 1(7):296–306, 2019.

R. Cabeza and N. Dennis. Frontal lobes and aging: deterioration and compensation. In *Principles of Frontal Lobe Function*, volume 2, pages 628–652. Oxford University Press: New York., 2013.

B. S. Cade and B. R. Noon. A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment*, 1(8):412–420, 2003.

V. D. Calhoun, T. Adali, G. D. Pearlson, and J. J. Pekar. A method for making group inferences from functional MRI data using independent component analysis. *Human Brain Mapping*, 16(2):131–131, jun 2002. ISSN 1065-9471.

H. Cardot, C. Crambes, and P. Sarda. Quantile regression when the covariates are functions. *Nonparametric Statistics*, 17(7):841–856, 2005.

J. C. Carr, R. K. Beatson, J. B. Cherrie, T. J. Mitchell, W. R. Fright, B. C. McCallum, and T. R. Evans. Reconstruction and representation of 3D objects with radial basis functions. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 67–76, 2001.

K. Chen and H.-G. Müller. Conditional quantile analysis when covariates are functions, with application to growth data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):67–89, 2012.

X. Chen, W. Liu, Y. Zhang, et al. Quantile regression under memory constraint. *The Annals of Statistics*, 47(6):3244–3273, 2019.

Y. Chen, W. K. Härdle, Q. He, and P. Majer. Risk related brain regions detection and individual risk classification with 3D image FPCA. *Statistics & Risk Modeling*, 35 (3-4):89–110, 2018.

V. Chernozhukov, I. Fernández-Val, and A. Galichon. Quantile and probability curves without crossing. *Econometrica*, 78(3):1093–1125, 2010.

I. H. Choo, D. Y. Lee, J. S. Oh, J. S. Lee, D. S. Lee, I. C. Song, J. C. Youn, S. G. Kim, K. W. Kim, J. H. Jhoo, et al. Posterior cingulate cortex atrophy and regional cingulum disruption in mild cognitive impairment and alzheimer's disease. *Neurobiology of aging*, 31(5):772–779, 2010.

M. K. Chung. *Computational neuroanatomy: The methods*. World Scientific, 2013.

M. K. Chung, K. J. Worsley, S. Robbins, T. Paus, J. Taylor, J. N. Giedd, J. L. Rapoport, and A. C. Evans. Deformation-based surface morphometry applied to gray matter deformation. *NeuroImage*, 18(2):198–213, 2003.

L. R. Clark, L. Delano-Wood, D. J. Libon, C. R. McDonald, D. A. Nation, K. J. Bangen, A. J. Jak, R. Au, D. P. Salmon, and M. W. Bondi. Are empirically-derived subtypes of mild cognitive impairment consistent with conventional subtypes? *Journal of the International Neuropsychological Society*, 19(6):635–645, 2013.

A. D. Cohen, Z. Chen, O. Parker Jones, C. Niu, and Y. Wang. Regression-based machine-learning approaches to predict task activation using resting-state fmri. *Human brain mapping*, 41(3):815–826, 2020.

J. H. Cole. Neuroimaging-derived brain-age: an ageing biomarker? *Aging (Albany NY)*, 9(8):1861, 2017.

J. H. Cole, R. P. Poudel, D. Tsagkrasoulis, M. W. Caan, C. Steves, T. D. Spector, and G. Montana. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage*, 163:115–124, 2017.

J. H. Cole, K. Franke, and N. Cherbuin. Quantification of the biological age of the brain using neuroimaging. In *Biomarkers of Human Aging*, pages 293–328. Springer, 2019.

C. M. Crainiceanu, A.-M. Staicu, and C.-Z. Di. Generalized multilevel functional regression. *Journal of the American Statistical Association*, 104(488):1550–1561, 2009.

C. Davatzikos. Machine learning in neuroimaging: Progress and challenges. *Neuroimage*, 197:652, 2019.

C. Davino, M. Furno, and D. Vistocco. *Quantile regression: theory and applications*. John Wiley & Sons, 2013.

C. De Boor. *A practical guide to splines*, volume 27. Springer-Verlag New York, 1978.

P. L. de Micheaux, B. Liquet, and M. Sutton. PLS for big data: A unified parallel algorithm for regularised group PLS. *Statistics Surveys*, 13:119–149, 2019.

A. Delaigle and P. Hall. Methodology and theory for partial least squares applied to functional data. *The Annals of Statistics*, 40(1):322–352, 2012.

N. A. Dennis and R. Cabeza. Neuroimaging of healthy cognitive aging. In *The handbook of aging and cognition*, pages 10–63. Psychology Press, 2011.

P. Denver and P. L. McClean. Distinguishing normal brain aging from the development of Alzheimer's disease: inflammation, insulin signaling and cognition. *Neural regeneration research*, 13(10):1719, 2018.

E. Dohmatob, H. Richard, A. L. Pinho, and B. Thirion. Brain topography beyond parcellations: local gradients of functional maps. *NeuroImage*, 229:117706, 2021.

G. E. Fasshauer and J. G. Zhang. On choosing "optimal" shape parameters for RBF approximation. *Numerical Algorithms*, 45(1-4):345–368, 2007.

M. Febrero-Bande, P. Galeano, and W. González-Manteiga. Functional principal component regression and functional partial least-squares regression: an overview and a comparative study. *International Statistical Review*, 85(1):61–83, 2017.

F. Ferraty and P. Vieu. *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media, 2006.

D. Firth. Multiplicative errors: log-normal or gamma? *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):266–268, 1988.

B. Fitzenberger, R. Koenker, and J. A. Machado. *Economic applications of quantile regression*. Springer Science & Business Media, 2013.

A. M. Fjell, K. B. Walhovd, C. Fennema-Notestine, L. K. McEvoy, D. J. Hagler, D. Holland, J. B. Brewer, and A. M. Dale. One-year brain atrophy evident in healthy aging. *Journal of Neuroscience*, 29(48):15223–15231, 2009.

A. M. Fjell, L. McEvoy, D. Holland, A. M. Dale, K. B. Walhovd, A. D. N. Initiative, et al. What is normal in normal aging? Effects of aging, amyloid and Alzheimer's disease on the cerebral cortex and the hippocampus. *Progress in Neurobiology*, 117:20–40, 2014.

K. Franke and C. Gaser. Ten years of brainAGE as a neuroimaging biomarker of brain aging: What insights have we gained? *Frontiers in Neurology*, 10:789, 2019. ISSN 1664-2295. doi: 10.3389/fneur.2019.00789.

K. Franke, E. Luders, A. May, M. Wilke, and C. Gaser. Brain maturation: predicting individual BrainAGE in children and adolescents using structural MRI. *Neuroimage*, 63(3):1305–1312, 2012.

D. Freedman and P. Diaconis. On the histogram as a density estimator: L_2 theory. *Probability theory and related fields*, 57(4):453–476, 1981.

K. J. Friston, A. P. Holmes, K. J. Worsley, J.-P. Poline, C. D. Frith, and R. S. Frackowiak. Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping*, 2(4):189–210, 1994.

J. E. Gellar, E. Colantuoni, D. M. Needham, and C. M. Crainiceanu. Cox regression models with functional covariates for survival data. *Statistical modelling*, 15(3): 256–278, 2015.

J. Goldsmith, L. Huang, and C. M. Crainiceanu. Smooth scalar-on-image regression via spatial bayesian variable selection. *Journal of Computational and Graphical Statistics*, 23(1):46–64, 2014.

J. J. Hanfelt, J. Wuu, A. B. Sollinger, M. C. Greenaway, J. J. Lah, A. I. Levey, and F. C. Goldstein. An exploration of subgroups of mild cognitive impairment based on cognitive, neuropsychiatric and functional features: analysis of data from the National Alzheimer's Coordinating Center. *The American Journal of Geriatric Psychiatry*, 19(11):940–950, 2011.

C. Happ. *MFPCA: Multivariate Functional Principal Component Analysis for Data Observed on Different Dimensional Domains*, 2018. R package version 1.2-2.

C. Happ and S. Greven. Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association*, 113(522):649–659, 2018.

C. Happ, S. Greven, and V. J. Schmid. The impact of model assumptions in scalar-on-image regression. *Statistics in medicine*, 37(28):4298–4317, 2018.

A. R. Hariri, A. Tessitore, V. S. Mattay, F. Fera, and D. R. Weinberger. The amygdala response to emotional stimuli: a comparison of faces and scenes. *Neuroimage*, 17(1):317–323, 2002.

T. J. Hastie and R. J. Tibshirani. *Generalized additive models*. Routledge, 2017.

G. Heinze, C. Wallisch, and D. Dunkler. Variable selection–a review and recommendations for the practicing statistician. *Biometrical Journal*, 60(3):431–449, 2018.

N. J. Higham. Cholesky factorization. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(2):251–254, 2009.

L. Horváth and P. Kokoszka. *Inference for functional data with applications*, volume 200. Springer Science & Business Media, 2012.

T. Hsing and R. Eubank. *Theoretical foundations of functional data analysis, with an introduction to linear operators*. John Wiley & Sons, 2015.

X. Hua, A. D. Leow, S. Lee, A. D. Klunder, A. W. Toga, N. Lepore, Y.-Y. Chou, C. Brun, M.-C. Chiang, M. Barysheva, et al. 3D characterization of brain atrophy in Alzheimer's disease and mild cognitive impairment using tensor-based morphometry. *Neuroimage*, 41(1):19–34, 2008.

X. Hua, D. P. Hibar, C. R. Ching, C. P. Boyle, P. Rajagopalan, B. A. Gutman, A. D. Leow, A. W. Toga, C. R. Jack Jr, D. Harvey, M. W. Weiner, P. M. Thompson, and the Alzheimer's Disease Neuroimaging Initiative. Unbiased tensor-based morphometry: improved robustness and sample size estimates for Alzheimer's disease clinical trials. *Neuroimage*, 66:648–661, 2013.

A. E. Ivanescu, A.-M. Staicu, F. Scheipl, and S. Greven. Penalized function-on-function regression. *Computational Statistics*, 30(2):539–568, 2015.

G. M. James, J. Wang, and J. Zhu. Functional linear regression that's interpretable. *The Annals of Statistics*, 37(5A):2083–2108, 2009.

M. J. Kane, J. Emerson, and S. Weston. Scalable strategies for computing with massive data. *Journal of Statistical Software*, 55(14):1–19, 2013.

J. Kang, B. J. Reich, and A.-M. Staicu. Scalar-on-image regression via the soft-thresholded Gaussian process. *Biometrika*, 105(1):165–184, 2018.

K. Kato. Estimation in functional linear quantile regression. *The Annals of Statistics*, 40(6):3108–3136, 2012.

A. Kneip, D. Poß, and P. Sarda. Functional linear regression with points of impact. *The Annals of Statistics*, 44(1):1–30, 2016.

R. Koenker and G. Bassett. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

R. Koenker and K. F. Hallock. Quantile regression. *Journal of economic perspectives*, 15(4):143–156, 2001.

P. Kokoszka and M. Reimherr. *Introduction to functional data analysis*. CRC Press, 2017.

D. Kong, J. G. Ibrahim, E. Lee, and H. Zhu. Flcrm: Functional linear cox regression model. *Biometrics*, 74(1):109–117, 2018.

J. L. Krichmar, J. L. Olds, J. V. Sanchez-Andres, and H. Tang. Explainable artificial intelligence and neuroscience: Cross-disciplinary perspectives. *Frontiers in Neurorobotics*, page 105, 2021.

J. K. Kueper, M. Speechley, and M. Montero-Odasso. The Alzheimer's disease assessment scale–cognitive subscale (ADAS-Cog): modifications and responsiveness in pre-dementia populations. a narrative review. *Journal of Alzheimer's Disease*, 63(2):423–444, 2018.

R. Leech and D. J. Sharp. The role of the posterior cingulate cortex in cognition and disease. *Brain*, 137(1):12–32, 2014.

A. D. Leow, I. Yanovsky, M.-C. Chiang, A. D. Lee, A. D. Klunder, A. Lu, J. T. Becker, S. W. Davis, A. W. Toga, and P. M. Thompson. Statistical properties of Jacobian maps and the realization of unbiased large-deformation nonlinear image registration. *IEEE transactions on medical imaging*, 26(6):822–832, 2007.

M. Li, A.-M. Staicu, and H. D. Bondell. Incorporating covariates in skewed functional data models. *Biostatistics*, 16(3):413–426, 2015.

E. Lila and J. A. Aston. Statistical analysis of functions on surfaces, with an application to medical imaging. *Journal of the American Statistical Association*, 115 (531):1420–1434, 2020.

F. Lin, P. Ren, M. Mapstone, S. P. Meyers, A. Porsteinsson, T. M. Baran, A. D. N. Initiative, et al. The cingulate cortex of older adults with excellent memory capacity. *Cortex*, 86:83–92, 2017.

M. A. Lindquist, J. M. Loh, and Y. R. Yue. Adaptive spatial smoothing of fMRI images. *Statistics and its Interface*, 3(1):3–13, 2010.

T. J. Littlejohns, J. Holliday, L. M. Gibson, S. Garratt, N. Oesingmann, F. Alfaro-Almagro, J. D. Bell, C. Boultwood, R. Collins, M. C. Conroy, et al. The uk biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nature communications*, 11(1):1–12, 2020.

S. N. Lockhart and C. DeCarli. Structural imaging measures of brain aging. *Neuropsychology review*, 24(3):271–289, 2014.

S. López-Pintado and J. Romo. On the concept of depth for functional data. *Journal of the American Statistical Association*, 104(486):718–734, 2009.

S. E. MacPherson and S. R. Cox. *The frontal ageing hypothesis: Evidence from normal ageing and dementia*, pages 139–159. Research Methods in Developmental Psychology: A Handbook Series. Routledge/Taylor & Francis Group, United Kingdom, Nov 2016.

S. L. Mann, E. A. Hazlett, W. Byne, P. R. Hof, M. S. Buchsbaum, B. H. Cohen, K. E. Goldstein, M. M. Haznedar, E. M. Mitsis, L. J. Siever, et al. Anterior and posterior cingulate cortex volume in healthy adults: effects of aging and gender differences. *Brain research*, 1401:18–29, 2011.

A. F. Marquand, I. Rezek, J. Buitelaar, and C. F. Beckmann. Understanding heterogeneity in clinical cohorts using normative models: beyond case-control studies. *Biological psychiatry*, 80(7):552–561, 2016a.

A. F. Marquand, T. Wolfers, M. Mennes, J. Buitelaar, and C. F. Beckmann. Beyond lumping and splitting: a review of computational approaches for stratifying psychiatric disorders. *Biological psychiatry: cognitive neuroscience and neuroimaging*, 1(5):433–447, 2016b.

A. F. Marquand, S. M. Kia, M. Zabihi, T. Wolfers, J. K. Buitelaar, and C. F. Beckmann. Conceptualizing mental disorders as deviations from normative functioning. *Molecular psychiatry*, 24(10):1415–1424, 2019.

B. D. Marx and P. H. Eilers. Multidimensional penalized signal regression. *Technometrics*, 47(1):13–22, 2005.

A. Mayr, T. Hothorn, and N. Fenske. Prediction intervals for future BMI values of individual children – a non-parametric approach by quantile boosting. *BMC Medical Research Methodology*, 12(1):6, 2012.

N. Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999, 2006.

A. F. Mejia, M. B. Nebel, Y. Wang, B. S. Caffo, and Y. Guo. Template independent component analysis: Targeted and reliable estimation of subject-level brain networks using big data population priors. *Journal of the American Statistical Association*, 115(531):1151–1177, 2020.

K. L. Miller, F. Alfaro-Almagro, N. K. Bangerter, D. L. Thomas, E. Yacoub, J. Xu, A. J. Bartsch, S. Jbabdi, S. N. Sotiropoulos, J. L. Andersson, et al. Multimodal population brain imaging in the uk biobank prospective epidemiological study. *Nature neuroscience*, 19(11):1523–1536, 2016.

A. C. Monti et al. A note on the estimation of the skew normal and the skew exponential power distributions. *Metron*, 61(2):205–219, 2003.

J. S. Morris. Functional regression. *Annual Review of Statistics and Its Application*, 2:321–359, 2015.

K. Mosler and Y. Polyakova. General notions of depth for functional data. *arXiv preprint arXiv:1208.1981*, 2012.

F. Mosteller and J. W. Tukey. *Data analysis and regression: a second course in statistics*. Addison-Wesley Series in Behavioral Science: Quantitative Methods. Pearson, 1977.

S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. R. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett. Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimer's & Dementia*, 1(1):55–66, 2005.

H.-G. Müller and U. Stadtmüller. Generalized functional linear models. *the Annals of Statistics*, 33(2):774–805, 2005.

J. Muschelli, E. Sweeney, M. Lindquist, and C. Crainiceanu. fslr: Connecting the FSL software with R. *The R journal*, 7(1):163, 2015.

J. L. Myers, A. D. Well, and R. F. Lorch Jr. *Research design and statistical analysis*. Routledge, 2013.

L. D. Nickerson, S. M. Smith, D. Öngür, and C. F. Beckmann. Using dual regression to investigate network shape and amplitude in functional connectivity analyses. *Frontiers in neuroscience*, 11:115, 2017.

N. L. Olsen, A. Pini, and S. Vantini. Local inference for functional data controlling the functional false discovery rate. In *International Workshop on Functional and Operatorial Statistics*, pages 205–211. Springer, 2020.

M. Palma, S. Tavakoli, J. Brettschneider, T. E. Nichols, and ADNI. Quantifying uncertainty in brain-predicted age using scalar-on-image quantile regression. *NeuroImage*, 219, 2020.

A. Y. Park, J. A. Aston, and F. Ferraty. Stable and predictive functional domain selection with application to brain images. *arXiv preprint arXiv:1606.02186*, 2016.

O. Parker Jones, N. Voets, J. Adcock, R. Stacey, and S. Jbabdi. Resting connectivity predicts task activation in pre-surgical populations. *NeuroImage: Clinical*, 13: 378–385, 2017.

G. Peng, J. Wang, Z. Feng, P. Liu, Y. Zhang, F. He, Z. Chen, K. Zhao, and B. Luo. Clinical and neuroimaging differences between posterior cortical atrophy and typical amnestic Alzheimer's disease patients at an early disease stage. *Scientific reports*, 6(1):1–11, 2016.

W. D. Penny, K. J. Friston, J. T. Ashburner, S. J. Kiebel, and T. E. Nichols. *Statistical parametric mapping: the analysis of functional brain images*. Elsevier, 2011.

A. Pini and S. Vantini. Interval-wise testing for functional data. *Journal of Nonparametric Statistics*, 29(2):407–424, 2017.

C. Preda and G. Saporta. PLS regression on a stochastic process. *Computational Statistics & Data Analysis*, 48(1):149–158, 2005.

C. Preda and J. Schiltz. Functional PLS regression with functional response: the basis expansion approach. In *Proceedings of the 14th Applied Stochastic Models and Data Analysis Conference*, pages 1126–1133. Università di Roma La Sapienza, 2011.

C. Preul, M. Hund-Georgiadis, B. U. Forstmann, and G. Lohmann. Characterization of cortical thickness and ventricular width in normal aging: a morphometric study at 3 tesla. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 24(3):513–519, 2006.

F. Privé, H. Aschard, A. Ziyatdinov, and M. G. Blum. Efficient analysis of large-scale genome-wide data with two r packages: bigstatsr and bigsnpr. *Bioinformatics*, 2018. doi: 10.1093/bioinformatics/bty185.

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer, 2005. ISBN 9780387400808.

P. T. Reiss and R. T. Ogden. Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association*, 102 (479):984–996, 2007.

P. T. Reiss and R. T. Ogden. Functional generalized linear models with images as predictors. *Biometrics*, 66(1):61–69, 2010.

P. T. Reiss, J. Goldsmith, H. L. Shang, and R. T. Ogden. Methods for Scalar-on-Function Regression. *International Statistical Review*, 85(2):228–249, 2017.

R. A. Rigby and D. M. Stasinopoulos. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554, 2005.

A. N. Ruigrok, G. Salimi-Khorshidi, M.-C. Lai, S. Baron-Cohen, M. V. Lombardo, R. J. Tait, and J. Suckling. A meta-analysis of sex differences in human brain structure. *Neuroscience & Biobehavioral Reviews*, 39:34–50, 2014.

D. Scheinost, S. Noble, C. Horien, A. S. Greene, E. M. Lake, M. Salehi, S. Gao, X. Shen, D. O'Connor, D. S. Barron, S. W. Yip, M. D. Rosenberg, and R. T. Constable. Ten simple rules for predictive modeling of individual differences in neuroimaging. *Neuroimage*, 193:35–45, 2019.

M.-A. Schulz, B. T. Yeo, J. T. Vogelstein, J. Mourao-Miranada, J. N. Kather, K. Kording, B. Richards, and D. Bzdok. Different scaling of linear models and deep learning in UK Biobank brain images versus machine-learning datasets. *Nature communications*, 11(1):1–15, 2020.

B. Sherwood and A. Maidman. *rqPen: Penalized Quantile Regression*, 2017. R package version 2.0.

M. Sklar. Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 8:229–231, 1959.

S. Smith, F. Alfaro-Almagro, and K. Miller. UK Biobank brain imaging documentation, 2018.

S. M. Smith and T. E. Nichols. Statistical challenges in "big data" human neuroimaging. *Neuron*, 97(2):263–268, 2018.

S. M. Smith, T. E. Nichols, D. Vidaurre, A. M. Winkler, T. E. Behrens, M. F. Glasser, K. Ugurbil, D. M. Barch, D. C. Van Essen, and K. L. Miller. A positive-negative

mode of population covariation links brain connectivity, demographics and behavior. *Nature neuroscience*, 18(11):1565–1567, 2015.

S. M. Smith, D. Vidaurre, F. Alfaro-Almagro, T. E. Nichols, and K. L. Miller. Estimation of brain age delta from brain imaging. *Neuroimage*, 200:528–539, 2019.

D. Sone, I. Beheshti, N. Maikusa, M. Ota, Y. Kimura, N. Sato, M. Koepp, and H. Matsuda. Neuroimaging-based brain-age prediction in diverse forms of epilepsy: a signature of psychosis and beyond. *Molecular psychiatry*, pages 1–10, 2019. doi: https://doi.org/10.1038/s41380-019-0446-9.

A.-M. Staicu, C. M. Crainiceanu, D. S. Reich, and D. Ruppert. Modeling functional data with spatially heterogeneous shape characteristics. *Biometrics*, 68(2):331–343, 2012.

C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, B. Liu, P. Matthews, G. Ong, J. Pell, A. Silman, A. Young, T. Sprosen, T. Peakman, and R. Collins. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

M. Sutton, R. Thiébaut, and B. Liquet. Sparse partial least squares with group and subgroup structure. *Statistics in medicine*, 37(23):3338–3356, 2018.

I. Tavor, O. P. Jones, R. B. Mars, S. Smith, T. Behrens, and S. Jbabdi. Task-free MRI predicts individual differences in brain activity during task performance. *Science*, 352(6282):216–220, 2016.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Y. Varatharajah, S. Baradwaj, A. Kiraly, D. Ardila, R. Iyer, S. Shetty, and K. Kohlhoff. Predicting brain age using structural neuroimaging and deep learning. *bioRxiv*, page 497925, 2018.

J. Wang, M. J. Knol, A. Tiulpin, F. Dubost, M. de Bruijne, M. W. Vernooij, H. H. H. Adams, M. A. Ikram, W. J. Niessen, and G. V. Roshchupkin. Gray matter age prediction as a biomarker for risk of dementia. *Proceedings of the National Academy of Sciences*, 116(42):21213–21218, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1902376116.

J.-L. Wang, J.-M. Chiou, and H.-G. Müller. Functional data analysis. *Annual Review of Statistics and Its Application*, 3:257–295, 2016.

L. Wang. The L_1 penalized LAD estimator for high dimensional linear regression. *Journal of Multivariate Analysis*, 120:135–151, 2013.

X. Wang, B. Nan, J. Zhu, and R. Koeppe. Regularized 3D functional regression for brain image data via Haar wavelets. *The Annals of Applied Statistics*, 8(2):1045, 2014.

X. Wang, H. Zhu, and ADNI. Generalized scalar-on-image regression models via total variation. *Journal of the American Statistical Association*, 112(519):1156–1168, 2017.

S. N. Wood. *Generalized additive models: an introduction with R*. CRC press, 2017.

M. W. Woolrich, B. D. Ripley, M. Brady, and S. M. Smith. Temporal autocorrelation in univariate linear modeling of FMRI data. *Neuroimage*, 14(6):1370–1386, 2001.

B. A. Yankner, T. Lu, and P. Loerch. The aging brain. *Annu. Rev. Path. Mech. Dis.*, 3:41–66, 2008.

F. Yao, H.-G. Müller, and J.-L. Wang. Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, pages 2873–2903, 2005.

F. Yao, S. Sue-Chee, and F. Wang. Regularized partially functional quantile regression. *Journal of Multivariate Analysis*, 156:39–56, 2017.

K. Yoo, M. D. Rosenberg, W.-T. Hsu, S. Zhang, C.-S. R. Li, D. Scheinost, R. T. Constable, and M. M. Chun. Connectome-based predictive modeling of attention: Comparing different functional connectivity features and prediction methods across datasets. *Neuroimage*, 167:11–22, 2018.

S. Yu, G. Wang, L. Wang, and L. Yang. Multivariate spline estimation and inference for image-on-scalar regression. *arXiv preprint arXiv:2106.01431*, 2021.

Y.-Q. Zheng, S.-R. Farahibozorg, W. Gong, H. Rafipoor, S. Jbabdi, and S. Smith. Accurate predictions of individual differences in task-evoked brain activity from resting-state fMRI using a sparse ensemble learner. *bioRxiv*, 2021.

K. Q. Zhou and S. L. Portnoy. Direct use of regression quantiles to construct confidence sets in linear models. *The Annals of Statistics*, 24(1):287–306, 1996.

Y. Zhou, L. Zhao, N. Zhou, Y. Zhao, S. Marino, T. Wang, H. Sun, A. Toga, and I. Dinov. Predictive big data analytics using the UK Biobank data. *Scientific Reports*, 9(1):6012, 2019.

Z. Zhou. Fast implementation of partial least squares for function-on-function regression. *Journal of Multivariate Analysis*, 185:104769, 2021.

V. Zipunnikov, B. Caffo, D. M. Yousem, C. Davatzikos, B. S. Schwartz, and C. Crainiceanu. Multilevel functional principal component analysis for high-dimensional data. *Journal of Computational and Graphical Statistics*, 20(4): 852–873, 2011.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2): 301–320, 2005.